Predicting soil organic matter and soil moisture content from digital camera images: comparison of regression and machine learning approaches

Authors: Taneja, Perry, Vasava, Hiteshkumar Bhogilal, Fathololoumi, Solmaz, Daggupati, Prasad, and Biswas, Asim

Source: Canadian Journal of Soil Science, 102(3) : 767-784

Published By: Canadian Science Publishing

URL: https://doi.org/10.1139/cjss-2021-0133

Canadian Journal of Soil Science

Article

Predicting soil organic matter and soil moisture content from digital camera images: comparison of regression and machine learning approaches

Perry Taneja[a], **Hiteshkumar Bhogilal Vasava**[b], **Solmaz Fathololoumi**[b], **Prasad Daggupati**[a], **and Asim Biswas**[b]

[a]School of Engineering, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada; [b]School of Environmental Sciences, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada

Corresponding author: **Asim Biswas** (email: biswas@uoguelph.ca)

Abstract

Appropriate soil management maintains and improves the health of the entire ecosystem. Soil appropriate administration necessitates proper characterization of its properties including soil organic matter (SOM) and soil moisture content (SMC). Image-based soil characterization has shown strong potential in comparison with traditional methods. This study compared the performance of 22 different supervised regression and machine learning algorithms, including support vector machines (SVMs), Gaussian process regression (GPR) models, ensembles of trees, and artificial neural network (ANN), in predicting SOM and SMC from soil images taken with a digital camera in the laboratory setting. A total of 22 image parameters were extracted and used as predictor variables in the models in two steps. First models were developed using all 22 extracted features and then using a subset of six best features for both SOM and SMC. Saturation index (redness index) was the most important variable for SOM prediction, and contrast (median S) for SMC prediction, respectively. The color and textural parameters demonstrated a high correlation with both SOM and SMC. Results revealed a satisfactory agreement between the image parameters and the laboratory-measured SOM (R^2 and root mean square error (RMSE) of 0.74 and 9.80% using cubist) and SMC (R^2 and RMSE of 0.86 and 8.79% using random forest) for the validation data set using six predictor variables. Overall, GPR models and tree models (cubist, RF, and boosted trees) best captured and explained the nonlinear relationships between SOM, SMC, and image parameters for this study.

Key words: digital camera images, image color and texture features, cubist, random forest, soil characterization, computer vision

Résumé

Une bonne gestion du sol gardera l'écosystème en santé et en rehaussera la vitalité. Pour cela, il faut d'abord en caractériser correctement les propriétés, notamment la concentration de matière organique (CMO) et la teneur en eau (TE). La caractérisation du sol par l'image laisse entrevoir un potentiel supérieur à celui des méthodes classiques. Les auteurs ont comparé la performance de 22 algorithmes de régression et d'apprentissage automatique supervisé, y compris des machines à vecteurs de support (MVS), des modèles de régression gaussiens (MRG), des forêts d'arbres décisionnels et des réseaux neuronaux artificiels (RNA) pour prédire la CMO et la TE à partir de photos du sol prises avec un appareil numérique en laboratoire. À cette fin, ils ont extrait 22 paramètres des images et s'en sont servi comme variables explicatives en deux étapes, dans les modèles. Tout d'abord, ils ont élaboré les modèles à partir de ces paramètres, puis ils y ont appliqué un sous-ensemble constitué des six meilleurs à la prévision de la CMO et de la TE. L'indice de saturation (rougeur du sol) est la variable la plus utile pour prédire la CMO, le contraste (médiane de S) ayant la même utilité pour la TE. La couleur et la texture sont des paramètres étroitement corrélés à la CMO et à la TE. Les résultats révèlent une concordance satisfaisante entre les paramètres de l'image et la CMO (R^2 et écart-type de 0,74 et 9,80 % respectivement avec cubist) ainsi que la TE (R^2 et écart-type de 0,86 et 8,79 %, respectivement avec la forêt d'arbres décisionnels) établies en laboratoire pour valider le jeu de données fondé sur les six variables explicatives retenues. Dans l'ensemble, les MRG et les modèles à arbre décisionnel (cubist, forêt d'arbres décisionnels, arbres décisionnels amplifiés) saisissent et expliquent mieux les relations non linéaires entre la CMO, la TE et les paramètres de l'image examinés dans le cadre de cette étude. [Traduit par la Rédaction]

Mots-clés : photos numériques, couleur et texture de l'image, cubist, forêt d'arbres décisionnels, caractérisation du sol, vision-ique

Introduction

Soil organic matter (SOM), an indicator of soil health and quality (Zhang et al. 2006), is a significant component of any ecosystem (Li et al. 2013) and influences agricultural sustainability, food security, and climate (Were et al. 2015). Organic carbon (OC), as a key element of soil, plays an essential role in the global carbon cycle, so it is critical to measure its content in the soil (Kumar and Lal 2011; Yang et al. 2016). Soil moisture content (SMC), another significant component, not only influences the growth of crops, but also is a key factor in any crop management decisions including precision agriculture practices (Chukalla et al. 2015; Feki et al. 2018). Therefore, quantification of the spatial and temporal distribution and dynamics of SOM and SMC provide critical information to authorities concerned with the management and policymaking regarding soil and climate (Meersmans et al. 2008), food production (Taghizadeh-Mehrjardi et al. 2016), ecosystem modeling (Li et al. 2003), agriculture, forestry, land degradation management, environment protection, and most importantly land-use planning (Li et al. 2013). However, for detailed characterization, traditional measurement approaches are expensive, often involve use of hazardous chemical reagents, and time- and labor-intensive (Sudarsan et al. 2016; Lazzaretti et al. 2020). This often leads to delays in making decisions or resorting to outdated data, which ultimately forces users to make wrong decisions.

With the advancement of machine learning techniques and increasing access to digital image acquisition systems, digital image processing has emerged as an inexpensive technique to deal with these problems (Sudarsan et al. 2016; Fu et al. 2020; Swetha et al. 2020). With digital image processing approaches, soil properties, including but not limited to SOM, SMC, soil texture, iron, and fine particle contents, can be quantitatively estimated by formulating relationships between laboratory-measured soil properties and readily measurable soil image color and texture features (Levin et al. 2005; Rossel et al. 2008; Zhu et al. 2011; Sudarsan et al. 2016). Generally, color and (or) reflectance of soil can be attributed to numerous properties of soil such as SMC, SOM, parent material, mineralogy, and texture (Hummel et al. 2001; Fu et al. 2020; Gholizadeh et al. 2020; Taneja et al. 2021). This association justifies developing relationships between soil reflectance and its properties to predict their content using modeling.

In developing predictive relationships with image parameters, regression-based methods have been used in many fields including soil science (Persson 2005; dos Santos et al. 2016; Wu et al. 2017; Sakti et al. 2018). While it showed variable performance, various linear and nonlinear regression-based methods are still the commonly and popularly used methods (Rossel et al. 2008; dos Santos et al. 2016; Swetha et al. 2020). Recently, with advances in data processing and computing power, several data-driven modeling and machine learning approaches, including support vector machine (SVM), ensembles of trees (cubist, random forest, boosted trees, bagged trees), and Gaussian process regression (GPR), have been utilized with variable performance and reasonable success in developing predictive relationships in many fields (Gill et al. 2006; Matei et al. 2017; Chen et al. 2019; Kotlar et al. 2019). However, image-related approaches and the details of models' performance in predicting soil properties are limited and need further studies. In addition, due to the variable performances in the variable image set, it is difficult to understand and compare the performance of these algorithms compared to conventional regression-based methods in predicting soil properties. However, some researchers have compared the performance of two to three different algorithms in creating a predictive relationship (Gregory et al. 2006; Rossel et al. 2008; Wu et al. 2018). A complete comparison can only provide a good justification for the choice of method for predicting soil properties, especially SOM and SMC.

In addition to modeling, the collection of soil images is another important component that determines the success of image-based soil characterization. With a focus on targeted laboratory applications, most of the studies collected soil images under defined enclosures illuminated by controlled sources of light (Rossel et al. 2008; Zhu et al. 2011; Gómez-Robledo et al. 2013; Sakti et al. 2018; Wu et al. 2018; Fu et al. 2020). However, with a focus on developing computer vision or image analysis-based proximal soil sensors for various in situ applications, including precision agriculture, research carried out in controlled environment may not provide as much useful information as required. Collection of images in natural conditions would improve universality of the relationship developed.

Studies on the direct use of machine learning algorithms on image-derived color and texture data for SOM and SMC prediction have not yet appeared. Therefore, the overall objective of this research was to comprehensively compare the performance of various commonly used regression-based methods with machine learning based methods in predicting SOM and SMC from soil images collected with an inexpensive digital camera. The specific objectives of this study were to

1. assess the feasibility and usability of digital images to predict SOM and SMC in soil;
2. optimize the image parameters for developing predictive relationships;
3. and comprehensively compare the performance of a range of regression and machine learning algorithms (22 in total) in predicting SOM and SMC.

Proper assessment and comparison of various modeling algorithms and an optimum set of image parameters will serve as an informative guide on the use of digital images for predicting SOM and SMC.

Materials and methods

The overall methodology is divided into three sections: data collection, image analysis, and data analysis (Fig. 1).

Study site description and sample collection

Soil samples were collected in an earlier study by Ji et al. (2016) from two agricultural fields, Field 26 (~11 ha) and Field 86 (~17 ha) located on the MacDonald Campus research farm of McGill University, Sainte Anne De Bellevue, Quebec, Canada (Fig. 2). These two fields exhibited high spatial variability in terms of soil texture, organic matter, and soil type (Ji et al. 2016). The landscape of this area has undergone numerous processes during the last deglaciation including land-level rise, invasion of saline water, lake formation, retreat of ice, and deposition of glaciers, leading to the formation of highly variable soil. For example, soils of Field 26 ranges from mineral to organic deposits (peat) with high variability in soil textures including clay loam, loam, silt loam, sandy loam, and sand.

Field 86 mostly includes mineral soils with sandy clay loam, loam, sandy loam, clay, and clay loam texture (Fig. 3). Soil samples from the depth of 0–15 cm were collected from Field 26 and Field 86, respectively, in late April and early May in 2015 before seeding following a stratified random sampling strategy.

The fields were under no-tillage practices and corn–soybean rotation with soybean and corn being the preceding crops in Field 26 and Field 86, respectively. A total of 25 soil samples (17 from Field 26 and 8 from Field 86) exhibiting a wide variation in SOM (3.30%–62.70%), representing the range of SOM present in these fields, were carefully selected for this study (Fig. 4). These 25 samples represented both organic (mainly from Field 26) and mineral soils (present in both fields). This was done deliberately to include universality and increase robustness in training models. However, in laboratory terms, by adding different amounts of moisture to the samples, the number of samples used in modeling has increased significantly for processing (125 samples).

Laboratory analysis and soil imaging

The samples were air-dried, ground, and sieved through a 2 mm sieve. The processed samples were then used to capture images as well as measure soil properties in the laboratory. SOM was measured using loss on ignition (LOI) method. Ignition conditions were 550 °C for three hours (Schulte and Hopkins 1996). Processed soil samples were evenly placed in Petri dishes (~8 mm thickness) and the surface of the samples were captured with a 12.1-megapixel digital camera (Canon PowerShot SX270 HS) mounted on a tripod (27 cm) with the lens facing downward toward the sample. The camera was set to a 3000 × 4000 pixel resolution and the "best" jpeg compression, thereby supplying smaller sized images in contrast to uncompressed tiff files, but of comparable quality. A camera lens aperture setting of f/3.5 was regarded as appropriate for image acquisition under the normal lighting conditions of the laboratory, which was determined by repetitive tests conducted on distinct group references (Fu et al. 2020; Taneja et al. 2021).

A total of five sets of pictures were captured on the same soil samples. Before starting image capture, the weight of the empty Petri dishes and dishes with air-dried soil samples were recorded. The first set of images was collected on these soil samples (set 1). Then, water was added carefully (without disturbing the soil surface) and gradually over a period of time to simulate saturation-like conditions. The second set of images was collected corresponding to this condition (set 2). The saturated soil samples were then permitted to dry in open air under laboratory conditions. Two more sets of images were captured during the natural drying process of samples corresponding to two different SMCs. Finally, the soil samples were oven-dried at 105 °C for 24 hours to get 0% SMC and the images were captured of the driest soils. The weight of the soil samples (including the Petri dish) was recorded during each stage of image acquisition to calculate the gravimetric SMC based on the loss of weight during each drying event. Thus, a total of five sets of images corresponding to five different levels of SMC were collected. These sets were then grouped into five categories in increasing order of SMC with the images of oven-dried soil samples forming the Group 1, while those corresponding to the highest SMC and simulating saturation conditions formed the Group 5. Two images were captured for each soil sample in each SMC level (250 initial images). To reduce the uncertainty in modeling different soil parameters using imaging, the imaging was repeated. Then, the average of the two images is used (125 final images used in next steps). Figure 5 shows the SMC (%) of 25 soil samples at five different SMC levels.

