Geoderma 385 (2021) 114863

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Multi-algorithm comparison to predict soil organic matter and soil moisture content from cell phone images

Perry Taneja [a], Hitesh Kumar Vasava [b], Prasad Daggupati [a], Asim Biswas [b,*]

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

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Mobile phone images
Image color and texture features
Cubist
Random forest
Gaussian process regression
Soil characterization
Computer vision

ABSTRACT

An accurate assessment of soil organic matter (SOM) and soil moisture content (SMC) is critical for applications in the fields of agriculture, environment, and engineering. However, characterization and measurement of these properties is costly, time-consuming, and labor-intensive. Research has demonstrated that soil spectral reflectance characteristics can be associated with various soil properties providing an indirect way of measurement. With advancements in technological and computational facilities, high resolution digital images and computer vision algorithms have shown potential to provide rapid and nondestructive characterization of soil properties. Additionally, acceptance of cell phones in everyday life made the digital photography easier and accessible. The objective of this study was to develop and compare various regression and machine learning algorithms to estimate SOM and SMC from cell phone images. A cell phone (LG G5 model) was used to capture images of 25 soil samples from two agricultural fields with highly variable SOM at 6 different soil moisture levels from over-dry to saturated. The images were preprocessed using contrast enhancement and segmentation techniques to deal with illumination inconsistencies and remove non-soil parts of the image including black cracks, leaf residues and specular reflection. A total of 22 color and texture features were extracted from images and predictive relationships were developed against laboratory measured soil properties. A set of 24 supervised regression and machine learning prediction models including six Linear Regression Models, three Decision/Regression Trees, six Support Vector Machines (SVM), four Gaussian Process Regression (GPR) Models, four Ensembles of Trees including random forest and cubist, and other models including Artificial Neural Network (ANN) were compared in this study to predict SOM and SMC. A z-score was used to identify a set of six optimum predictors (subset of 22). Exponential GPR and Cubist model performed the best for SMC prediction, with coefficients of determination (R^2) values of 0.84 and 0.86, and RMSE of 10.18% and 10.43%, respectively, (internal validation with 10-fold cross-validation) when all (22) and a subset of 6 predictors were used. For SOM, ANN and Cubist produced satisfactory prediction accuracy with R^2 values of 0.91 and 0.72 and RMSE values of 5.45% and 9.90%, respectively, when 22 and 6 predictors were used. The external validation results exhibited reasonable predictive ability with Exponential GPR and Matern 5/2 GPR producing R^2 values of 0.92 and 0.95 and RMSE of 5.79% and 5.04%, respectively using 22 and 6 predictors for SMC. Medium Gaussian SVM and Squared Exponential GPR produced R^2 values of 0.56 and 0.53 and RMSE of 8.59% and 8.27%, respectively using 22 and 6 predictors for SOM. This shows potential in fabricating an efficient proximal soil sensor using computer vision and machine learning which can be used to provide quick, accurate and nondestructive predictions of soil properties.

1. Introduction

Soil is a fundamental part of any ecosystem. Inappropriate management of the inherently spatially variable soil can lead to deterioration of its health and thus the health of the entire ecosystem (Bezdicek et al., 1996). For example, uniform application of crop inputs such as fertilizers and chemicals to a spatially variable field may result in both over and under application of chemicals at various locations within an agricultural field leading to a gap in production and deterioration of environmental health from loss from extra inputs (Basso et al., 2011).

* Corresponding author.
E-mail addresses: ptaneja@uoguelph.ca (P. Taneja), hvasava@uoguelph.ca (H.K. Vasava), pdaggupa@uoguelph.ca (P. Daggupati), biswas@uoguelph.ca (A. Biswas).

https://doi.org/10.1016/j.geoderma.2020.114863
Received 24 August 2020; Received in revised form 13 November 2020; Accepted 18 November 2020
Available online 9 December 2020

Precision agriculture approaches account for this spatial variability of soils and can recommend varying amount of input application depending on the specific characteristics of the field (Blackmer and White, 1998). Soil organic matter (SOM) plays a critical role determining variable rates, which necessitates for its accurate assessment (Nordmeyer, 2015). In addition, SOM regulates various physical, chemical, and biological processes and properties. It also contributes towards reducing soil erosion and influences the water holding capacity of the soil (FAO, 2019). Soil moisture content (SMC) is another important component and is the amount of water present in soil. There are countless benefits associated with knowing the amount of moisture present in the soil including but not limited to quantifying the need for irrigation, enabling efficient and economic irrigation through need-based application, availability of nutrients and chemicals and their movements, biological activity, erosion and compaction potential are too few to mention. Therefore, the knowledge of soil properties, such as SOM and SMC, gives an indication of soil health which effects the capability of soil to produce sustainably. Consequently, it also helps land managers and farmers to make informed decisions to ameliorate soil conditions.

Few challenges associated with the assessment of SOM and SMC involve:

a) laborious, expensive and time intensive laboratory analysis (O'Halloran et al., 2004);
b) variability in space (Conant et al., 2011), which necessitates a requirement for spatially dense soil sampling (10 m or less);
c) the effect of a number of elements for instance, type of soil and land (Martin et al., 2010; van Wesemael et al., 2011), that may be worked out using stratified sampling technique;
d) monitoring variation in SOM and SMC over time (Chapman et al., 2013) along with
e) the necessity to acquire estimations quickly and easily to maintain both environmental and economical sustainability.

This terminal challenge has garnered significant attention from researchers lately in developing and finding cheap, cost effective, and faster ways to measure and characterize soil properties. Among the potential techniques, correlating spectral signatures with SOM and SMC showed strong potential due to their contribution towards soil color (Escadafal et al., 1988; Webster and Butler, 1976).

Soil color, an important characteristic, has not only long been used for soil identification (Dudley, 1975), but also, for quantitative and qualitative estimations of soil properties (e.g. (Webster and Butler, 1976)). This may be attributed to the fact that different properties of soil exhibit spectral reflectance characteristics in the near-infrared (NIR) and visible part of the electromagnetic spectrum. Soil color was found to be significantly correlated to spectral reflectance properties (Escadafal et al., 1988) and to some soil properties including SOM content and soil color (Ben-Dor et al., 1997; Krishnan et al., 1980; Lindbo et al., 1998; Schulze et al., 1993; Steinhardt and Franzmeier, 1979) and SMC and soil color (dos Santos et al., 2016; Persson, 2005; Sakti et al., 2018a; Zhu et al., 2011). The darker color in soils has principally been linked to high SOM and SMC and high intrinsic soil fertility (Gelder et al., 2011; Liles et al., 2013). The association between soil reflectance and its moisture and organic matter content can thus, be utilized to estimate their content through modeling.

Traditionally, Munsell colorimetric system (Munsell, 1994) was used for soil color comparison, which required visual matching of the standard color chips to the soil sample. However, subjective visual matching, dependence on light and limited standard color chips of the Munsell color charts render this technique inappropriate when precise measurements of soil color are required (Melville and Atkinson, 1985). These restrictions urged the scientists to look for alternate methods to quantify soil color which are more accurate and objective (Barrett, 2002; Viscarra Rossel et al., 2006). Optical techniques such as image-based soil characterization from digital cameras received strong attention (dos Santos et al., 2016; Persson, 2005; Viscarra Rossel and Awlter, 2002; Wu et al., 2018; Zhu et al., 2011), since they permit greater physically-based measurements soil color. More recently, with technological advances, cell phones became increasingly popular in taking digital images and demonstrated the ability to capture images as good as a digital camera. Sharp resolution cameras in cell phones can effortlessly acquire the red, green, and blue values of intensity of the soil surface pertaining to the RGB color model. Red, green, and blue and the primary colors and they can produce other colors present in the nature. Thus, these three bands in an image by a camera, are able to effortlessly detect color of soil (Doi and Ranamukhaarachchi, 2007; Gómez-Robledo et al., 2013; Levin et al., 2005; Moonrungsee et al., 2015; Stiglitz et al., 2016). Also, color sensors attached to cell-phones have been utilized to carry out soil classification (Aitkenhead et al., 2016; Fu et al., 2019; Han et al., 2016; Swetha et al., 2020). As cellphones get portable and affordable, high resolution cameras can allow the observation of fractions of soil at a remarkable extent. These cameras exhibit an enormous potential to be utilized as a proximal soil sensor (which can be employed to quantify the soil properties) provided they are positioned correctly, either in contact with or at a considerably short distance and the acquired images are analyzed with correct algorithms. Therefore, it is necessary to examine the feasibility of cellphone cameras to acquire good quality images of soil and test if those can be used as an alternative to the digital camera images in predicting SOM and SMC. As these cameras are already being used to take digital images, this research does not focus on improving the camera or its image acquisition power but explores the processing and predicting capability of soil properties particularly SOM and SMC from these images. Additionally, while attachment such as separate lens with cell phone (Lu, 2016) brings added complication in the image acquisition system, inbuilt camera in cell phone should be tested individually for its capability in predicting SOM and SMC provided a suitable modeling algorithm is selected.

Furthermore, variable reduction or data compression techniques have been rarely studied for image-based soil characterization. However, the development of a unified index or scoring system can help identify useful features which would provide reasonable approximation without comprising the accuracy of results, while at the same time enhancing computational speed.

Among the most commonly used modelling techniques to develop predictive relationship between SOM and/or SMC with soil color either from cell-phone or digital camera images, various forms of linear regression have received the most attention (dos Santos et al., 2016; Persson, 2005; Sakti et al., 2018b; Wu et al., 2017). Although, some researches have tried two to three different models to develop predictive relationships (Gregory et al., 2006; Viscarra Rossel et al., 2008; Wu et al., 2018), a comprehensive comparison of the popularly used algorithms and models is scarce. Additionally, there are some models which have been used in some instances to develop predictive relationships between SOM and SMC with reasonable performance. These models include support vector machines (SVM), ensembles of trees (cubist, random forest, boosted trees, bagged trees), and Gaussian Processes Regression (GPR) (Chen et al., 2019; Gill et al., 2006; Kotlar et al., 2019; Matei et al., 2017). However, the capability of these advanced machine learning techniques on image data (except for the use of neural networks by (Donnelly et al., 2013) has not been explored yet.

Therefore, the objectives of this study were to (1) identify the important features extracted from color space models such as RGB, HSV and grayscale for SOM and SMC prediction; (2) calibrate, validate and compare various supervised regression and machine learning models for developing predictive relationships between image features and laboratory measured SOM and SMC.

