Using an ensemble learning approach in digital soil mapping of soil pH for the Thompson-Okanagan region of British Columbia

Authors: Zhang, Jin, Schmidt, Margaret G., Heung, Brandon, Bulmer, Chuck E., and Knudby, Anders

Source: Canadian Journal of Soil Science, 102(3) : 579-596

Published By: Canadian Science Publishing

URL: https://doi.org/10.1139/cjss-2021-0091

Canadian Journal of Soil Science

OPEN ACCESS | Article

Using an ensemble learning approach in digital soil mapping of soil pH for the Thompson-Okanagan region of British Columbia

Jin Zhang[a], Margaret G. Schmidt[b], Brandon Heung[c], Chuck E. Bulmer[d], and Anders Knudby[e]

[a]Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada; [b]Department of Geography and School of Environmental Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada; [c]Department of Plant, Food, and Environmental Sciences, Dalhousie University, 21 Cox Road, Truro, NS B2N 5E3, Canada; [d]British Columbia Ministry of Forests, Lands, Natural Resource Operations & Rural Development, Vernon, BC V1B 2C7, Canada; [e]Department of Geography, Environment and Geomatics, University of Ottawa, 60 University, Ottawa, ON K1N 6N5, Canada

Corresponding author: **Margaret G. Schmidt** (email: Margaret_Schmidt@sfu.ca)

Abstract

Information on the spatial distribution of soil pH is essential for assessing soil quality and soil productivity. Digital soil mapping (DSM) is commonly used to predict soil characteristics over various types of landscapes. Over the past decade, researchers have made progress using machine learning techniques to provide reliable predictions of soil properties with limited data. DSM studies often use a single learning approach, which is constructed with a machine learner that systematically extracts soil–environment relationships from a large database, whereby a fitted model is used to predict soil information in an unmapped area. The practice of using an ensemble learning approach, especially one that combines several base learners, has rarely been tested in DSM. We developed a workflow for using an ensemble learning algorithm to predict soil properties for the Thompson-Okanagan region of British Columbia, Canada. Here, we focused on soil pH and tested a variety of base learners. Base learners with high prediction accuracies were then used to construct a SuperLearner (SL) to extract the complex relationships between soil properties and environmental variables. The fitted SL was then used to predict soil properties at 25 m spatial resolution at three depth intervals (0–5, 5–15, and 15–30 cm). Prediction accuracies were assessed using an independent test dataset, which indicated that the SL had a similar prediction accuracy to the best individual base learners. Using the heterogeneous ensemble learning approach with a weighted average stacked generalization process eliminated the need to choose the best base learner.

Key words: ensemble learning, machine learning, SuperLearner, stacked generalization, digital soil mapping, pH, loss-based estimation

Résumé

On a absolument besoin d'informations sur la répartition du pH du sol dans l'espace pour évaluer la qualité du sol et sa productivité. Des cartes numériques du sol (CNS) servent couramment à prédire les particularités d'un sol en fonction du relief. Au cours de la dernière décennie, des chercheurs fait progresser cette technique en recourant à l'apprentissage automatique pour prédire de façon fiable les propriétés du sol à partir de données restreintes. Les travaux sur les CNS appliquent souvent une seule méthode d'apprentissage, s'appuyant sur un algorithme unique qui extrait de façon systématique les liens entre le sol et l'environnement d'une vaste base de données, et en vertu de laquelle le modèle ajusté prédit les particularités du sol dans les régions non cartographiées. On a rarement testé une méthode réunissant plusieurs algorithmes d'apprentissage pour les CSN. Les auteurs ont élaboré un flux de tâches reposant sur un algorithme d'apprentissage combinatoire en vue de prédire les propriétés du sol dans la région de Thompson-Okanagan, en Colombie-Britannique (Canada). À cette fin, ils se sont concentrés sur le pH du sol et ont testé plusieurs algorithmes de base. Ceux qui enregistraient les prévisions les plus exactes ont ensuite servi à bâtir un super algorithme dont on s'est servi pour extraire les relations complexes entre les propriétés du sol et les variables environnementales. Après ajustement, les auteurs ont utilisé leur super algorithme pour prédire les propriétés du sol à une résolution spatiale de 25 m, à trois intervalles de profondeur (0 à 5 cm, 5 à 15 cm et 15 à 30 cm). Ensuite, ils ont vérifié

l'exactitude des prévisions grâce à un jeu de données indépendant, ce qui a révélé que le super algorithme était aussi précis que le meilleur algorithme de base. L'apprentissage combinatoire avec des algorithmes hétérogènes associé à un procédé de généralisation séquentielle fondé sur le calcul de la moyenne a permis de passer outre la sélection du meilleur algorithme de base. [Traduit par la Rédaction]

Mots-clés : apprentissage combinatoire, apprentissage automatique, super algorithme d'apprentissage, généralisation séquentielle, cartes du sol numériques, pH, estimation selon les pertes

Introduction

Digital soil mapping (DSM) has increasingly applied novel machine learning techniques to predict the spatial distribution of soil properties and types (Brungard et al. 2015; Heung et al. 2016; Khaledian and Miller 2020). Machine learning algorithms have the potential to quantify the high-dimensional and nonlinear relationships between the environmental predictors and soil response variables over diverse ecosystem types. With improvements in computer technology (Rossiter 2018) and machine learning algorithms over the past decade, more powerful learners were designed to process larger datasets using a larger number of environmental variables. Examples of such algorithms have included, but are not limited to, generalized linear regression (GLM; Hastie and Pregibon 1992), stepwise regression (STEP; Hastie and Pregibon 1992), and lasso and elastic net regularized generalized linear regression (GLMNET; Friedman et al. 2010), which are capable of processing nonlinear relationships for both categorical and continuous data (Simon et al. 2011). Furthermore, the use of tree-based learners, such as the classification and regression trees (CART; Breiman et al. 1984), has led to the development of predictive modelling techniques that are effective in capturing the hierarchical relationships between predictors. The CART approaches also form the basis of more advanced, tree-based learners such as CART with bagging (Breiman 1996*a*), the cubist learner (Quinlan 1992, 1993), and the random forest (RF) model (Breiman 2001).

The availability of numerous machine learning algorithms has encouraged model comparison studies, and these studies have shown that by using the same input data, different learners could generate digital soil maps that are drastically different from one another (Brungard et al. 2015; Heung et al. 2016). Hence, it has been recommended that model comparison should be carried out as part of best practice in DSM (Heung et al. 2016). In addition to a diverse array of machine learners, DSM practitioners have also investigated the use of ensemble models by extending the application of the "bagging" concept proposed in Breiman (1996*b*). Here, multiple models are built on bootstrapped samples of the training data and integrated into a single predictive model to improve the model predictions in comparison to predictions made using only one model (Rokach 2010).

Building an ensemble of models using a single type of learner (i.e., homogeneous ensemble learning) has been of interest in the DSM literature when predicting the spatial distribution of soil categorical data and continuous data. Studies such as Heung et al. (2017) have applied a bootstrapping routine for *k*-nearest neighbours, multinomial logistic regression, and logistic model trees for mapping soil classes, and Padarian et al. (2017) used a bootstrapping of CART for predicting a variety of soil properties across six depth intervals. Furthermore, polygon disaggregation approaches such as DSMART, which uses the See5 tree-based algorithm (Odgers et al. 2014), and its subsequent implementation using the RF algorithm (Chaney et al. 2016), both operate on a similar principle.

Although homogeneous ensemble learning methods have been tested to some extent in DSM (e.g., Heung et al. 2017; Padarian et al. 2017), there has been considerably less attention on modelling approaches that combine multiple types of learners (i.e., heterogeneous ensemble learning). Within the DSM literature, Malone et al. (2014) tested a variety of model averaging approaches using equal weight averaging, Bates–Granger averaging, Bayesian model averaging, and Granger–Ramanathan averaging, which showed that model averaging had the potential to improve map accuracy. Subsequently, O'Rourke et al. (2016) applied a similar model averaging approach and evaluated the improved accuracy of portable visible, near-infrared, and X-ray fluorescence spectrometers. More recently, multiple studies in France have evaluated the use of model averaging techniques in DSM (Román Dobarco et al. 2017; Caubet et al. 2019; Chen et al. 2020).

Within the machine learning literature, stacked generalization is a type of ensemble learning and model averaging approach. As with other model averaging techniques, stacked generalization operates on the concept that multiple predictive learners (i.e., "base learners") are aggregated into a combined learner, using a combiner algorithm (i.e., "meta-learner"), whereby the expectation is that the combined model has a higher predictive performance (Wolpert 1992). Here, the meta-learner evaluates the predictive performance of the individual base learners and builds an optimal combination. Stacked generalization was first proposed and tested for categorical data (Wolpert 1992) and later adapted into regression stacking for continuous data (Breiman 1996*a*).

An example of an ensemble learning approach that uses stacked generalization is the SuperLearner (SL), which was first proposed by van der Laan et al. (2007) and further evaluated by Polley and van der Laan (2010). The SL is unique in that it uses a variety of different base learners and a cost function based on cross-validation to create a heterogeneous ensemble. For example, its first implementation (van der Laan et al. 2007) combined modelling methods such as regression trees, RF, least angle regression, logistic regression, and adaptive regression splines. The construction of an SL includes two steps. First, the ensemble learning algorithm uses cross-validation to evaluate the performance of the base learners. In the second step, a cost function is applied, based on the cross-validation results, to calculate a weighted average prediction from the base learners. Whereas model averag-

ing incorporates the predictions of all base learners, the SL also calculates a weighted combination function using least squares regression but includes the additional constraint that all model weights are positive — a constraint that is ensured by removing all learners that have a negative weight (i.e., non-negative least squares). Subsequently, all model weights are normalized. Using an SL can mitigate bias, noise, and uncertainties from individual base learners, and may improve the overall accuracy of the prediction (Polley and van der Laan 2010).

Although the use of SL has rarely been demonstrated in DSM applications, its use was suggested by Hengl and MacMillan (2019). Applying the stacked generalization approach to ensemble learning and the mapping of continuous soil attributes is relatively uncommon in DSM, with the recent exception of Taghizadeh-Mehrjardi et al. (2021), who used a stacked generalization process to predict soil organic carbon. Hence, the objectives of this study were to (1) evaluate and compare a set of base learners, (2) test the potential of using the ensemble learning approach with stacked generalization to extract the relationships between soil properties and environmental variables derived from a digital elevation model (DEM), and (3) compare and assess the use of the ensemble learner with the individual base learners for mapping the spatial distribution of soil pH at multiple depth increments for the Thompson-Okanagan region of British Columbia, Canada.

Materials and methods

This study modelled soil pH by using multiple base learners, e.g., GLM, STEP, generalized linear model with lasso or elastic net regularization (GLMNET_lasso, GLMNET_ridge, and GLMNET enet), support vector machine–radial (SVMR), *k*-nearest neighbours (kNN), RF, and extreme gradient boosting (XGBoost), and compared them against an ensemble learning approach with stacked generalization. In all cases, point data were spatially intersected with raster layers representing soil–environmental variables to create the training data. The training data were then used to fit the base learners and a stacked generalization process was applied using the SL algorithm to create the ensemble prediction of soil pH. As the goal of this study was to examine the effectiveness of the ensemble learning process with a continuous soil attribute, we statistically compared the performance of SL to the base learners. Figure 1 outlines the general workflow of using an ensemble learning algorithm with stacked generalization for DSM purposes.

Study area

This study was carried out in the Thompson-Okanagan region of British Columbia, Canada (map sheet NTS 092INE; Fig. 2). This area is located in the southern interior of British Columbia and includes the Thompson Plateau, the Fraser Plateau, and the Shuswap Highland physiographic subdivisions (Young et al. 1992). It spans latitudes 50.5°N to 51.0°N and longitudes 120.0°W to 121.0°W, and is approximately 4350 km^2 in size. The elevation ranges from 318 m to 2088 m above mean sea level, and the area includes the following biogeoclimatic zones: Bunchgrass Zone, Interior Douglas-fir Zone, Montane Spruce Zone, Sub-Boreal Spruce Zone, Interior Cedar–Hemlock Zone, Engelmann Spruce–Subalpine Fir Zone, and Alpine Tundra Zone (Lloyd et al. 1990). The combination of these biogeoclimatic zones results in a variety of ecosystems, and the soil maps of the Thompson-Okanagan region include 99 soil associations for the study area. Each soil association was linked to one or more soil subgroups from the Canadian System of Soil Classification, of which there were 31 in total (Young et al. 1992). Digitized conventional soil polygon maps for the study area were accessed and downloaded from the BC Soil Information Finder Tool (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018).