In this study, the SMCs of samples in the same group were not kept constant. It was different from the SMC settings of other studies in which soil samples had the same SMC at the same wetting level and abrupt bi- or tri-modal soil moisture distributions were generated (Nocita et al. 2013; Rienzi et al. 2014; Rodionov et al. 2014). However, SMC in a field is likely to follow a normal or quasi-normal distribution. Moreover, soil samples with varying levels of SOM have different water-holding capacities and, thus, have varying drying characteristics. As an example, the sample with 3.3% SOM had a saturation SMC of 36.91% while that with 62.7% SOM had a saturation SMC of 119.60% (Fig. 5). Therefore, setting up a fixed SMC would have biased the image acquisition process. Consequently, the setting of varying soil moisture in this study (not controlling the SMC and allowing it to vary) had advantages to well simulate the continuous variation of soil moisture through space in the field.

In various studies that use digital images to obtain information about soil color, soil samples are confined to defined enclosures that are illuminated by a fixed light source (Rossel et al. 2008; Zhu et al. 2011; Gómez-Robledo et al. 2013; Sakti et al. 2018; Wu et al. 2018). But, in this study, there was not any such limitation laid down during imaging. This was done intentionally to simulate field conditions since variations in lighting conditions in actual field conditions are abrupt, variable, and uncontrolled. Moreover, it is necessary to avoid such restrictions in developing proximal soil sensors that must be used in field conditions and not in controlled laboratory environment.

Fig. 1. Overview of the framework used for this study. [Color online]

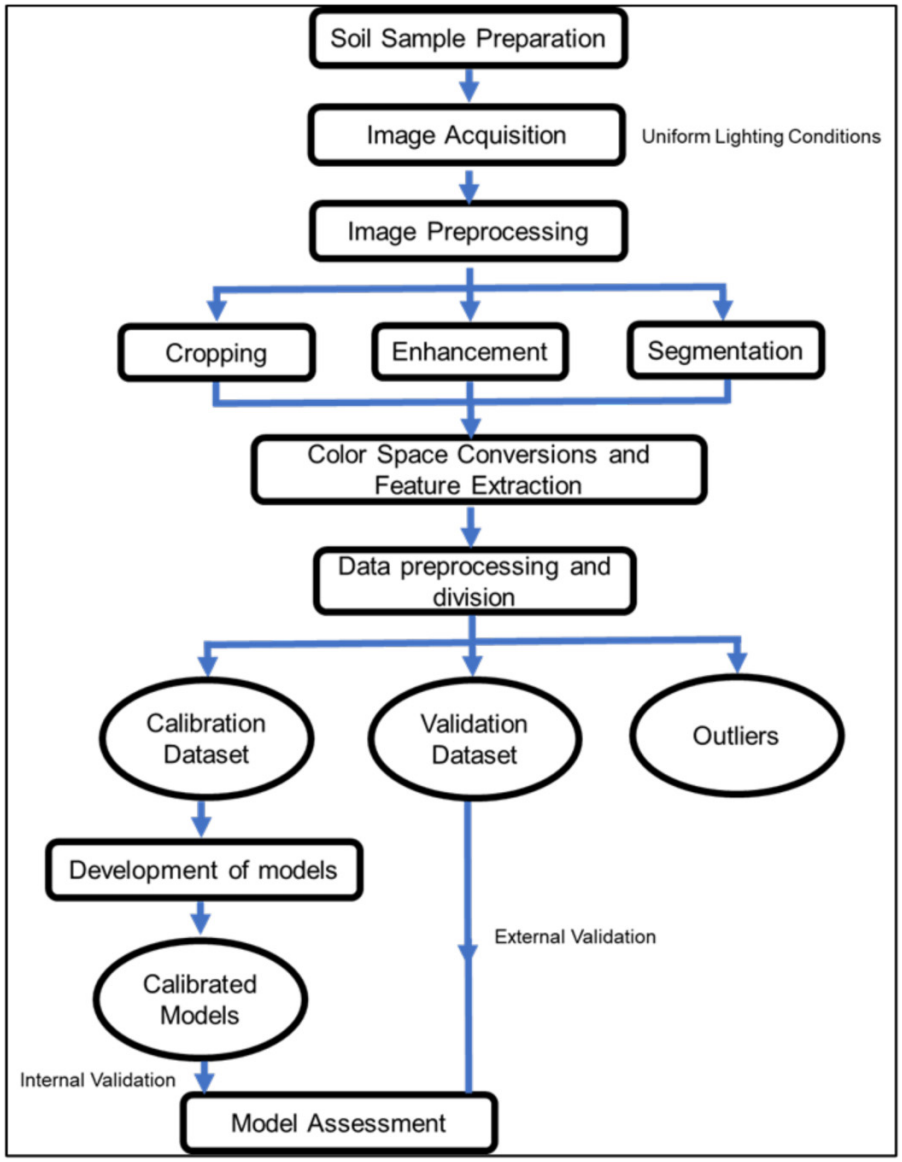


Image analysis

A proficient image acquisition system tries to capture quality images and appropriate image analysis approaches help to derive useful information from the images and make a substantial contribution to the computer vision applications. Similar to other disciplines, it is necessary to exercise prudence when images captured using digital cameras must be processed. Numerous elements, for instance, reflection from water (in case of high SMC), nonuniform lighting conditions, and foreign particles (plant litter, residues of roots, and white colored powder such as that of lime or fertilizer) present on the surface of the soil, affect the quality of the image (Gonzalez et al. 2004). Thus, corrections must be made before useful information can be extracted from the images.

Image preprocessing-cropping

Images were cropped to a square area of 950 × 950 pixels roughly from the geometric center of the image. This was done to remove the white background as well as to reduce the effects caused by the edges of the Petri dishes (Fig. 6).

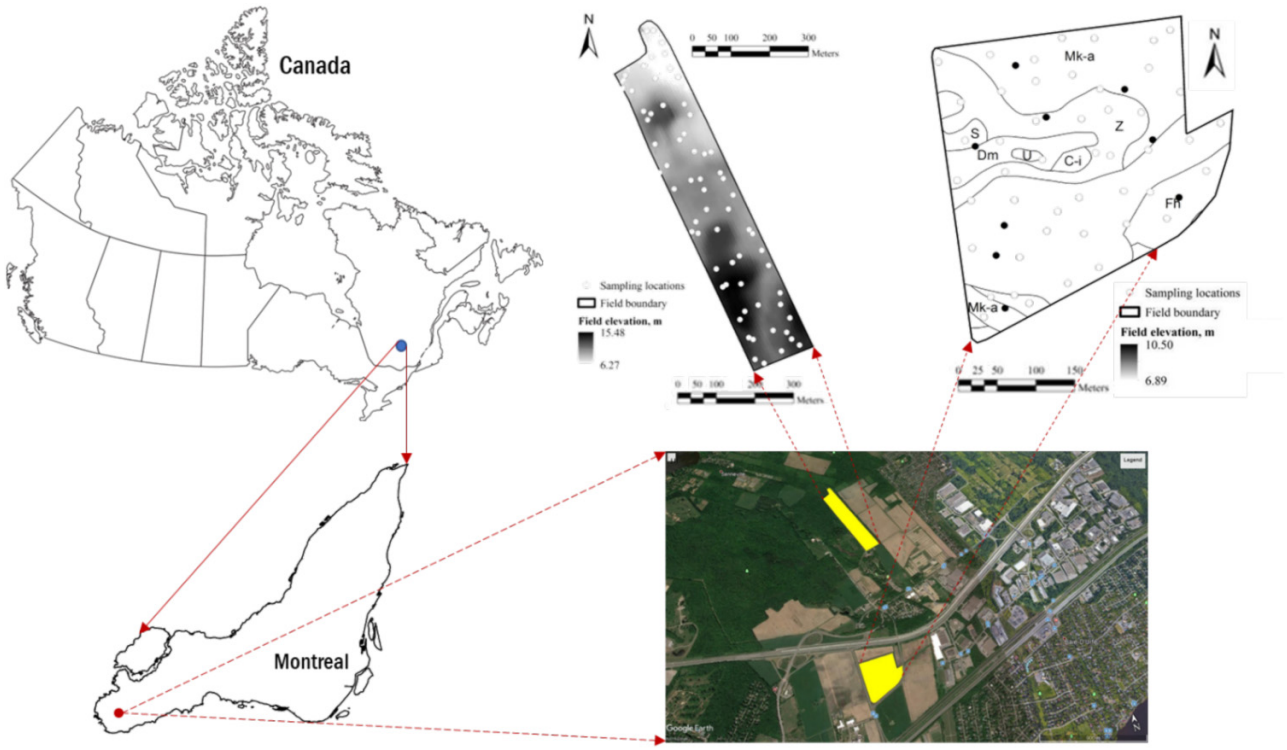
Image preprocessing enhancement

To enhance the images, contrast adjustment was performed using "imadjust" function of MATLAB (MathWorks 2017). This assisted in segmentation (next step) through exclusion of noise and prevention of useful information from fading into noise (Fig. 6).

Image segmentation

For this study, image segmentation denoted identification and retention of the pixels that represented soil in the images. For instance, certain parts of some images were visibly occupied by residues of small leaves and black cracks (only detected after careful examination) or a film of water which gave rise to exceptionally bright reflections (Fig. 6). Irrespective of the area occupied by the gloss or the foreign particles,

Fig. 2. Geographic location of the study area, Field 86 (left) and Field 26 (right) of Macdonald Campus Farm, McGill University, Quebec, Canada, as well as field elevation maps for Field 26 and Field 86 along with the soil map. The letters in the map represent various soil series. The base map is downloaded from Google Earth and processed in ArcGIS, the projection used in NAD84 with UTM zone 18. [Color online]



it was considered essential to eliminate them to avoid inaccurate calculations. Because the image intensity values corresponding to these areas do not depict the image intensity values of the actual soil surface, the mean value would not denote the mean of the soils' pixels.

Therefore, an experiential-based segmentation technique was developed based on the image histogram to distinguish the pixels covered by soil from nonsoil particles. To segment the image, noticeable dissimilarities in the intensity values of the pixels of soil and nonsoil were used. Because the nonsoil pixels occupied a small portion in contrast with the whole image, a value was ascertained after several trials. This assumption was made with the conviction that image intensity values whose counts were lower than or equivalent to the defined value were regarded as those belonging to nonsoil matter and subsequently discarded. A value of 3000 was chosen for this study; the value may differ for soil from different regions and parent materials (Taneja et al. 2021). In addition, the pixels analogous to the water film were white-colored. In such cases, the histogram was examined to obtain the "highest count" of the image intensity values falling in the range 248–255 Gy scale values (illustrating a range of values demonstrating the color white). In this situation, different threshold values were set rather than previous for nonsaturated soils. For example, the images were converted to grayscale and the pixels with gray-scale values between 248 and 255 were examined for the "highest count." The "highest count" was then compared with the threshold set for nonsaturated soils (i.e., 3000) and the greater value was set as the threshold. For example, for a saturated soil sample, the "highest count" of 4518 was recorded on gray-scale value of 251. Then, the 4518 value was compared with the previously optimized value of 3000 and the higher value, which means 4518, was set as the threshold.

Color space conversions and feature extraction

The RGB images (the color of pixel was made up of red, green, and blue components (Kumar and Verma 2010)) were converted to HSV (the colors were represented by hue (tones), saturation (purity), and value (brightness)) and grayscale images using color space conversions (Fig. 6); then, color features such as redness index (RI), coloration index (CI), hue index (HI), and saturation index (SI), as well as texture features including entropy, contrast, energy, and homogeneity, were the sum of squared elements in the gray-level co-occurrence matrix (GLCM). "Homogeneity" was extracted and indices were derived using MATLAB. The list of extracted features and derived indices is presented in Fig. 7.