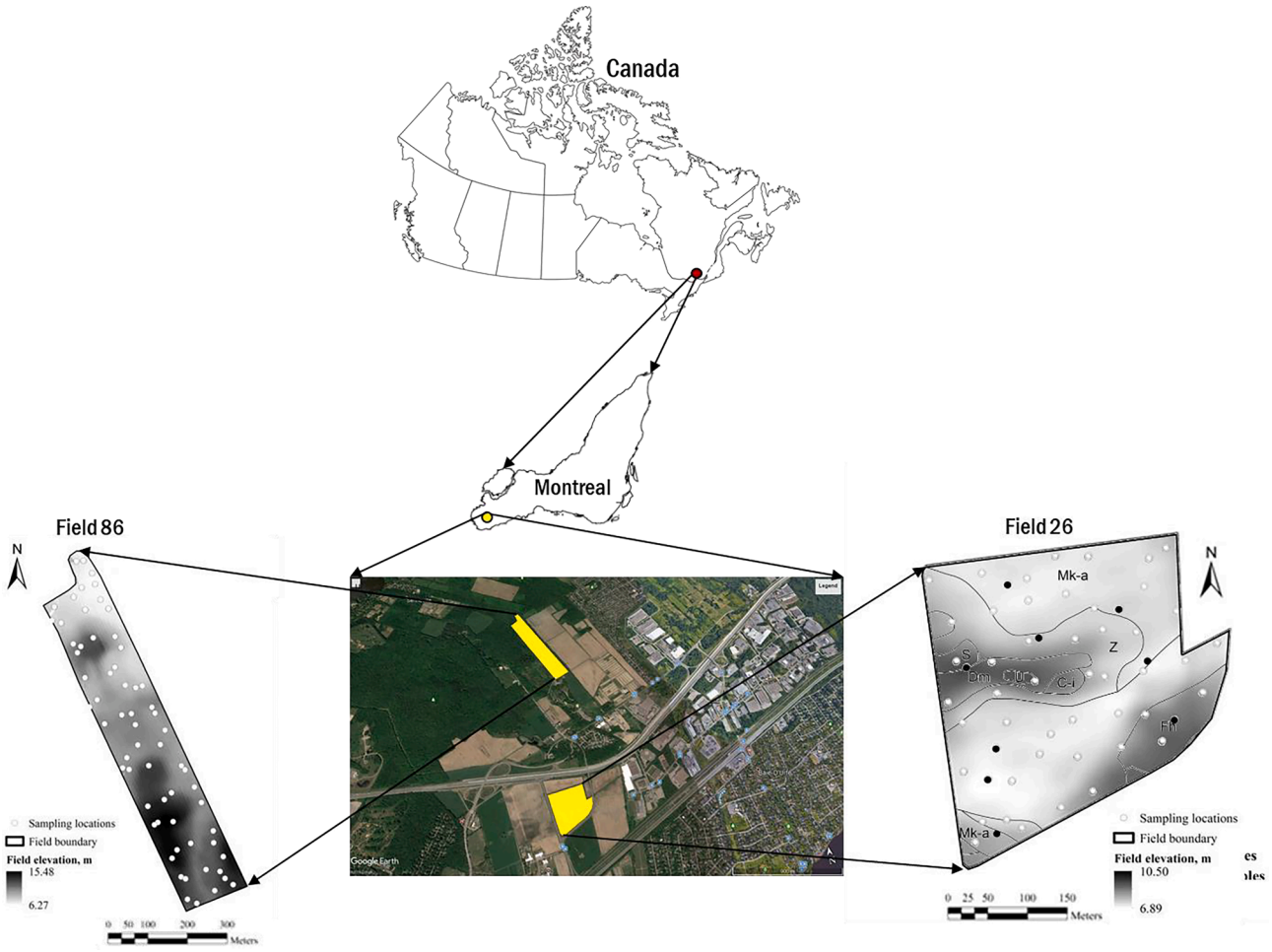

Fig. 1. Geographic location of the study area, Field 86 (left) and Field 26 (right) of Macdonald Campus Farm, McGill University, Quebec, Canada as well as field elevation maps for Field 26 and Field 86 along with the soil map. The letters in the map represent various soil series. Soils in Field 26 are classified into multiple soil series including Muck, ST-Zotique, Soulanges, Chateauguay, Farmington, Uplands, ST-Damase, Chicot and Field 86 into Courval, Ste-Rosalie, St-Amable, Macdonald, St-Bernard, Dalhousie, and Chicot.

2. Materials and methods

2.1. Data collection

2.1.1. Study area and soil sample collection

The soil samples were collected from two agricultural fields namely, Field 26 (~11 ha) and Field 86 (~17 ha) from the MacDonald Campus research farm of McGill University, Sainte Anne De Bellevue, Quebec, Canada (Fig. 1). Both the fields demonstrated a high spatial (within the field) variability in soil types (Ji et al., 2016). The landscape of this area has undergone numerous processes during 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, Field 26 carries soils ranging from mineral to organic deposits (peat) with high variability in soil textures including clay loam, loam, silt loam, sandy loam and sand, whereas, Field 86 mostly includes mineral soils with sandy clay loam, loam, sandy loam, clay and clay loam texture (Fig. 2). Both fields were under no-tillage practices and corn-soybean rotation with soybean and corn being the preceding crop in Field 26 and Field 86, respectively.

While, a previous study (Ji et al., 2016) collected (April to early May in 2015 before seeding), 56 and 64 top (0–10 cm) soil samples from Field 26 and Field 86, respectively following a stratified random sampling strategy, a total of twenty-five soil samples (17 from Field 26 and 8 from Field 86) were selected for this study representing the range of SOM present in these fields. SOM varied strongly within the field with ranges from 3.3 to 62.7% and the 25 samples were selected representing the range (Fig. 3). These 25 samples represented both organic (mainly in Field 26) and mineral soils (present in both fields). This was done deliberately to include universality and increase robustness in training models.

2.1.2. Laboratory analysis and soil imaging

The field collected samples were air dried, ground and sieved through 2 mm sieve. The processed soil samples were used to collect images as well as for measuring soil properties in laboratory. A low chart of the overall methodology is presented in Fig. 3. The SOM analysis was carried out following loss on ignition (LOI) method (Schulte and Hopkins, 1996).

Soil images were collected with a cellphone (Model LG G5) with fixed flash for constant lighting from a fixed distance (32 cm) using a stand. The images were collected at a resolution of 2322 × 4128 pixel and saved as JPEG format. A petri dish (Fig. 4) was filled with processed soil samples and a total of six Groups of images were collected in laboratory conditions at six different soil moisture conditions. Weight of the blank petri dish and dish with soil were recorded. First group (Group 1) of images was collected on the air-dried soil samples. Then water was sprayed carefully without disturbing the soil surface gradually over a period until the water stopped infiltrating (or disappearing from the surface) and show saturation. Soil images (Group 2) were captured at

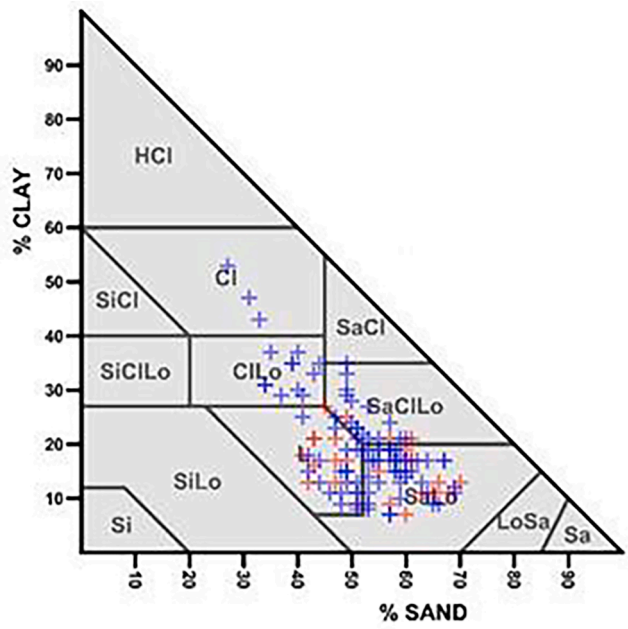

Fig. 2. Soil texture classification (following Canadian System of Soil Classification) of soil samples collected from Field 26 and Field 86. 25 samples selected for this study are represented by red colored signs while blue color signs represent the remaining 95 samples (out of the total 120 samples). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this condition and weights were recorded. The petri dishes were then allowed to dry, and soil images and weights were collected for 3 more times before the petri dishes were oven dried to get the final group of images and weights. Soil moisture levels were then calculated based on weight loss from each drying event. Finally, six Groups of images were grouped into six categories according to the moisture levels with Group 1 containing images of oven-dried soil samples to Group 6 comprising of images analogous to highest SMC, simulating saturation conditions. There was a total of 146 images: 23 images for Group 1, 24 images for Group 2 and Group 6 and 25 images for the rest of the groups. The images were captured under normal lighting conditions of the laboratory. Fig. 5 (right) shows the SMC of 25 soil samples at six different moisture levels.

In this study, the SMC of samples belonging to a single group was not held constant during the experiment, unlike the soil moisture content settings of other studies where all samples SMC were kept at a fixed level leading to abrupt bi- or tri-modal SMC distributions (Nocita et al., 2013; Rienzi et al., 2014; Rodionov et al., 2014). However, SMC usually follows a normal or quasi-normal distribution in a field. Additionally,

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

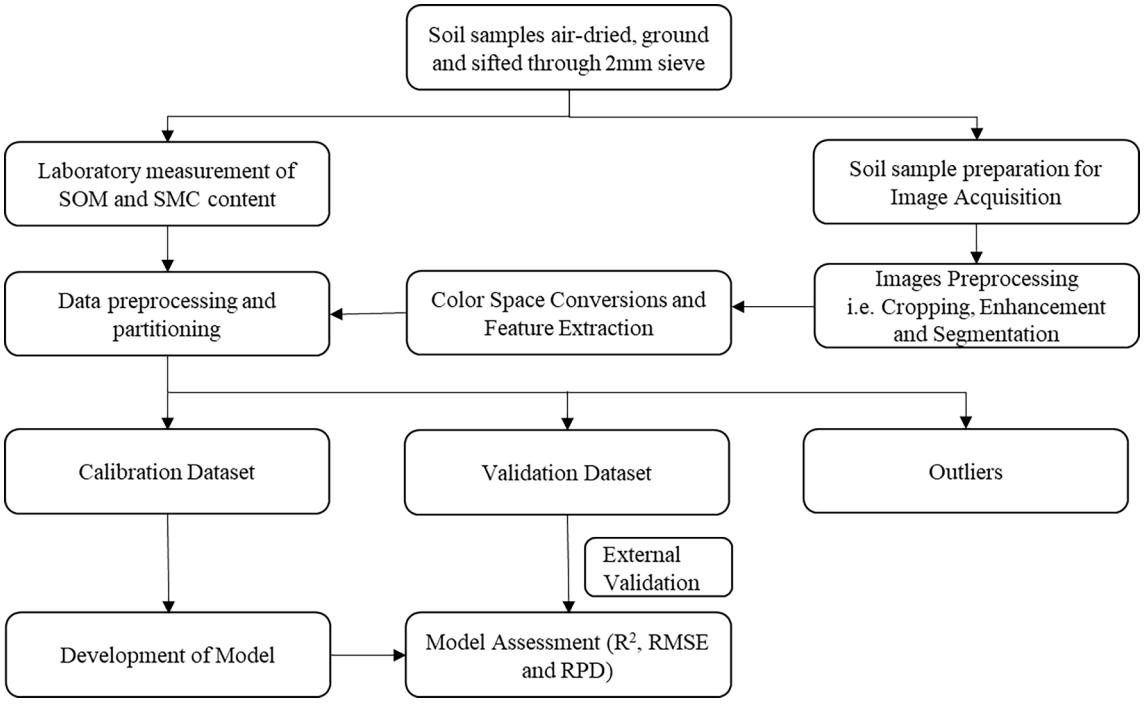

Fig. 3. Workflow diagram of the methodology for this research.

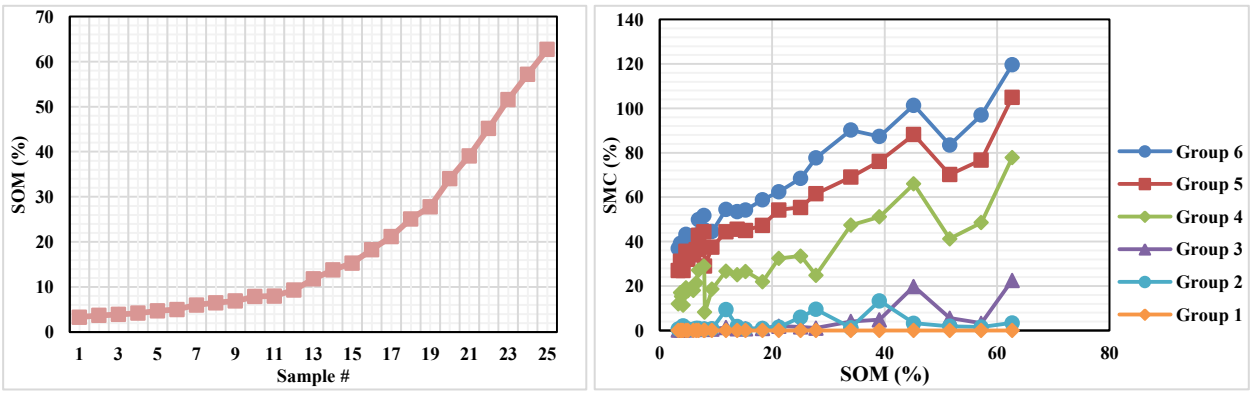

Fig. 5. (left): Soil Organic Matter (SOM) content for the 25 soil samples used in this study ranging from 3.3% to 62.7% collected from Field 26 and Field 86; **(right):** SMC (%) vs SOM (%) for the 25 soil samples corresponding to six different levels of moisture represented as six groups.