Due to the climatic conditions and topography, the landscape features in the study area represent a combination of grassland regions, transitional regions, and dry interior forested regions (Klenner et al. 2008). The majority of the grasslands occupy catchment areas near Kamloops Lake and the Thompson River at lower elevations. Grasslands cover 40% of the region and occur between elevations of 230 m and 800 m. The climate is warm and dry in the grassland region; the dominant vegetation consists of grasses and sedges, and the dominant soils are Chernozems. At higher elevations, away from the water basins, the climate is colder and wetter; vegetation transitions into forests dominated by ponderosa pine (*Pinus ponderosa*), Douglas fir (*Pseudotsuga menziesii*), lodgepole pine (*Pinus contorta*), white spruce (*Picea glauca*), and Engelmann spruce (*Picea engelmannii*) at increasingly higher elevations (>800 m; Lloyd et al. 1990; Moore et al. 2010). In the grassland region, the mean annual temperature is 7.9 °C and the mean annual precipitation is 285 mm, whereas in the forest region the mean annual temperature is 5.0 °C and the mean annual precipitation is 476 mm (Lloyd et al. 1990). Most of the grasslands are maintained as pasture to support livestock, and forestry is the other primary industry in the area.

Soil sampling

A conditioned Latin hypercube sampling approach (Minasny and McBratney 2006) was used to select 300 sample locations based on topographic variables (Fig. 2). To ensure accessibility, locations were constrained to lie within 200 m of the road network, which included paved, logging, and gravel roads. There were 15 of the selected locations that were not sampled, as 10 locations were inaccessible due to road conditions and five locations had either exposed bedrock or shallow soils over bedrock, not meeting the minimum thickness of soil for the first depth interval. Thus, 285 profiles were sampled, with total depths ranging from 10 cm to 45 cm. Among the 285 sampled profiles, 278 had soil depth greater than 30 cm, four had soil depth between 15 cm and 30 cm, and three were shallower than 15 cm.

Fieldwork was carried out in the summer of 2015 using a field sheet based on the second edition of the "Field Manual for Describing Terrestrial Ecosystems" (B.C. Ministry of Forests and Range and B.C. Ministry of Environment 2010).

Fig. 1. The workflow used to predict soil pH (0–5 , 5–15, and 15–30 cm) in the Thompson–Okanagan region. The output map is in 25 m raster format. [Colour online.]

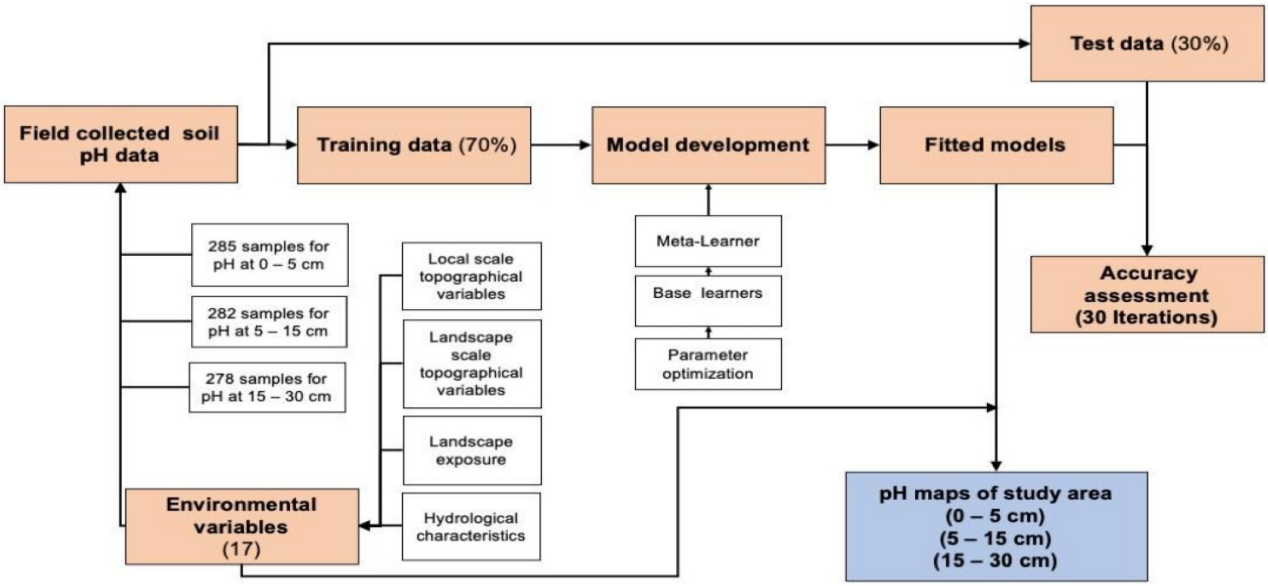
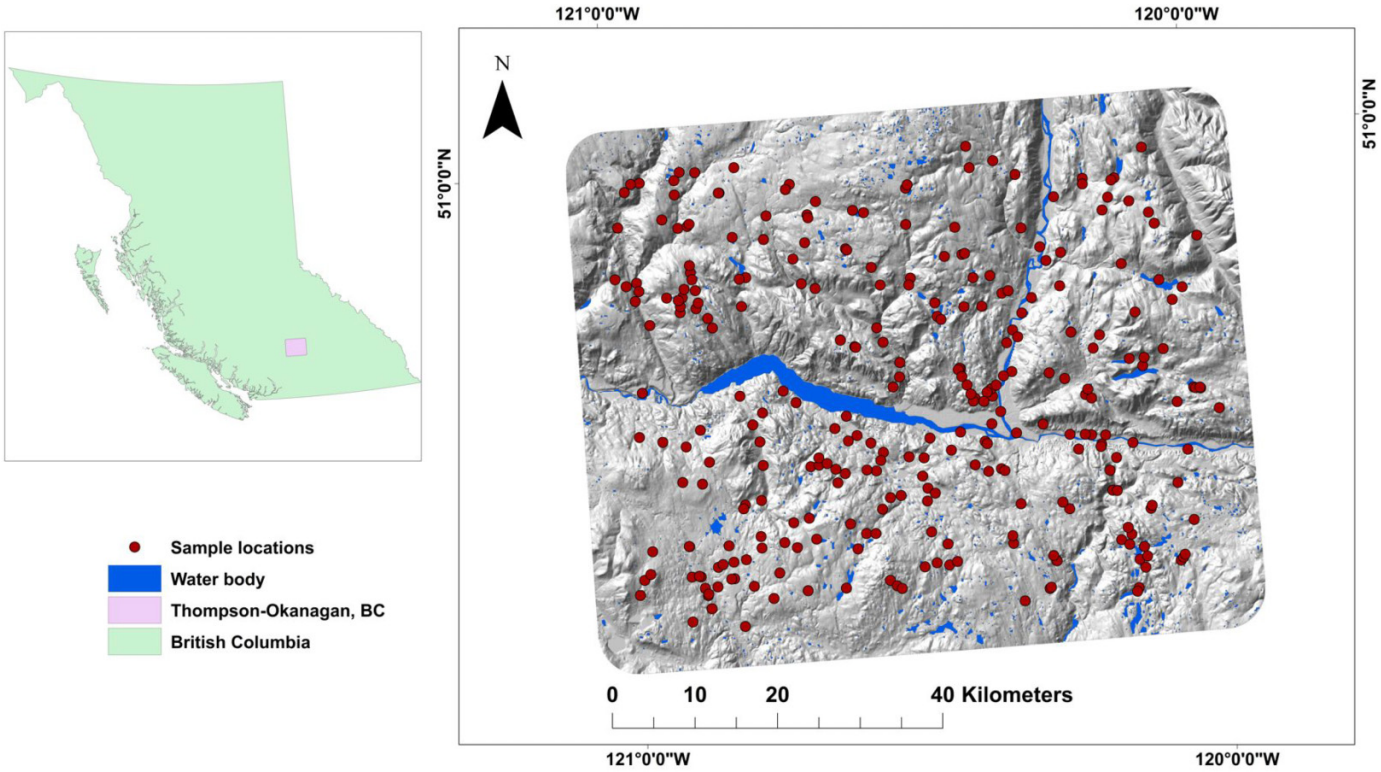


Fig. 2. Study area in the Thompson-Okanagan region, BC (1:25 000 map grid, composed of map sheets: 92I090, 92I010, 92I150, and 92I160). The red dots are the sample points. The coordinates refer to UTM zone 10N and the projection is NAD83/BC Albers. ArcGIS 10.3. software was used to produce the map with a hillshade underlain. [Colour online.]



At each field location, mineral samples were collected from individual horizons. In total, 845 soil samples were collected from 285 field locations. All mineral soil samples were air dried and passed through a 2 mm sieve. The fine fraction of each air-dried soil sample was analysed for pH (water). Lab analyses were carried out in the B.C. Ministry of Environment Analytical Chemical Research Laboratory. The pH was measured with a pH/ion conductivity meter using a water solution at a ratio of 1:2 (Kalra and Maynard 1991).

Because the soil samples were collected on a horizon basis, soil pH data were converted into standard depth incre-

ments (0–5, 5–15, and 15–30 cm), based on the specifications of GlobalSoilMap.net products (Arrouays et al. 2014), using the equal-area spline function (Bishop et al. 1999) from the ithir package in the R statistical language (Malone 2017). The 0–5, 5–15, and 15–30 cm depth increments had pH values for 285, 282, and 278 sample locations, respectively. In addition, descriptive statistics were calculated for the soil pH values using the JMP 13.0 software. Analysis of variance was also carried out at the three standard depth increments to compare the pH values with respect to the samples acquired from the forest-dominated (F), mixed forest and grass (FG), and grass-dominated (G) landscapes. Surface vegetation type (F, FG, or G) was classified for each sample location based on the observation of vegetation cover in the field.

Environmental predictors

To build the training dataset and predict the soil pH, 17 environmental variables (Table 1) were derived at a 25 m spatial resolution from a DEM (B.C. Ministry of Sustainable Resource Management 2002) and used as predictors. The DEM was preprocessed using sequential mean filters with window sizes of 3×3, 3×3, and 5×5 cells to reduce noise and anomalies in the rasters (Heung et al. 2014). Environmental variables were then calculated to represent local-scale topography (e.g., elevation, slope, aspect, and curvature), landscape-scale topography (e.g., multiresolution index of valley bottom flatness), climatic characteristics (e.g., diurnal anisotropic heating and diffused insolation), and hydrological characteristics (e.g., topographic wetness index). All variables were calculated using the System for Automated Geoscientific Analysis (SAGA) software (SAGA Development Core Team 2011; Conrad et al. 2015) and projected using the Albers equal-area conic projection system using the NAD83 datum.

Predictive models

The following sections provide a brief description of the base learners; however, readers are encouraged to refer to the references provided in Table 2 for detailed descriptions.

The GLM assumes that the regression function is linear in its inputs, which are comprised of independent environmental variables, and takes the following form:

$$(1) \quad y_i = \beta_0 + \beta_i x_{ij} + \varepsilon_{ij}, \quad \varepsilon\left(0, \sigma^2\right)$$

$$(2) \quad \varepsilon_{ij} = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}$$

$$(3) \quad L_{\mathrm{OLS}}(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

where y is the dependent response variable (soil attribute), x is the independent predictor variable, N is the number of predictor variables, β_0 is the intercept, β_i is the partial regression coefficient for each predictor variable, and ε is the error term. The ordinary least squares (OLS) method determines the coefficients of the independent variables as well as the intercept value by minimizing the sum of squared residuals (eq. 3). As a result of its structure, GLM often has an interpretable description of how the predictor variables influence the target variable (Hastie et al. 2009). The R package glm was used to develop the GLM learner (Dobson 2002).

