Journal of the Indian Society of Remote Sensing
https://doi.org/10.1007/s12524-022-01524-9

RESEARCH ARTICLE

Detecting Soil pH from Open-Source Remote Sensing Data: A Case Study of Angul and Balangir Districts, Odisha State

Pranuthi Gogumalla[1] · Srikanth Rupavatharam[1] · Aviraj Datta[1] · Rohan Khopade[1] · Pushpajeet Choudhari[1] · Ramkiran Dhulipala[1] · Sreenath Dixit[1]

Received: 25 June 2021 / Accepted: 19 February 2022

Abstract

Soil sampling, collection, and analysis are a costly and labor-intensive activity that cannot cover the entire farmlands; hence, it was conceived to use high-speed open-source platforms like Google Earth Engine in this research to estimate soil characteristics remotely using high-resolution open-source satellite data. The objective of this research was to estimate soil pH from Sentinel-1, Sentinel-2, and Landsat-8 satellite-derived indices; data from Sentinel-1, Sentinel-2, and Landsat-8 satellite missions were used to generate indices and as proxies in a statistical model to estimate soil pH. Step-wise multiple regression (SWMR), artificial neural networks (ANN), and random forest (RF) regression were used to develop predictive models for soil pH, SWMR, ANN, and RF regression models. The SWMR greedy method of variable selection was used to select the appropriate independent variables that were highly correlated with soil pH. Variables that were retained in the SWMR are B2, B11, Brightness index, Salinity index 2, Salinity index 5 of Sentinel-2 data; VH/VV index of Sentinel 1 and TIR1 (thermal infrared band1) Landsat-8 with p-value < 0.05. Among the four statistical models developed, the class-wise RF model performed better than other models with a cumulative correlation coefficient of 0.87 and RMSE of 0.35. The better performance of class-wise RF models can be attributed to different spectral characteristics of different soil pH groups. More than 70% of the soils in Angul and Balangir districts are acidic soils, and therefore, the training of the dataset was affected by that leading to misclassification of neutral and alkaline soils hindering the performance of single class models. Our results showed that the spectral bands and indices can be used as proxies to soil pH with individual classes of acidic, neutral, and alkaline soils. This study has shown the potential in using big data analytics to predict soil pH leading to the accurate mapping of soils and help in decision support.

Keywords Soil pH · GEE · Sentinel · Landsat-8 · ANN · Random forest · Odisha

Introduction

Soil pH is defined as the negative logarithm of the hydrogen ion concentration. Soil pH is an important indicator of soil health that affects crop yields, crop suitability, plant nutrient availability, and soil micro-organism activity. Soil pH is an excellent indicator of a soil's suitability for plant growth. For most crops, soil pH a range of 6–7.5 is the best. When implementing different precision agriculture practices, site-specific management of soil pH is one of the most promising strategies in fields with substantial variability in soil pH. Soil pH influences the effectiveness and use efficiency of fertilizers (von Tucher et al., 2018; Wang et al., 2018), herbicides (Buerge et al., 2019; Liu et al., 2018) and insecticides, and solubility of heavy metals depend on pH (Kah et al., 2007; Spadotto & Hornsby, 2003). Therefore, it is quite necessary to measure soil pH to make effective decisions regarding sowing, fertilization, and other crop management practices.

Currently, a variety of techniques are being used to investigate the soil pH status, including traditional soil sampling methods and other novel methods with soil sensors. In-situ measurements can directly obtain steady and accurate soil pH but cannot represent a large area spatially. Furthermore, these ground measurements consume time

✉ Pushpajeet Choudhari
P.Choudhari@cgiar.org

[1] International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Hyderabad 502324, India

Published online: 14 March 2022

and labor, and it is expensive to maintain both the quality and dense network of the observations (Chang & Islam, 2000; Elshorbagy & Parasuraman, 2008). Among these novel methods, digital soil mapping using remote sensing data has emerged as a promising and reliable new technique (Eisele et al., 2015; McBratney et al., 2003).

Remote sensing (RS) is well established as a cost-effective, rapid, and reproducible means of providing quantitative and spatially distributed data on soil properties. The increasing power of RS technologies (e.g., global positioning systems, airborne and satellite platforms, unmanned aerial vehicles, and ground-based sensors), geographic information systems (GIS), and spatial data models (e.g., DEM-Digital Elevation Model) is offering new ways forward in soil science (Eli-Chukwu, 2019; Grishin & Timirgaleeva, 2020; Rodrigo-Comino et al., 2020).

Digital soil mapping is being employed to assess the spatial distribution of soil properties in agricultural areas and other land resources (Forkuor et al., 2017; Minasny et al., 2013; Taghizadeh-Mehrjardi et al., 2016). Recently, in several studies, soil properties such as soil pH (Pahlavan-Rad & Akbarimoghaddam, 2018), soil organic matter (Byrne & Yang, 2016), electrical conductivity (Ranjbar & Jalali, 2016), and phosphorus (Wilson et al., 2016) have been predicted and mapped.

SoilGrids 2.0 (De Sousa et al., 2020; Hengl et al., 2017) provides global estimates of some basic soil properties such as organic carbon, bulk density, cation exchange capacity (CEC), pH, soil texture fractions, and coarse fragments at seven standard depths (0 cm, 5 cm, 15 cm, 30 cm, 60 cm, 100 cm, and 200 cm) with 250 m resolution. Estimates are made from the previously collected soil data which is used for training the models and with 158 covariates (primarily derived from MODIS land products, SRTM DEM derivatives, climatic images and global landform and lithology maps), which were used to fit an ensemble of machine learning methods—random forest and gradient boosting and/or multinomial logistic regression. However, these estimates are coarser in resolution and cannot explain the within field variability. The availability of better resolution satellite images (10–30 m resolution) helps us to improve the accuracy of soil information estimated from the remotely sensed data.

The Department of Agriculture, Government of Odisha and the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) are implementing a developmental project initiative called "Bhoochetana"(Wani et al., 2016). Under this project soil analysis, nutrient management recommendations and treatment are being shared with farmers. This will help increase productivity through improved practices. To fulfill this objective, ICRISAT has collected and analyzed soil samples from all the villages of Angul and Balangir districts of Odisha state. In this research, we have used this ground truth data to test whether the satellite-derived indices can act as proxies to predict soil pH through models.