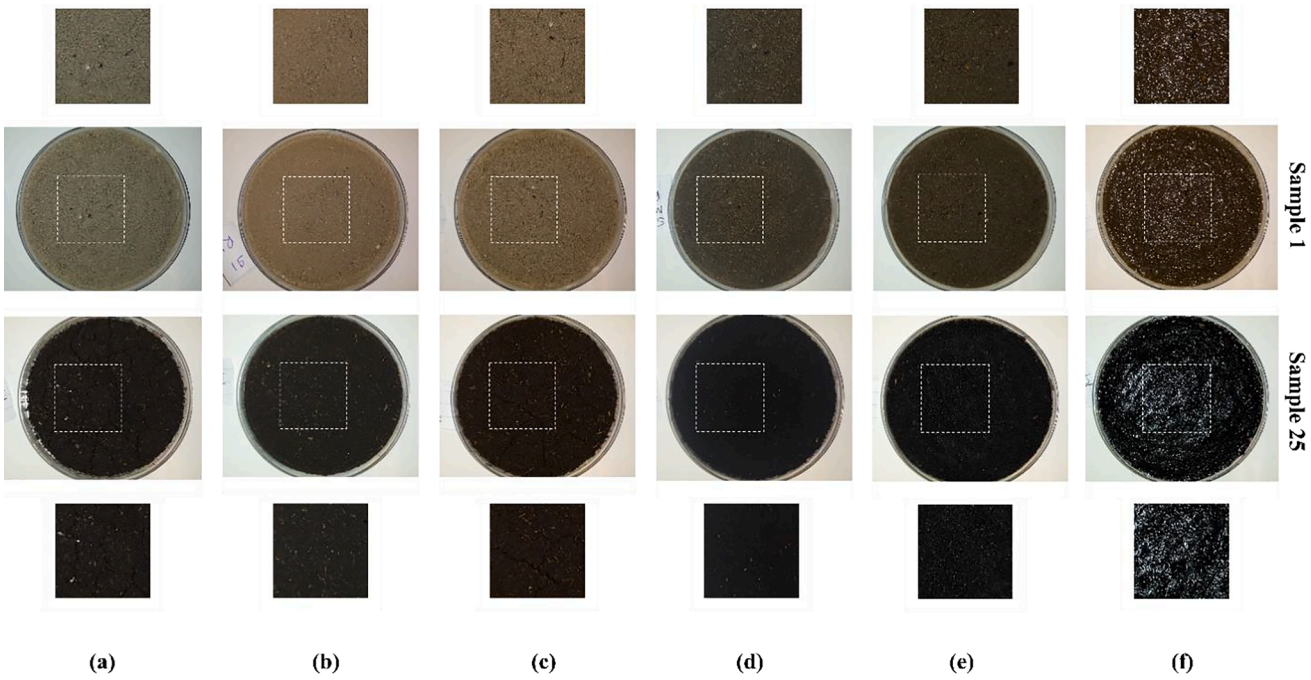

Fig. 6. Images showing two soil samples (sample 1 and sample 25 with 3.3% and 62.7% SOM, respectively) in Petri dishes under six different soil moisture conditions and the corresponding cropped regions. (a) "Group 1"; (b) "Group 2"; (c) "Group 3"; (d) "Group 4"; (e) "Group 5"; and (f) "Group 6".

keeping the SMC constant for each group would bias the imaging process, since samples with different SOM values tend to have varying water holding capacities and thus different drying pattern. For instance, 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 (right)). Keeping this in mind, this study was designed to capture images at six different moisture levels (not controlling the SMC and allowing it to vary) within its natural capacity to hold water and was beneficial to simulate the manner in which soil moisture varies continually through space in a field.

Additionally, numerous studies which used digital image acquisition systems to obtain information about soil color were restricted to controlled sources of light in the laboratory by placing the samples in defined enclosures illuminated by a fixed light source (Gómez-Robledo et al., 2013; Sakti et al., 2018b; Wu et al., 2018; Zhu et al., 2011). However, there was not any such restriction imposed on the image acquisition process in this study since variation in lighting conditions in actual field conditions are abrupt, variable, and uncontrolled. This was done deliberately so that the information obtained from this study can be used in future to develop a proximal soil sensor that can be deployed in field for in situ analysis under highly variable conditions.

2.2. Image analysis

While a good image acquisition system ensures quality of the images captured, suitable image analysis techniques contribute to the derivation of important information from the images and play a critical role in the computer vision applications. Like other fields, caution must be exercised when processing images captured using cell phones. Uncontrolled variables such as the presence of foreign particles on the surface of the soil, non-uniform illumination, reflection from water (in case of high soil moisture content) and other foreign materials can influence the quality of the images (Gonzalez et al., 2004) and must be corrected or taken care of before further analysis.

2.2.1. Image preprocessing

2.2.1.1. Image cropping. A square area of 950 × 950 pixels was cropped

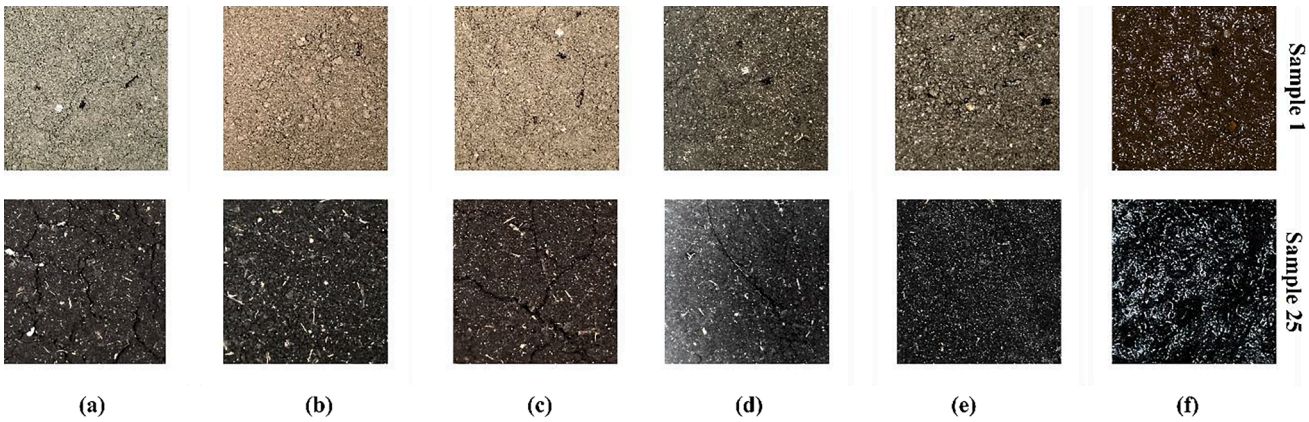

Fig. 7. Images obtained after contrast adjustment of two soil samples (sample 1 and sample 25 with 3.3% and 62.7% SOM, respectively). (a) "Group 1"; (b) "Group 2"; (c) "Group 3"; (d) "Group 4"; (e) "Group 5"; and (f) "Group 6".

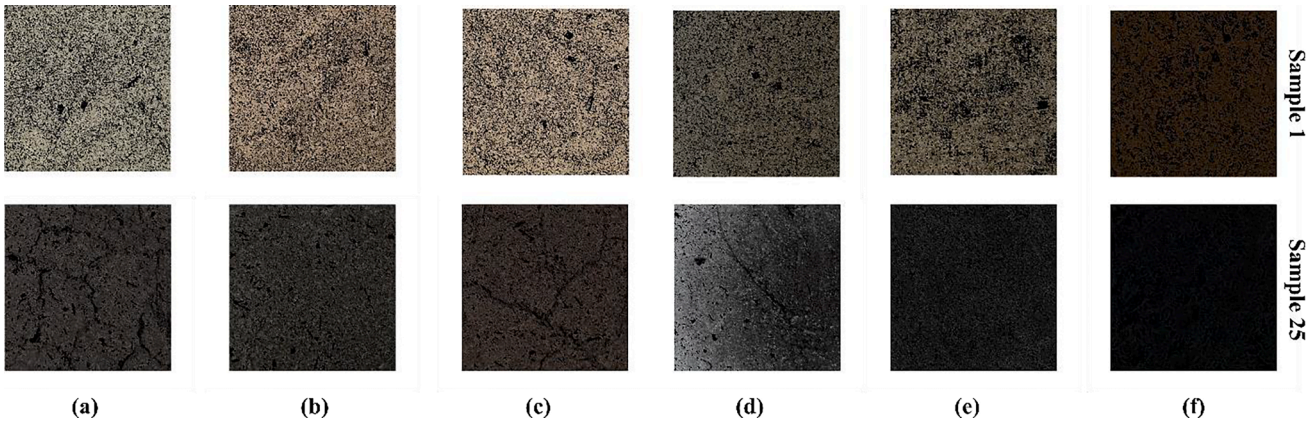

Fig. 8. Images obtained after segmentation of two soil samples (sample 1 and sample 25 with 3.3% and 62.7% SOM respectively). (a) "Group 1"; (b) "Group 2"; (c) "Group 3"; (d) "Group 4"; (e) "Group 5"; and (f) "Group 6".

approximately from the center of the image for the removal of the white background and to minimize the edge effects from the petri dishes (Fig. 6).

2.2.1.2. Image enhancement. Contrast adjustment was carried out using 'imadjust' function in the MATLAB ((MathWorks, 2017) to enhance the images. This facilitated segmentation (next step) by eliminating noise from the image and preventing useful information from fading into noise (Fig. 7).

2.2.1.3. Image segmentation. Image segmentation, as presented here, represented identifying and retaining pixels representing soil from non-soils. For example, certain portions of some images were covered with residues of small leaves, black cracks (only visible after close observation) or film of water producing bright reflections (Fig. 8) and must be identified and removed from pixels representing soil. Despite occupying a small area of the whole image, they were often distributed all over the image and must be excluded from further calculations to reduce erroneous calculations. For example, image intensity values corresponding to these pixels do not represent the image intensity values of the actual soil surface and thus an average value will not represent the true average of the soils' pixels.

An experiential-based segmentation criterion dependent on the histogram of the image was thus developed to differentiate the pixels representing soil from non-soils. There was a significant difference between the intensity values of the pixels of soil from non-soil and was used to segment the image. Since the non-soil pixels covered a small portion in relation to the entire image, a value was determined after various trials. This was assumed with the belief that image intensity values whose counts would be lower than or equal to the defined value, will be considered as those belonging to non-soil materials and thus will be disregarded. A value of 3000 was selected for this study. Additionally, the pixels with water film reflection were almost white in color. In this situation, the histogram was scanned to determine the "highest count" of the image intensity values lying in the range of 248–255 Gy scale values (depicting range of values exhibiting white color). In this situation, the threshold values were set differently than previous with non-saturated 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 non-saturated soils (i.e. 3000) and the greater value was set as the final threshold. For example, for a saturated soil sample, the 'highest count' of 4518 was recorded on gray scale value of 251. Now, the 4518 value was compared with the previously optimized value of 3000 and the higher value, means 4518 was set as the final threshold.

2.2.2. Color space conversions and feature extraction

After image segmentation, color space conversions were first used to convert an RGB image to HSV and monochrome images (Fig. 9). Then color and textural features were extracted from the RGB, HSV and monochrome images. The list of features and their brief explanation is given in Table 1.

A total of 22 features were extracted including Mean R, Mean G, Mean B, Median R, Median G, Median B, Redness Index, Coloration Index, Hue Index, Saturation Index (from RGB images), Mean H, Mean S,

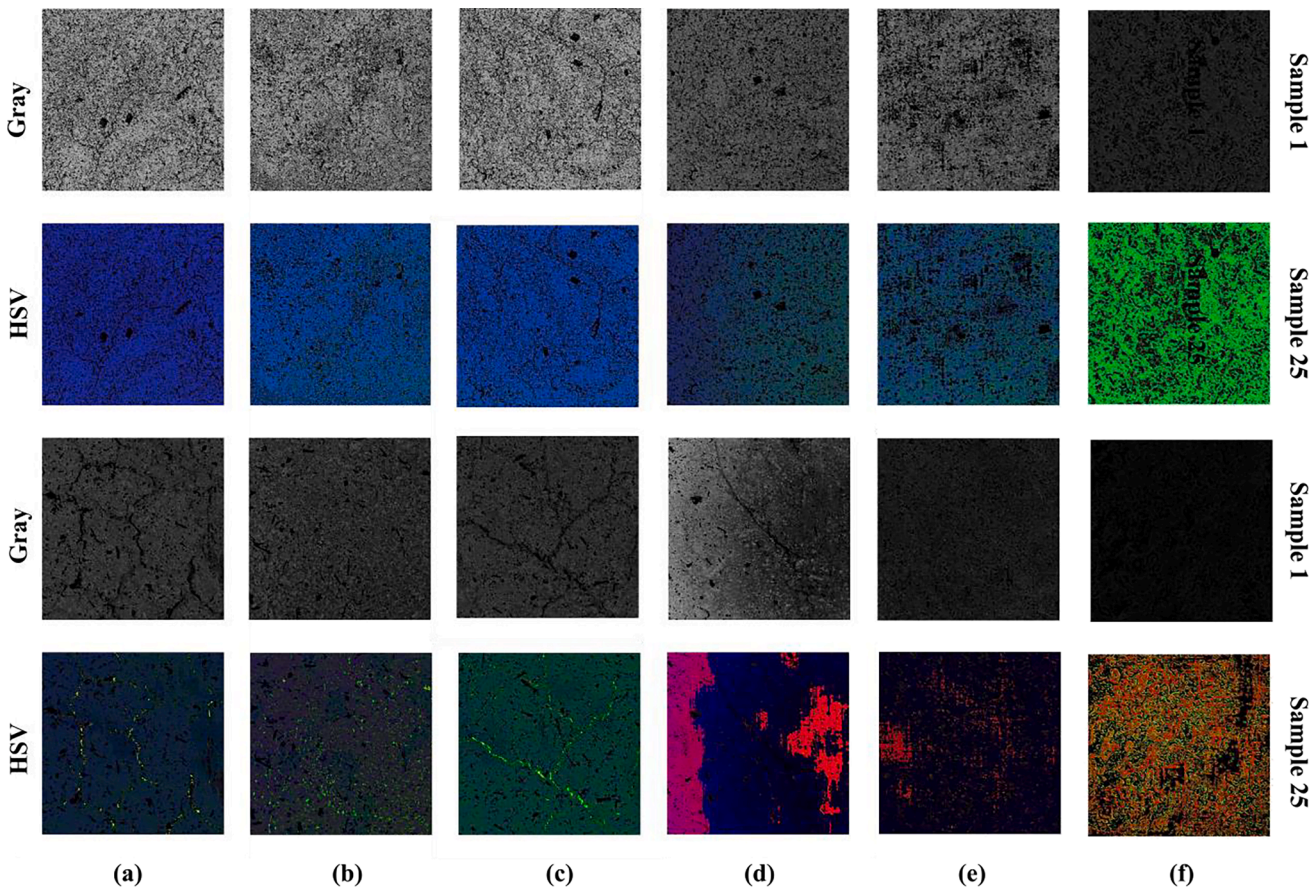

Fig. 9. Images obtained after color space conversions of two soil samples (sample 1 and sample 25 with 3.3% and 62.7% SOM respectively). (a) "Group 1"; (b) "Group 2"; (c) "Group 3"; (d) "Group 4"; (e) "Group 5"; and (f) "Group 6".