The STEP is a type of multiple regression technique that selects the best-fitted combination of independent variables to predict the dependent variable. The process includes forward addition and backward removal of predictors based on the Akaike information criterion. The R package step with backward removal was used to develop the STEP learner (Venables and Ripley 2002).

The GLMNET is an extension of the GLM model; however, it applies a shrinkage and/or regularization approach to minimize the number of predictors within the model (Hastie and Qian 2016; Hastie et al. 2016). The shrinkage method used by GLMNET is controlled by the alpha hyperparameter: when alpha = 0, ridge regression (Ridge) is employed; when alpha = 1, lasso regression (LASSO) is employed; and when 0 < alpha < 1, elastic net regression (ENET), a hybrid (i.e., mixing) of ridge and lasso, is employed. A full description of these methods is beyond the scope of this study, but readers may refer to Friedman et al. (2010) for more details. The use of GLMNET is commonly seen in medical studies and biological science, and it has been especially popular in epigenome-wide association studies (Horvath 2013; Knight et al. 2016), but it is less common in DSM. The R package glmnet was used to develop the GLMNET learner (Friedman et al. 2010). Previously, soil organic carbon (r^2 = 0.50) and four other important soil nutrients were predicted using GLMNET in India (Sirsat et al. 2018), and Li et al. (2020) used GLMNET with multiple environmental variables to estimate soil thickness (concordance correlation coefficient (CCC) = 0.76) in Henan Province, China. Most recently, Taghizadeh-Mehrjardi et al. (2021) used GLMNET to predict 13 soil properties in Iran.

The SVMR was proposed by Vapnik et al. (1997) to use a nonlinear transformation technique to project the original input into hyperspace and then generate a linear regression in this newly developed multidimensional feature space. In this study, a radial basis function kernel was used to create the regression function. kNN is a supervised learner that uses a nonparametric method to predict the value in the target cell based on the values of the k closest neighbouring observations in feature space (Kuhn 2008). The RF learner uses an ensemble of individual tree-based models and is derived from the CART model (Breiman 2001). The individual trees are trained using a bootstrap sample of the training data and additional randomness is incorporated into the model because the variables used to generate the binary splits at each node of each tree are drawn using a random subset of the predictor variables (Breiman 2001). XGBoost uses a gradient boosting framework to build a strong learner from several weak learners and uses many decision trees to make the prediction. The uniqueness of XGBoost is the construction of new decision trees, which are based on the prediction errors

Table 1. Covariates derived from a 25 m spatial resolution DEM used for modelling soil pH.

Covariate type	Covariate	Data format	Reference
Local-scale topography	Elevation	Raster	
	Slope	Raster	Zevenbergen and Thorne (1987)
	Aspect	Raster	Zevenbergen and Thorne (1987)
	Plan curvature	Raster	Zevenbergen and Thorne (1987)
	Profile curvature	Raster	Zevenbergen and Thorne (1987)
	Convergence index	Raster	Koethe and Lehmeier (1996)
Landscape-scale topography	Multiresolution index of valley bottom flatness	Raster	Gallant and Dowling (2003)
	Multiresolution ridge top flatness	Raster	Gallant and Dowling (2003)
	Total catchment area	Raster	SAGA Development Core Team (2011)
	Slope length factor	Raster	Moore et al. (1993)
Climatic characteristics	Diffused insolation	Raster	Böhner and Antonić (2009); Oke (2002); Wilson and Gallant (2000); Hofierka and Suri (2002)
	Directed insolation	Raster	Böhner and Antonić (2009); Oke (2002); Wilson and Gallant (2000); Hofierka and Suri (2002)
	Analytical hillshading	Raster	SAGA Development Core Team (2011)
	Diurnal anisotropic heating	Raster	Böhner and Antonić (2009)
Hydrological characteristics	Channel network base level	Raster	SAGA Development Core Team (2011)
	Topographic wetness index	Raster	Beven and Kirkby (1979)
	Altitude above channel network	Raster	SAGA Development Core Team (2011)

Table 2. Summary of all the learners and corresponding hyperparameters.

Learners	Definition	Hyperparameters	R package	Reference
GLM	General linear regression	None	glm	Dobson (2002)
STEP	Stepwise linear regression	None	step	Venables and Ripley (2002)
Ridge	GLMNET—ridge regression	lambda	caret	Kuhn (2008); Hastie and Qian (2016); Hastie et al. (2016)
LASSO	GLMNET—the least absolute shrinkage and selection operator	fraction	caret	Kuhn (2008); Hastie and Qian (2016); Hastie et al. (2016)
ENET	GLMNET—the eleastic net	lambda, fraction	caret, enet	Kuhn (2008); Hastie and Qian (2016); Hastie et al. (2016)
SVMR	Support vector machine—radial	*C*, sigma	caret	Vapnik et al. (1997); Kuhn (2008)
kNN	*k*-nearest neighbours	*K*	caret	Kuhn (2008)
RF	Random forest	mtry, ntree	caret	Breiman (2001); Kuhn (2008)
XGBoost	Extreme gradient boosting	e.g., booster, nrounds, max_depth, gamma, eta	caret	Kuhn (2008); Chen et al. (2015)
SuperLearner	Model ensemble	weights	superlearner	Polley and van der Laan (2010)

of the previous tree model to minimize the prediction error of the final prediction. Therefore, the final predictions are an ensemble of several decision trees (Chen et al. 2015; Chen and Guestrin 2016).

Stacked generalization using SL

The SL algorithm is an ensemble learner that uses the stacked generalization concept (Polley and van der Laan 2010). Here, a model intercept is not included, and the coefficients, which represent the weights of the weighted combination of the learners, cannot be negative and must sum to 1. The following equation describes the weighted combination function:

$$(4) \quad Y_{\text{obs}} = \alpha_1 \widehat{Y}_1 + \cdots + \alpha_k \widehat{Y}_k$$

where Y_{obs} represents the observed value, $\widehat{Y}_k$ represents the predicted value from base learner k, and α_k represents the weight of that base learner's predicted value. When estimating the weights, a non-negative least squares regression approach is applied with the aim to minimize the mean square error (MSE). To ensure the non-negative coefficient constraint, all base learners that have negative coefficients following the MSE minimization are removed from the SL model. Then, to ensure that all coefficients sum to 1, the

remaining coefficients are normalized. The SL and stacked generalization approach is distinguished from model averaging approaches based on these additional constraints. This approach differs from Granger–Ramanathan model averaging, which does not include the non-negative coefficient constraint.

All the base learners were available using the caret package (Kuhn 2008) and the SuperLearner package (Polley et al. 2019) in the R statistical software (R Development Core Team 2012). The caret package was used to facilitate optimization of the hyperparameters of the base learners, whereas the SuperLearner package integrated the stacked generalization process of the predictions (van der Laan and Dudoit 2003). The SuperLearner package is particularly useful for the ensemble learning approach as the package compiles a library of base learners from the existing R packages, including the caret package. The SL has been previously evaluated and tested in biostatistical studies (van der Laan and Dudoit 2003; van der Laan et al. 2007). Using the outputs predicted by the base learners and their associated weights, the SL may be used to generate a map of a target soil variable.

Model training and testing

Every sample location was spatially intersected with the 17 environmental variable layers to create the full observational dataset. Random holdback cross-validation was used and thus the full dataset was partitioned into a training dataset (70%) and a test dataset (30%). The training data were fed into each base learner to evaluate the relationship between soil pH and environmental variables; this relationship was then used to predict soil pH for all locations in the study area. A nested cross-validation was applied to build and test the SL and base learners. Within the training dataset (70%), a 10-fold cross-validation procedure was used to optimize the model hyperparameters and to determine the weights of the individual base learners used to build the SL (Polley et al. 2019). The independent test dataset (30%) was used to calculate the accuracy metrics.

The prediction accuracies of both the base learners and the SL were quantified using MSE, Lin's CCC, and bias. Here, the accuracy metrics were calculated using only the independent test data (30%) from the nested cross-validation procedure, which were not used for fitting the SL or optimizing the model hyperparameters. Mean square error is defined as the mean of the square of the difference between the observed values and predicted values, and is a measure of global model uncertainty (Schluchter 2005); CCC measures the agreement between the observed values and the predicted values, and is a measure of model accuracy (Lin 1989); and bias is calculated as the difference between the mean of the predictions and the mean of the observed values (Bellon-Maurel et al. 2010).

We repeated the nested cross-validation procedure 30 times to ensure the stability and reliability of the results, and we reported the mean value and standard deviation for each accuracy metric (Engebretsen and Bohlin 2019; Fig. 1). Furthermore, the accuracy metrics from these 30 repeats were used as the basis for statistical comparisons between the individual base learners and the SL using a comparison of means with a control using Dunnett's test with $\alpha = 0.05$.

Spatial prediction using SL

After 30 repeats of the nested cross-validation procedure, the process resulted in 30 fitted SL models with the corresponding 30 sets of fitted base learners. The SL that yielded the highest CCC value was then used in the spatial prediction, where the weighted combination of selected base learners was calculated during the training process. The spatial prediction process with SL had two steps. The first was to individually produce the spatial predictions of all the fitted base learners with the 17 topographic raster layers. In the second step, the outputs of the base learners were then used as inputs to the weighted combination to produce the final SL output. All maps were produced at a 25 m spatial resolution.

Results and discussion

Soil pH

The mean value of soil pH was highest at the 15–30 cm depth increment and lowest at the 0–5 cm increment (Table 3). Vegetation type, which had been determined by observation in the field, had a strong influence on pH (Table 4). At both the 0–5 and 5–15 cm depth increments, pH was significantly different between each of the vegetation types, with pH being highest for grass (G), intermediate for forest intermixed with grass (FG), and lowest for forest (F). At the 15–30 cm depth increment, there were no significant differences in pH among the three vegetation types. These significant effects in the surface horizons can be partly attributed to the increase in effective moisture and leaching in the forested environments, which tend to occur at higher elevations in the study area (Young et al. 1992; Jobbágy and Jackson 2003). Deeper in the profile, where parent material is expected to exert a greater influence on soil properties, the effect of vegetation is muted, and fewer significant effects due to vegetation are expected.

Topography can influence soil pH by controlling water flow, material redistribution, and microclimate (Moore et al. 1993). In our study, pH values were significantly correlated with two topographic variables (channel network base level and elevation) at all three standard depth increments, and the correlations were all negative (Table 5): pH tended to be higher at lower channel network base level values and at lower elevations. These results also reflect the increased leaching intensity at higher elevations. Similar trends of pH being related to topographic variables have been observed in several previous studies (Moore et al. 1993; Smith et al. 2002; Zhang et al. 2019). Several studies have found an increase in soil pH at downslope positions (Brubaker et al. 1993; Chen et al. 1997; Zhang et al. 2019) and Chen et al. (1997) found that topographic variables, such as aspect and slope, were controlling factors of the spatial distribution of soil pH in the mountainous area of east Taiwan.

The effects of vegetation and topography on soil pH are likely interdependent. Overall, soil pH was higher in the grass-covered area, which was at lower elevations and was

Table 3. Summary of descriptive statistics for the soil pH data at three standard depth intervals.

Soil properties (layers)	*N*	Min	Max	Mean	SD	Q25	Q50	Q75	CV	Skewness
pH (0–5 cm)	285	4.75	8.60	6.73a	0.91	6.13	6.78	7.44	13.60	–0.21
pH (5–15 cm)	278	4.83	8.55	6.78ab	0.93	6.12	6.77	7.55	13.66	–0.09
pH (15–30 cm)	269	4.83	9.63	6.92b	1.01	6.13	6.77	7.77	14.62	0.06

Note: pH, soil pH [log(H^+/OH^+)]; Min, minimum; Max, maximum; SD, standard deviation; CV, coefficient of variation (%) is defined as the ratio of the SD to the mean; Q25, Q50, and Q75 refer to the 25% quartile, median, and 75% quartile, respectively. Letters refer to differences in mean pH between different depths based on analysis of variance, $\alpha = 0.5$.

Table 4. Analysis of variance and analysis of mean results for the difference in pH among vegetation types.