A total of 22 image parameters were extracted. "Mean" represented average of values of all the pixels in an image. "Median" represented the middle pixel value after all the pixels were sorted in numerical order. "Entropy" was the statistical measure of randomness. "Contrast" was the measure of intensity contrast between a pixel and its neighbor over the whole image. "Energy" was the sum of squared elements in the GLCM. "Homogeneity" was the closeness of distribution

Fig. 3. Soil texture classification (following Canadian System of Soil Classification) of soil samples collected from Field 26 and Field 86. The 25 samples selected for this study are represented by red colored signs while blue color signs represent the remaining 95 samples (out of the total 120 samples). The triangle was prepared using "soiltexture" package in R. [Color online]

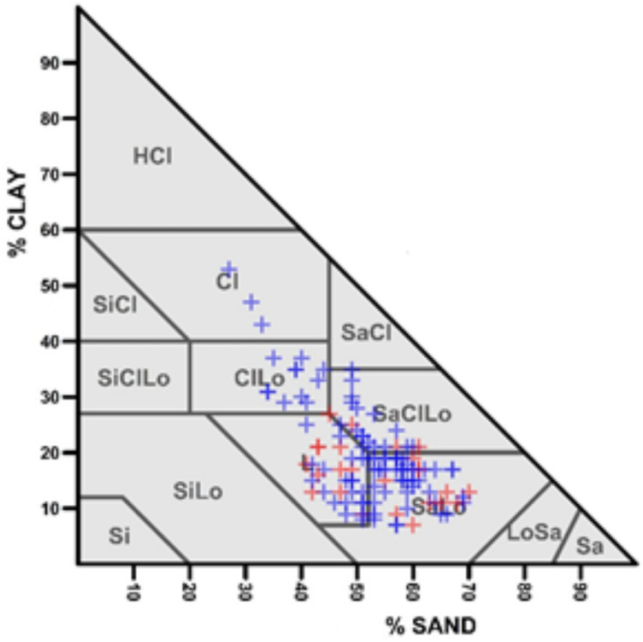
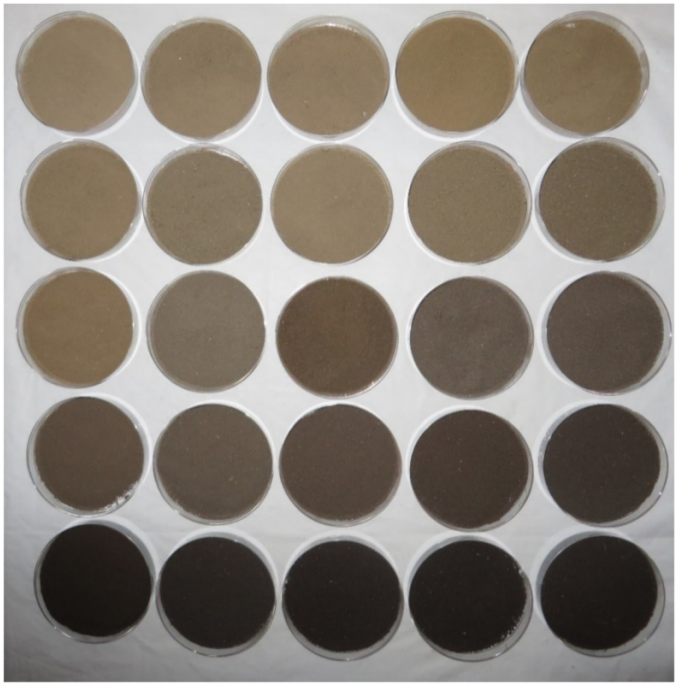


Fig. 4. The 25 soil samples selected for this study collected from Field 26 and Field 86.



of elements in the GLCM to the GLCM. RI, CI, HI, and SI were calculated as:

$$\text{RI} = \frac{R^2}{B \times G^3} \quad (1)$$

$$\text{CI} = \frac{R - G}{R + G} \quad (2)$$

$$\text{HI} = \frac{2 \times R - G - B}{G - B} \quad (3)$$

$$\text{SI} = \frac{R - B}{R + B} \quad (4)$$

Both mean and median values were used as predictors in the modeling due to inconsistent information in the literature. For example, while some researchers used mean values (Rossel et al. 2008; Sudarsan et al. 2016), others employed median values (Persson 2005; Rossel et al. 2008) in their research. In fact, Persson (2005) advocates the application of median values to handle the deviations resulting from the shading of the microrelief developed on the surfaces of the samples of soil. Depending on the viewing angle with respect to the direction of the incident light, there might exist bidirectional reflectance distribution function (BRDF) and shading influences (King 1995; Lillesand et al. 2015). The indices derived from the images (RI, CI, HI, SI) were expected to reduce these influences. Being the ratio indices, a view effectively balances the abnormalities in brightness arising from the disparities in the topography and emphasizes the color content of the samples.

Data preprocessing and division

Multivariate outliers were determined based on the Mahalanobis distance (De Maesschalck et al. 2000). Regression approaches were employed to decide if a specific sample from a sample population was an outlier through the combination of $\geq$2 variable scores. Following this, data obtained from five images were detected as outliers and were not included in further calculations. The data were split to calibration and validation sets randomly; 70% of the data were used as calibration data (84 images) and 30% of the data were used as validation data (36 images). Statistical distribution of calibration and validation samples was normal. These data also include a wide range of SOM and SMC values. All the necessary standards are considered in the selection and classification of calibration and validation data sets (Table 1).

Model development

The image color and texture-related features were used to develop predictive relationships against laboratory-measured SOM and SMC. Under six broad types: (1) linear regression (LR), (2) regression trees, (3) SVMs, (4) GPR, (5) ensemble of trees, and (6) artificial neural network (ANN), a total 22 models were developed. Codes were written in MATLAB to run

Fig. 5. Soil moisture content (SMC; %) for the 25 soil samples corresponding to five different levels of moisture represented as five groups. [Color online]

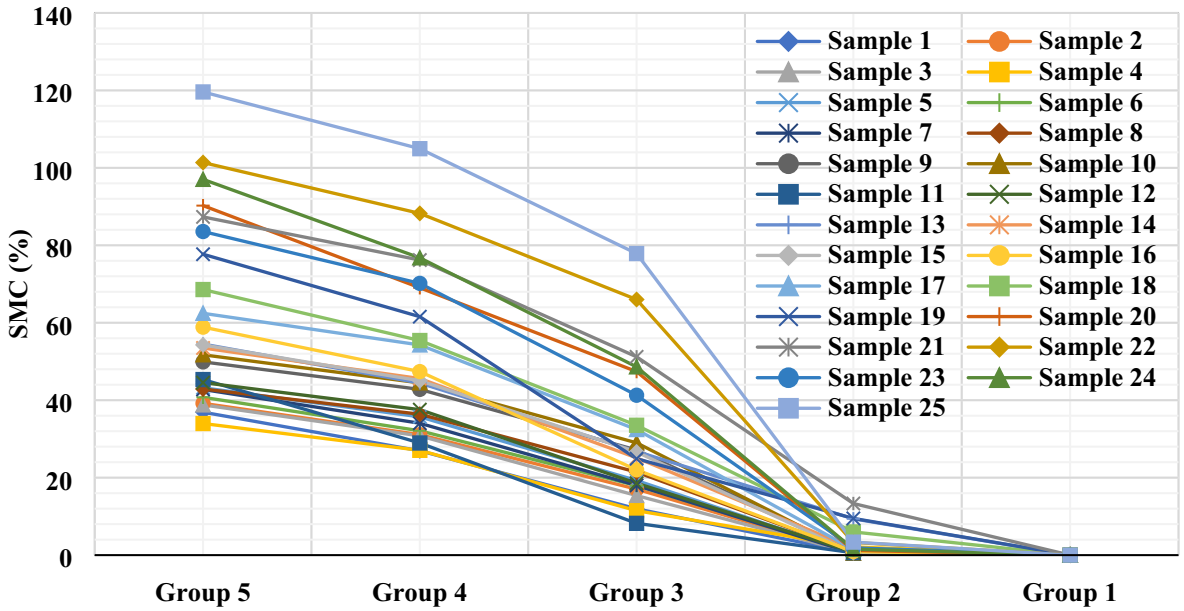
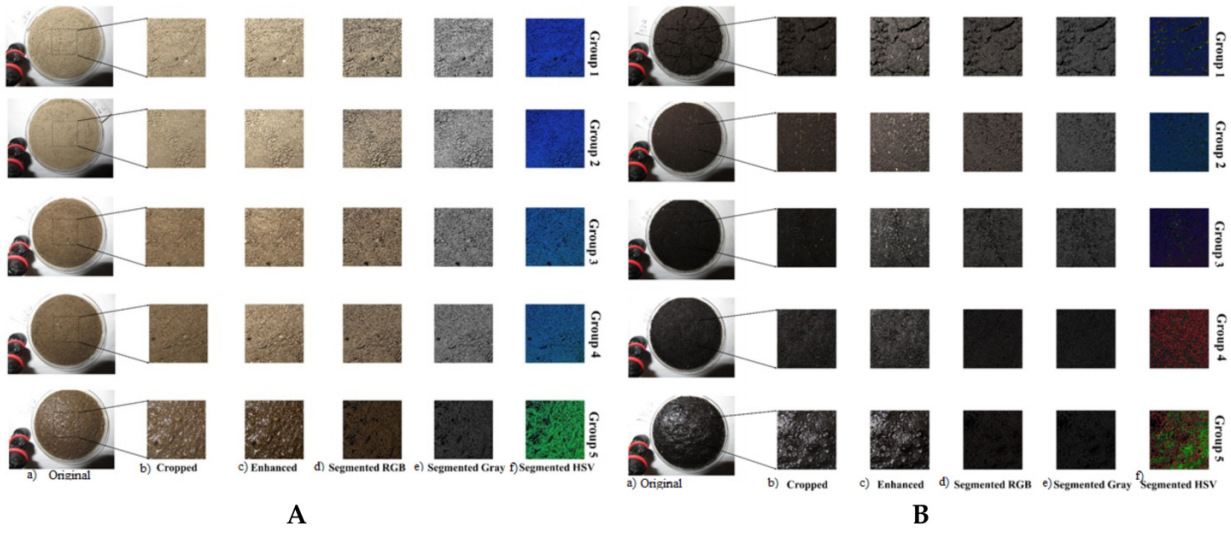


Fig. 6. Images of soil sample with (A) 3.3% and (B) 62.7% soil organic matter (SOM) under five different soil moisture condition, while the columns represent (*a*) original images, (*b*) corresponding cropped regions, (*c*) enhanced images, (*d*) segmented images, (*e*) color space converted gray images, and (*f*) color space-converted hue saturation value (HSV) images, respectively. [Color online]



these models on data sets except for cubist model, which was executed in R program (version 3.5.3) on RStudio (Team 2015).

Model performance assessment

Several statistical parameters were computed to assess the accuracy of the models.

1. Coefficient of determination (R^2): it represents the percentage of total variation in dependent variable. Its value can vary from 0 to 1. Large values imply higher prediction accuracies (eq. 5).
2. Root mean square error (RMSE): it represents the mean absolute error between the measured and observed values. Lower values are desirable (eq. 6).
3. Lin's concordance correlation coefficient (LCCC): the LCCC was employed for model quality evaluation since it represents the fit of 1:1 line of the predicted and observed values. Also, because of it being unitless in nature, it is advantageous to compare different models of the same soil property and (or) comparison of models for different soil properties (Sorenson et al. 2017). Large values represent higher prediction accuracy (eq. 7).

Fig. 7. Overview of extracted features and indices derived from the images. RGB, red, green, and blue; HSV, hue, saturation, and value. [Color online]

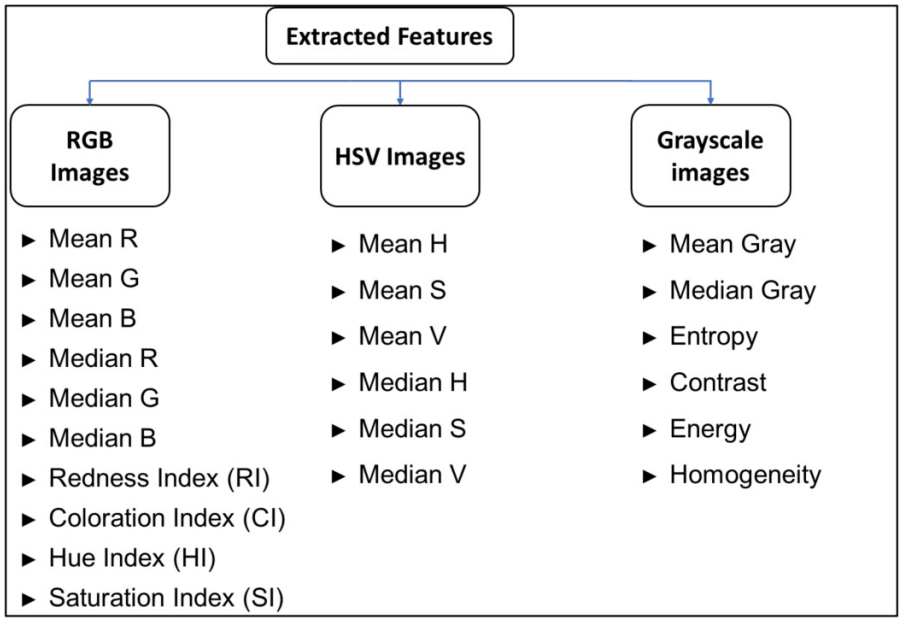


Table 1. Descriptive statistics of the whole, calibration, and validation data set for soil organic matter (SOM) and soil moisture content (SMC).