Table 1
Overview of features extracted from the images. R, G and B denote the Red, Green and Blue Plane of the RGB color space, respectively.

S. No.	Feature	Description
1.	Mean	Average of values of all pixels in an image
2.	Median	Middle pixel value after all the pixels are sorted in numerical order
3.	Entropy	Statistical measure of randomness
4.	Contrast	Measure of intensity contrast between a pixel and its neighbor over the whole image
5.	Energy	Sum of squared elements in the gray level co-occurrence matrix (GLCM)
6.	Homogeneity	Closeness of distribution of elements in the GLCM to the GLCM
7.	Redness Index, RI	$\frac{R^2}{B \times G^3}$
8.	Colouration Index, CI	$\frac{R-G}{R+G}$
9.	Hue Index, HI	$\frac{2 \times R - G - B}{G - B}$
10.	Saturation Index, SI	$\frac{R-B}{R+B}$

Mean V, Median H, Median S and Median V (from HSV images); Mean Gray, Median Gray, Entropy, Contrast, Energy and Homogeneity (from monochrome images). Some studies utilized mean values (Viscarra Rossel et al., 2008) of channels while others employed median values (dos Santos et al., 2016; Persson, 2005; Viscarra Rossel et al., 2008) in their calculations. This study used both. Persson (2005) recommends utilization of median as a technique to deal with the deviations brought by shading of the microrelief developed on the surfaces of the samples of soil. More scientifically, the effects of Bidirectional Reflectance Distribution Function (BRDF) and shading influences (for these indices are mainly ratio indices) that may exist on the images from the viewing angle and the direction of incident light (King, 1995; Lillesand et al., 2015) can be brought down with indices like RI, CI, HI, and SI (Levin et al., 2005). This may be possible as these indices constructively balance the variation in brightness and highlights the color content of the samples.

2.3. Data analysis

2.3.1. Data preprocessing and division

Multivariate outliers in the data were identified following the Mahalonobis distance approach (De Maesschalck et al., 2000). Under this technique, regression approaches are employed to decide if a specific case belonging to a sample population is an outlier through the combination of $\geq$2 variable scores. A total of 12 images were identified as outliers and were not considered in further analysis. The remaining dataset of 134 images were then split into calibration (70%, 94 images) and validation set (30%, 40 images) using the Kennard-Stone algorithm (Kennard and Stone, 1969) (Table 2).

2.3.2. Model development

The image features (a total of 22 color and texture features) were then used to develop predictive relationship against laboratory measured SOM and SMC. A total of 24 predictive models under six broad groups: **I) Linear Regression Models** attempt to build the relationship between response variable and observed variables whose model parameters are linear in nature (Myers, 1990). **II) Regression Trees** are decision trees which have binary splits for regression wherein the

Table 2
Descriptive statistics of the whole, calibration, and validation dataset for SOM (%) and SMC (%).

Scope		SOM (%)				SMC (%)			
	Count	Min	Max	Mean	SD	Min	Max	Mean	SD
All	146	3.3	62.7	19.53	18.2	0	119.6	24.35	28.08
Without Outliers	134	3.3	62.7	18.01	16.74	0	101.37	21.05	24.51
Calibration	94	3.3	62.7	20.26	17.91	0	101.37	24.7	25.35
Validation	40	3.3	51.6	12.71	12.23	0	77.69	12.46	20.21

response variable can acquire continuous values (Breiman, 2017). **III) Support vector machine (SVM)** is a common supervised machine learning algorithm for regression and classification. It was introduced by Boser, Guyon and Vapnik in 1992 (Vapnik, 1995). Since it is dependent on kernel functions, it is classified as a non-parametric technique. Using MATLAB, we can execute linear epsilon-insensitive SVM (ε-SVM) regression, also referred as L1 loss. It works on the principle of finding a function which deviates from observed response values by a number smaller than ε for each point of training dataset, while trying to stay as flat as it can. **IV) Gaussian Process Regression Models** are nonparametric kernel-based probabilistic models wherein the model responses are based on employing a probability distribution over a space of functions (Rasmussen and Nickisch, 2010a). **V) Ensembles of Trees** combines several decision trees to improve the prediction performance including boosted trees, bagged trees. Random forest (Breiman, 2001) introduced the concept of random forest, which tries to take benefit from random feature selection in addition to bagging. When growing a tree in a random forest, each node is split utilizing a best selection amongst a subset of features picked randomly at that node. Decision trees are grown until a specific number of nodes is reached which can be predetermined by the user (Douglas et al., 2018). The cubist method, which is prediction-oriented rule–based regression model. It is a combination of ideas of Quinlan's M5 model tree wherein the prediction depends on terminating leaves consisting of linear regression models (Minasny and McBratney, 2008b). **VI) Artificial Neural Network (ANN)** approaches are inspired the architecture of the neural network in the human brain. It is basically dependent on "learning by doing" approach which utilizes patterns of activity spread over a large system consisting of processing components like that of a neuron locally communicating by means of a group of unidirectional weighted connections (Zurada, 1992). A brief description of each sub-type of these models is given in supplementary Table S1, while the details can be found elsewhere. Codes were written in MATLAB to run these models except for Cubist model, which was run in R program (Version 3.5.3) on RStudio (Team, 2015).

2.3.3. Model performance assessment

Initially, 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 (%). A 10-fold cross-validation was performed as internal validation. 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 the development of pedotransfer functions. Several statistical parameters including R^2 (coefficient of determination), RMSE (Root Mean Square Error), LCCC (Lin's Concordance Correlation Coefficient), bias (mean of the residuals), RPD (Ratio of Performance to Deviation) and RPIQ (Ratio of Performance to Interquartile Distance) were computed. The RPD (Chang et al., 2001) is defined as the ratio of standard deviation of observed or measured values to the standard error of prediction and is used as an indicator of quality of the model. Models with RPD > 2 are often considered to represent good quantitative models. Since, RPD is closely associated to R^2 (Minasny and McBratney, 2013), especially for data with a normal distribution and having a large sample size, it was thus, suggested to use RPIQ instead (as reported by (Bellon-Maurel et al., 2010), which provides a better representation of the spread of values in the dataset. The values of RPIQ are generally inside a factor of 2 of RPD values, and hence, it is feasible (but not to be encouraged) to use RPIQ > 2 to denote good quantitative model (Aitkenhead et al., 2016). Also, since SOM and SMC are continuous in nature and can attain numerous values, using ordinary R^2 along with other statistical parameters was deemed suitable for this study. Small values of Root Mean Square Error (RMSE), bias and large values of coefficient of determination (R^2), Lin's Concordance Correlation Coefficient (LCCC) represent higher prediction accuracies.

Table 3
Descriptive statistics of SOM, SMC, and soil color measurements. SD and CV refer to standard deviation and coefficient of variation, respectively.

Parameter	Mean	SD	Range	Skewness	CV (%)
Mean R	0.25	0.12	0.02–0.48	−0.05	46.94
Mean G	0.24	0.11	0.02–0.43	−0.12	45.74
Mean B	0.20	0.09	0.02–0.39	−0.17	45.19
Mean H	0.11	0.07	0.04–0.54	4.05	63.36
Mean S	0.16	0.09	0–0.45	0.63	53.76
Mean V	0.25	0.12	0.03–0.48	−0.04	46.55
Mean gray	0.24	0.11	0.03–0.43	−0.11	45.89
Median R	0.29	0.15	0.02–0.63	0.18	52.49
Median G	0.27	0.14	0–0.56	0.16	51.60
Median B	0.23	0.12	0.02–0.47	0.15	51.73
Median H	0.11	0.09	0–0.67	4.39	78.08
Median S	0.18	0.11	0–0.58	0.58	62.06
Median V	0.29	0.15	0–0.63	0.19	52.24
Median Gray	0.27	0.14	0.02–0.56	0.16	51.66
Entropy	4.80	0.45	3.24–5.62	−0.73	9.42
Contrast	1.33	1.07	0–4.96	0.97	80.70
Energy	0.24	0.23	0.09–1	2.08	96.63
Homogeneity	0.81	0.07	0.69–1	0.93	9.27
RI	81.53	152.04	6.2–1030.62	3.26	186.49
CI	0.03	0.03	−0.04–0.19	1.05	102.42
HI	1.73	2.18	−17.74–5.06	−5.85	126.20
SI	0.10	0.08	−0.09–0.45	0.79	77.92
SOM (%)	18.01	16.74	3.3–62.7	1.23	92.97
SMC (%)	21.05	24.51	0–101.37	1.01	116.46

2.3.4. Variable screening to identify optimum predictors

A z-score was defined based on the performance of each predictor following six different analysis: ANOVA (Analysis of Variance), RF (Random Forest), Cubist, PCA (Principal Component Analysis), Vtreat variable reduction and Correlation analysis. Each predictor was rated on a scale of 0 (least important) to 100 (most important) and then averaged to get a z-score. While Cubist and RF provide variable importance on a scale of 0–100, we converted the result to the same scale range for the analysis that did not produce variable importance on a scale 0 to 100. For example, the coefficient for each principle component from PCA was multiplied by the proportion of variance explained by that specific component. The sum of values for all the principal components was computed for each predictor variable and scaled at 0 to 100 with the minimum value being assigned to 0 and the maximum value being assigned to 100. Correlation analysis was simply a 1:1 correlation between the dependent variable and each predictor. The absolute values of the correlation coefficients were first calculated and were scaled at 0 to 100, with 0 and 100 being assigned to the lowest and the highest absolute correlation coefficient, respectively. For ANOVA, the p-value for each predictor variable was scaled to 0–100 with 0 and 100 being

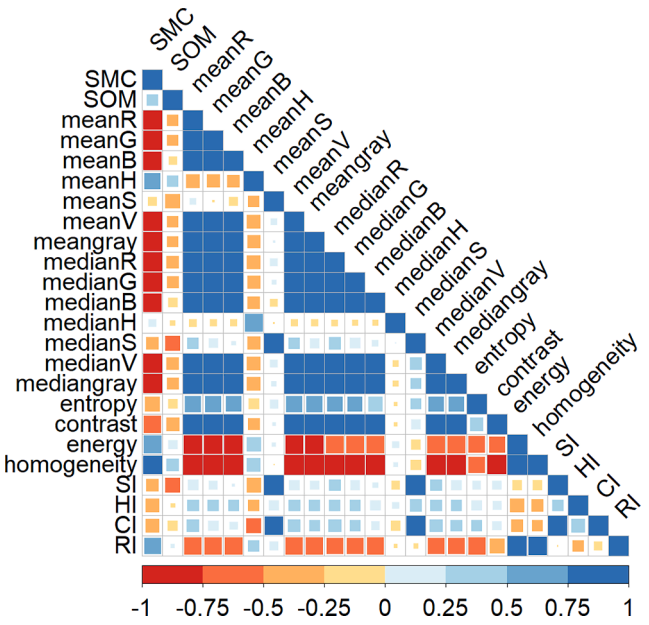

Fig. 10. Correlation plot for SOM, SMC, color space model features and indices derived from them.

assigned to the lowest and the highest p-value, respectively. 'Vtreat' is a R package for looking at the variable importance/significance. The values of R^2 were scaled at 0 to 100, with the lowest and highest value being assigned 0 and 100, respectively. These 0 to 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 particular predictor. The top six (6) predictor variables were then identified as the optimum predictors for both SOM and SMC. Like before, all the models were developed using these 6 predictors as independent variables and the model assessments were carried out.