Materials and Methods

The Study Region

The District of Angul situated at the heart of Odisha. The district lies within the geographical limits of 20° 42′ 08.15″ N latitude and 83° 28′ 49.43″ E longitude at an average altitude of 142 m (Fig. 1). The total geographical area of the district is 6790 km^2; total cultivated area of 3460 km^2 and a forest area of 1540 km^2. Out of the total cultivated area, only 16% of are is under irrigation and the rest is rainfed. Soils that are predominant in the district are red and black soils. The area receives an annual rainfall of 1290 mm, and the crops that are majorly grown are rice and mung bean occupying 80% of total cultivated area.

Balangir district is one of the less developed districts of the Odisha state with severe agrarian crisis (https://rcdcindia.org/places/regional-offices/bolangir/) (Fig. 1). The district is located within the geographic limits of 20° 09' N, 21° 05″ N latitudes and 82° 41″ E to 83° 42' E longitudes. The percent of cultivated area is more than 50% with rice, mung bean, and cotton as major crops. Out of the total cultivated area of 346,000 ha, only 53,920 ha is irrigated which accounts to 15% of total cultivated area. Soils of Balangir are predominantly mixed red and yellow soils followed by red and black soils.

Soil Data Collection and Analysis

In May–June 2018, the ICRISAT team collected and analyzed 2244 soil samples from the districts of Angul (766) and Balangir (1478), Odisha under the Bhoochetana project (Wani et al., 2016). Soil pH was analyzed in the soil laboratory using standard operating methods. Data needed to be processed before performing any analysis. The data with incorrect lat/long locations were omitted, and after that, the entire data were corrected for outliers using the nearest neighbor method. The data with distance > 0.01 (mean $\approx$ median) from the nearest cluster were omitted. Finally, the number of soil datasets that remained are 2073 (634 for Angul and 1438 for Balangir districts). This soil data is partitioned into training, validation, and test datasets for model building. The details of the dataset are given in Table 1.

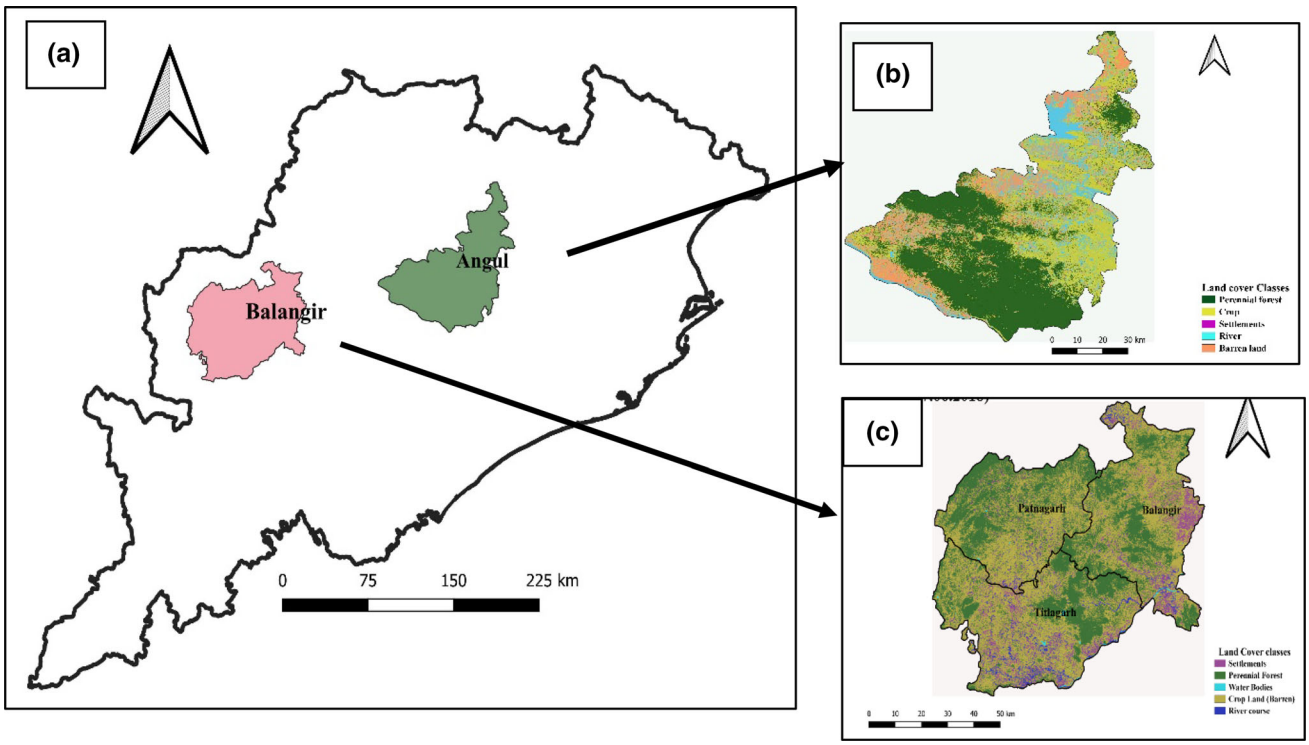

Fig.1 **a** Geographical map of Odisha state with Angul district (green) and Balangir district (pink). **b** Land cover classified Sentinel 2 image of Angul **c** Land cover classified Sentinel 2 image of Angul

Table 1 Partitioning of soil data for calibration, validation and testing of soil pH prediction models

Dataset	Percentage	No.of datasets
Training	60% of Balangir data	834
Validation	20% of Balangir data	285
Test	20% of Balangir data + 100% Angul data	279 + 634