		SOM (%)				SMC (%)			
Scope	Count	Min	Max	Mean	SD	Min	Max	Mean	SD
All	125	3.3	62.7	19.0	17.6	0	119.6	29.2	29.2
Without outliers*	120	3.3	62.7	18.4	17.2	0	119.6	25.2	27.5
Calibration	84	3.3	62.7	17.9	16.6	0	119.6	29.2	0.2
Validation	36	3.3	62.7	19.7	18.6	0	87.3	15.9	22.9

Note: SD, standard deviation. An asterisk (*) indicates that the number of outliers is five.

4. Mean of the residuals (Bias): it is used to analyze the underfitting or overfitting of the model predictions. Value of bias = 0 implies unbiased predictions (eq. 8).
5. Ratio of performance to deviation (RPD): it is the ratio of standard deviation of observed or measured values to the standard error of prediction (Chang et al. 2001). The RPD values >2 are often considered to represent good model performance.
6. Ratio of performance to interquartile distance (RPIQ): it is the ratio of interquartile range of the observed values to the RMSE of prediction. The RPIQ takes into consideration both the variation of measured values and the prediction error, thereby being an indicator of model quality, which is more objective than the RMSE of prediction and, thus, it can be easily used for the comparison of different models. The greater the value of RPIQ, higher is the model's capacity to predict.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_{\text{observed}} - Y_{\text{predicted}}\right)^2}{\sum_{i=N}^{N}\left(\overline{Y}_{\text{observed}} - Y_{\text{predicted}}\right)^2} \quad (5)$$

$$\text{RMSE} = \sqrt{\left(\frac{\sum_{i=1}^{N}(\text{observed}_i - \text{predicted}_i)^2}{N}\right)^2} \quad (6)$$

$$\text{LCCC} = \frac{2\rho\sigma_{\text{predicted}}\sigma_{\text{observed}}}{\sigma^2_{\text{predicted}} + \sigma^2_{\text{observed}} + \left(\mu_{\text{predicted}} - \mu_{\text{observed}}\right)^2} \quad (7)$$

$$\text{Bias} = \frac{\sum_{i=1}^{N}\text{predicted}_i - \text{observed}_i}{N} \quad (8)$$

where N was the number of samples, $Y_{\text{predicted}}$ was the predicted values, Y_{observed} was the observed values and $\overline{Y}_{\text{observed}}$ was the mean of observed values.

All these statistics were tested on both the calibration and validation data sets. At first, all the 22 extracted features (color and texture characteristics) were treated as predictor variables and were used to develop the models for SOM (%) and SMC (%) prediction. Later, a subset of six optimum predic-

tors (optimization described in the next section) were used for model development. A 10-fold cross-validation was performed on the calibration data set as a means of internal validation (IV). Models were also externally validated using an independent validation data set. The residuals (difference between observed and predicted) were also tested for the presence of normality and the absence of autocorrelation and were found satisfactory for regression relations development.

Variable screening to identify optimum predictors

To study the relative importance of predictor variables in predicting SOM and SMC, a *z*-score was defined following six different analysis: analysis of variance (ANOVA), random forest (RF), cubist, principal component analysis (PCA), Vtreat variable reduction, and correlation analysis. Under each analysis, all the image parameters (predictor variables) were rated on a scale of 0–100 (most important) and then averaged to get a *z*-score. While some analysis techniques, such as cubist and RF, by default, provided variable importance on a scale of 0–100. Correlation analysis was simply 1:1 correlation between dependent variable and each predictor. The absolute values of the correlation coefficients were first calculated and were scaled at 0–100, with the lowest and highest absolute correlation coefficients being assigned a value of 0 and 100, respectively. For ANOVA, the *p* value for each predictor variable was scaled to 0–100 with 0 and 100 being assigned to the lowest and highest *p* value, respectively. "Vtreat" is an R package for looking at the variable importance/significance. The values of R^2 were scaled to 0–100, with the lowest and highest value being assigned 0 and 100, respectively. These 0–100 scaled values were then added and averaged to get the final scaled values at the range of 0 and 100 and was named *z*-score for that predictor. The top six predictor variables were then identified as the optimum predictors for both SOM and SMC. All the models were then developed using these six predictors as independent variables and the model performance statistics were recalculated.

Results

Descriptive statistics of soil properties

Table 2 presents the descriptive statistics for SOM, SMC, various image color, and texture features and derived indices. The soil properties and thus image parameters showed a high degree of variation with the coefficient of variation, CV (%) varying between 10.14 and 168.20. The SOM content varied between 3.30 (%) and 62.70 (%) with an average of 18.44 (%) and a standard deviation of 17.23 (%). The SMC also varied between 0.00 (%) and 119.60 (%) with a mean of 25.16 (%) and a standard deviation of 27.48 (%). With an acceptable approximation, all image parameters except mean H, energy, and RI were normally distributed (kurtosis approximately between −3 and 3). The RI had a very large CV of about 168.20%. On the other hand, homogeneity comparatively varied less significantly, with a CV of around 10.14%. The high variability of SOM and SMC presented an opportunity to test the prediction capability of the developed models.

Table 2. Descriptive statistics of soil organic matter (SOM), soil moisture content (SMC), and soil color measurements.

Parameter	Mean	SD	Range	Kurtosis	CV (%)
Mean R	0.31	0.15	0.06–0.59	−1.24	47.90
Mean G	0.28	0.14	0.05–0.53	−1.21	49.06
Mean B	0.24	0.11	0.04–0.44	−1.19	47.86
Mean H	0.08	0.03	0.04–0.22	6.75	38.35
Mean S	0.20	0.08	0.05–0.48	−0.27	41.88
Mean V	0.31	0.15	0.06–0.59	−1.24	47.82
Mean gray	0.28	0.14	0.05–0.54	−1.22	48.43
Median R	0.37	0.20	0.00–0.74	−1.26	53.56
Median G	0.32	0.18	0.00–0.67	−1.23	55.60
Median B	0.27	0.15	0.00–0.54	−1.20	55.57
Median H	0.08	0.02	0.00–0.11	2.24	22.87
Median S	0.22	0.10	0.00–0.51	−0.13	45.94
Median V	0.37	0.20	0.00–0.74	−1.26	53.56
Median gray	0.33	0.18	0.00–0.67	−1.24	54.75
Entropy	5.47	0.60	2.85–6.40	1.32	11.00
Contrast	2.14	1.64	0.09–6.48	−0.84	76.98
Energy	0.17	0.15	0.06–0.80	3.64	86.31
Homogeneity	0.76	0.08	0.66–0.96	−0.43	10.14
Redness index	73.26	123.23	5.30–734.50	8.44	168.20
Coloration index	0.07	0.07	0.01–0.19	1.14	49.98
Hue index	3.38	3.38	2.11–7.47	1.77	30.25
Saturation index	0.14	0.14	0.03–0.44	1.56	52.76
SOM (%)	18.44	17.23	3.30–62.70	0.36	93.45
SMC (%)	25.16	27.48	0.00–119.60	0.46	109.19

Note: SD, standard deviation; CV, coefficient of variation.

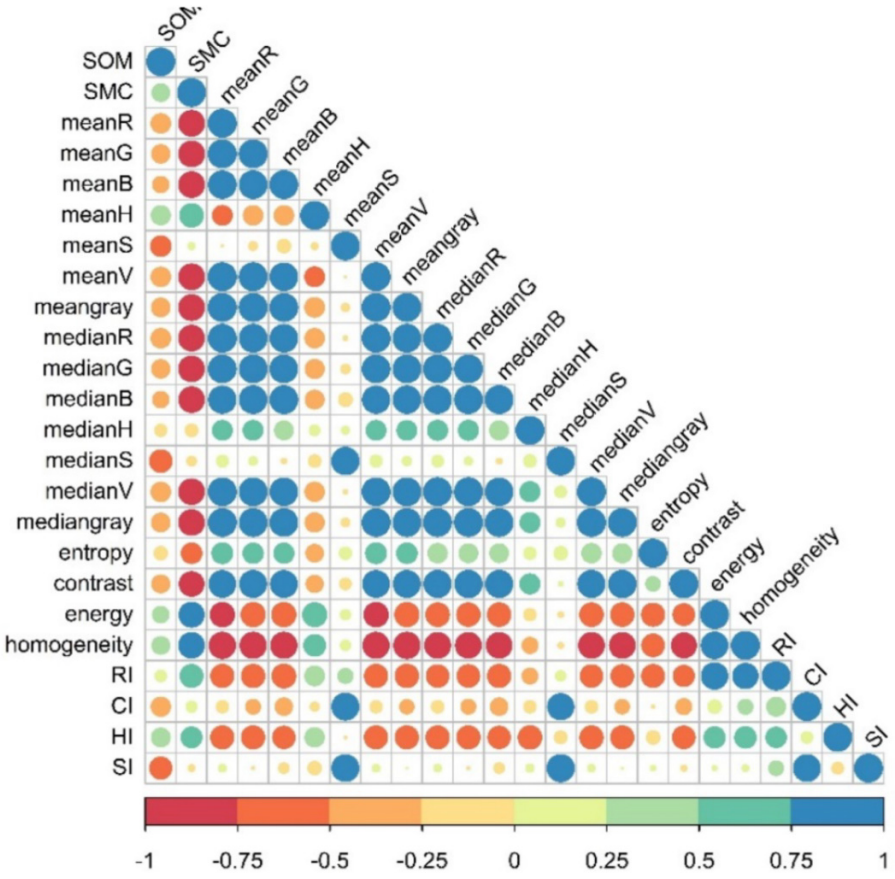
Linear correlation between SOM, SMC, and soil color

Several soil color parameters showed high correlations with SMC, although comparatively weaker correlations were observed with SOM (Fig. 8). Soil moisture content showed a high negative correlation with mean gray with a correlation coefficient of –0.85. In addition, SMC was also negatively correlated with mean B (–0.84), mean G (–0.84), and mean R (–0.84). SOM content was negatively correlated with median S values (–0.65) followed by SI (–0.62) and mean S (–0.54). SMC was weakly correlated with SI (–0.06) while SOM was weakly correlated with RI (0.16). In general, the reflection intensity decreased with the increase in organic matter and moisture content. Significant correlation was also observed among color and texture parameters to some extent.

Identification of optimum predictors

To underline which explanatory variables were mainly important for the prediction of SOM and SMC, radial plots were studied following six different analysiss (Figs. 9–12). Color fea-

Fig. 8. Correlation plot for soil organic matter (SOM), soil moisture content (SMC), color space model parameters, and indices derived from them. RI, redness index; CI, coloration index; HI, hue index; SI, saturation index. [Color online]



tures were more important than textural features in SOM prediction. Also, the impact of mean values in SOM prediction accuracy was greater than that of median values. Whereas, for SMC, the impact of median values and textural features in SOM prediction accuracy were greater than that of mean values and color features.

Saturation index was the most important variable for SOM prediction followed by mean H, median R, mean R, mean V, and median S. The least important variable was RI (Fig. 10).

For SMC, contrast was the most important predictor variable followed by median B, median R, mean B, homogeneity, and energy. The least important variable was median S (Fig. 12). Several soil color parameters showed high correlations with SMC, although comparatively weaker correlations were observed with SOM (Fig. 8). Soil moisture content showed a high negative correlation with mean gray with a correlation coefficient of –0.85. In addition, SMC was also highly negatively correlated with mean B (–0.84), mean G (–0.84), and mean R (–0.84). SOM content was negatively correlated with median S values (–0.65) followed by SI (–0.62) and mean S (–0.54). Soil moisture content was weakly correlated to SI (–0.06) while SOM was weakly correlated to RI (0.16). In general, the reflection intensity decreased with the increase in organic matter and moisture content. Significant correlation was also observed among color and texture parameters to some extent.

Predictive accuracy of the models

Saturation index models developed with 22 and 6 image color- and texture-related features were calibrated and validated against laboratory-measured SOM and SMC. Descriptive regression statistics of the predicted vs. laboratory measured values of soil properties are presented in Tables 3 and 5 for SOM using 22 and 6 predictor variables, respectively, and Tables 4 and 6 for SMC using 22 and 6 predictor variables, respectively.

Prediction of SOM using 22 predictor variables

10-fold cross (internal) validation

In general, the GPR-based models yield the higher predictive accuracy, while the LR-based models were least accurate for the SOM prediction. From the results (Table 3), it was evident that the most accurate predictions were obtained using ANN. The R^2, RMSE, LCCC, bias, RPD, and RPIQ values were 0.86, 6.32%, 0.91, –0.13, 2.63, and 3.18, respectively. Also, the second accurate predictions were obtained using exponential GPR model with R^2, RMSE, LCCC, bias, RPD, and RPIQ values of 0.79, 7.60%, 0.88, –0.08, 2.19, and 2.65, respectively. The least accurate predictions were obtained using interactions linear model with R^2, RMSE, LCCC, bias, RPD, and RPIQ values of 0.01, 256.99%, 0.01, 6.78, 0.06, and 0.08, respectively (Table 3).