3. Results

3.1. Descriptive statistics of the soil properties

Table 3 documents descriptive statistics for soil properties (SOM and SMC) and various image features. A relatively high degree of variation with the coefficient of variation, CV (%) varying between 9.27 and 186.49 was observed for image features and soil properties (Table 3). Organic matter ranged between 3.3% and 62.7% with a mean value of 18.01% and a standard deviation of 16.74%. These highly variable soil samples were chosen to ensure the universality of the results we get from this study. Owing from high SOM, a vast distribution was observed for SMC varying between 0% and 101.37% with a mean of 21.05% and a standard deviation of 24.51%. With an acceptable approximation, all image features except Mean H, Median H, Energy, RI, and HI were normally distributed (skewness approximately between −1 and 1) (Table 3). The RI exhibited a very large CV of about 186.49%. On the other hand, homogeneity comparatively varied less significantly, with a CV of around 9.27%.

3.2. Linear correlation between SOM, SMC, and soil color

Soil color alone is not a functional characteristic of soil. Thus, its utility was assessed by considering the association between digital measurements of soil color, SOM, and SMC. Soil color was highly correlated to SMC, while exhibited relatively weaker correlations to SOM (Fig. 10). The correlation between SMC and median R values was high (r = −0.84) followed by median V (r = −0.83) and median gray (r = −0.82). Median H values were weakly correlated to SMC (r = 0.13). SOM exhibited the strongest correlation with Median S (r = −0.53),

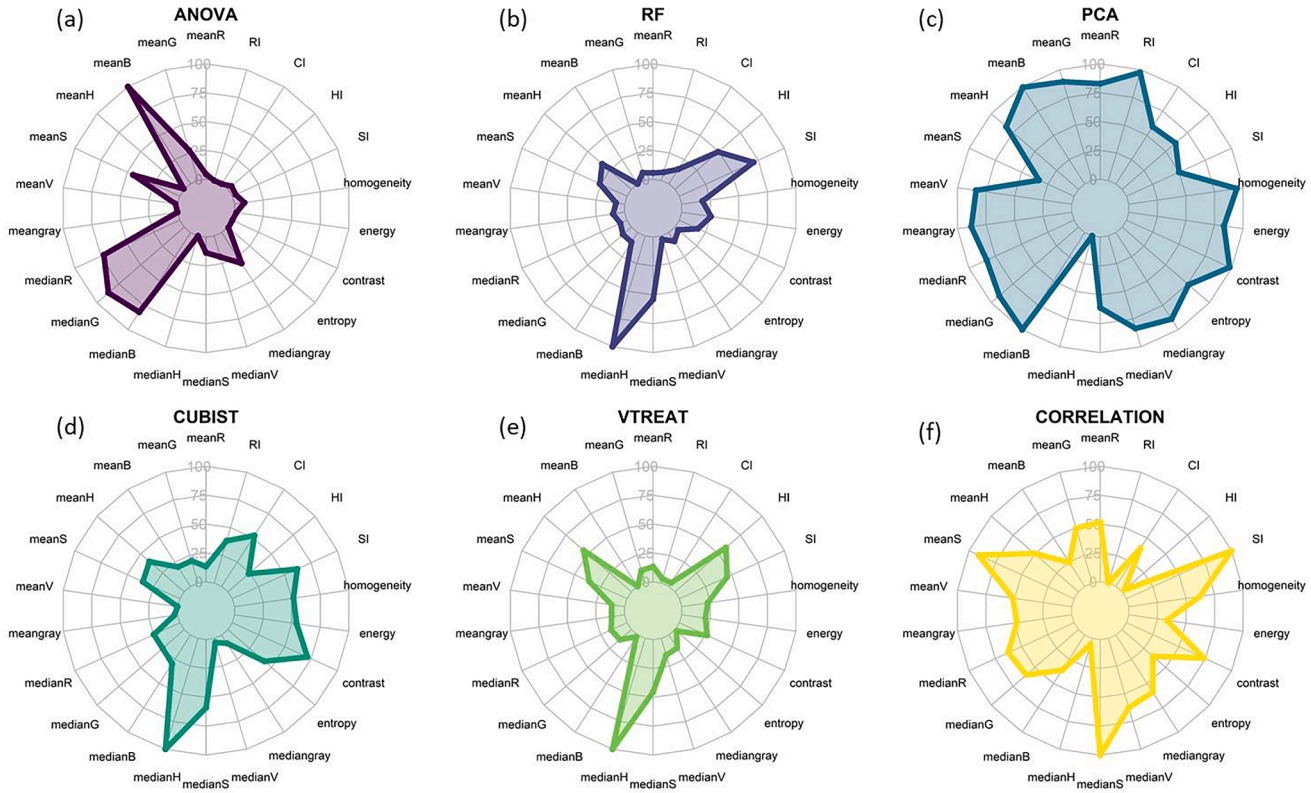

Fig. 11. Relative significance of each individual image feature as a predictor variable for SOM prediction corresponding to (a) ANOVA; (b) RF; (c) PCA; (d) Cubist; (e) Vtreat; (f) Correlation.

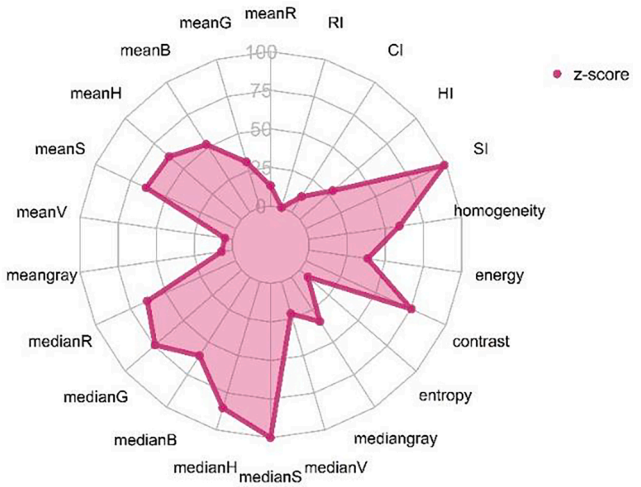

Fig. 12. z-Score of each individual image feature representing its contribution towards SOM prediction.

followed by SI (r = −0.52) and mean S (r = −0.47). RI values exhibited the lowest correlation (r = 0.03). In general, the reflection intensity decreased with the increase in SOM and SMC. Correlation between and among color and texture features were also significant in many cases.

3.2.1. Identification of optimum predictors

Relative importance of predictor variables in predicting SOM and SMC following 6 different analysis are presented in radial plots (Figs. 11–14). In general, color features played a critical role in predicting SOM and SMC than textural features. HSV and RGB channels were more important in predicting SOM and SMC, respectively than other channels. In addition, median values of the channels were more important in the prediction than mean values.

Median S was identified as the most important variable in predicting SOM followed by SI, Median H, Contrast, Median G and Mean S. The least important variable was RI (Fig. 12). Median R and Median V were identified as the most important predictor variables in predicting SMC followed by Mean G, Median B, Median Gray and Mean B. The least important variable was Mean S (Fig. 14).

3.3. Predictive accuracy of the models

Models developed with 22 and 6 image features were regression

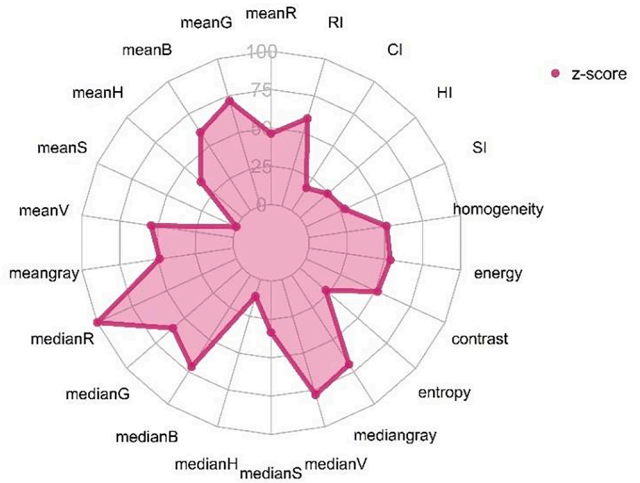

Fig. 14. z-Score of each individual image feature representing its contribution towards SMC prediction.

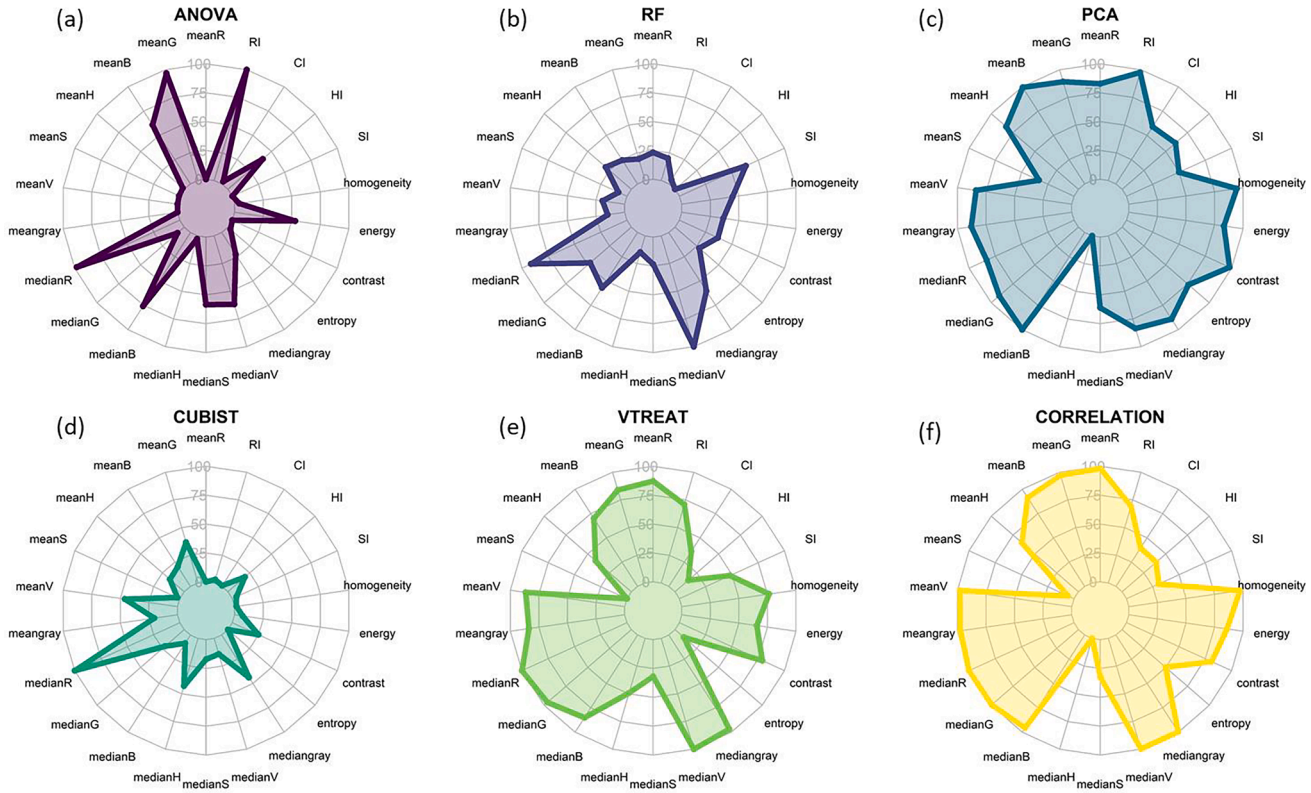

Fig. 13. Relative significance of each individual image feature as a predictor variable for SMC prediction corresponding to (a) ANOVA; (b) RF; (c) PCA; (d) Cubist; (e) Vtreat; (f) Correlation.

Table 4
The highest prediction accuracy for SOM resulting in each model category using 22 and 6 predictor variables. These are results of 10-fold cross-validation internal validation (IV) and external validation (EV). RPD: Ratio of Prediction to Deviation, and RPIQ: Ratio of Performance to Interquartile Distance. Values in bold indicate the best performing result.