Satellite Data

Open-source satellite data Sentinel-1(Potin et al., 2012; Torres et al., 2012), Sentinel-2 (Drusch et al., 2012; Gascon et al., 2014), and Landsat-8 (Loveland & Irons, 2016; Roy et al., 2014) data have been used to estimate soil pH. The Sentinel-1 mission provides data from a dual-polarization C-band Synthetic Aperture Radar (SAR) instrument at 5.405 GHz (C band) which consists of ground range detected (GRD) scenes. These images are processed using the Sentinel-1 Toolbox to generate a calibrated, ortho-rectified products. Sentinel-1 image of 15th June, 2018 along with its two bands VV & VH have been used in developing soil pH model (https://code.earthengine.google.com/2649fcc9747730a8e234d126b012af96). The Sentinel-2 mission carries the multispectral instrument which measures the reflected solar spectral radiances in 13 spectral bands ranging from the visible to the shortwave infrared (SWIR) bands with 5-day revisit time and a spatial resolution of 10–60 m over land and coastal areas (Drusch et al., 2012). Out of the 13 spectral bands, only 10 spectral bands in different spectral regions namely Blue (B2), Green (B3), Red (B4), Red Edge (B5, B6 & B7), NIR (B8 & B8A), SWIR(B11 &B12) were relevant to this study. The Sentinel-2 L2 data are obtained by rectifying the L1 images using sen2cor model, and these datasets are provided through GEE repository. However, we have very limited cloud-free images and also the soil should be free from the crop. To select a cloud-free image with the possible nearest date of soil sample collection, the Sentinel-2 image of 17th June, 2018 covered by 4 tiles of Sentinel-2 image were used in this study (https://code.earthengine.google.com/8ab3197dac35ef60e7a49fc969594329). Similarly, the land surface temperature retrieved from the brightness temperature of thermal bands 10 & 11 of Landsat-8 (Roy et al., 2014) using the algorithm given by (Parastatidis et al., 2017) which uses different emissivity sources (https://code.earthengine.google.com/5964230990 8906db1bb599fce7e1cb50).

Soil and vegetation indices (Table 2) were generated using satellite data with the aid of Google Earth Engine (GEE) (Gorelick, 2013; Gorelick et al., 2017) which is a

Table 2 Indices developed from Sentinel-1, Sentinel-2 and Landsat-8 satellite data

Index	Acronym	Formula	Reference
Advanced vegetation index	AVI	$\sqrt[3]{(B4+1)*(256-B3)*(B4-B3)}$	(Banerjee et al., 2014)
Normalized differential vegetation index	NDVI	$\frac{B8-B4}{B8+B4}$	(Tucker et al., 1979)
Normalized differential salinity index	NDSI	$\frac{B4-B8}{B4+B8}$	(Khan et al., 2001)
Normalized moisture stress index 1	NMSI1	$\frac{B8-B11}{B8+B11}$	(Gao, 1996)
Red edged inflection point	REIP	$700+(40*\frac{\left(\frac{B4+B7}{2}\right)-B5}{B6-B5}$	(Vogelmann et al., 1993)
Advanced vegetation index	AVI	$\sqrt[3]{B8*(1-B4)*(B8-B4)}$	(Rikimaru et al., 2002)
Bare soil index	BSI	$\frac{(B11+B4)-(B8+B2)}{(B11+B4)+(B8+B2)}$	(Li & Chen, 2014)
Brightness index	BI	$\frac{(B6-B4)-(B5-B2)}{(B6-B4)+(B5-B2)}*100+100$	(Todd & Hoffer, 1998)
Salinity index 1	SI1	$\sqrt[2]{B2*B4}$	(Douaoui et al., 2006)
Salinity index 2	SI2	$\sqrt[2]{B3*B4}$	
Salinity index 3	SI3	$\sqrt[2]{B3^2+B4^2+B8^2}$	
Salinity index 4	SI4	$\sqrt[2]{B3^2+B4^2}$	
Salinity index 5	SI5	$\frac{B2}{B4}$	
Salinity index 6	SI6	$\frac{B2-B4}{B2+B4}$	
Soil Salinity and sodicity index	SSSI	$B11-B12$	(Bannari et al., 2016)

freely available cloud-based platform for processing geospatial datasets. Using GEE JavaScript API, various indices were estimated from Sentinel-1, Sentinel-2, and Landsat-8 data and were extracted for each point of soil sampling. Backscatter of Sentinel-1 mission, Reflectance of 10 spectral bands combined with soil indices developed from the Sentinel-2 spectral bands, and LST retrieved from thermal bands of Landsat-8 were used as proxies to soil pH. The list of the soil indices/ vegetation indices used with the reference and formula are presented in Table 2.

Developing Statistical Models for Predicting Soil pH

To use the satellite-derived soil indices as proxies to pH, a proper fitting model is required. Collinearity exists between spectral bands and soil indices so to eliminate collinearity variance inflation factor (VIF) is employed and the variables with VIF value less than 4 are selected and in the third step in SWMR through forward and backward selection the variables have been selected to develop the soil pH estimation models. Generally, the linear and non-linear regression methods are used to develop a model with predictors that have probability ($p < 0.05$). Deep Learning and Machine Learning techniques such as ANN and Random forest, respectively, are also used to develop a model to predict pH from the soil indices developed from remotely sensed data. For the model, building the predictor being pH while the satellite-derived band reflectance and indices are taken as predictands.

The models developed in this study are:

Step-Wise Multiple Regression Model (SWMR)

SWMR is a combination of the forward and backward selection techniques. SWMR is a modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a non-significant variable is found, it is removed from the model. Step-wise regression requires two significance levels: one for adding variables and one for removing variables (Breaux, 1967). In this study for both forward and backward regression, we have used a significant probability level of 0.05. The variables or the indices have been selected in three step process: in the first step Pearson's correlation of 0.2 was used to select variables; in the second step, the VIF with < 5 were used to retain the.

ANN Regression (ANN)

Neural networks belong to deep learning methods. ANN is a complicated form of nonlinear regression designed to be able to model complex structures in the data. ANN studies the relationship of the independent variable with each of

the dependent variables and develops hidden layers of various regression models and ultimately which are summed up to finally predict the predictor. These hidden layers perform various types of mathematical computation on the input data and recognize the patterns that are part of. This process is quite complex but we have built-in algorithms for these models which eases the analysis (Kartalopoulos & Kartakapoulos, 1997). ANN model was developed using Jmp 14.0 statistical software (J. Li & Mocko, 2020), which develops hidden layers of the model using three transformation functions (TanH, Linear, and Gaussian) and a learning rate of 0.1. ANN model developed in the study had nine hidden nodes with three linear, three tangential, and three Gaussian transformations.