	Mean pH value	*P* value comparison with total mean	*P* value comparison with F	*P* value comparison with FG	*P* value comparison with G
pH (0–5 cm)					
F	6.4	<0.0001	–	<0.0001	<0.0001
FG	7.0	0.0279	<0.0001	–	0.0018
G	7.5	<0.0001	<0.0001	0.0018	–
pH (5–15 cm)					
F	6.4	<0.0001	–	<0.0001	<0.0001
FG	7.1	0.0093	<0.0001	–	0.0005
G	7.6	<0.0001	<0.0001	0.0005	–
pH (15–30 cm)					
F	6.5	0.69	–	0.72	0.73
FG	7.3	087	0.76	–	1
G	8.0	0.87	0.77	1	–

Note: The confidence interval is $\alpha = 0.5$. F, a forest-dominated landscape; FG, a forest and grass mixed landscape; G, a grass-covered landscape. BEC database and field observation were used to define the classification of surface vegetation units. The points located in the Bunchgrass Zone (BG) were classified as G. *P* value, comparison with the mean value in that depth interval. $\alpha = 0.5$.

Table 5. Pearson correlation coefficients between pH at three depths and environmental variables.

Environmental variables	pH (0–5 cm)	pH (5–15 cm)	pH (15–30 cm)
Analytical hillshading	0.09	0.09	0.08
Altitude above channel network	–0.15*	–0.14*	–0.14*
Aspect	0.02	0.004	–0.02
Total catchment area	–0.01	–0.01	–0.02
Channel network base level	–0.67***	–0.69***	–0.69***
Convergence index	–0.005	0.001	0.02
Diurnal anisotropic heating	0.13*	0.12*	0.10
Elevation	–0.68***	–0.71***	–0.70***
Slope length factor	0.11	0.11	0.10
Multiresolution ridge top flatness	0.02	0.006	–0.009
Multiresolution index of valley bottom flatness	0.001	–0.010	–0.01
Plan curvature	0.01	0.01	0.02
Profile curvature	–0.20**	–0.21**	–0.20**
Slope	0.10	0.11	0.10
Topographic wetness index	0.03	0.02	0.01
Directed insolation	–0.02	–0.03	–0.04
Diffused insolation	–0.67***	–0.69***	–0.68***

*Correlation is significant at $P < 0.05$.
**Correlation is significant at $P < 0.005$
***Correlation is significant at $P < 0.0001$.

Table 6. Overall error rate of the ensemble learners and base learners using 30 repeats of random holdback cross-validation for soil pH at three depth intervals.

Depth increment	Learners	CCC			MSE			Bias		
		Mean	SD	*P* value	Mean	SD	*P* value	Mean	SD	*P* value
pH (0–5 cm)	GLM	0.49	0.21	0.16	1.41	1.8	<0.01	−0.01	0.10	1
	STEP	0.53	0.18	0.90	0.91	0.91	0.57	0.00	0.10	1
	GLMNET_RIDGE	0.49	0.20	0.27	1.28	1.47	0.02	−0.01	0.10	1
	GLMNET_LASSO	0.49	0.21	0.18	1.35	1.64	0.01	−0.01	0.10	1
	GLMNET_ENET	0.56	0.11	1	0.59	0.35	1	0.00	0.10	1
	SVMR	0.58	0.05	0.99	0.47	0.04	1	−0.02	0.10	1
	kNN	0.27	0.07	<0.01	0.76	0.10	0.96	0.01	0.12	0.88
	RF	0.60	0.04	0.77	0.47	0.04	1	−0.02	0.10	1
	XGBoost	0.60	0.04	0.92	0.48	0.06	1	−0.06	0.10	0.65
	SL	0.56	0.10	–	0.55	0.21	–	−0.02	0.10	–
pH (5–15 cm)	GLM	0.62	0.05	1	0.49	0.04	0.39	−0.01	0.10	1
	STEP	0.63	0.04	0.78	0.47	0.04	1	0.00	0.02	1
	GLMNET_RIDGE	0.63	0.05	0.97	0.48	0.05	0.88	−0.01	0.10	1
	GLMNET_LASSO	0.62	0.05	1	0.49	0.05	0.48	−0.01	0.10	1
	GLMNET_ENET	0.62	0.04	0.99	0.46	0.04	0.99	0.00	0.10	1
	SVMR	0.61	0.04	1	0.46	0.04	1	−0.02	0.10	1
	kNN	0.34	0.08	<0.01	0.74	0.09	<0.01	0.01	0.12	0.88
	RF	0.62	0.03	1	0.48	0.04	0.98	−0.02	0.10	1
	XGBoost	0.61	0.04	1	0.48	0.05	0.72	−0.06	0.10	0.65
	SL	0.61	0.04	–	0.46	0.04	–	−0.02	0.10	–
pH (15–30 cm)	GLM	0.57	0.03	0.27	0.89	0.88	0.09	0.03	0.02	0.98
	STEP	0.59	0.14	0.88	0.79	0.76	0.43	0.02	0.11	0.99
	GLMNET_RIDGE	0.58	0.15	0.60	0.86	0.86	0.17	0.02	0.11	1
	GLMNET_LASSO	0.57	0.15	0.31	0.88	0.87	0.10	0.03	0.12	0.99
	GLMNET_ENET	0.62	0.07	1	0.55	0.12	1	0.01	0.10	1
	SVMR	0.62	0.06	1	0.55	0.08	1	−0.04	0.09	0.61
	kNN	0.29	0.07	<0.01	0.88	0.10	0.10	0.02	0.13	1
	RF	0.63	0.05	0.99	0.54	0.07	1	0.02	0.10	1
	XGBoost	0.62	0.05	1	0.56	0.08	1	−0.04	0.10	0.65
	SL	0.62	0.07	–	0.56	0.14	–	0.01	0.10	–

Note: Concordance (CCC), mean square error (MSE), and bias (Bias) with their associated standard deviations (SD) were calculated based on the results of 30 iterations. Mean comparison analysis was carried out between each base learner with SL, $\alpha = 0.05$.

generally dry and hot, and soil pH was lower in the forested area, which was at higher elevations with cooler temperatures and humid conditions. Chytrý et al. (2007) observed that soil pH decreased with increasing precipitation in the mountain area in southern Siberia. The higher amount of precipitation at higher elevations, where the forest is, increases the rate of leaching in the soil, which increases the concentration of H^+ ions and thus decreases the pH. Lower soil pH was also reported at higher elevations in an oak woodland–conifer forest in the western United States (Dahlgren et al. 1997) and pine forests in the eastern United States.

Evaluation of base learners

The prediction performance and accuracy assessments for each base learner and SL are summarized in Table 6. Mean square error, CCC, and bias were calculated based on the 30 repeats for each depth interval. Random forest and XGBoost performed consistently well with mean CCC ranging from 0.60 to 0.63 for all three depth increments; furthermore, the standard deviation values were consistently low as they ranged from 0.03 to 0.05, which indicates stability in the models.

At the 0–5 cm depth increment, RF and XGBoost were the two best performing base learners with CCC = 0.60; however, SVMR and GLMNET_ENET performed similarly with CCC = 0.58 and 0.56, respectively. Furthermore, the MSE and bias were also similar, ranging from MSE = 0.47 to 0.59 and bias = –0.06 to –0.02. With the exception of the kNN base learner, all other learners performed similarly at the 5–15 cm depth increment where the CCC ranged from 0.61 to 0.63 and the MSE and bias shared similar values. At the 15–30 cm depth increment, a similar pattern was observed in which the kNN learner performed the worst with CCC = 0.29 in comparison to the other learners, where the CCC ranged from 0.57 to 0.63.

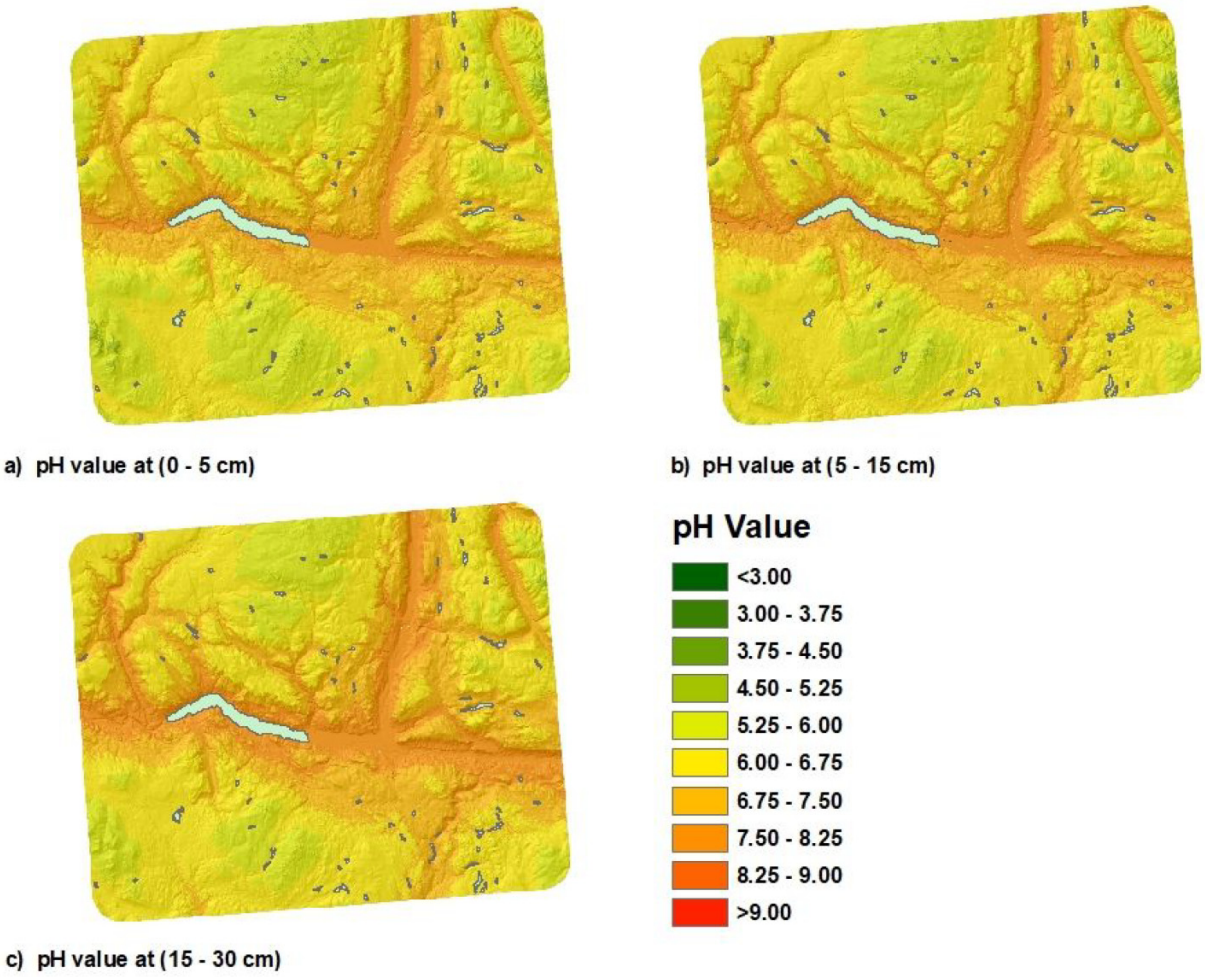
In Taghizadeh-Mehrjardi et al. (2021), a higher performance of the RF and XGBoost learners was similarly observed. It is possible that this may be attributed to the fact that both

Table 7. Summary of weights for each base learner over 10 nested cross-validations.