Model	Sub-Type	R^2		LCCC		RMSE		Bias		RPD		RPIQ	
		IV	EV	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV
Using 22 predictor variables													
LR	Stepwise	0.51	0.08	0.70	0.26	14.21	14.49	0.66	−3.93	1.26	0.84	1.53	0.68
RT	Medium Tree	0.54	0.39	0.71	0.60	12.13	9.93	−0.77	−0.37	1.48	1.23	1.79	1.00
SVM	Medium Gaussian	0.57	**0.56**	0.64	0.60	12.35	**8.59**	−2.65	−1.75	1.45	**1.42**	1.75	1.16
GPR	Rational Quadratic	0.65	0.49	0.79	0.65	10.64	8.79	0.29	−1.27	1.68	1.39	2.04	1.13
ET	Cubist	0.73	0.49	0.73	0.59	9.62	8.95	−0.33	−2.00	2.27	1.37	2.31	1.11
Other	ANN	**0.91**	0.54	0.94	0.72	**5.45**	8.84	−0.04	−1.15	**3.29**	1.38	3.98	1.12
Using 6 predictor variables													
LR	Pure Quadratic	0.53	0.46	0.70	0.65	11.58	9.22	0.58	1.19	1.28	1.33	1.33	1.08
RT	Medium Tree	0.37	0.40	0.59	0.61	12.73	9.78	1.53	−0.38	1.16	1.25	1.21	1.01
SVM	Medium Gaussian	0.54	0.50	0.70	0.64	10.05	8.53	−0.42	−0.38	1.48	1.43	1.54	1.16
GPR	Matern 5/2	0.51	**0.53**	0.69	0.68	10.65	**8.27**	−0.16	0.10	1.39	**1.48**	1.45	1.20
ET	Cubist	**0.72**	0.42	0.72	0.61	**9.90**	9.59	0.37	−1.78	**2.49**	1.28	2.24	1.03
Other	ANN	0.69	0.43	0.80	0.62	9.89	9.43	0.68	1.09	1.81	1.30	2.19	1.05

LR: Linear Regression; RT: Regression Trees; SVM: Support Vector Machines; GPR: Gaussian Process Regression; ET: Ensembles of Trees

Table 5
The highest prediction accuracy for SMC resulting in each model category using 22 and 6 predictor variables. These are results of 10-fold cross-validation internal validation (IV) and external validation (EV). RPD: Ratio of Prediction to Deviation, and RPIQ: Ratio of Performance to Interquartile Distance. Values in bold indicate the best performing result.

Model	Sub-Type	R^2		LCCC		RMSE		Bias		RPD		RPIQ	
		IV	EV	IV	EV	IV	EV	IV	EV	IV	EV	IV	EV
Using 22 predictor variables													
LR	Linear	0.71	0.76	0.82	0.84	14.30	10.13	−1.71	−1.39	1.83	1.99	2.22	1.86
RT	Fine Tree	0.81	0.80	0.88	0.87	11.34	8.88	−0.97	−0.39	2.31	2.27	2.80	2.13
SVM	Coarse Gaussian	0.78	0.85	0.82	0.89	13.07	7.72	−2.83	0.81	2.01	2.62	2.43	2.45
GPR	Exponential	**0.84**	**0.92**	0.90	0.93	**10.18**	**5.79**	0.05	−0.65	2.49	3.49	4.27	3.26
ET	Random Forest	0.81	0.92	0.88	0.92	10.99	6.01	−0.54	−0.74	2.31	3.36	3.95	3.14
Other	ANN	0.75	0.77	0.78	0.81	18.00	12.43	11.12	6.88	1.41	1.63	2.41	1.52
Using 6 predictor variables													
LR	Interactions Linear	0.80	0.91	0.88	0.93	11.35	6.24	0.34	−0.54	2.23	3.24	3.83	3.03
RT	Medium Tree	0.68	0.84	0.80	0.88	14.86	8.14	−0.69	−0.78	1.76	2.48	2.13	2.32
SVM	Quadratic	0.76	0.93	0.83	0.94	13.19	5.56	−3.78	1.94	1.99	3.63	2.41	3.39
GPR	Matern 5/2	0.74	**0.95**	0.84	0.94	13.36	**5.04**	−1.13	−0.95	1.96	4.01	2.37	3.75
ET	Cubist	**0.86**	0.93	0.80	0.93	**10.43**	5.68	−0.74	−0.64	2.92	3.56	4.72	3.32
Other	ANN	0.80	0.94	0.87	0.93	12.00	5.93	3.92	3.17	2.11	3.41	3.62	3.19

LR: Linear Regression; RT: Regression Trees; SVM: Support Vector Machines; GPR: Gaussian Process Regression; ET: Ensembles of Trees

calibrated and validated against the laboratory measured SOM and SMC. The detail prediction statistics obtained using 22 and 6 predictor variables for SOC is presented in supplementary Tables S2 and S3, respectively. The supplementary Tables S4 and S5 presents the prediction statistics for SMC. The summary of best performing model in each category for the prediction of SOC and SMC are presented in Tables 4 and 5, respectively.

3.3.1. Prediction of SOM

3.3.1.1. Predictive accuracy of the models using 22 predictor variables

3.3.1.1.1. 10-fold cross (internal) validation. The accuracy of the models varied greatly in predicting SOM (Table 4 and S2). The most accurate predictions were obtained using ANN (Fig. 15a and Table 4). The R^2, RMSE, LCCC, bias, RPD and RPIQ values were 0.91, 5.45%, 0.94, −0.04, 3.29 and 3.98 respectively (Table 4). The next best predictions were produced by Cubist model with R^2, RMSE, LCCC, bias, RPD and RPIQ values of 0.73, 9.62%, 0.73, −0.33, 2.27 and 2.315, respectively. However, the Support vector machines, Regression trees and Linear regression performed poorly ($R^2 > 0.60$).

3.3.1.1.2. External validation. The R^2 value for the model trained using Medium Gaussian SVM producing best predictions was 0.56, the RMSE was 8.59%, the LCCC was 0.60, the bias was −1.75, the RPD was 1.42 and the RPIQ was 1.16 (Fig. 15b and Table 4). The performance of ANN for the test dataset was comparable but relatively weaker, with R^2 of 0.54 and RMSE of 8.84%. The LCCC was 0.72, the bias was −0.04, the RPD was 1.38 and the RPIQ was 1.12. On the other hand, the poorest predictions were produced by Stepwise Linear Regression model giving an R^2, RMSE, LCCC, bias, RPD and RPIQ values of 0.08, 14.49%, 0.26, −3.93, 0.84 and 0.68, respectively.

3.3.1.2. Predictive accuracy of the models using 6 predictor variables

3.3.1.2.1. 10-fold cross (internal) validation. After variable reduction, the R^2 values for the calibration dataset decreased for 20 out of 24 models (Tables S2 and S3). The Cubist model produced the most accurate predictions with R^2, RMSE, LCCC, bias, RPD and RPIQ of 0.72, 9.90%, 0.72, 0.37, 2.49 and 2.24, respectively (Fig. 15c and Table 4). The second highest prediction accuracy for calibration dataset was obtained using ANN model with R^2, RMSE, LCCC, bias, RPD and RPIQ of 0.69, 9.89%, 0.80, 0.68, 1.81 and 2.19. On the contrary, the Regression Tree (Medium Tree) trained model produced poor predictions with R^2, RMSE, LCCC, bias, RPD and RPIQ of 0.37, 12.73%, 0.59, 1.53, 1.16 and 1.21, respectively.

3.3.1.2.2. External validation. Reduction in the predictor variables from 22 to 6 resulted in an increase in predictive accuracy for 19 out of

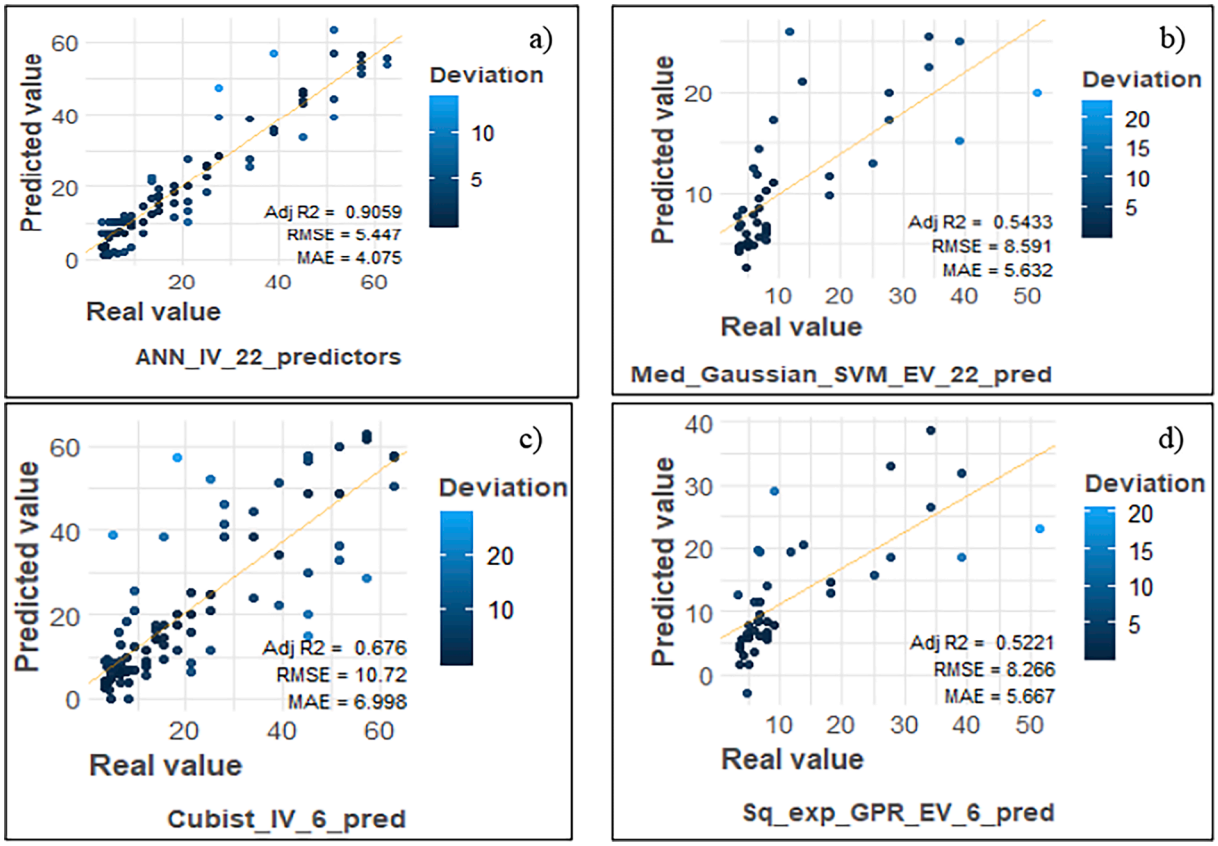

Fig. 15. The best prediction accuracy for SOM using 22 predictor variables: a) Calibration b) Validation dataset, and 6 predictor variables - c) Calibration d) Validation dataset.

24 models, since the R^2 values for them showed an increase or remained the same (Tables S2 and S3). The most accurate predictions were obtained by Squared Exponential GPR, with R^2, RMSE, LCCC, bias, RPD and RPIQ of 0.53, 8.27%, 0.69,-0.09, 1.48 and 1.20, respectively (Fig. 15d and Table 4). The least accurate predictions were those produced by Medium Tree model with R^2, RMSE, LCCC, bias, RPD and RPIQ of 0.40, 9.78%, 0.61, −0.38, 1.5 and 1.01, respectively (Table 4). The SVM and GPR models performed relatively constant for calibration and validation dataset ($R^2 > 0.50$). However, ANN and SVM are the only two model types which did not improve the prediction accuracy for validation dataset when predictor variable reduced to 6 from 22 (Table 4).

3.3.2 Prediction of SMC

3.3.2.1 Predictive accuracy of the models using 22 predictor variables

3.3.2.1.1. 10-fold cross (internal) validation. All the model trained using calibration dataset predicted the SMC with high accuracy ($R^2 > 0.70$). The Exponential GPR model produced the best predictive relationship between SMC and soil color and texture features with$R^2 = 0.84$, RMSE = 10.18%, LCCC = 0.90, bias = 0.05, RPD = 2.49 and RPIQ = 4.27 (Fig. 16a and Table 5). The Linear Regression model obtained relatively low accuracy with R^2 of 0.71 and an RMSE of 14.30% while the LCCC, bias, RPD and RPIQ were 0.82, −1.71, 1.83 and 2.22, respectively.