Random Forest (RF)

A Random Forest (RF) is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees (Breitenbach et al., 2003). The random forest developed in the study has 100 trees with boot strap rate of 1 and with minimum split of 5 trees per sample and maximum split of 500 trees per sample.

Class-wise RF

Different soil types with different soil pH values will interact differently with the electromagnetic spectrum. Therefore, individual RF models for every soil pH class were developed using Balangir data and tested for Angul soil data. Random forest models for each soil pH class RF-Acidic, RF-Alkaline, and RF-Neutral were developed and integrated into a single-model Class-wise RF to be able to compare it with SWMR, ANN, and RF models. The class-wise RF classified every single point into the probable class by using K-means clustering method within the algorithm.

First, we compare the integrated Class-wise RF model with SWMR, ANN, and RF, and later we tried to separately study each model (RF-Acidic, RF-Alkaline, and RF-Neutral) in detail.

Pearson's r of the correlation, coefficient of determination (R^2) (Ozer, 1985) and root square mean error (RMSE) (Fichter, 1984) were used as measures of model performance and to compare between models. The effect summary of each variable in the models was described in terms of contribution percentage. All statistical analyses were carried out using JMP ® software version 14.0 (SAS Institute Inc., USA). Coefficient of determination (R^2) (Ozer, 1985) and root square mean error (RMSE) (Fichter, 1984) were used as measures of model performance and to compare between models. The effect summary of each variable in the models was described in terms of contribution percentage. All statistical analyses were carried out using JMP ® software version 14.0 (SAS Institute Inc., USA) (Sall et al., 2017). Accuracy percentage was calculated by estimating the error between the measured soil pH and the estimated soil pH. Cohen's Kappa (Cohen, 1960) was calculated to see how accurately soil pH estimation models were able to estimate soil pH.

Results

General Statistics of Soil pH in Balangir District

The soil was collected from 8 blocks and 93 villages of Angul district; 14 blocks and 170 villages of Balangir district, from each village at least five soil samples, were collected. From the frequency distribution graph of soil pH of Angul, it is evident that more than 75% of soils are acidic and less than 2% soils are alkaline (Fig. 2).

Almost 60% of soils of Balangir are acidic, 30% soils are neutral, and only 10% soils are alkaline (Fig. 2). The summary statistics of the soil pH data collected from Angul and Balangir districts is given in Table 3 from which it is evident that the soil pH ranged between 4.06 and 8.16 for Balangir district and 4.0 and 7.8 for Angul districts. The coefficient of variation is 17% and 16% for Angul and Balangir districts, respectively. From skewness, the Balangir soil pH data is left skewed whereas Angul soil pH data is right skewed. From the kurtosis, it is seen that both Angul and Balangir soil pH data are platykurtic (Table 3). A simple Pearson's correlation was calculated between soil pH and spectral bands and indices; the reflectance of B11, B12 and B5 has shown a higher correlation of -0.46, -0.45 and -0.44, respectively, with the soil pH in comparison with other spectral bands. Similarly, Salinity index-6 (SI6) has shown a higher correlation of 0.39 with the soil pH (Fig. 3a). Very familiar vegetation indices NDVI and NMSI were 0.2 and 0.3, respectively. The Sentinel-2 spectral signatures of acidic, alkaline and neutral soils are shown in Fig. 3b which clearly indicates that the soils with different pH can be identified with B4, B5 and B11 and B12 spectral bands.

Soil pH Prediction Models

Among the ANN and RF models, the class-wise RF model was found to perform better than the other three models with 0.97, 0.88 and 0.77 coefficient of correlation (r) for calibration, validation and test datasets respectively (Table 4). The class-wise RF models performed far better than

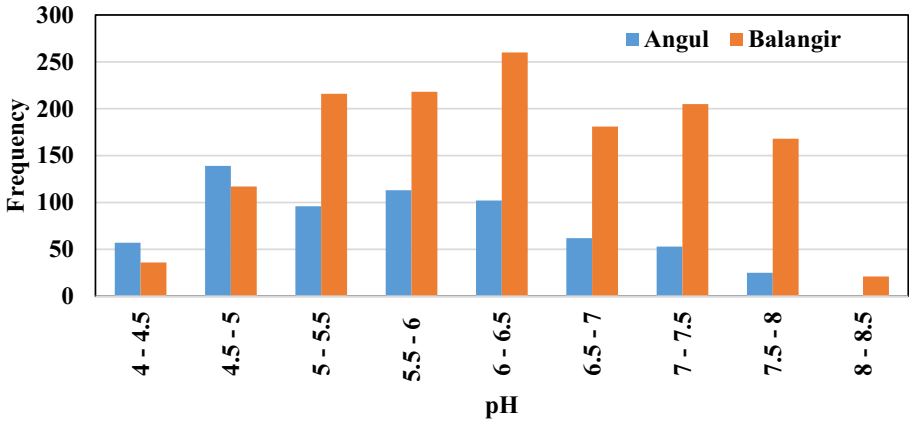

Fig. 2 Frequency distribution of soil pH at Angul and Balangir districts

Table 3 Descriptive statistics of the soil pH data collected from Angul and Balangir districts in the year 2018

S.No	Descriptive statistics	Balangir	Angul
1	Number of observations	1422	647
2	Blocks	14	8
3	Villages	170	93
4	Mean	6.25	5.65
5	Minimum	4.03	4.00
6	Maximum	8.16	7.80
8	Standard deviation	0.98	0.96
9	Coefficient of Variation (%)	16	17
10	Skewness (Fisher)	-0.02	0.31
11	Kurtosis (Fisher)	-1.00	-0.92

SWMR, ANN and RF models. R^2 for class-wise RF models is 0.94, 0.87 and 0.54 for calibration, validation and test datasets, respectively (Fig. 4). Even RMSE is quite lower than other models with 0.23, 0.48 and 0.63 for calibration, validation and test datasets, respectively (Table 4). The other three models SWMR, ANN and RF performed almost similarly; however, the RF model performed slightly better than SWMR and ANN with 0.89, 0.57 and 0.46 Pearson's correlation coefficient for calibration, validation and test datasets, respectively (Table 4). R^2 and RMSE are the measures that indicate the higher model performance of class-wise RF models, Cohen's kappa and accuracy percentage were also estimated to test the ability of models to classify.