Base learners	pH (0–5 cm) weight (%)		pH (5–15 cm) weight (%)		pH (15–30 cm) weight (%)	
	Mean	SD	Mean	SD	Mean	SD
GLM	0	0	0	0	0	0
STEP	1	0.04	3	0.08	0	0
GLMNET_RIDGE	0	0	0	0	0	0
GLMNET_LASSO	0	0	0	0	2	0.06
GLMNET_ENET	24	0.32	72	0.17	19	0.29
SVMR	39	0.27	4	0.10	35	0.26
kNN	1	0.02	0	0	1	0.01
RF	28	0.16	7	0.11	21	0.14
XGBoost	7	0.11	14	0.12	22	0.20

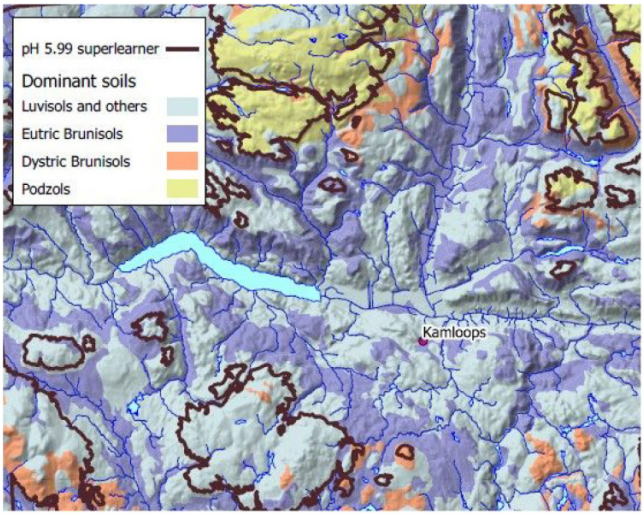
Note: SD, standard deviation.

Fig. 3. Predicted soil pH at three depth intervals in the study area with the optimized ensemble learning model (SL). The coordinates refer to UTM zone 10N and the map projection is NAD83/BC Albers. ArcGIS 10.3. software was used to produce the map with a hillshade underlain. [Colour online.]



models are ensemble machine learning models themselves—RF uses a bagging framework while XGBoost uses a boosting framework, and furthermore both models are tree-based models. Comparing the other learners, GLMNET and STEP showed a better prediction accuracy than GLM, which indicated that variable selection and the regularization process, such as the one used in the GLMNET model, may have reduced the prediction errors of the linear regression

Fig. 4. Contour line of $pH_{H_2O} = 5.99$ (equivalent to $pH_{CaCl_2} = 5.5$) and soil classification of Eutric Brunisol, Dystric Brunisol, and Podzol in the Thompson–Okanagan region. Based on the description in CSSS (Soil Classification Working Group 1998), $pH_{CaCl_2} = 5.5$ in the B horizon (estimated as 5–15 cm in our study) is the defining criterion to separate Dystric Brunisol and Eutric Brunisol. The shapefile of the water bodies in the region was obtained from the B.C. data catalogue (B.C. Ministry of Agriculture and Land 2008). The digitized soil map was obtained from the British Columbia Soil Information Finder Tool (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018). The coordinates refer to UTM zone 10N and the map projection is NAD83/BC Albers. ArcGIS 10.3. software was used to produce the map with a hillshade underlain. [Colour online.]

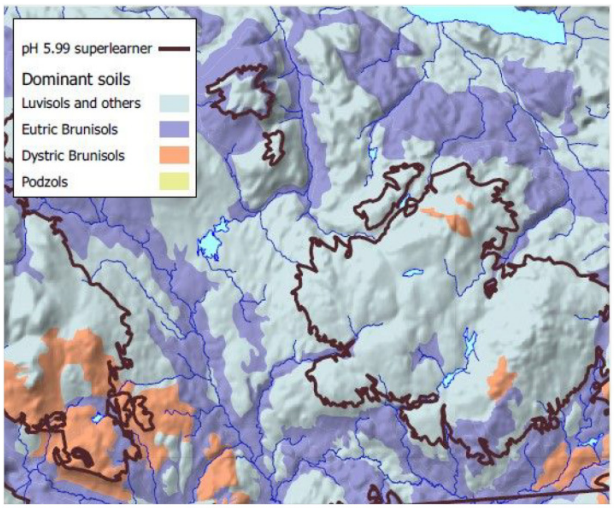


learners. Previous studies showed that when comparing both linear and nonlinear models, the linear models were the least effective while the decision tree learners were the most effective (Khaledian and Miller 2020; Taghizadeh-Mehrjardi et al. 2021). We observed this to be the case only at the 0–5 cm depth increment and not at the 5–15 cm and 15–30 cm depth increments, thus suggesting that nonlinear relationships may be present between soil pH and the environment for surficial soils.

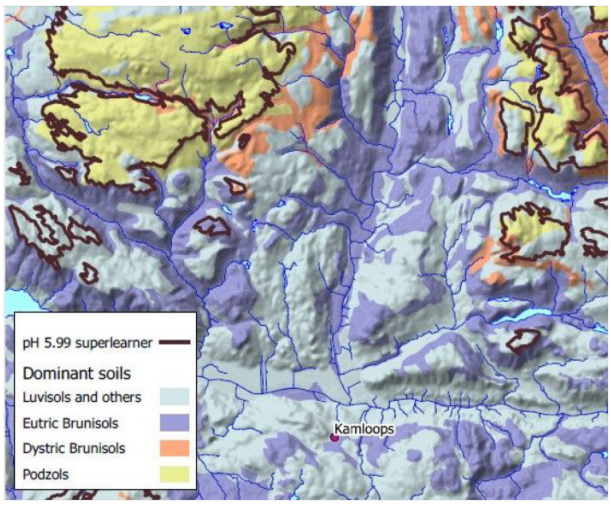
Evaluation of SL

In fitting the SL, the base learners were weighted for each depth increment and are summarized in Table 7. Here, we reported the mean model weight from the 30 repeats and the corresponding standard deviation values for each base learner. It is important to note that the SL applies a nonnegative least squares framework in estimating the model weights and hence the models that do not meet the nonnegative condition were assigned a weight of 0. Overall, GLMNET_ENET, SVMR, RF, and XGBoost were consistently used in fitting the SL. Based on the low standard deviation values for the model weights, the SL appears to be stable when selecting the base learners as well as when calculating their corresponding weights.

Fig. 5. Close-ups showing the contrasting relationship between the $pH_{H_2O} = 5.99$ contour and the boundaries of soil units in the study area. (A) In the southwest portion of the study area, the contour line is relatively close to or slightly above the upper elevation boundary of polygons with Eutric Brunisols present (shown in purple). This approximates the expected relationship based on the criteria in CSSS (Soil Classification Working Group 1998). (B) In the north northeast portion, the $pH_{H_2O} = 5.99$ contour line often occurs at a higher elevation than expected. In this portion of the study area, soil units mapped as Dystric Brunisols (shown in orange) often appear at lower elevations than the contour line. In this portion of the map, either the pH 5.99 contour is being predicted at a higher elevation than expected or the soil mapping for Dystric Brunisols is inaccurate. The shapefile of the water bodies in the region was obtained from the B.C. data catalogue (B.C. Ministry of Agriculture and Land 2008). Digitized soil maps were obtained from the British Columbia Soil Information Finder Tool (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018). The coordinates refer to UTM zone 10N and the map projection is NAD83/BC Albers. ArcGIS 10.3. software was used to produce the map with a hillshade underlain. [Colour online.]

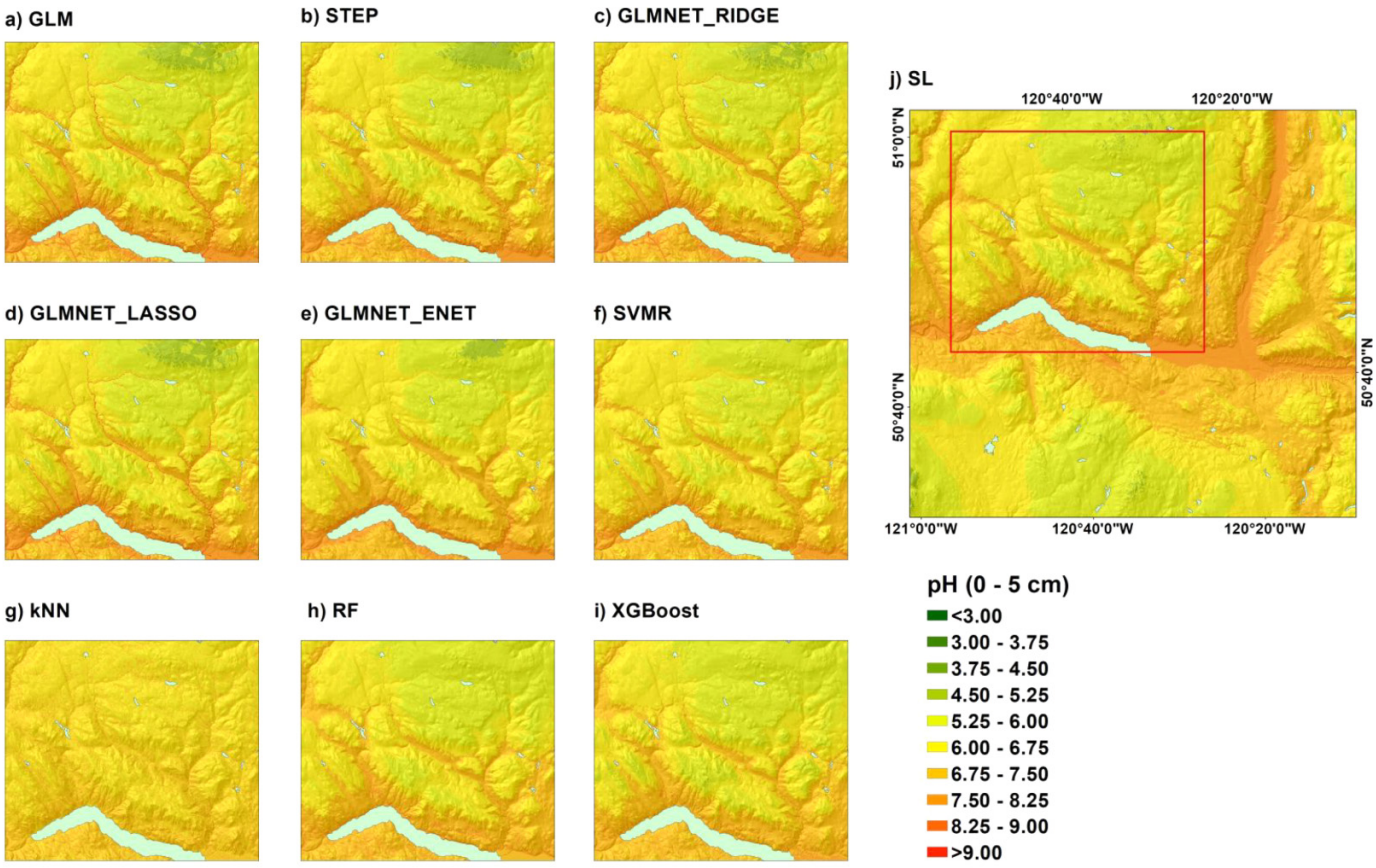


A



B

Fig. 6. Close-up showing the difference between the prediction results from different base learners and the results from the optimized ensemble learning model (SL) at a depth increment of 0–5 cm. (*a*) GLM, CCC = 0.09; (*b*) STEP, CCC = 0.71; (*c*) GLMNET_RIDGE, CCC = 0.16; (*d*) GLMNET_LASSO, CCC = 0.11; (*e*) GLMNET_ENET, CCC = 0.68; (*f*) SVMR, CCC = 0.66; (*g*) kNN, CCC = 0.28; (*h*) RF, CCC = 0.69; (*i*) XGBoost, CCC = 0.66; and (*j*) SL, CCC = 0.70. The shapefile of the water bodies in the region was obtained from the B.C. data catalogue (B.C. Ministry of Agriculture and Land 2008). Digitized soil maps were obtained from the British Columbia Soil Information Finder Tool (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018). The coordinates refer to UTM zone 10N and the map projection is NAD83/BC Albers. ArcGIS 10.3. software was used to produce the map with a hillshade underlain. [Colour online.]