3.3.2.1.2. External validation. Nearly all the model types (Linear, Regression Trees, SVM, GPRs, Ensemble of Trees, and ANN) predicted SMC using the validation dataset with higher accuracy than the calibration dataset. The highest R^2 for calibration dataset was 0.84 and further improved to 0.92 (for validation) using the exponential GPR model. The RMSE, LCCC, bias, RPD and RPIQ were 5.79%, 0.93, −0.65, 3.49 and 3.26, respectively (Fig. 16a-b and Table 5). Similar to SOM, Linear models again account for the relatively poor prediction accuracy. The Interaction Linear Model was with R^2, RMSE, LCCC, bias, RPD and RPIQ of 0.43, 619.12%, −0.04, −120.25, 0.03 and 0.03 respectively (Table S4).

3.3.2.2 Predictive accuracy of the models using 6 predictor variables

3.3.2.2.1. 10-fold cross (internal) validation. There was a slight decrease in the predictive accuracy for 12 out of 24 models, for the calibration dataset, as was evident from the decrease in R^2 values (Table S5). The Cubist model produced the best predictive relationship between SMC and soil color and texture features with$R^2 = 0.86$, RMSE = 10.43%, LCCC = 0.80, bias = −0.74, RPD = 2.92 and RPIQ = 4.72 (Fig. 16c and Table 5). However, the Regression Tree (Medium Tree) model exhibited relatively poor predictive performance with an R^2 of 0.68 and an RMSE of 14.86% while the LCCC, bias, RPD and RPIQ were 0.80, −0.69, 1.76 and 2.13, respectively (Table 5).

3.3.2.2.2. External validation. Overall, excellent predictions were obtained for most of the models. Utilizing R^2 and RPD to evaluate the model performance also produced similar results, with validation $R^2 > 0.80$ and RPD > 2, for all the calibrated models (Table 5). The R^2 value for the model trained using Matern 5/2 GPR produced best predictions were 0.95, the RMSE was 5.04%, the LCCC was 0.94, the bias was −0.95, the RPD was 4.01 and the RPIQ was 3.75 (Fig. 16d and Table 5).

4. Discussion

For SOM, there was no significant reduction in the prediction accuracy obtained following the removal of insignificant predictor variables from all the models. The new brief models with reduced predictor variables were as accurate as the ones developed with all the variables, suggesting that the entire set of 22 predictor variables was not needed. Thus, it also necessary to identify few useful image features for

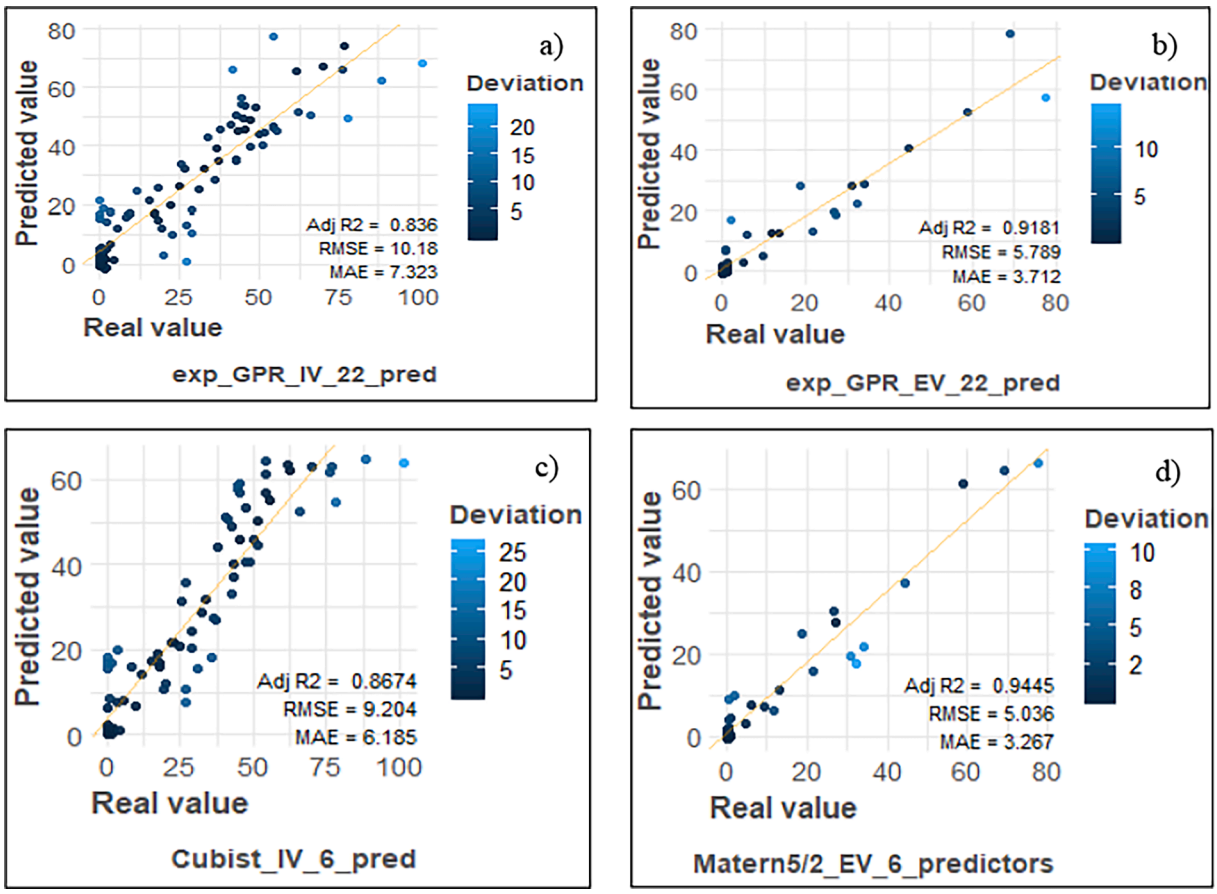

Fig. 16. The best prediction accuracy for SMC using 22 predictor variables: a) Calibration b) Validation dataset, and 6 predictor variables - c) Calibration d) Validation dataset.

reasonable approximation of SOM and SMC. This redundancy also forms the rationale to investigate the data compression techniques required to analyze the data.

There was a huge difference in the accuracies of prediction provided by the regression and machine learning methods for both SOM and SMC. For prediction of SOM, ANN provided the best predictions using calibration dataset containing 22 predictor variables. However, it could not sustain its performance for the test dataset prediction whereas SVM performed well in that case (Table 4). This could be because SVMs with proficient pattern recognition deal with local minima which are ordinary issues associated with the training process of ANNs (Haghverdi et al., 2014; Lamorski et al., 2008). In another terms, structural risk optimization facilitated in SVMs helps in the reduction of size of the models and errors in the prediction, whereas ANNs possess a predefined structure directed only towards minimizing error on training data (Elbisy, 2015; Vapnik et al., 1997).

Amongst all types of trees (or rule-based decision methods), Cubist outperformed individual and ensembles of trees. The reason for its success may be attributed to the fact that it separates data well and forms multivariate linear models at terminal nodes rather than distinct values (unlike individual trees). Its performance was also superior to RF. This could be because Cubist produces more continuous predicted values, since in Cubist, each linear model at the terminal node permits a smooth transition in the prediction among trees (Dangal et al., 2019). Whereas, a subset of input data is chosen randomly via bagging, in RF, and the final prediction is based on the mean of outputs from all discrete trees (Breiman, 2001). Even though, Cubist provided a good model fit for the calibration dataset (for both SOM and SMC with 6 predictor variables), it did not perform well for the validation dataset. After careful inspection, it was noticed that Cubist tends to predict poorly at higher values (Doetterl et al., 2015; Minasny and McBratney, 2008a).

Gaussian process regression models demonstrated excellent predictive capability for both calibration datasets (for SOM with 22 predictor variables) and validation datasets (for SMC with 22 predictor variables and for both SOM and SMC with 6 predictor variables). GPRs presume that adjoining observations should impart information about each other (Pal and Deswal, 2010). They are alike in their performance to SVMs since they use kernel functions in addition to nonparametric basic. However, they yield reliable responses to the input data provided, which increases their reliability as a probabilistic model (Rasmussen and Nickisch, 2010b). Nevertheless, one thing should be pointed out here about these methods. In this study using a small dataset, we found certain methods performed better. However, the best method identified for this dataset may not be the best one for data from other sites. Thus, exploration of methods should be beneficial for future examples of this work (and that automated searching for the best option using R or other software would be an important area of future research) (Aitkenhead et al., 2016).

An implication of this study is that it can help in selecting promising image features and appropriate regression/ machine learning model which can together potentially be extended to and implemented in a new software environment (for instance, a standalone app or an online platform) such as SOCiT (Aitkenhead et al., 2013; Donnelly et al., 2013) or improve the available ones through collaboration (Aitkenhead et al., 2016; Donnelly et al., 2013; Swetha et al., 2020). The long-term goal is to provide a tool to farmers and land managers for in situ analysis of SOM and SMC content and possibly along with soil texture. The app would rely on color information of soil sample alone obtained using inbuilt camera of a cell phone to provide for reasonable predictions of SOM and SMC both in situ and ex situ regardless of the need to use any

supporting device or possess background knowledge about the soil profile or without the requirement of any sample preparation in the laboratory (for in situ application). In this study, SMC was measured alongside with SOM and predicted using the selected algorithms to estimate them separately, SMC and SOM both has joint influence on color. Thus, their joint impact (Fu et al., 2019) should be a focus of future studies.

High prediction accuracies also implied that images collected were of good/reasonable quality and image preprocessing for the removal of noise (deduction of non-soil pixels) in the form of image cropping, enhancement and segmentation proved to be successful and facilitated model development.

Cell phones, nowadays, have become increasingly popular and have revolutionized and replaced many devices including digital cameras. In this study, cell phones exhibited the potential to be used for image-based soil property characterization. They demonstrated the ability to capture images as good as a digital camera because similar SOM and SMC prediction accuracies were obtained with cell phone images as that were obtained with digital cameras images.

5. Conclusions

This study evaluated several models in terms of their performance towards rapid and reasonable estimation of SOM and SMC from a set of cell-phone images collected in the laboratory. The aim is to contribute towards development of a proximal soil sensor using cell phone for rapid and cost-effective characterization of soil properties without analyzing them in laboratory following traditional techniques that are costly, time-consuming, and labor-intensive. A laboratory experiment was carried out with soil samples from two agricultural fields with highly variable SOM. Images were captured using a cell phone at variable soil moisture contents to simulate the continuous variation of soil moisture in the field. Models were developed based on color and texture features derived from the images first using all 22 extracted features and then using 6 best features for both SOM and SMC separately. The dark color of soils can be attributed to both high SOM content as well as high moisture content in the soil. Color features demonstrated high correlation with both SOM and SMC. For the models trained using only 6 best predictor variables, both SOM and SMC were again modelled with approximately the same accuracy as compared to that with all the 22 predictors. The only exception was for internal validation predictions of SOM which considerably reduced prediction accuracy. Overall, Gaussian Process Regression and Cubist models best captured and explained the non-linear relationships between SOM, SMC, and image features. However, this study only considered SOM and SMC prediction separately. Future studies should focus on teasing out interactive relationships between SOM and SMC on soil color to improve the prediction of SOM. Similarly, soil color may also be contributed from the difference in soil type and soil texture. Though we included various soil types with large variations in SOM, impact of soil type differences or soil textural differences were not explicitly considered in this study and must be tested in future studies. Variations in soil color can also be related to other factors like topography, geology, climate and others and must be included in future studies, Nevertheless, the success of predicting SOM and SMC from images provides an opportunity to develop a proximal soil sensor to be used further for easy, rapid, and cost-effective analysis characterization of soil properties including SOM and SMC.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This project was funded by grants to Asim Biswas from NSERC (Natural Sciences and Engineering Research Council of Canada, RGPIN-2014-04100). The authors would also like to thank Dr. Viacheslav Adamchuk and Dr. Wenjun Ji for involvement in the broader project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2020.114863.