Sentinel-2, Sentinel-1 and Landsat-8 data and their derived spectral indices were used to develop soil pH, prediction models. Three different regression models (SWMR. ANN, RF and Class-wise RF models) were developed to identify the best method to predict soil pH from satellite data. Step-wise multiple linear regression (SWMR) model was built to relate soil pH with remote sensing variables and it yielded an R^2 of 0.26, 0.20 and 0.17 for calibration, validation and test datasets, respectively (Figs. 4, 5 and 6). The multi-collinear variables are removed before developing SWMR, ANN and RF models using the VIF method, and variables with $p < 0.05$ are also removed in the SWMR method which retains only the significant variables in the model. The SWMR model found variables B2, B11, Brightness index, SI2, SI5, T11 and VH/VV to significantly affect the soil pH.

Among the statistical models, the class-wise RF model was found to perform better than the other three models with 0.97, 0.88 and 0.77 coefficient of correlation (r) for calibration, validation and test datasets, respectively (Table 4). The class-wise RF models performed far better than SWMR, ANN and RF models. R2 for class-wise RF models is 0.94, 0.87 and 0.54 for calibration, validation and test datasets, respectively (Figs. 4, 5 and 6). Even RMSE is quite lower than other models with 0.23, 0.48 and 0.63 for calibration, validation and test datasets, respectively (Table 4). The other three models SWMR, ANN and RF performed almost similarly; however, the RF model performed slightly better than SWMR and ANN with 0.89, 0.57 and 0.46 Pearson's correlation coefficient for calibration, validation and test datasets, respectively (Table 4). R2 and RMSE are the measures that indicate the higher model performance of class-wise RF models, Cohen's kappa and accuracy percentage were also estimated to test the ability of models to classify. The derived soil pH for all sites is classified into three categories viz., alkaline, acidic and neutral. Accuracy percentage (Ac) and Cohen's Kappa (K) (Cohen, 1960) indicate the efficiency of the model to identify different soil, pH classes. The higher the accuracy percentage higher is the performance of the model. Similarly, Cohen's Kappa > 0.5 is required for a good and reliable classification (Vieira et al., 2010). Based on the classification SWMR, ANN, RF and class-wise RF models showed an overall accuracy of 67%, 68%, 74% and 98%, respectively (Table 4). Similarly, Cohen's Kappa for all the

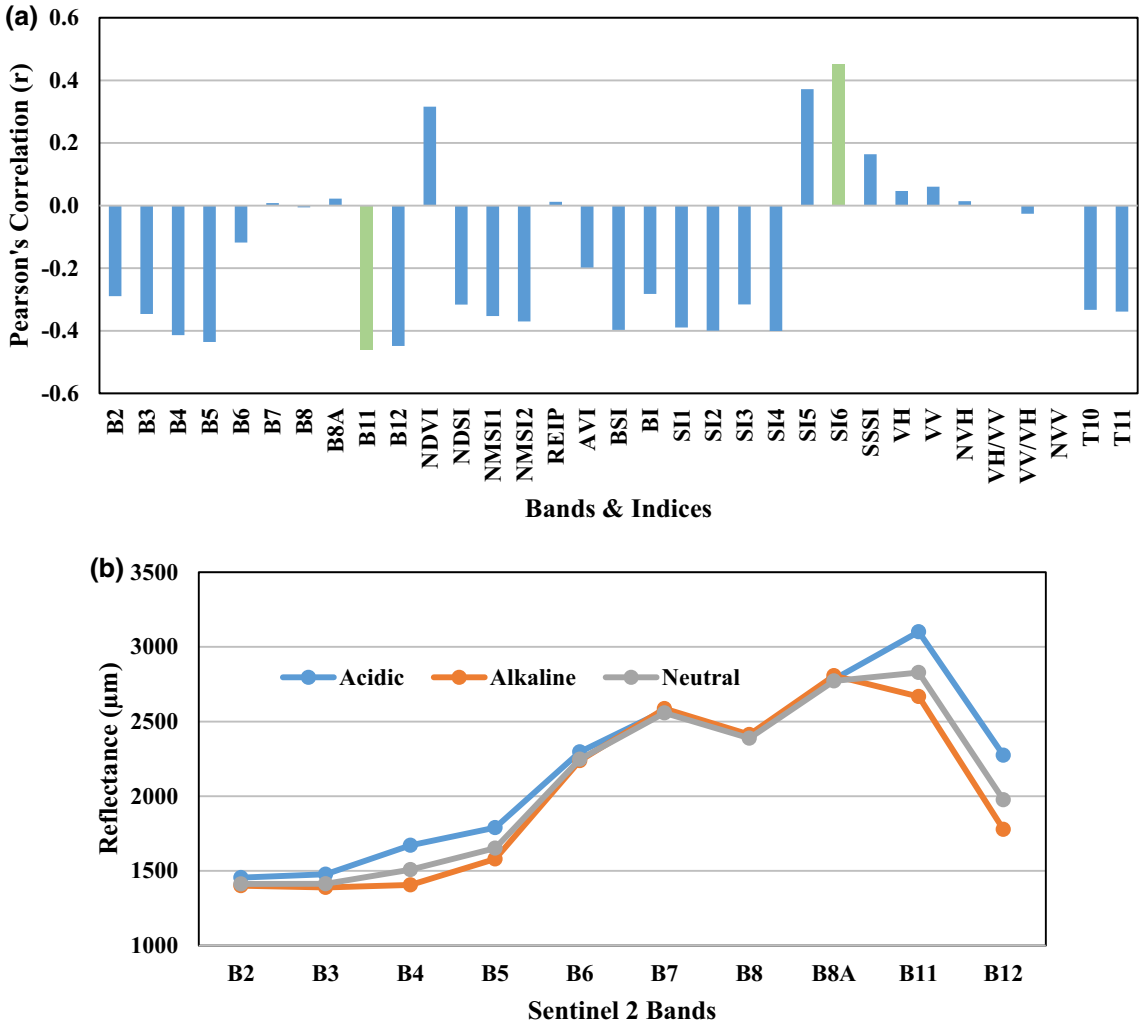

Fig. 3 **a** Pearson's correlation coefficient estimated between measured soil pH and spectral bands and satellite indices of Angul and Balangir districts soil pH data **b** Average of Sentinel-2 Spectral signatures of acidic, neutral and alkaline group of soils of Angul and Balangir districts