Fig. 9. Relative significance of each individual image parameter as a predictor variable for soil organic matter prediction corresponding to (*a*) analysis of variance (ANOVA), (*b*) random forest (RF), (*c*) principal component analyses (PCA), (*d*) cubist, (*e*) Vtreat, and (*f*) correlation. [Color online]

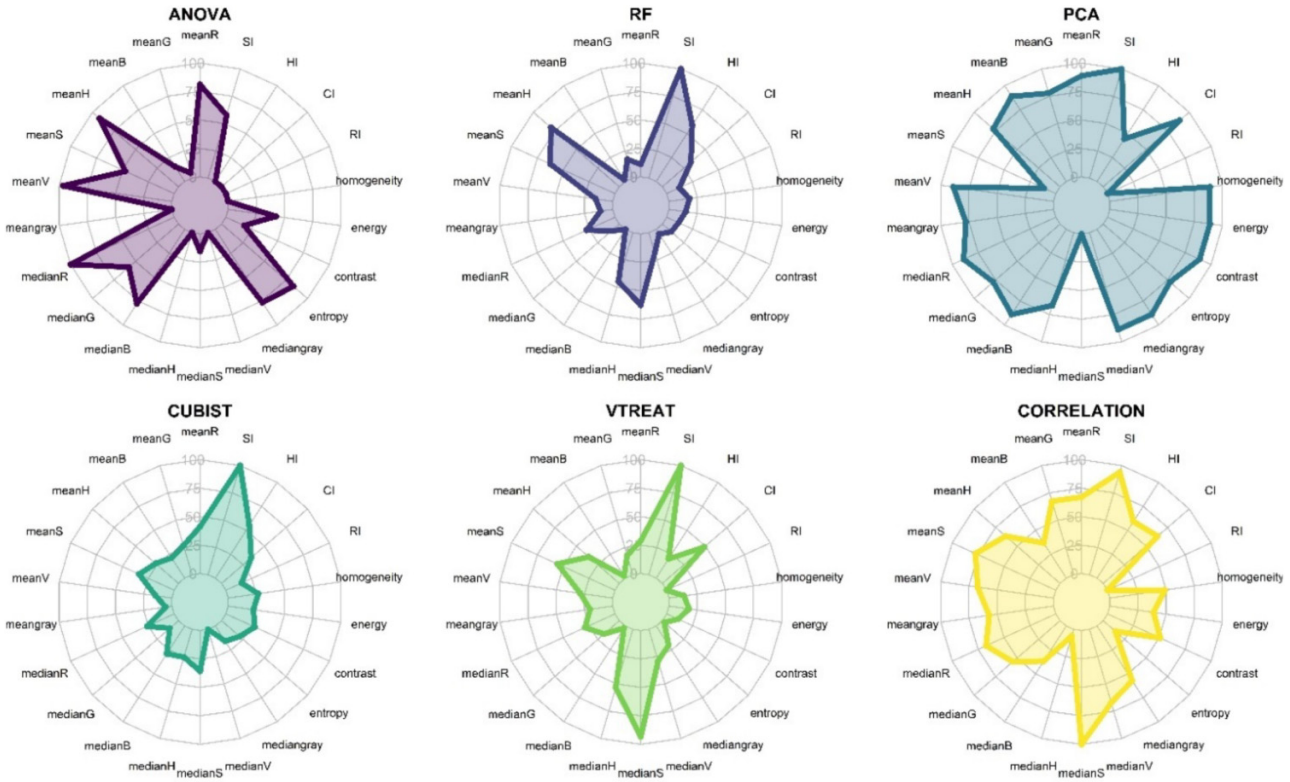
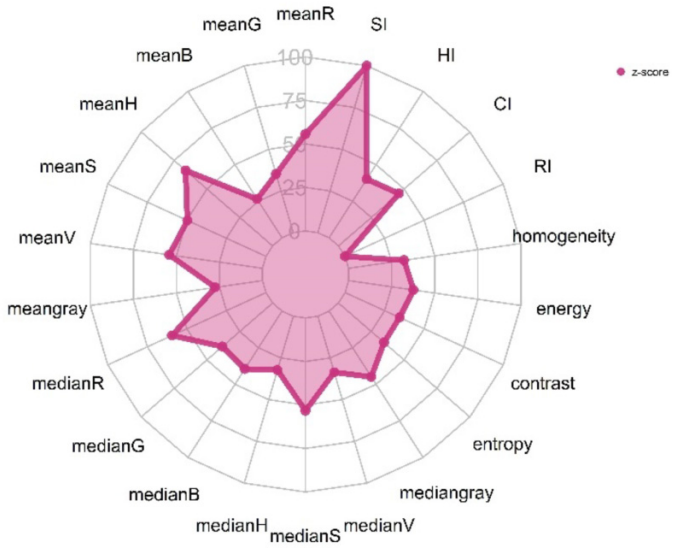


Fig. 10. *z*-score of each individual image parameter representing its contribution toward soil organic matter prediction. RI, redness index; CI, coloration index; HI, hue index; SI, saturation index. [Color online]



External validation

The GPR-based models yield the highest accuracy with an average R^2 higher than 0.70, followed by SVM and regression tree-based models. The R^2 value for the model trained using squared exponential GPR producing best predictions was 0.77, the RMSE was 8.87%, the LCCC was 0.85, the bias was –0.68, the RPD was 2.09, and the RPIQ was 2.46 (Table 3). The performance of ANN for the test data set was comparable but relatively weaker with R^2 of 0.74 and RMSE of 9.88%. The LCCC was 0.80, the bias was –1.31, the RPD was 1.88, and the RPIQ was 2.21. On the other hand, the poorest predictions were produced by interactions linear model giving an R^2, RMSE, LCCC, bias, RPD, and RPIQ values of 0.14, 69.64%, 0.17, 4.19, 0.27, and 0.31, respectively (Table 3).

Prediction of SMC using 22 predictor variables

10-fold cross (internal) validation

The SMC was predicted with higher accuracy than SOM. Except the interaction LR approach, all other modeling approaches predicted SMC with high accuracy. GPR approaches, however, outperformed other models with consistent higher prediction. The exponential GPR model produced the best predictive relationship between SMC and soil color and texture features with R^2 = 0.89, RMSE = 9.40%, LCCC = 0.93, bias= –0.10, RPD = 3.02, and RPIQ = 4.78 (Table 4). The inter-

Fig. 11. Relative significance of each individual image parameter as a predictor variable for soil moisture content prediction corresponding to (*a*) analysis of variance (ANOVA), (*b*) random forest (RF), (*c*) principal component analyses (PCA), (*d*) cubist, (*e*) Vtreat, and (*f*) correlation. [Color online]

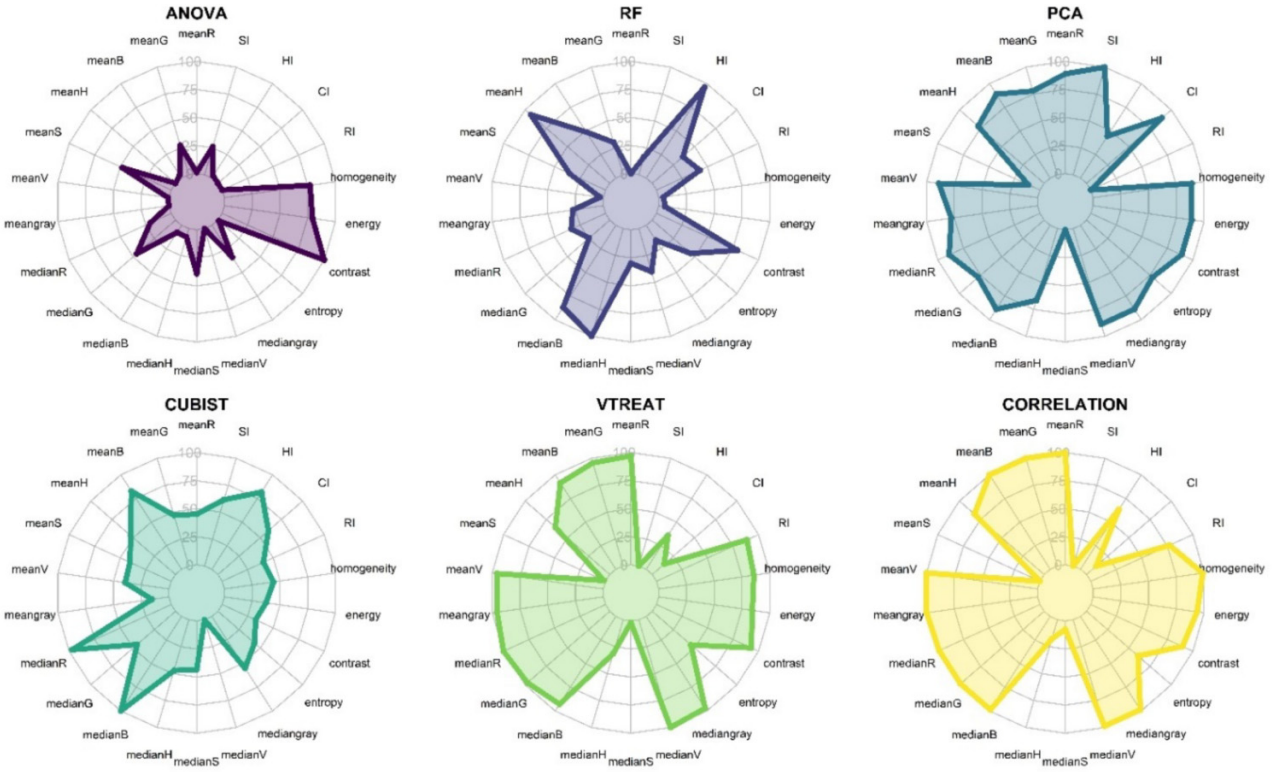
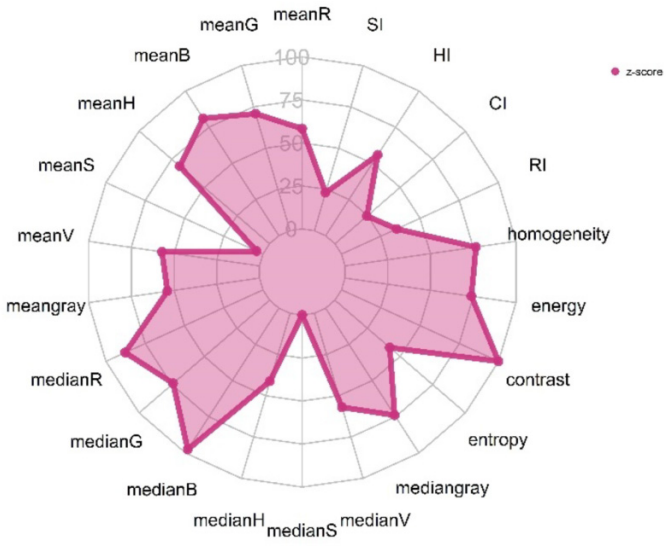


Fig. 12. *z*-score of each individual image parameter representing its contribution toward soil moisture content prediction. RI, redness index; CI, coloration index; HI, hue index; SI, saturation index. [Color online]



actions linear model exhibited poor predictive performance with an R^2 of 0.00 and an RMSE of 308.97%, while the LCCC, bias, RPD, and RPIQ were –0.01, –47.91, 0.09, and 0.15, respectively (Table 4).

External validation

Excellent prediction was observed using all the models with the R^2 > 0.80 and RPD values >2 except for ANN, interactions linear, and PLSR. The R^2 value for the model trained using exponential GPR, which produced best predictions, was 0.95, the RMSE was 5.21%, the LCCC was 0.96, the bias was 1.12, the RPD was 4.39, and the RPIQ was 4.82 (Table 4). The worst predictions were produced by interactions linear model with R^2, RMSE, LCCC, bias, RPD, and RPIQ of 0.01, 75.87%, 0.06, −5.65, 0.30, and 0.33, respectively (Table 4).

Prediction of SOM using six predictor variables

10-fold cross (internal) validation

ANN produced the most accurate predictions with R^2, RMSE, LCCC, bias, RPD, and RPIQ of 0.74, 8.51%, 0.84, –0.26, 1.96, and 2.36, respectively (Table 5). Overall, the ensemble tree and GPR modeling approaches predicted SOM with higher accuracy (R^2 > 0.65). The LR, regression trees and SVM yield inconsistent prediction accuracy. The model calibrated using cubic SVM produced the worst predictions with R^2, RMSE, LCCC, bias, RPD, and RPIQ of 0.47, 13.52%, 0.68, –1.15, 1.23, and 1.49, respectively (Table 5).