References

Aitkenhead, M., Donnelly, D., Coull, M., Black, H., 2013. E-smart: environmental sensing for monitoring and advising in real-time. In: International Symposium on Environmental Software Systems. Springer, pp. 129–142.
Aitkenhead, M., Coull, M., Gwatkin, R., Donnelly, D., 2016. Automated soil physical parameter assessment using Smartphone and digital camera imagery. J. Imaging 2 (4), 35.
Barrett, L.R., 2002. Spectrophotometric color measurement in situ in well drained sandy soils. Geoderma 108 (1–2), 49–77.
Basso, B., Ritchie, J.T., Cammarano, D., Sartori, L., 2011. A strategic and tactical management approach to select optimal N fertilizer rates for wheat in a spatially variable field. Eur. J. Agron. 35 (4), 215–222.
Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC, Trends Anal. Chem. 29 (9), 1073–1081.
Ben-Dor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. Remote Sens. Environ. 61 (1), 1–15.
Bezdicek, D.F., Papendick, R.I., Lal, R., 1996. Introduction: Importance of soil quality to health and sustainable land management. in: Doran, J.W., Jones, A.J. (eds.) Methods for assessing soil quality SSSA Spec. Publ. 49, Madison. 1-18.
Blackmer, A.M., White, S.E., 1998. Using precision farming technologies to improve management of soil and fertiliser nitrogen. Aust. J. Agric. Res. 49 (3), 555–564.
Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.
Breiman, L., 2017. Classification and Regression Trees. CRC Press.
Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65 (2), 480–490.
Chapman, S.J., Bell, J.S., Campbell, C.D., Hudson, G., Lilly, A., Nolan, A.J., Robertson, A. H.J., Potts, J.M., Towers, W., 2013. Comparison of soil carbon stocks in Scottish soils between 1978 and 2009. Eur. J. Soil Sci. 64 (4), 455–465.
Chen, D., Chang, N., Xiao, J., Zhou, Q., Wu, W., 2019. Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. Sci. Total Environ. 669, 844–855.
Conant, R.T., Ogle, S.M., Paul, E.A., Paustian, K., 2011. Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. Front. Ecol. Environ. 9 (3), 169–173.
Dangal, S.R.S., Sanderman, J., Wills, S., Ramirez-Lopez, L., 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Syst. 3 (1), 11.
De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L., 2000. The mahalanobis distance. Chem. Intelligent Lab. Syst. 50 (1), 1–18.
Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Pinto, M.C., Casanova-Katny, A., Muñoz, C., Boudin, M., Venegas, E.Z., 2015. Soil carbon storage controlled by interactions between geochemistry and climate. Nat. Geosci. 8 (10), 780.
Doi, R., Ranamukhaarachchi, S., 2007. Soil colour designation using Adobe PhotoshopTM in estimating soil fertility restoration by Acacia auriculiformis plantation on degraded land. Curr. Sci. 92.
Donnelly, D., Aitkenhead, M.J., Coull, M.C., 2013. SOCiT Soil Carbon App for iPhone/Android.
dos Santos, J.F.C., Silva, H.R.F., Pinto, F.A.C., Assis, I.R.d., 2016. Use of digital images to estimate soil moisture. Revista Brasileira de Engenharia Agrícola e Ambiental 20 (12), 1051-1056.
Douglas, R.K., Nawar, S., Alamar, M.C., Mouazen, A.M., Coulon, F., 2018. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. Sci. Total Environ. 616, 147–155.
Dudley, R.J., 1975. The use of colour in the discrimination between soils. J. Forensic Sci. Soc. 15 (3), 209–218.
Elbisy, M.S., 2015. Support Vector Machine and regression analysis to predict the field hydraulic conductivity of sandy soil. KSCE J. Civ. Eng. 19 (7), 2307–2316.
Escadafal, R., Girard, M.-C., Courault, D., 1988. La couleur des sols: appréciation, mesure et relations avec les propriétés spectrales. Agronomie 8 (2), 147–154.
FAO, 2019. The Multi-Faced rôle of Soil in the Near East and North Africa Region- Policy Brief. Rome. 26pp. Licence:CC BY-NC-SA 3.0 IGO.
Fu, Y., Taneja, P., Lin, S., Adamchuk, V., Ji, W., Daggupati, P., Biswas, A., 2019. Predicting soil organic matter from cellular phone images under varying soil moisture. Geoderma. In Press.

Gelder, B.K., Anex, R.P., Kaspar, T.C., Sauer, T.J., Karlen, D.L., 2011. Estimating soil organic carbon in Central Iowa using aerial imagery and soil surveys. Soil Sci. Soc. Am. J. 75, 1821–1828.
Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines 1. J. Am. Water resour. Assoc. 42 (4), 1033–1046.
Gómez-Robledo, L., López-Ruiz, N., Melgosa, M., Palma, A.J., Capitán-Vallvey, L.F., Sánchez-Maraňón, 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., Eddins, S.L., 2004. Digital Image Processing using MATLAB. Pearson Education India.
Gregory, S.D., Lauzon, J.D., O'Halloran, I.P., 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 (3), 573–584.
Haghverdi, A., Öztürk, H.S., Cornelis, W.M., 2014. Revisiting the pseudo continuous pedotransfer function concept: impact of data quality and data mining method. Geoderma 226, 31–38.
Han, P., Dong, D., Zhao, X., Jiao, L., Lang, Y., 2016. A smartphone-based soil color sensor: for soil type classification. Comput. Electron. Agric. 123, 232–241.
Ji, W., Adamchuk, V.I., Biswas, A., Dhawale, N.M., Sudarsan, B., Zhang, Y., Rossel, R.A. V., Shi, Z., 2016. Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. Biosyst. Eng. 152, 14–27.
Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11 (1), 137–148.
King, D.J., 1995. Airborne multispectral digital camera and video sensors: a critical review of system designs and applications. Can. J. Remote Sens. 21 (3), 245–273.
Kotlar, A.M., Iversen, B.V., 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 (1).
Krishnan, P., Alexander, J.D., Butler, B.J., Hummel, J.W., 1980. Reflectance technique for predicting soil organic matter 1. Soil Sci. Soc. Am. J. 44 (6), 1282–1285.
Lamorski, K., Pachepsky, Y., Sławiński, C., Walczak, R.T., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Sci. Soc. Am. J. 72 (5), 1243–1247.
Levin, N., Ben-Dor, E., 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 (24), 5475–5492.
Liles, G.C., Beaudette, D.E., O'Geen, A.T., Horwath, W.R., 2013. Developing predictive soil C models for soils using quantitative color measurements. Soil Sci. Soc. Am. J. 77 (6), 2173–2181.
Lillesand, T., Kiefer, R.W., Chipman, J., 2015. Remote Sensing and Image Interpretation. John Wiley & Sons.
Lindbo, D.L., Rabenhorst, M.C., Rhoton, F.E., 1998. Soil color, organic carbon, and hydromorphology relationships in sandy epipedons. Quantifying Soil Hydromorphol. (quantifyingsoil) 95–105.
Lu, M., 2016. "A smartphone-based device for measuring soil organic matter" (2016). Leopold Center Completed Grant Reports. 517. http://lib.dr.iastate.edu/leopold_grantreports/517.
Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2010. Spatial distribution of soil organic carbon stocks in France: discussion paper. Biogeosci. Discuss.
Matei, O., Rusu, T., Petrovan, A., Mihuț, G., 2017. A data mining system for real time soil moisture prediction. Procedia Eng. 181, 837–844.
MathWorks, I., 2017. MATLAB 2017b. The MathWorks, Inc., Natick, MA, USA.
Melville, M.D., Atkinson, G., 1985. Soil colour: its measurement and its designation in models of uniform colour space. J. Soil Sci. 36 (4), 495–512.
Minasny, B., McBratney, A.B., 2008a. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chem. Intelligent Lab. Syst. 94 (1), 72–79.
Minasny, B., McBratney, A.B., 2008b. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemo. Intelli. Lab. Sys. 94 (1), 72–79.
Minasny, B., McBratney, A., 2013. Why you don't need to use RPD. Pedometron 33, 14–15.
Moonrungsee, N., Pencharee, S., Jakmunee, J., 2015. Colorimetric analyzer based on mobile phone camera for determination of available phosphorus in soil. Talanta 136, 204–209.
Munsell, A.H., 1994. Soil color charts, revised edn. Macbeth Division of Kollmorgan Instruments, New Windsor.
Nocita, M., Stevens, A., Noon, C., van Wesemael, B., 2013. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. Geoderma 199, 37–42.
Nordmeyer, H., 2015. Herbicide application in precision farming based on soil organic matter. Am. J. Exp. Agric. 8 (3), 144–151.
O'Halloran, I.P., von Bertoldi, A.P., Peterson, S., 2004. Spatial variability of barley (Hordeum vulgare) and corn (Zea mays L.) yields, yield response to fertilizer N and soil N test levels. Can. J. Soil Sci. 84 (3), 307–316.
Pal, M., Deswal, S., 2010. Modelling pile capacity using Gaussian process regression. Comput. Geotech. 37 (7–8), 942–947.
Persson, M., 2005. Estimating surface soil moisture from soil color using image analysis. Vadose Zone J. 4 (4), 1119–1122.
Rasmussen, C.E., Nickisch, H., 2010a. Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learning Res. 11 (Nov), 3011–3015.
Rasmussen, C.E., Nickisch, H., 2010b. Gaussian processes for machine learning (GPML) toolbox. J. Machine Learning Res. 11, 3011–3015.
Rienzi, E.A., Mijatovic, B., Mueller, T.G., Matocha, C.J., Sikora, F.J., Castrignanò, A., 2014. Prediction of soil organic carbon under varying moisture levels using reflectance spectroscopy. Soil Sci. Soc. Am. J. 78 (3), 958–967.
Rodionov, A., Pätzold, S., Welp, G., Damerow, L., 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 (3), 949–957.
Sakti, M.B.G., Komariah, Ariyanto, Sumani, D.P., 2018a. Estimating soil moisture content using red-green-blue imagery from digital camera. IOP Conf. Series: Earth Environ. Sci. 200.
Sakti, M.B.G., Komariah, K., Ariyanto, D.P., Sumani, 2018b. Estimating soil moisture content using red-green-blue imagery from digital camera. IOP Conf. Series: Earth Environ. Sci. 200.
Schulte, E.E., Hopkins, B.G., 1996. Estimation of soil organic matter by weight loss-on-ignition. In Magdoff, F. R. et al. (eds.) Soil Organic Matter: Analysis and Interpretation. SSSA Spec. Pub. No. 46. SSSA, Madison. pp. 21–31.
Schulze, D.G., Nagel, J.L., Van Scoyoc, G.E., Henderson, T.L., Baumgardner, M.F., Stott, D.E., 1993. Significance of organic matter in determining soil colors. in: Bigham, J. M., Ciolkosz, E.J. (eds.) Soil color. SSSA Spec. Publ. 31. SSSA, Madison, WI.
Steinhardt, G.C., Franzmeier, D.P., 1979. Comparison of organic matter content with soil color for silt loam soils of Indiana. Commun. Soil Sci. Plant Anal. 10 (10), 1271–1277.
Stiglitz, R., Mikhailova, E., Post, C., Schlautman, M., Sharp, J., 2016. Evaluation of an inexpensive sensor to measure soil color. Comput. Electron. Agric. 121, 141–148.
Swetha, R.K., Bende, P., Singh, K., Gorthi, S., Biswas, A., Li, B., Weindorf, D.C., Chakraborty, S., 2020. Predicting soil texture from smartphone-captured digital images and an application. Geoderma 376.
Team, R., 2015. RStudio: integrated development for R. RStudio, Inc., Boston, MA URL http://www. rstudio. com 42, 14.
van Wesemael, B., Paustian, K., Andrén, O., Cerri, C.E.P., Dodd, M., Etchevers, J., Goidts, E., Grace, P., Kätterer, T., McConkey, B.G., 2011. How can soil monitoring networks be used to improve predictions of organic carbon pool dynamics and CO 2 fluxes in agricultural soils? Plant Soil 338 (1–2), 247–259.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.
Vapnik, V., Golowich, S.E., Smola, A.J., 1997. Support vector method for function approximation, regression estimation and signal processing, pp. 281-287.
Viscarra Rossel, R., Walter, C., 2002. Towards a quantitative assessment of soil organic carbon using proximally sensed digital imagery. in: 17th World Congress of Soil Science, Bangkok, Thailand, 14-20 August 2002, 1523-1523.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1–2), 59–75.
Viscarra Rossel, R.A., Fouad, Y., Walter, C., 2008. Using a digital camera to measure soil organic carbon and iron contents. Biosyst. Eng. 100 (2), 149–159.
Webster, R., Butler, B.E., 1976. Soil classification and survey studies at Ginninderra. Soil Res. 14 (1), 1–24.
Wu, C., Yang, Y., Xia, J., 2017. A simple digital imaging method for estimating black-soil organic matter under visible spectrum. Arch. Agron. Soil Sci. 63 (10), 1346–1354.
Wu, C., Xia, J., Yang, H., Yang, Y., Zhang, Y., Cheng, F., 2018. Rapid determination of soil organic matter content based on soil colour obtained by a digital camera. Int. J. Remote Sens. 39 (20), 6557–6571.
Zhu, Y., Wang, Y., Shao, M., Horton, R., 2011. Estimating soil water content from surface digital image gray level measurements under visible spectrum. Can. J. Soil Sci. 91 (1), 69–76.
Zurada, J.M., 1992. Introduction to Artificial Neural Systems, 8. West Publishing Company, St Paul.