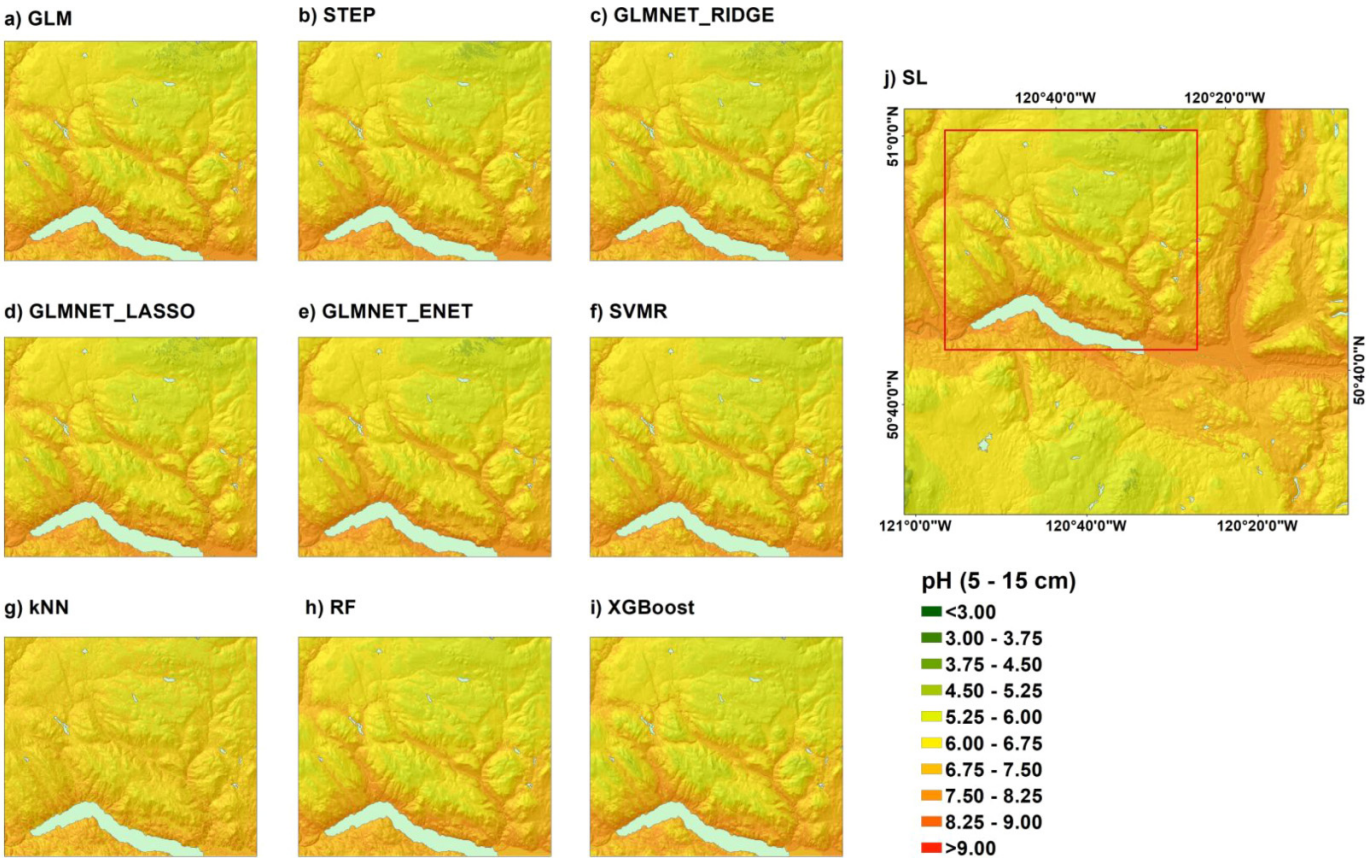


It is also interesting to note that of the linear models, GLMNET_ENET performed the best and was included in the SL while all other models were assigned weights close to 0. In comparison, models with drastically different structures, such as SVMR, RF, and XGBoost, were consistently included in the SL. This observation shows that the SL selects diverse models rather than models with similar structures, such as the linear models. Although kNN is also distinct from the other models, it was weighted close to 0, which is due to the fact that it was consistently the poorest performing model across all depth increments. These findings were consistent with Polley et al. (2019), who suggested that a diverse set of base learners, including both linear and nonlinear base learners, should be used to fit the ensemble learning algorithm instead of using only similar base learners or testing only a few base learners.

The external test data showed that, based on the mean of the 30 iterations, the SL had CCC values of 0.56, 0.61, and 0.62 for the 0–5, 5–15, and 15–30 cm depth increments, respectively. Based on the CCC, the SL performed 14.3% better than the GLM learner at the 0–5 cm depth increment while performing approximately twice as effectively as kNN (Table 6). With respect to MSE, the SL had a significantly lower MSE of 0.55 when compared to the GLM model (1.41) at the 0–5 cm depth increment and had significantly lower MSE of 0.46 when compared to the kNN model (0.74) at the 5–15 cm increment. When using bias to assess the difference between the mean predictions of all the learners and the mean of the observed values, there were no significant differences. In general, the SL did not show significant improvements in accuracy, global uncertainty, and bias, which contradicted our original expectations based on Polley and van der Laan (2010) and Taghizadeh-Mehrjardi et al. (2021).

In a recent study, Taghizadeh-Mehrjardi et al. (2021) used 14 models to construct an ensemble learner and their overall finding was that the ensemble learner outperformed all base

Fig. 7. Close-up showing the difference between the prediction results from different base learners and the results from the optimized ensemble learning model (SL) at a depth increment of 5–15 cm. (*a*) GLM, CCC = 0.70; (*b*) STEP, CCC = 0.71; (*c*) GLMNET_RIDGE, CCC = 0.71; (*d*) GLMNET_LASSO, CCC = 0.70; (*e*) GLMNET_ENET, CCC = 0.71; (*f*) SVMR, CCC = 0.69; (*g*) kNN, CCC = 0.49; (*h*) RF, CCC = 0.68; (*i*) XGBoost, CCC = 0.71; and (*j*) SL, CCC = 0.71. The shapefile of the water bodies in the region was obtained from the B.C. data catalogue (B.C. Ministry of Agriculture and Land 2008). The digitized soil maps were obtained from the British Columbia Soil Information Finder Tool (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018). The coordinates refer to UTM zone 10N and the map projection is NAD83/BC Albers. ArcGIS 10.3. software was used to produce the map with a hillshade underlain. [Colour online.]



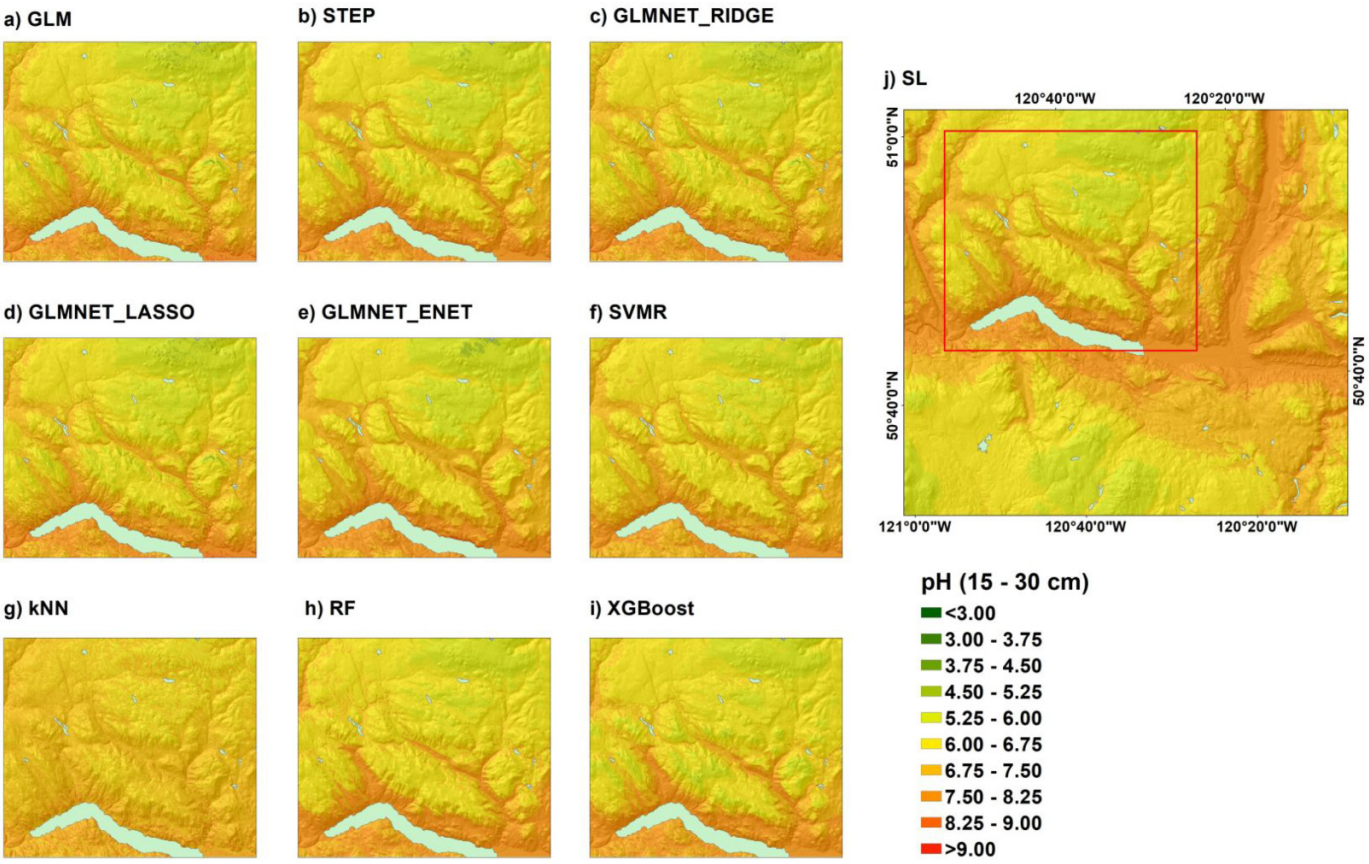
learners. Whereas this study showed that except for the kNN model, the range of accuracy metrics was fairly similar across all models, the range of accuracy metrics varied far more in Taghizadeh-Mehrjardi et al. (2021). A possible recommendation would be that DSM practitioners should first perform a comprehensive comparison of base learners, and if the results are inconsistent, the application of SL may be warranted despite the cost of additional computation and model interpretability.

Visual assessment

We used the SL to predict the spatial distribution of soil pH at three depth intervals over the Thompson-Okanagan region (Fig. 3). The greatest spatial variation in pH occurred at the 0–5 cm depth increment (Fig. 3*a*), in which it is clear that soil pH is highest near the Thompson River and Kamloops Lake, and lowest at the boundary of the mountain regions of the study area. A similar spatial pattern was also observed at the 5–15 and 15–30 cm depths (Figs. 3*b* and 3*c*), in which the soil pH decreases with increasing elevation, and with distance from the stream network. This is also revealed by the correlation coefficient analysis, which showed that soil pH had a strong negative relationship with elevation (Table 5), and the predicted map of soil pH shows that soil has lower pH in the forest at higher elevations (Fig. 3). The low pH in the forest region could partly be related to the high organic matter content in the forest floor. A second pattern shows that soil pH is lower at the 0–5 cm depth and higher at the 15–30 cm depth. This could be the effect of accumulated base cations from the parent material.

In Fig. 4, we present a contour line for the pH value measured in water (pH_{H_2O}) of 5.99, which closely approximates a pH measured in $CaCl_2$ (pH_{CaCl_2}) of 5.5, for the 5–15 cm depth interval. The contour line is shown in relation to the soil mapping boundaries for the study area. In the Canadian System of Soil Classification, a pH_{CaCl_2} value of 5.5 is a diagnostic criterion for the separation of the Eutric Brunisol great group ($pH_{CaCl_2} > 5.5$) from the Dystric Brunisol great group ($pH_{CaCl_2} < 5.5$). Because of the declining pH values with increasing elevation in the study area, the contour line at pH_{H_2O} of 5.99 was

Fig. 8. Close-up showing the difference between the prediction results from different base learners and the results from the optimized ensemble learning model (SL) at a depth increment of 15–30 cm. (*a*) GLM, CCC = 0.64; (*b*) STEP, CCC = 0.68; (*c*) GLMNET_RIDGE, CCC = 0.64; (*d*) GLMNET_LASSO, CCC = 0.64; (*e*) GLMNET_ENET, CCC = 0.66; (*f*) SVMR, CCC = 0.61; (*g*) kNN, CCC = 0.30; (*h*) RF, CCC = 0.67; (*i*) XGBoost, CCC = 0.66; and (*j*) SL, CCC = 0.66. The shapefile of the water bodies in the region was obtained from the B.C. data catalogue (B.C. Ministry of Agriculture and Land 2008). The digitized soil maps were obtained from the British Columbia Soil Information Finder Tool (B.C. Ministry of Agriculture and B.C. Ministry of Environment 2018). The coordinates refer to UTM zone 10N and the map projection is NAD83/BC Albers. ArcGIS 10.3. software was used to produce the map with a hillshade underlain. [Colour online.]



expected to occur near the upper elevation boundary for map units where Eutric Brunisols occur and near the lower elevation boundary for map units containing Dystric Brunisols.