Table 3. Accuracy of different models for the prediction of soil organic matter in the calibration and validation data sets using 22 predictor variables.

	R^2		RMSE		Concordance		Bias		RPD		RPIQ	
Models	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV
Random forest	0.68	0.74	9.40	9.81	0.79	0.80	−0.52	−1.42	1.77	1.89	2.14	2.22
Cubist	0.77	0.74	8.02	9.49	0.87	0.83	0.07	−1.08	2.07	1.96	2.50	2.30
Aritificial neural network	**0.86**	0.74	**6.32**	9.88	0.91	0.80	−0.13	−1.31	2.63	1.88	3.18	2.21
Linear regression	0.55	0.64	11.42	11.29	0.73	0.74	0.19	−1.12	1.46	1.65	1.76	1.93
Interactions linear	0.01	0.14	256.99	69.64	0.01	0.17	6.78	4.19	0.06	0.27	0.08	0.31
Robust linear	0.31	0.52	15.26	13.49	0.55	0.62	−1.36	−4.20	1.09	1.38	1.32	1.62
Linear regression-pure quadratic	0.48	0.65	13.56	11.04	0.69	0.75	0.93	−1.24	1.23	1.68	1.48	1.97
Stepwise linear	0.68	0.73	10.09	9.75	0.82	0.82	1.02	−1.19	1.65	1.90	1.99	2.24
Fine tree	0.60	0.69	10.87	10.40	0.76	0.80	−0.35	−1.68	1.53	1.79	1.85	2.10
Medium tree	0.58	0.65	10.93	10.94	0.75	0.79	−0.24	−0.16	1.52	1.70	1.84	1.99
Coarse tree	0.43	0.54	12.55	12.63	0.62	0.68	−0.07	−1.89	1.33	1.47	1.60	1.73
Linear SVM	0.59	0.58	10.84	13.08	0.72	0.62	−2.17	−4.11	1.54	1.42	1.85	1.67
Quadratic SVM	0.70	0.75	9.70	9.65	0.83	0.82	−0.50	−1.92	1.72	1.92	2.07	2.26
Cubic SVM	0.24	0.73	25.52	9.83	0.42	0.84	0.57	−0.35	0.65	1.89	0.79	2.22
Fine Gaussian SVM	0.57	0.68	11.65	11.53	0.61	0.71	−1.47	−2.86	1.43	1.61	1.73	1.89
Medium Gaussian SVM	0.70	0.69	9.29	10.96	0.79	0.74	−1.27	−2.02	1.79	1.69	2.16	1.99
Coarse Gaussian SVM	0.61	0.59	11.65	14.40	0.63	0.52	−3.66	−5.66	1.43	1.29	1.73	1.51
Boosted tree	0.71	0.72	8.97	10.06	0.82	0.80	−0.99	−2.30	1.86	1.85	2.24	2.17
Bagged tree	0.70	0.69	9.15	10.75	0.80	0.75	−0.15	−1.63	1.82	1.73	2.20	2.03
Squared exponential GPR	0.75	**0.77**	8.24	**8.87**	0.86	0.85	−0.07	−0.68	2.02	2.09	2.44	2.46
Matern 5/2 GPR	0.76	0.76	8.13	9.02	0.86	0.85	−0.10	−0.87	2.05	2.06	2.47	2.42
Exponential GPR	0.79	0.75	7.60	9.30	0.88	0.83	−0.08	−0.93	2.19	2.00	2.65	2.35
Rational quadratic GPR	0.77	0.74	7.98	9.47	0.87	0.83	−0.15	−1.33	2.09	1.96	2.52	2.30
Partial least square regression	0.33	0.40	13.59	14.78	0.49	0.48	0.00	−2.60	1.23	1.26	1.48	1.48

Note: RMSE, root mean square error; RPD, ratio of prediction to deviation; RPIQ, ratio of performance to interquartile distance; IV, internal validation; EV, external validation; SVM, support vector machine; GPR, Gaussian process regression. These are results of 10-fold cross-validation IV, EV, RPD, and RPIQ, respectively. The bold numbers in the rows refer to the best models.

External validation

For the external validation data set, the ensemble tree having edge to GPR-based model with consistent higher accuracy. The most accurate predictions were obtained by cubist, with R^2, RMSE, LCCC, bias, RPD and RPIQ of 0.74, 9.80%, 0.81, −2.02, 1.90 and 2.23, respectively (Table 5). However, using other methods, the RMSE of the cubist model prediction lowered by approximately 9%–44%. The least accurate predictions were those produced by linear SVM model with R^2, RMSE, concordance, bias, RPD and RPIQ of 0.51, 14.08%, 0.57, −4.78, 1.32 and 1.55, respectively (Table 5).

Prediction of SMC using six predictor variables

10-fold cross (internal) validation

The ensemble tree modeling approaches constantly yield prediction accuracy with $R^2 > 0.82$, while all the GPR-based models resulted in same prediction accuracy. The other model approaches also predicted SMC with an average accuracy with $R^2 > 0.70$. The most accurate predictions were obtained using boosted trees, with R^2, RMSE, LCCC, bias, RPD, and RPIQ values of 0.86, 10.86%, 0.91, −1.63, 2.61, and 4.13, respectively (Table 6). The next best predictions were produced by Cubist model with R^2, RMSE, LCCC, bias, RPD, and RPIQ values of 0.85, 10.88%, 0.92, 0.67, 2.61, and 4.13, respectively. The least accurate predictions were from coarse tree model with R^2, RMSE, LCCC, bias, RPD, and RPIQ values of 0.69, 15.76%, 0.82, −0.54, 1.80, and 2.85, respectively (Table 6).

External validation

Overall, excellent predictions were obtained with all the models showing RPD > 2 apart from few models (LR, Robust Linear, Linear SVM, and Coarse Gaussian SVM) showing an RPD < 2. Utilizing R^2 to evaluate the model performance also produced similar results, with validation $R^2 \geq 0.69$ for all the calibrated models (Table 6). The R^2 value for the model trained using RF producing best predictions was 0.86, the RMSE was 8.79%, the LCCC was 0.91, the bias was 1.73, the RPD was 2.60, and the RPIQ was 2.86 (Table 6). On the other hand, the poorest predictions were produced by linear SVM model giving an R^2, RMSE, LCCC, bias, RPD, and RPIQ values of 0.73, 12.29%, 0.81, 3.64, 1.86, and 2.04, respectively (Table 6).

Table 4. Accuracy of different models for the prediction of soil moisture content in the calibration and validation data sets using 22 predictor variables.

	R^2		RMSE		Concordance		Bias		RPD		RPIQ	
Models	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV
Random forest	0.87	0.87	10.07	8.44	0.93	0.91	−0.03	1.62	2.82	2.71	4.46	2.97
Cubist	0.88	0.93	9.62	6.20	0.93	0.95	0.22	1.27	2.95	3.68	4.66	4.05
Aritficial neural network	0.81	0.76	14.33	12.50	0.81	0.78	−2.15	3.36	1.98	1.83	3.13	2.01
Linear regression	0.82	0.86	11.94	8.84	0.90	0.90	0.38	2.79	2.38	2.59	3.76	2.84
Interactions linear	0.00	0.01	308.97	75.87	−0.01	0.06	−47.01	5.65	0.09	0.30	0.15	0.33
Robust linear	0.83	0.86	11.67	8.90	0.90	0.90	−0.43	2.94	2.43	2.57	3.85	2.82
Linear regression-pure quadratic	0.78	0.86	14.30	8.71	0.87	0.90	0.64	1.38	1.99	2.63	3.14	2.88
Stepwise linear	0.85	0.90	11.27	7.31	0.92	0.93	0.41	0.92	2.52	3.13	3.99	3.43
Fine tree	0.84	0.85	11.61	9.06	0.91	0.91	0.32	1.63	2.45	2.52	3.87	2.77
Medium tree	0.83	0.83	11.83	9.78	0.90	0.89	−0.18	2.21	2.40	2.34	3.80	2.57
Coarse tree	0.66	0.80	16.43	10.25	0.80	0.87	−0.56	1.97	1.73	2.23	2.73	2.45
Linear SVM	0.85	0.82	11.16	10.40	0.91	0.87	0.23	3.86	2.54	2.20	4.02	2.42
Quadratic SVM	0.86	0.90	10.94	7.26	0.92	0.93	−0.09	1.46	2.60	3.15	4.11	3.46
Cubic SVM	0.55	0.82	21.66	9.68	0.73	0.89	−1.10	0.34	1.31	2.36	2.07	2.59
Fine Gaussian SVM	0.68	0.90	17.69	7.87	0.70	0.92	−1.98	1.78	1.61	2.91	2.54	3.19
Medium Gaussian SVM	0.87	0.91	10.21	7.15	0.92	0.93	−0.67	1.66	2.78	3.20	4.40	3.51
Coarse Gaussian SVM	0.85	0.78	11.47	11.47	0.90	0.84	−0.51	4.08	2.47	1.99	3.91	2.19
Boosted trees	0.87	0.87	10.33	8.10	0.92	0.92	−1.13	0.70	2.75	2.82	4.35	3.10
Bagged trees	0.88	0.87	9.99	8.50	0.93	0.91	−0.14	2.05	2.84	2.69	4.49	2.96
Squared exponential GPR	0.88	0.93	9.78	6.32	0.93	0.94	−0.15	1.37	2.90	3.62	4.59	3.97
Matern 5/2 GPR	0.88	0.93	9.82	5.94	0.93	0.95	−0.08	1.24	2.89	3.85	4.57	4.23
Exponential GPR	**0.89**	**0.95**	**9.40**	**5.21**	0.93	0.96	−0.10	1.12	3.02	4.39	4.78	4.82
Rational quadratic GPR	0.88	0.93	9.88	6.18	0.93	0.95	−0.21	1.35	2.87	3.70	4.54	4.07
Partial least square regression	0.72	0.69	14.90	13.26	0.83	0.80	0.00	3.61	1.91	1.72	3.01	1.90

Note: RMSE, root mean square error; RPD, ratio of prediction to deviation; RPIQ, ratio of performance to interquartile distance; IV, internal validation; EV, external validation; SVM, support vector machine; GPR, Gaussian process regression. These are results of 10-fold cross-validation IV, EV, RPD, and RPIQ, respectively. The bold numbers in the rows refer to the best models.

Discussion

Identification of important predictors

Reasonable and similar prediction accuracies were obtained for both soil properties (i.e., SOM and SMC), even after the removal of insignificant predictors compared to that obtained using the full set of predictor variables. This suggested that a lot of parameters explained only a very little portion of the variation and, hence, their identification and removal was necessary. In addition, removal of redundant parameters also facilitated reduction in processing power and time without compromising the accuracy. Other researchers have shown that the large number of model inputs does not necessarily increase its accuracy, and the removal of additional and ineffective parameters improves the model's performance in predicting SOM and SMC (Zhao et al. 2020; Fathololoumi et al. 2021*b*).

Model performance

The independently validated statistics also showed that both SMC and SOM content of samples could be predicted with high accuracy using appropriate modeling techniques. Overall, SMC was predicted with greater accuracy than SOM content, and the choice of different models had a clear impact on the prediction quality for both SMC and SOM content (Tables 4-6). This result is in line with some previous research (Paloscia et al. 2008; Fang et al. 2020; Zhou et al. 2020). Fu et al. (2020) quantified the effects of soil moisture on the relationship between SOM and the color parameters derived from mobile phone images using univariate LR models. However, in the present study, various neural network and machine learning algorithms were used to evaluate the impact of soil properties on SOM and SMC. The results showed different performance of these algorithms.

A closer look at the results showed that the GPR models demonstrated excellent prediction ability, as compared to all other models, for both calibration data sets and validation data sets (for both SOM and SMC with 22 predictor variables). Its superior performance can be attributed to the fact that it yields reliable responses to the provided input data, thereby increasing its reliability as a probabilistic model (Rasmussen and Nickisch 2010).

Table 5. Accuracy of different models for the prediction of soil organic matter in the calibration and validation data sets using six predictor variables.