datasets for SWMR, ANN, RF and class-wise RF models showed a cumulative Kappa of 0.24, 0.26, 0.43 and 0.96, respectively (Table 4). Class-wise RF models showed exceptionally high accuracy and a perfect score of Cohen's Kappa with 97%, 99% and 99% accuracy and 0.97, 0.97 and 0.99 Kappa coefficient for calibration, validation and test datasets, respectively (Table 4). All the single-class models (SWMR, ANN and RF) showed more than 60% accuracy in estimating soil pH correctly for different classes; however, the RF model had an accuracy of 77%, 63% and 74% for calibration, validation and test datasets, respectively (Table 4). Kappa coefficient was less than 0.5 for all the single class models (SWMR, ANN and RF) with RF slightly better than other models with 0.58, 0.26 and 0.24 for calibration, validation and test datasets respectively.

The deviation % calculated between the measured soil pH and the model estimated soil pH by SWMR, ANN, RF and class-wise RF models for Angul and Balangir districts is presented in Figs. 7 and 8. The deviation percentage was calculated for each location and it is spatially interpolated in QGIS 3.8 software using inverse distance weighted (IDW) method of interpolation. Spatially interpolated deviation % for Balangir district ranged between − 29.8–57.7%, − 29.4–55.7%, − 22.6–38.7% and − 14.9–28.5% for SWMR, ANN, RF and class-wise RF models, respectively (Figs. 7 and 8). Spatially interpolated deviation % for Angul district ranged between − 31.3–40.3%, − 37.5–56.9%, − 24.0–42.5% and − 16.5–29.9% for SWMR, ANN,RF and class-wise RF models, respectively (Figs. 7 and 8). As Balangir district soil pH data are used as calibration, the percentage error is less than ± 5% except for few places which have more than 10–15% error, whereas for Angul district data which is used as test most of the locations had more than 15% error particularly for SWMR and ANN and comparatively

Table 4 Pearson's correlation coefficient (r), RMSE, Accuracy (Ac) and Cohen's Kappa coefficient (K) for SWMR, ANN, RF and class-wise RF models

Models	Datasets	r	RMSE	Accuracy	Cohen's Kappa
SWMR	Cumulative	0.50	0.88	0.67	0.24
	Calibration	0.51	0.86	0.63	0.28
	Validation	0.45	0.85	0.59	0.17
	Test	0.42	0.91	0.74	0.18
ANN	Cumulative	0.48	0.89	0.68	0.26
	Calibration	0.58	0.81	0.64	0.29
	Validation	0.51	0.82	0.61	0.21
	Test	0.30	0.98	0.74	0.20
RF	Cumulative	0.70	0.74	0.74	0.43
	Calibration	0.89	0.53	0.77	0.58
	Validation	0.57	0.78	0.63	0.26
	Test	0.46	0.88	0.74	0.24
Class-wise RF	Cumulative	0.87	0.35	0.98	0.98
	Calibration	0.97	0.23	0.97	0.97
	Validation	0.88	0.33	0.97	0.97
	Test	0.77	0.50	0.99	0.99

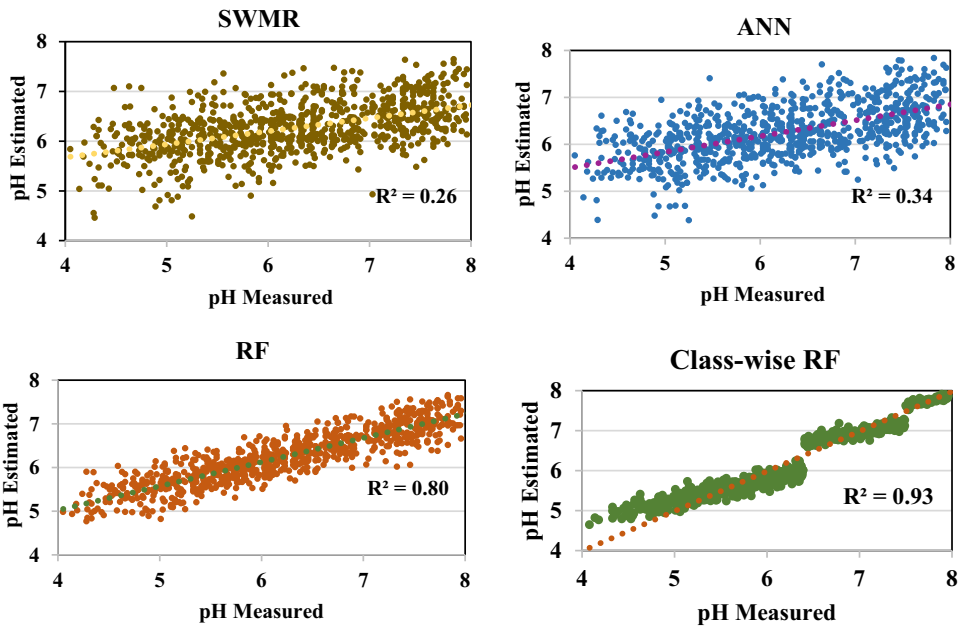

Fig. 4 Scatterplot between measured and estimated soil pH by SWMR, ANN and RF models for calibration dataset

less for RF model. The IDW interpolation of class-wise RF models showed that for Balangir the deviation percentage for most of the locations is < ± 5%; for Angul district, the deviation percentage is in the limits of ± 10% but for the northern part of the district for some locations the deviation is more than ± 20%.