There was considerable variability throughout the study area, and Brunisols did not occur in all map units; in the southwestern portion of the study area, the location of the pH_{H_2O} = 5.99 contour line generally followed the expected behaviour by occurring near or slightly above the upper elevation where Eutric Brunisols were mapped (Fig. 5). In the north and northeastern portions of the study area, the pH_{H_2O} = 5.99 contour at times appeared at a higher elevation than expected. Despite the variability observed in our map results, information from the Canadian Soil Information System (Agriculture and Agri-Food Ottawa, Ontario 2000) confirms that the soil units mapped in the vicinity of our contour line have pH values in the B horizon that generally agree with the placement of the contour line. These results were consistent with our pedological understanding of the region and accuracy and reliability of our pH predictions, but also point to areas where the pH predictions could be improved.

To further investigate the predicted outputs of the SL in comparison to its constituent base learners, Figs. 6–8 show a close-up region of the study area for each base learner with respect to the SL and for each depth increment. It should be noted that these figures show the predictions for the single repeat of the nested cross-validation that resulted in the highest CCC, in which the CCC values were 0.70, 0.71, and 0.66 for the 0–5, 5–15, and 15–30 cm depth increments, respectively. The spatial patterns are similar between the base learners and SL within each depth increment, except for kNN, where spatial patterning was less obvious, and a relatively uniform distribution of pH was predicted across the study area, which may account for the model's lower accuracy. Similar to when the accuracy metrics were investigated, when the spatial patterns between the base learners were similar, there appeared to be a limited influence on the spatial patterns when using the SL. This, again, suggests that a soil mapper should carry out a visual assessment of the base learners prior to investing additional time and computation into the SL.

General discussion

For the surface horizon (0–5 cm), this study showed that the variation in topography had a direct influence on the spatial distribution of pH, and the linear regression learners with variable reduction achieved a prediction accuracy similar to the regression tree learners. As shown in the results, more powerful learners such as RF and GLMNET were more effective than GLM and kNN. On the other hand, this study only used topographic variables as environmental variables. Adding additional variables to represent vegetation, climate, and parent material may have improved the prediction accuracy.

Other studies have shown that increasing sample size and using additional environmental variables derived from hyperspectral images can improve the accuracy of prediction (Lagacherie et al. 2019). We recognize the potential limitation of using only topographic variables as predictors, and we considered leveraging satellite imagery, such as Landsat and Sentinel 2 images, to represent vegetative patterns in the study area. However, in the study area, vegetative and climatic patterns are largely controlled by topography; lower elevations of the region are dominated by grasslands where it is warmer and drier, and higher elevations are dominated by dry interior forests where moisture is higher. Furthermore, an important limitation of satellite data is that disturbance to natural vegetation patterns, primarily caused by fire and pine beetles, but also by clear-cutting, is substantial in this area. Satellite images show present vegetation patterns; considering that the time frame during which the soils developed, and as the result of the pH of the original parent material being altered by pedogenic processes, is over 200 times as long as the period of significant human activity in the area, we believe that the topography serves as a more reliable indicator of the distribution of soil forming processes across the landscape.

Previous studies (Seibert et al. 2007; Tu et al. 2018) showed that the use of topographic indices alone, as in this study, had the potential to effectively map the spatial distribution of soil properties, including pH. Tu et al. (2018) suggested that at a local scale, soil pH in the upper profile has the strongest correlation with topographic indices compared to the lower profile. However, in this study, both the individual base learners and a stacked ensemble learner had higher prediction accuracies in the lower profile. Other researchers have found that parent material and vegetation-related indices, such as rooting depth, have a stronger influence on soil pH than topography (Reuter et al. 2008; Gruba and Socha 2016; Zhang et al. 2019).

Conclusion

This paper presented the use of SL, an ensemble learning algorithm with stacked generalization, to map the spatial distribution of soil pH at three depth intervals. The approach was applied to examine the use of topographic indices to map the spatial distribution of soil pH in the dry forest ecosystem in the Thompson-Okanagan region of British Columbia, Canada. The DSMs of soil pH at three depth intervals provide the first full-coverage map for the area, and the workflow may be further applied to map the spatial distribution of other soil properties. Elevation and surface vegetation types have a strong influence on the distribution of soil pH, with pH around 7.5 near the valley basin on the grassland and around 5.5 near the mountain tops in the forested area. Soil pH was higher in the 15–30 cm depth increment, compared to the shallower depth increments, which is likely the result of leaching, seasonal fluctuations of the water table, and sodium-rich parent material. Additional spatial information, such as vegetation, water, or parent material data, should be considered as additional predictor variables in future studies.

This study demonstrated how to use an ensemble learning algorithm with stacked generalization in DSM studies. When using a machine learning approach in data analysis and prediction, learner selection is often a challenge, especially when more than two learners show promising preliminary results. The ensemble learning approach with stacked generalization provides the option to combine the results from multiple learners to create an integrated mapping output with relatively stable performance. The SuperLearner package provides a solution that allows DSM practitioners to test many learners at the same time and to help them identify the most effective learners. Contrary to results from other studies, the SL does not necessarily outperform all the base learners, but it does provide near-optimal performance. We suggest that DSM practitioners should first carry out a comprehensive comparison of base learners, and if the model outputs yield substantially different results, the SL may provide the means for improving predictions.

Acknowledgements

The authors are thankful for the field and lab assistance provided by Carson Li and M.J. Jeon. The authors are also grateful for the support from the B.C. Ministry of Forests, Lands, Natural Resource Operations and Rural Development, and Simon Fraser University.

Article information

History dates

Received: 20 July 2021
Accepted: 28 November 2021
Accepted manuscript online: 4 February 2022
Version of record online: 18 July 2022

Notes

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

Copyright

Author information

Author notes

Authors Brandon Heung and Chuck E. Bulmer served as Guest Editors at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by Daniel Saurette.

References

Agriculture and Agri-Food Ottawa, Ontario 2000. Canadian soil information service[online]. Available from https://sis.agr.gc.ca/cansis/index.html. [accessed 22 June 2016].

Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., and McBratney, A.B. 2014. GlobalSoilMap: basis of the global spatial soil information system[online]. Taylor & Francis, Milton Park, Abingdon-on-Thames, Oxfordshire United Kingdom. Availablefrom https://books.google.ca/books?id=S5ClAgAAQBAJ. [accessed August 2018].

B.C. Ministry of Agriculture and B.C. Ministry of Environment. 2018. British Columbia Soil Information Finder Tool[online]. Available-from https://governmentofbc.maps.arcgis.com/apps/MapSeries/index.html?appid=cc25e43525c5471ca7b13d639bbcd7aa. [accessed September 2019].

B.C. Ministry of Agriculture and Land. 2008. Freshwater atlas[online]. B.C. Ministry of Agriculture and Land, British Columbia. Available from https://catalogue.data.gov.bc.ca/dataset/freshwater-atlas-watersheds.

B.C. Ministry of Forests and Range and B.C. Ministry of Environment. 2010. Field manual for describing terrestrial ecosystems. B.C. Ministry of Forests and Range and B.C. Ministry of Environment, British Columbia.

B.C. Ministry of Sustainable Resource Management. 2002. Gridded digital elevation model product specification. 2nd ed.[online]. Digital Elevation Model, B.C. Ministry of Sustainable Resource Management, British Columbia. Available from https://www2.gov.bc.ca/gov/content/data/geographic-data-services/topographic-data/elevation/digital-elevation-model. [accessed January 2015].

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., and McBratney, A. 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. Trends Anal. Chem. **29**: 1073–1081. doi:10.1016/j.trac.2010.05.006.

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. null. **24**: 43–69. Taylor & Francis. doi: 10.1080/02626667909491834.

Bishop, T.F.A., McBratney, A.B., and Laslett, G.M. 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma **91**: 27–45. doi: 10.1016/S0016-7061(99)00003-8

Böhner, J., and Antonić, O. 2009. Chapter 8 Land-Surface Parameters Specific to Topo-Climatology. In Developments in Soil Science, edited by-Hengl and Reuter.**33**: 195–226. Elsevier, doi: https://doi.org/10.1016/S0166-2481(08)00008-1. [accessed 15 September 2014].

Breiman, L. 1996a. Bagging predictors[online]. Kluwer Academic Publishers, Boston, Massachusetts, United States. doi: https://doi.org/10.1023/A:1018054314350.

Breiman, L. 1996*b*. Stacked regressions. Mach. Learn. **24**: 49–64. doi:10.1007/BF00117832.

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

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. 1984. Classification and regression trees. Taylor & Francis, Milton Park, Abingdon-on-Thames, Oxfordshire, United Kingdom.

Brubaker, S.C., Jones, A.J., Lewis, D.T., and Frank, K. 1993. Soil properties associated with landscape position. Soil Sci. Soc. Am. J. **57**: 235–239. doi:10.2136/sssaj1993.03615995005700010041x.

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.

Caubet, M., Román Dobarco, M., Arrouys, D., Minasny, B., and Saby, N.P.A. 2019. Merging country, continental and global predictions of soil texture: lessons from ensemble modelling in France. Geoderma, **337**: 99–110. doi:10.1016/j.geoderma.2018.09.007.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., and Odgers, N.P. 2016. POLARIS: a 30-meter probabilistic soil series map of the contiguous United States. Geoderma, **274**: 54–67. doi:10.1016/j.geoderma.2016.03.025.

Chen, S., Mulder, V.L., Heuvelink, G.B.M., Poggio, L., Caubet, M., Román Dobarco, M., Walter, C., and Arrouays, D. 2020. Model averaging for mapping topsoil organic carbon in France. Geoderma, **366**: 114237. doi:10.1016/j.geoderma.2020.114237.

Chen, T., and Guestrin, C. 2016. Xgboost: a scalable tree boosting system. *In* Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York. pp. 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., and Cho, H. 2015. XGBoost: extreme gradient boosting. R package version 0.4.2.

Chen, Z.-S., Hsieh, C.-F., Jiang, F.-Y., Hsieh, T.-H., and Sun, I.-F. 1997. Relations of soil properties to topography and vegetation in a subtropical rain forest in southern Taiwan. Plant Ecol. **132**: 229–241. doi:10.1023/A:1009762704553.

Chytrý, M., Danihelka, J., Ermakov, N., Hájek, M., Hájková, P., Kočí, M., et al. 2007. Plant species richness in continental southern Siberia: effects of pH and climate in the context of the species pool hypothesis. Global Ecol. Biogeogr. **16**: 668–678. doi:10.1111/j.1466-8238.2007.00320.x.

Conrad, O., Bechtel, B., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., et al. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. doi:10.5194/gmd-8-1991-2015.

Dahlgren, R.A., Boettinger, J.L., Huntington, G.L., and Amundson, R.G. 1997. Soil development along an elevational transect in the western Sierra Nevada, California. Geoderma, **78**: 207–236. doi:10.1016/S0016-7061(97)00034-7.

Dobson, A.J. 2002. An introduction to generalized linear models. Chapman & Hall/CRC, Boca Raton, FL.

Engebretsen, S., and Bohlin, J. 2019. Statistical predictions with glmnet. Clin. Epigenet. **11**: 123. doi:10.1186/s13148-019-0730-1.

Friedman, J.H., Hastie, T., and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**: 1–22. doi:10.18637/jss.v033.i01.

Gallant, J.C., and Dowling, T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resource Research. **39**: 1347–1359. doi:10.1029/2002WR001426. [accessed 15 September 2014].