	R^2		RMSE		Concordance		Bias		RPD		RPIQ	
Models	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV
Random forest	0.66	0.68	9.72	10.66	0.79	0.77	−0.61	−1.86	1.71	1.74	2.07	2.04
Cubist	0.71	**0.74**	8.94	**9.80**	0.83	0.81	−0.40	−2.02	1.86	1.90	2.25	2.23
Artificial neural network	**0.74**	0.62	**8.51**	11.57	0.84	0.72	−0.26	−1.23	1.96	1.61	2.36	1.88
Linear regression	0.58	0.57	10.83	12.57	0.74	0.64	−0.13	−1.37	1.54	1.48	1.86	1.73
Interactions linear	0.68	0.62	9.61	11.52	0.82	0.72	0.11	−0.92	1.73	1.61	2.09	1.89
Robust linear	0.42	0.55	13.07	13.38	0.63	0.61	−1.88	−4.06	1.27	1.39	1.54	1.63
Linear regression-pure quadratic	0.66	0.61	9.80	11.71	0.80	0.71	−0.08	−1.03	1.70	1.59	2.05	1.86
Stepwise linear	0.63	0.61	10.19	11.70	0.79	0.71	−0.05	−1.07	1.63	1.59	1.97	1.86
Fine tree	0.58	0.52	11.17	13.06	0.75	0.69	−0.26	−2.39	1.49	1.42	1.80	1.67
Medium tree	0.50	0.65	11.96	10.94	0.69	0.77	−0.34	−1.51	1.39	1.70	1.68	1.99
Coarse tree	0.44	0.54	12.49	12.67	0.62	0.65	−0.13	0.32	1.33	1.47	1.61	1.72
Linear SVM	0.57	0.51	11.00	14.08	0.72	0.57	−1.93	−4.78	1.51	1.32	1.83	1.55
Quadratic SVM	0.60	0.58	10.53	12.95	0.76	0.63	−1.19	−3.39	1.58	1.43	1.91	1.68
Cubic SVM	0.47	0.58	13.52	12.11	0.68	0.71	−1.15	−2.06	1.23	1.53	1.49	1.80
Fine Gaussian SVM	0.54	0.66	11.51	11.52	0.65	0.73	−1.56	−3.13	1.45	1.61	1.75	1.89
Medium Gaussian SVM	0.65	0.61	9.96	12.47	0.77	0.66	−1.53	−3.22	1.67	1.49	2.02	1.75
Coarse Gaussian SVM	0.61	0.55	10.96	13.87	0.70	0.57	−2.74	−4.75	1.52	1.34	1.83	1.57
Boosted trees	0.66	0.69	9.69	10.75	0.80	0.77	−0.86	−2.76	1.72	1.73	2.07	2.03
Bagged trees	0.63	0.68	10.12	11.12	0.76	0.74	−0.26	−2.31	1.65	1.67	1.99	1.96
Squared exponential GPR	0.67	0.62	9.52	11.53	0.80	0.72	−0.07	−0.89	1.75	1.61	2.11	1.89
Matern 5/2 GPR	0.67	0.63	9.47	11.34	0.80	0.73	−0.07	−0.86	1.76	1.64	2.12	1.92
Exponential GPR	0.68	0.68	9.42	10.75	0.80	0.76	0.00	−1.13	1.77	1.73	2.13	2.03
Rational quadratic GPR	0.67	0.63	9.61	11.46	0.80	0.73	−0.10	−0.87	1.73	1.62	2.09	1.90
Partial least square regression	0.64	0.85	10.02	10.78	0.77	0.92	0.00	0.00	1.66	2.63	2.01	4.16

Note: RMSE, root mean square error; RPD, ratio of prediction to deviation; RPIQ, ratio of performance to interquartile distance; IV, internal validation; EV, external validation; SVM, support vector machine; GPR, Gaussian process regression. These are results of 10-fold cross-validation IV, EV, RPD and RPIQ, respectively. The bold numbers in the rows refer to the best models.

Artificial neural network models were observed to perform well during the SOM calibration phase (under both the cases of utilization of 22 and 6 predictor variables). However, it could not sustain its performance as far as the prediction of SOM was concerned during the validation phase. This could be due to the reason that ANNs possess a predefined structure directed only toward minimizing errors on the training data set. Zhao et al. (2020) and Fathololoumi et al. (2020) presented a similar result in their research.

Apart from these, tree models provided satisfactory prediction accuracies (cubist and RF for the prediction of SOM and SMC, respectively using six predictor variables during the validation phase and boosted trees for SMC using six predictor variables and during the calibration phase). The reason for their success could be linked to the several benefits associated with the utilization of tree models (or rule-based decision methods) such as insusceptibility to outliers, insensitiveness to irrelevant predictors, managing the provided data of varying measurement scale and level, instinctive structure of the models, etc. Similar results have been provided by Heung et al. (2016), Dharumarajan et al. (2017), and Hajdu et al. (2018).

Interactions linear model exhibited the poorest performance when 22 predictor variables were used, for both calibration and validation data sets for both soil properties. On the other hand, when six predictor variables were used, its performance was relatively better. On paying closer attention to the structure of the developed model, it was observed that utilization of 22 predictor variables resulted in a huge number of model parameters (interaction terms) as compared to fewer terms when only six predictor variables were used.

Linear SVM showed poor prediction ability during the validation phase for SOM and SMC using six predictors. This is simply because linear SVM does not yield reasonable results on data which are not linearly separable. This issue is dealt by choosing the right kernel, which is why other types of SVM used in this study performed somewhat better but not exceptionally good. In modeling based on regression models, the use of the optimal number of predictor variables is very important. To reduce processing volume and fieldwork, the op-

Table 6. Accuracy of different models for the prediction of soil moisture content in the calibration and validation data sets using six predictor variables.

	R^2		RMSE		Concordance		Bias		RPD		RPIQ	
Models	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV
Random forest	0.84	**0.86**	11.20	**8.79**	0.91	0.91	−0.07	1.73	2.54	2.60	4.01	2.86
Cubist	0.85	0.83	10.88	9.83	0.92	0.88	0.67	2.67	2.61	2.33	4.13	2.56
Artificial neural network	0.86	0.82	11.27	9.89	0.90	0.87	−3.24	0.97	2.52	2.31	3.98	2.54
Linear regression	0.79	0.76	12.91	11.65	0.88	0.84	0.03	3.07	2.20	1.96	3.48	2.16
Interactions linear	0.82	0.82	12.15	9.99	0.90	0.88	0.26	2.61	2.34	2.29	3.69	2.52
Robust linear	0.79	0.74	13.09	11.94	0.87	0.82	−0.79	2.88	2.17	1.91	3.43	2.10
Linear regression-pure quadratic	0.80	0.77	12.79	11.32	0.88	0.85	0.17	3.25	2.22	2.02	3.51	2.22
Stepwise linear	0.83	0.81	11.74	10.31	0.90	0.88	0.15	2.81	2.42	2.22	3.83	2.44
Fine tree	0.82	0.81	11.91	10.54	0.90	0.88	−0.31	2.44	2.38	2.17	3.77	2.38
Medium tree	0.85	0.82	11.04	9.84	0.91	0.89	−0.02	1.58	2.57	2.32	4.07	2.55
Coarse tree	0.69	0.80	15.76	10.25	0.82	0.87	−0.54	1.97	1.80	2.23	2.85	2.45
Linear SVM	0.80	0.73	12.84	12.29	0.88	0.81	0.18	3.64	2.21	1.86	3.50	2.04
Quadratic SVM	0.81	0.77	12.34	11.40	0.89	0.85	0.02	3.08	2.30	2.01	3.64	2.20
Cubic SVM	0.77	0.85	13.68	9.03	0.87	0.90	0.31	1.87	2.08	2.53	3.28	2.78
Fine Gaussian SVM	0.78	0.81	13.53	10.18	0.85	0.87	−0.21	1.60	2.10	2.25	3.32	2.47
Medium Gaussian SVM	0.81	0.76	12.59	11.49	0.88	0.84	−0.90	2.90	2.26	1.99	3.57	2.19
Coarse Gaussian SVM	0.78	0.74	13.59	12.04	0.85	0.81	−0.92	2.80	2.09	1.90	3.30	2.09
Boosted trees	**0.86**	0.83	**10.86**	9.49	0.91	0.89	−1.63	0.80	2.61	2.41	4.13	2.65
Bagged trees	0.82	0.85	11.97	8.95	0.89	0.90	−0.08	1.62	2.37	2.56	3.75	2.81
Squared exponential GPR	0.82	0.79	11.85	10.80	0.90	0.86	0.01	2.88	2.40	2.12	3.79	2.33
Matern 5/2 GPR	0.82	0.80	12.01	10.56	0.90	0.87	0.09	2.71	2.37	2.16	3.74	2.38
Exponential GPR	0.82	0.82	12.04	9.80	0.89	0.88	0.20	1.93	2.36	2.33	3.73	2.56
Rational Quadratic GPR	0.82	0.79	11.87	10.80	0.90	0.86	0.04	2.88	2.39	2.12	3.78	2.33
Partial least square regression	0.80	0.76	12.77	11.50	0.88	0.84	0.00	2.55	2.22	1.99	3.52	2.18

Note: RMSE, root mean square error; RPD, ratio of prediction to deviation; RPIQ, ratio of performance to interquartile distance; IV, internal validation; EV, external validation; SVM, support vector machine; GPR, Gaussian process regression. These are results of 10-fold cross-validation IV, EV, RPD and RPIQ, respectively. The bold numbers in the rows refer to the best models.

timal mode is to use the least number of predictor variables with the highest modeling accuracy. In this study, we reduced the number of 22 variables to 6 variables if the modeling accuracy did not change significantly. This shows that these six variables have been the most important and effective parameters in the modeling process. Although the modeling accuracy did not change significantly, the processing volume was significantly reduced.

Overall, the nonlinear models performed well than the linear ones, it was inferred that there exists a nonlinear relationship between the SOM, SMC, and image parameters. The efficiency of nonlinear models such as RF and cubist for SOM and SMC prediction has shown in some other studies (Taghizadeh-Mehrjardi et al. 2020; Fathololoumi et al. 2021*a*; Zeraatpisheh et al. 2022).

Conclusions

The SMC and SOM are known to influence the soil color; soil high in humus appears dark black to brown and along with high moisture content even 5% SOM is sufficient for darker appearance. The darker appearance with higher moisture content is attributed to higher light absorbance. However, the long-term higher moisture content also affects the soil color by enhancing anaerobic conditions and affecting state of iron oxides in soil (Jackson 2008). This study reports the calibration and validation of 22 supervised regression and machine learning algorithms to evaluate the potential of soil images captured by a digital camera to predict SOM and SMC. These models developed prediction relationships among SOM and SMC (measured in the laboratory) and various color- and texture-related features derived from images. Color parameters demonstrated high correlation with both SOM and SMC. Overall, the predicted SMC with greater accuracy than SOM implied that SMC exerts a considerable influence in imparting color to the soil. Results revealed a satisfactory agreement between the image parameters and the laboratory-measured SOM (R^2 and RMSE of 0.74 and 9.80% using cubist) and SMC (R^2 and RMSE of 0.86 and 8.79% using RF) for the validation data set using six predictor variables. Overall, GPRs and tree models (cubist, RF, and boosted trees) best captured and explained the nonlinear relationships between

SOM, SMC, and image parameters for this study. The soil color was also affected by temperature, climate, and mineral content; therefore, more research involving real field condition across different soil type and climatic regions was needed to establish a standard methodology for predicting SMC, SOM, and other soil properties using digital images. The advantage of this methodology over the traditional method would be rapid estimation of soil properties at a much reduced cost and be environmentally safe. Taken together, digital image-based soil characterization provides an opportunity to be used for proximal soil sensing.

Article information

History dates

Received: 16 September 2021
Accepted: 26 February 2022
Accepted manuscript online: 31 March 2022
Version of record online: 31 August 2022

Copyright

Author information

Author notes

Author Asim Biswas served as an Associate Editor at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by Mervin St. Luce.

Author contributions

P.T. contributed to formal analysis, investigation, data curation, and writing (original draft preparation); H.B.V. was responsible for writing, review and editing, and project administration; S.F. contributed to review and editing; P.D. contributed to resources, supervision, and writing (review and editing); A.B. was responsible for conceptualization, funding acquisition, investigation, methodology, project administration, resources, validation, visualization, writing (review and editing). All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no conflict of interest.

Funding information

This research was funded by Ontario Ministry of Agriculture, Food and Rural Affairs: UofG-2016-2600 and Natural Sciences and Engineering Research Council of Canada: RGPIN-2014-04100.

References

Chang, C.-W., Laird, D.A., Mausbach, M.J., and Hurburgh, C.R. 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. **65**: 480–490. doi: 10.2136/sssaj2001.652480x

Chen, D., Chang, N., Xiao, J., Zhou, Q., and Wu, W. 2019. Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. Sci. Total Environ. **669**: 844–855. doi: 10.1016/j.scitotenv.2019.03.151. PMID: 30897441

Chukalla, A.D., Krol, M.S., and Hoekstra, A.Y. 2015. Green and blue water footprint reduction in irrigated agriculture: effect of irrigation techniques, irrigation strategies and mulching. Hydrol. Earth Syst. Sci. **19**: 4877–4891. doi: 10.5194/hess-19-4877-2015

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D.L. 2000. The Mahalanobis distance. Chemom. Intell. Lab. Syst. **50**: 1–18. doi: 10.1016/S0169-7439(99)00047-7

Dharumarajan, S., Hegde, R., and Singh, S. 2017. Spatial prediction of major soil properties using random forest techniques—A case study in semi-arid tropics of South India. Geoderma Reg. **10**: 154–162. doi: 10.1016/j.geodrs.2017.07.005

dos Santos, J.F., Silva, H.R., Pinto, F.A., and Assis, I.R.d. 2016. Use of digital images to estimate soil moisture. Rev. Bras. de Eng. Agrícola e Ambient. **20**: 1051–1056.