Though the upper and lower range of error depicts the extent of error in the predicted soil pH, it is also misleading if only one data point has a very high error. Therefore, the error of predicted soil pH is partitioned into 11 error classes with a class interval of 5. The proportion of data partitioned into different deviation percentage classes is shown in Fig. 9. For SWMR models, only 22.7% of predicted soil pH dataset has an error ± 5%, 35.2% of data set error is the range of ± 15–20%, and 18.8% of dataset error is the range of ± > 20% (Fig. 9). For ANN models, only 25.3% of predicted soil pH dataset has an error ± 5%, 32.9% of dataset error is the range of ± 15–20%, 20.3% of dataset error is the range of ± > 20%. For RF models, only 32.9% of predicted soil pH dataset has an error ± 5%, 29.2% of dataset error is the range of ± 15–20%, and 13.7% of dataset error is the range of ± > 20% (Fig. 9). For class-wise RF models, 67.2% of predicted soil pH dataset has an error ± 5%, 10.2% of data set error is the

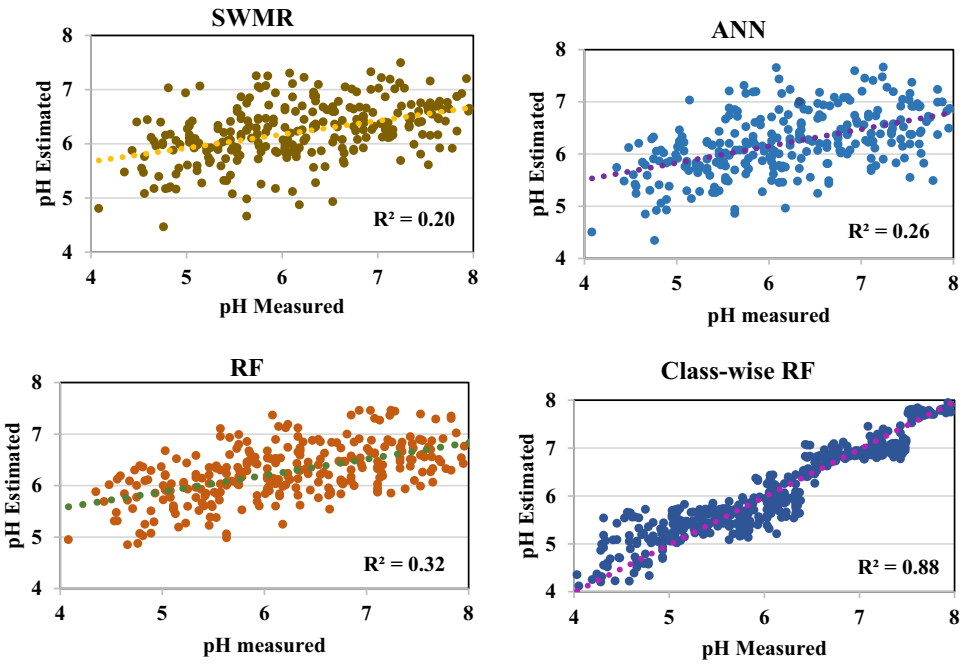

Fig. 5 Scatterplot between measured and estimated soil pH by SWMR, ANN and RF models for validation dataset

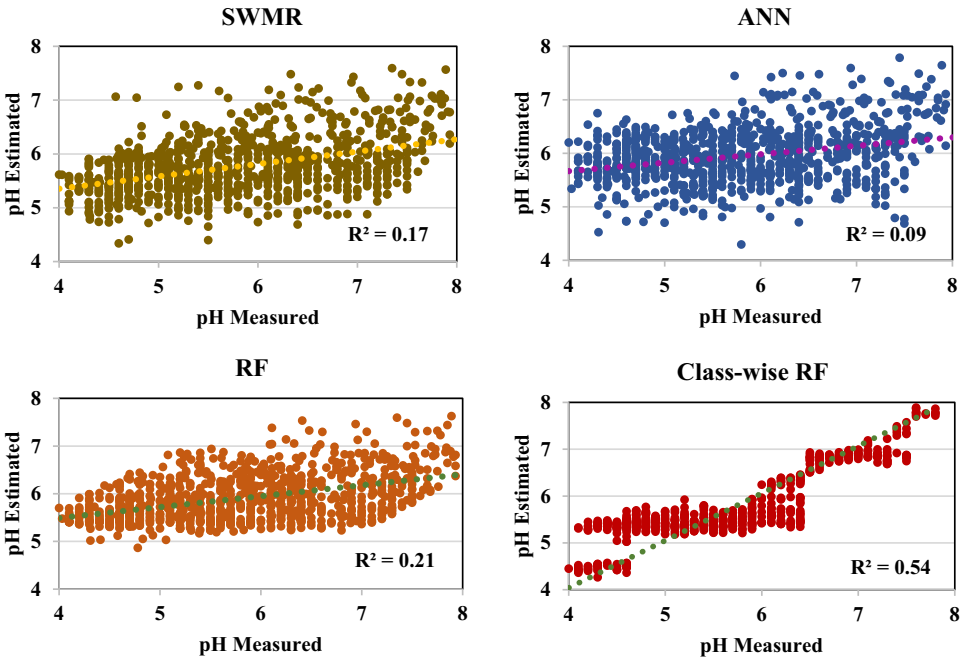

Fig. 6 Scatterplot between measured and estimated soil pH by SWMR, ANN and RF models for test datasets

range of ± 15–20%, and 2.4% of dataset error is the range of ± > 20% (Fig. 9).