Gruba, P., and Socha, J. 2016. Effect of parent material on soil acidity and carbon content in soils under silver fir (*Abies alba* Mill.) stands in Poland. Catena, **140**: 90–95. doi:10.1016/j.catena.2016.01.020.

Hastie, T., and Qian, J. 2016. Glmnet vignette. R - package. [Online] Available: https://cran.microsoft.com/snapshot/2018-03-30/web/packages/glmnet/vignettes/glmnet_beta.pdf. [accessed June 2019].

Hastie, T., Qian, J., and Tay, K. 2016. An introduction to glmnet[online]. Available from https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf. [accessed 13 June 2021].

Hastie, T.J., and Pregibon, D. 1992. Generalized linear models. *In* Statistical models in S. *Edited by* J.M. Chambers and T.J. Hastie. Wadsworth & Brooks/Cole, Belmont, California, United States. pp. 183–208.

Hastie, T.J., Tibshirani, R.J., and Friedman, J. 2009. The elements of statistical learning: data mining, inference and prediction. 2nd ed. Springer, New York.

Hengl, T., and MacMillan, R.A. 2019. Predictive soil mapping with R. OpenGeoHub Foundation, Wageningen, the Netherlands.

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., and Schmidt, M. 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. 2017. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma, **290**. doi:10.1016/j.geoderma.2016.12.001.

Hofierka, J., and Suri, M. 2002. The solar radiation model for Open source GIS: Implementation and applications. Page in Proceedings of the Open Source GIS-GRASS Users Conference. [Online] Available: http://skagit.meas.ncsu.edu/~jaroslav/trento/Hofierka_Jaroslav.pdf. [accessed 15 September 2014].

Horvath, S. 2013. DNA methylation age of human tissues and cell types. Genome Biol. **14**: 3156. doi:10.1186/gb-2013-14-10-r115.

Jobbágy, E.G., and Jackson, R.B. 2003. Patterns and mechanisms of soil acidification in the conversion of grasslands to forests. Biogeochemistry, **64**: 205–229. doi:10.1023/A:1024985629259.

Kalra, Y.P., and Maynard, D.G. 1991. Methods manual for forest soil and plant analysis. Forestry Canada Information Report NOR-S-319. Victoria, BC. pp. 31, 25, 42.

Khaledian, Y., and Miller, B.A. 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. **81**: 401–418. doi:10.1016/j.apm.2019.12.016.

Klenner, W., Walton, R., Arsenault, A., and Kremsater, L. 2008. Dry forests in the southern interior of British Columbia: historic disturbances and implications for restoration and management. For. Ecol. Manag. **256**: 1711–1722. doi:10.1016/j.foreco.2008.02.047.

Knight, A.K., Craig, J.M., Theda, C., Bækvad-Hansen, M., Bybjerg-Grauholm, J., Hansen, C.S., et al. 2016. An epigenetic clock for gestational age at birth based on blood methylation data. Genome Biol. **17**: 206. doi:10.1186/s13059-016-1068-z.

Koethe, R, and Lehmeier, F. 1996. SARA - System Zur Automatischen Relief-Analyse. User Manual, 2. Edition. Department of Geography, University of Göettingen,Göttingen. [accessed 15 September 2014].

Kuhn, M. 2008. Building predictive models in R using the caret package. J. Stat. Softw. **28**: 1–26. doi:10.18637/jss.v028.i05.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., and Saby, N.P.A. 2019. How far can the uncertainty on a digital soil map be known? A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma, **337**: 1320–1328. doi:10.1016/j.geoderma.2018.08.024.

Li, X., Luo, J., Jin, X., He, Q., and Niu, Y. 2020. Improving soil thickness estimations based on multiple environmental variables with stacking ensemble methods. Remote Sens. **12**: 3609 . doi:10.3390/rs12213609.

Lin, L.I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics, **45**: 255–268. doi:10.2307/2532051.

Lloyd, D., Angove, K., Hope, G., and Thompson, C. 1990. A guide to site identification and interpretation for the Kamloops Forest Region. B.C. Ministry of Forest, Canadian Cataloguing in Publication Data, Research Branch Ministry of Forests, Victoria, BC.

Malone, B.P. 2017. Ithir: functions and algorithms specific to pedometrics version 1.0[online]. Available from https://rdrr.io/rforge/ithir/. [accessed November 2017].

Malone, B.P., Minasny, B., Odgers, N.P., and McBratney, A.B. 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma, **232–234**: 34–44. doi:10.1016/j.geoderma.2014.04.033.

Minasny, B., and McBratney, A. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. **32**: 1378–1388. doi:10.1016/j.cageo.2005.12.009.

Moore, I.D., Gessler, P.E., Nielsen, G.A., and Peterson, G.A. 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. **57**: 443–452. doi:10.2136/sssaj1993.03615995005700020058x.

Moore, R.D., Spittlehouse, D.L., Whitfield, P.H., and Stahl, K. 2010. Weather and climate. *In* Compendium of forest hydrology and geomorphology in British Columbia[online]. *Edited by* R.G. Pike, T.E. Redding, R.D. Moore, R.D. Winkler, and K.D. Bladon, B.C. Ministry of Forest and Range, Forest Science Program, Victoria, BC and FORREX Forum for Research and Extension in Natural Resources, Kamloops, BC.pp. 47–84. Available from https://www.for.gov.bc.ca/hfd/pubs/docs/lmh/Lmh66.htm. [accessed November 2017].

O'Rourke, S.M., Stockmann, U., Holden, N.M., McBratney, A.B., and Minasny, B. 2016. An assessment of model averaging to improve predictive power of portable vis–NIR and XRF for the determination of agronomic soil properties. Geoderma, **279**: 31–44. doi:10.1016/j.geoderma.2016.05.005.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., and Clifford, D. 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma, **214–215**: 91–100. doi:10.1016/j.geoderma.2013.09.024.

Oke, T.R., 2002. Boundary layer climates, Routledge, Milton Park, Abingdon-on-Thames, Oxfordshire, United Kingdom.

Padarian, J., Minasny, B., and McBratney, A.B. 2017. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Reg. **9**: 17–28. doi:10.1016/j.geodrs.2016.12.001.

Polley, E., LeDell, E., Kennedy, C., Lendle, S., and van der Laan, M.J. 2019. SuperLearner: super learner prediction R packages [online]. Available from https://github.com/ecpolley/SuperLearner[accessed 2 January 2019].

Polley, E.C., and van der Laan, M.J. 2010. Super learner in prediction. Working Paper Series 266[online]. Division of Biostatistics, University of California, Berkeley. Available: http://biostats.bepress.com/ucbbiostat/paper266. [accessed December 2018].

Quinlan, J.R. 1992. Learning with continuous classes, 5th Australian joint conference on artificial intelligence. 343–348. in A. Adams and L. Sterling, eds. AI '92 Proceedings of the 5th Australian Joint Conference on Artificial Intelligence World Sci.Available: https://www.worldscientific.com/doi/abs/10.1142/9789814536271. [accessed 28 September 2016].

Quinlan, J.R. 1993. Combining instance-based and model-based learning. *In* Proceedings of the 10th international conference on international conference on machine learning. Morgan Kaufmann Publishers Inc., Amherst, MA. pp. 236–243.

R Development Core Team 2012. R: a language and environment for statistical computing[online]. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.r-project.org/. [accessed 26 February 2016].

Reuter, H.I., Lado, L.R., Hengl, T., and Montanarella, L. 2008. Continental-scale digital soil mapping using European soil profile data: soil pH. Hamb. Beitr. Phys. Geogr. Landschaftsökol. **19**: 91–102.

Rokach, L. 2010. Ensemble-based classifiers. Artif. Intell. Rev. **33**: 1–39. doi:10.1007/s10462-009-9124-7.

Román Dobarco, M., Arrouys, D., Lagacherie, P., Ciampalini, R., and Saby, N.P.A. 2017. Prediction of topsoil texture for region centre (France) applying model ensemble methods. Geoderma, **298**: 67–77. doi:10.1016/j.geoderma.2017.03.015.

Rossiter, D.G. 2018. Past, present & future of information technology in pedometrics. Geoderma, **324**: 131–137. doi:10.1016/j.geoderma.2018.03.009.

SAGA Development Core Team. 2011. System for Automated Geoscientific Analyses (SAGA)[online]. Availablefrom http://www.saga-gis.org/en/index.heml. [accessed September 2018].

Schluchter, M.D. 2005. Mean square error. *In* Encyclopedia of biostatistics. American Cancer Society, Atlanta, Georgia United States. doi: https://doi.org/10.1002/0470011815.b2a15087.

Seibert, J., Stendahl, J., and Sørensen, R. 2007. Topographical influences on soil properties in boreal forests. Geoderma, **141**: 139–148. doi:10.1016/j.geoderma.2007.05.013.

Simon, N., Friedman, J.H., Hastie, T., and Tibshirani, R. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. **39**: 1–13. doi:10.18637/jss.v039.i05.

Sirsat, M.S., Cernadas, E., Fernández-Delgado, M., and Barro, S. 2018. Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. Comput. Electron. Agric. **154**: 120–133. doi:10.1016/j.compag.2018.08.003.

Smith, J.L., Halvorson, J.J., and Bolton Jr,, H. 2002. Soil properties and microbial activity across a 500 m elevation gradient in a semi-arid environment. Soil Biol. Biochem. **34**: 1749–1757. doi:10.1016/S0038-0717(02)00162-1.

Soil Classification Working Group, Canadian Agricultural Services Coordinating Committee Soil Classification Working Group, National Research Council Canada, Canada Agriculture, and Agri-Food Canada Research Branch. 1998. The Canadian system of soil classification. 3rd ed. NRC Research Press, Ottawa, Ontario.

Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Ghebleh Goydaragh, M., Schmidt, K., and Scholten, T. 2021. Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. Geoderma, **399**: 115108. doi:10.1016/j.geoderma.2021.115108.

Tu, C., He, T., Lu, X., Luo, Y., and Smith, P. 2018. Extent to which pH and topographic factors control soil organic carbon level in dry farming cropland soils of the mountainous region of

Southwest China. Catena, **163**: 204–209. doi:10.1016/j.catena.2017.12.028.

van der Laan, M.J., and Dudoit, S. 2003. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples[online]. Technical Report 130, Division of Biostatistics, University of California, Berkeley. Available from: http://www.bepress.com/ucbbiostat/paper130/. [accessed December 2018].

van der Laan, M.J., Polley, E.C., and Hubbard, A.E. 2007. Super learner, UC Berkeley Division of Biostatistics Working Paper Series, 222.

Vapnik, V., Golowich, S.E., and Smola, A. 1997. Support vector method for function approximation, regression estimation, and signal processing. *In* Advances in neural information processing systems. *Edited by* M.C. Mozer, M. Jordan and T. Petsche Morgan Kaufmann Publishers, Cambridge. pp. 281–287.

Venables, W.N., and Ripley, B.D. 2002. Random and mixed effects. *In* Modern applied statistics with S. *Edited by* W.N. Venables and B.D. Ripley, Springer, New York. pp. 271–300. doi: 10.1007/978-0-387-21706-2_10.

Wilson, J.P., and Gallant, J.C. 2000. Secondary topographic attributes. *In* Terrain analysis - Principles and applications. *Edited by*Wilson and Gallant eds., Wiley, New York, New York, United States. 87–132.

Wolpert, D.H. 1992. Stacked generalization. Neural Netw. **5**: 241–259. doi:10.1016/S0893-6080(05)80023-1.

Young, G., Fenger, M.A., and Luttmerding, H.A. 1992. Soils of the Ashcroft map area. British Columbia Soil Survey. Integrated Management Branch Victoria, British Columbia[online]. Available from: https://sis.agr.gc.ca/cansis/publications/surveys/bc/bc26/index.html. [accessed September 2018].

Zevenbergen, L.W., and Thorne, C.R. 1987. Quantitative analysis of land surface topography. Earth surface processes and landforms **12**: 47–56. Wiley Online Library. doi:10.1002/esp.3290120107. [accessed 15 September 2014].

Zhang, Y.-Y., Wu, W., and Liu, H. 2019. Factors affecting variations of soil pH in different horizons in hilly regions. PLoS One, **14**: e0218563. doi:10.1371/journal.pone.0218563.