Fang, L., Zhan, X., Yin, J., Liu, J., Schull, M. Walker, J.P., et al. 2020. An intercomparison study of algorithms for downscaling SMAP radiometer soil moisture retrievals. J. Hydrometeorol. **21**: 1761–1775. doi: 10.1175/JHM-D-19-0034.1

Fathololoumi, S., Vaezi, A.R., Firozjaei, M.K., and Biswas, A. 2021a. Quantifying the effect of surface heterogeneity on soil moisture across regions and surface characteristic. J. Hydrol. **596**: 126132. doi: 10.1016/j.jhydrol.2021.126132

Fathololoumi, S., Vaezi, A.R., Alavipanah, S.K., Ghorbani, A., and Biswas, A. 2020. Comparison of spectral and spatial-based approaches for mapping the local variation of soil moisture in a semi-arid mountainous area. Sci. Total Environ. 138319.doi: 10.1016/j.scitotenv.2020.138319. PMID: 32408464

Fathololoumi, S., Vaezi, A.R., Alavipanah, S.K., Ghorbani, A., Saurette, D., and Biswas, A. 2021b. Effect of multi-temporal satellite images on soil moisture prediction using a digital soil mapping approach. Geoderma, **385**: 114901. doi: 10.1016/j.geoderma.2020.114901

Feki, M., Ravazzani, G., Ceppi, A., and Mancini, M. 2018. Influence of soil hydraulic variability on soil moisture simulations and irrigation scheduling in a maize field. Agric. Water Manag. **202**: 183–194. doi: 10.1016/j.agwat.2018.02.024

Fu, Y., Taneja, P., Lin, S., Ji, W., Adamchuk, V., Daggupati, P., and Biswas, A. 2020. Predicting soil organic matter from cellular phone images under varying soil moisture. Geoderma, **361**: 114020.doi: 10.1016/j.geoderma.2019.114020

Gholizadeh, A., Saberioon, M., Rossel, R.A.V., Boruvka, L., and Klement, A. 2020. Spectroscopic measurements and imaging of soil colour for field scale estimation of soil organic carbon. Geoderma, **357**: 113972. doi: 10.1016/j.geoderma.2019.113972

Gill, M.K., Asefa, T., Kemblowski, M.W., and McKee, M. 2006. Soil moisture prediction using support vector machines 1. J. Am. Water Resour. Assoc. **42**: 1033–1046. doi: 10.1111/j.1752-1688.2006.tb04512.x

Gómez-Robledo, L., López-Ruiz, N., Melgosa, M., Palma, A.J., Capitán-Vallvey, L.F., and Sánchez-Marañón, M. 2013. Using the mobile phone as Munsell soil-colour sensor: an experiment under controlled illumination conditions. Comput. Electron. Agric. **99**: 200–208.

Gonzalez, R.C., Woods, R.E., and Eddins, S.L. 2004. Digital image processing using MATLAB. Pearson Education India.

Gregory, S.D., Lauzon, J.D., O'Halloran, I.P., and Heck, R.J. 2006. Predicting soil organic matter content in southwestern Ontario fields using imagery from high-resolution digital cameras. Can. J. Soil Sci. **86**: 573–584. doi: 10.4141/S05-043

Hajdu, I., Yule, I., and Dehghan-Shear, M.H. 2018. Modelling of near-surface soil moisture using machine learning and multi-temporal sentinel 1 images in New Zealand. IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE. pp. 1422–1525.

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

Hummel, J.W., Sudduth, K.A., and Hollinger, S.E. 2001. Soil moisture and organic matter prediction of surface and subsurface soils using an

NIR soil sensor. Comput. Electron. Agric. **32**: 149–165. doi: 10.1016/S0168-1699(01)00163-6

Jackson, R.S. 2008. Wine science: principles and applications. Academic press, New York.

Ji, W., Adamchuk, V.I., Biswas, A., Dhawale, N.M., Sudarsan, B. Zhang, Y., et al. 2016. Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. Biosyst. Eng. **152**: 14–27. doi: 10.1016/j.biosystemseng.2016.06.005

King, D.J. 1995. Airborne multispectral digital camera and video sensors: a critical review of system designs and applications. Can. J. Remote Sens. **21**: 245–273. doi: 10.1080/07038992.1995.10874621

Kotlar, A.M., Iversen, B.V., and de Jong van Lier, Q. 2019. Evaluation of parametric and nonparametric machine-learning techniques for prediction of saturated and near-saturated hydraulic conductivity. Vadose Zone J. **18**. doi: 10.2136/vzj2018.07.0141

Kumar, S., and Lal, R. 2011. Mapping the organic carbon stocks of surface soils using local spatial interpolator. J. Environ. Monit. **13**: 3128–3135.doi: 10.1039/c1em10520e. PMID: 22009220

Kumar, T., and Verma, K. 2010. A theory based on conversion of RGB image to gray image. Int. J. Comp. Appl. **7**: 7–10. doi: 10.5120/777-1099

Lazzaretti, B.P., Silva, L.S.d., Drescher, G.L., Dotto, A.C., Britzke, D., and Nörnberg, J.L. 2020. Prediction of soil organic matter and clay contents by near-infrared spectroscopy-NIRS. Ciênc. Rural, **50**. doi: 10.1590/0103-8478cr20190506

Levin, N., Ben-Dor, E., and Singer, A. 2005. A digital camera as a tool to measure colour indices and related properties of sandy soils in semi-arid environments. Int. J. Remote Sens. **26**: 5475–5492. doi: 10.1080/01431160500099444

Li, C., Zhuang, Y., Frolking, S., Galloway, J., Harriss, R. Iii, Moore, et al. 2003. Modeling soil organic carbon change in croplands of China. Ecol. Appl. **13**: 327–336. doi: 10.1890/1051-0761(2003)013%5b0327:MSOCCI%5d2.0.CO;2

Li, Q., Yue, T., Wang, C.-Q., Zhang, W.-J., Yu, Y. Li, B., et al. 2013. Spatially distributed modeling of soil organic matter across China: an application of artificial neural network approach. Catena, **104**: 210–218. doi: 10.1016/j.catena.2012.11.012

Lillesand, T., Kiefer, R.W., and Chipman, J. 2015. Remote sensing and image interpretation. John Wiley & Sons, Hoboken, NJ.

Matei, O., Rusu, T., Petrovan, A., and Mihuţ, G. 2017. A data mining system for real time soil moisture prediction. Proc. Eng. **181**: 837–844. doi: 10.1016/j.proeng.2017.02.475

MathWorks, I. 2017. MATLAB 2017b. The MathWorks Inc. Natick, MA.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., and Van Molle, M. 2008. A multiple regression approach to assess the spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma, **143**: 1–13. doi: 10.1016/j.geoderma.2007.08.025

Nocita, M., Stevens, A., Noon, C., and van Wesemael, B. 2013. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. Geoderma, **199**: 37–42. doi: 10.1016/j.geoderma.2012.07.020

Paloscia, S., Pampaloni, P., Pettinato, S., and Santi, E. 2008. A comparison of algorithms for retrieving soil moisture from ENVISAT/ASAR images. IEEE Trans. Geosci. Remote Sens. **46**: 3274–3284. doi: 10.1109/TGRS.2008.920370

Persson, M. 2005. Estimating surface soil moisture from soil color using image analysis. Vadose Zone J. **4**: 1119–1122. doi: 10.2136/vzj2005.0023

Rasmussen, C.E., and Nickisch, H. 2010. Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. **11**: 3011–3015.

Rienzi, E.A., Mijatovic, B., Mueller, T.G., Matocha, C.J., Sikora, F.J., and Castrignanò, A. 2014. Prediction of soil organic carbon under varying moisture levels using reflectance spectroscopy. Soil Sci. Soc. Am. J. **78**: 958–967. doi: 10.2136/sssaj2013.09.0408

Rodionov, A., Pätzold, S., Welp, G., Damerow, L., and Amelung, W. 2014. Sensing of soil organic carbon using visible and near-infrared spectroscopy at variable moisture and surface roughness. Soil Sci. Soc. Am. J. **78**: 949–957. doi: 10.2136/sssaj2013.07.0264

Rossel, R.A.V., Fouad, Y., and Walter, C. 2008. Using a digital camera to measure soil organic carbon and iron contents. Biosyst. Eng. **100**: 149–159. doi: 10.1016/j.biosystemseng.2008.02.007

Sakti, M.B.G., Komariah, Ariyanto, D.P., and Sumani. 2018. Estimating soil moisture content using red-green-blue imagery from digital camera. IOP Conf. Ser. Earth Environ. Sci. **200**: 012004. doi: 10.1088/1755-1315/200/1/012004

Schulte, E.E., and Hopkins, B.G. 1996. Estimation of soil organic matter by weight loss-on-ignition. *In* Soil organic matter: analysis and interpretation. *Edited by* F.R. Magdoff. SSSA Spec. Pub. No. 46. SSSA, Madison. pp. 21–31.

Sorenson, P.T., Small, C., Tappert, M.C., Quideau, S.A., Drozdowski, B., Underwood, A., and Janz, A. 2017. Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy. Can. J. Soil Sci. **97**: 241–248.doi: 10.1139/cjss-2016-0116

Sudarsan, B., Ji, W., Biswas, A., and Adamchuk, V. 2016. Microscope-based computer vision to characterize soil texture and soil organic matter. Biosyst. Eng. **152**: 41–50. doi: 10.1016/j.biosystemseng.2016.06.006

Swetha, R., Bende, P., Singh, K., Gorthi, S., Biswas, A., Li, B., et al. 2020. Predicting soil texture from smartphone-captured digital images and an application. Geoderma, **376**: 114562. doi: 10.1016/j.geoderma.2020.114562

Taghizadeh-Mehrjardi, R., Nabiollahi, K., and Kerry, R. 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma, **266**: 98–110. doi: 10.1016/j.geoderma.2015.12.003

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., et al. 2020. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. Remote Sens. **12**: 1095. doi: 10.3390/rs12071095

Taneja, P., Vasava, H.K., Daggupati, P., and Biswas, A. 2021. Multi-algorithm comparison to predict soil organic matter and soil moisture content from cell phone images. Geoderma, **385**: 114863. doi: 10.1016/j.geoderma.2020.114863

Team R. 2015. RStudio: integrated development for R. Vol. 42. RStudio, Inc., Boston, MA. pp. 14. Available from http://www.rstudio.com.

Were, K., Bui, D.T., Dick, Ø.B., and Singh, B.R. 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. **52**: 394–403. doi: 10.1016/j.ecolind.2014.12.028

Wu, C., Yang, Y., and Xia, J. 2017. A simple digital imaging method for estimating black-soil organic matter under visible spectrum. Arch. Agron. Soil Sci. **63**: 1346–1354. doi: 10.1080/03650340.2017.1280728

Wu, C., Xia, J., Yang, H., Yang, Y., Zhang, Y., and Cheng, F. 2018. Rapid determination of soil organic matter content based on soil colour obtained by a digital camera. Int. J. Remote Sens. **39**: 6557–6571. doi: 10.1080/01431161.2018.1460511

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

Zeraatpisheh, M., Garosi, Y., Owliaie, H.R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., and Xu, M. 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: a comparison of a group of environmental covariates. Catena, **208**: 105723.doi: 10.1016/j.catena.2021.105723

Zhang, F., Li, C., Wang, Z., and Wu, H. 2006. Modeling impacts of management alternatives on soil carbon storage of farmland in Northwest China. Biogeosciences, **3**: 451–466. doi: 10.5194/bg-3-451-2006

Zhao, Z., Yang, Q., Sun, D., Ding, X., and Meng, F.-R. 2020. Extended model prediction of high-resolution soil organic matter over a large area using limited number of field samples. Comput. Electron. Agric. **169**: 105172. doi: 10.1016/j.compag.2019.105172

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., and Lausch, A. 2020. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Sci. Total Environ. **729**: 138244.doi: 10.1016/j.scitotenv.2020.138244. PMID: 32498148

Zhu, Y., Wang, Y., Shao, M., and Horton, R. 2011. Estimating soil water content from surface digital image gray level measurements under visible spectrum. Can. J. Soil Sci. **91**: 69–76. doi: 10.4141/cjss10054