Class-wise RF Models

Already in the earlier paragraphs, the class-wise RF models are compared with single class models (SWMR, ANN, and RF), here we study each class model, i.e., RF-Acidic, RF-Alkaline and RF-Neutral models in detail. From Figs. 4, 5 and 6 and Table 4, it is observed that class-wise RF models for each soil pH class performed far better with high R^2 (0.94, 0.77 & 0.59 for calibration, validation and test datasets, respectively) and low RMSE (0.23, 0.33 & 0.50) for calibration, validation and test datasets, respectively) than RF model. An in-depth study of each model will provide more insights into the relation of soil pH with the satellite spectral data (Table 4). The coefficient of determination (R^2) for RF-acidic, RF-neutral and RF-alkaline soil class for calibration data is 0.86, 0.79 and 0.66, respectively (Table 4). RMSE for RF-acidic, RF-neutral

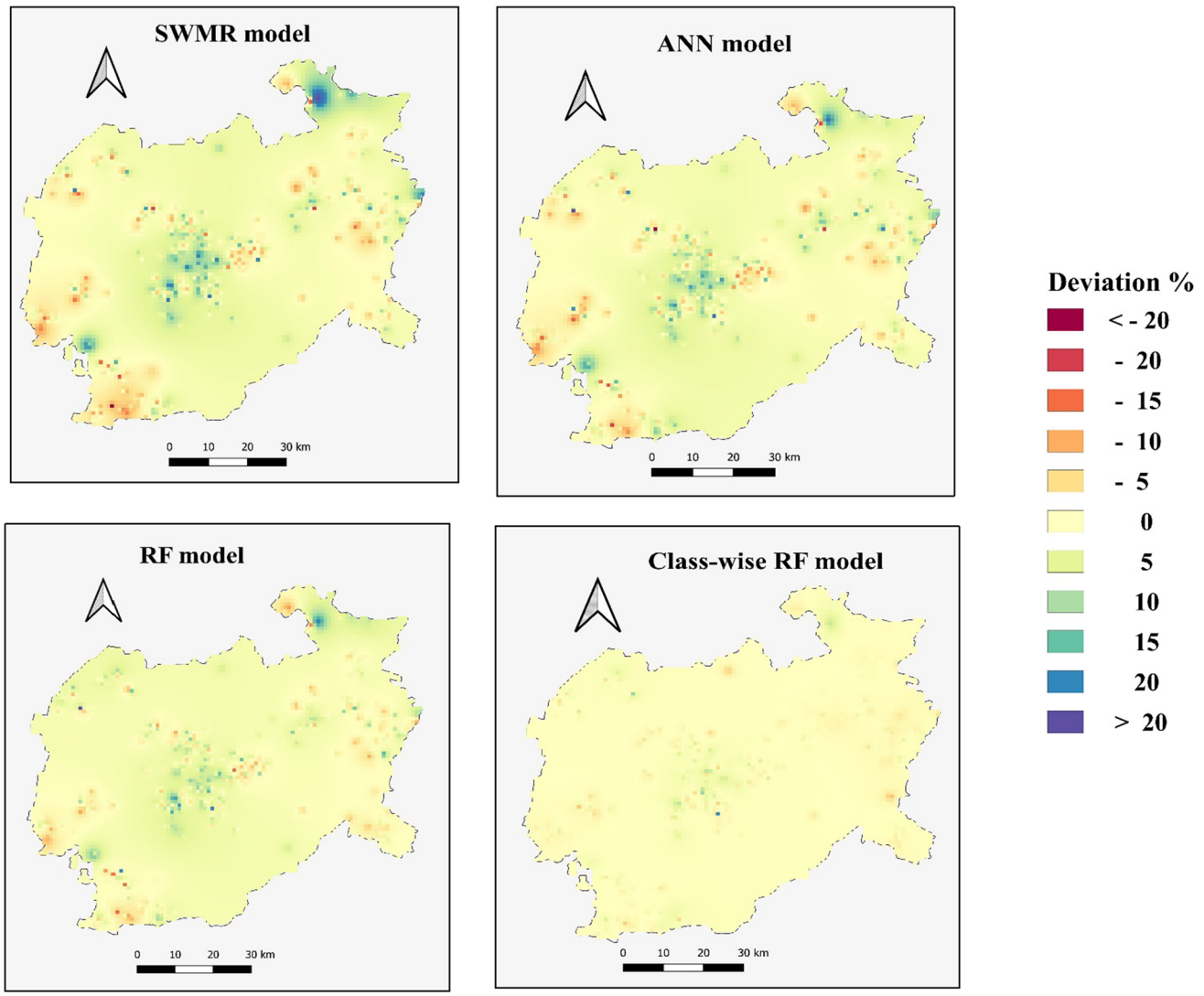

Fig. 7 Interpolated map of deviation percentage calculated between measured and estimated soil pH for SWMR, ANN, RF and class-wise RF models for Balangir district

and RF-alkaline soil pH prediction models are 0.27, 0.18 and 0.11 for the calibration dataset (Table 4). R^2 for validation is 0.60, 0.44 and 0.33 and RMSE of 0.38, 0.27 and 0.14 for RF-Acidic, RF-Neutral and RF-Alkaline models, respectively. The test data R^2 for RF-acidic and RF-Neutral is 0.41 and 0.25, but for RF-Alkaline the datasets have very few data points due to which the R^2 and RMSE for RF-alkaline models cannot be calculated. RMSE for test data is 0.54 and 0.29 for RF-acidic and RF-neutral soil pH models (Table 4). The higher R^2 values of RF-acidic, RF-neutral and RF-alkaline and lower RMSE indicate that class-wise RF models perform far better than single class models.

To study the spectral characteristics of different soil pH classes, the major spectral bands and Indices that influenced the models and their contributions are plotted in a graph (Fig. 7). The spectral bands and indices that help to identify acidic and neutral soil pH classes are similar: B5, B11/B12, SI6, T10 and T11. But for alkaline soils, the spectral bands that influence the soil pH are AVI, B8, B8A, VH/VV and SSSI (Fig. 7). Scatterplot of RF-acidic, RF-neutral and RF-alkaline model predicted soil pH against measured soil pH of Angul and Balangir districts (Fig. 4). For the calibration dataset, the R^2 value is 0.93 and RMSE is 0.23, with a clear distinction between acidic, neutral and alkaline classes. The estimated soil pH is very close to the measured soil pH. But for validation and test datasets, we observe an overlap between the classes indicating the misclassification of the model. However, the classes are more distinct when compared with all the datasets of single class models.

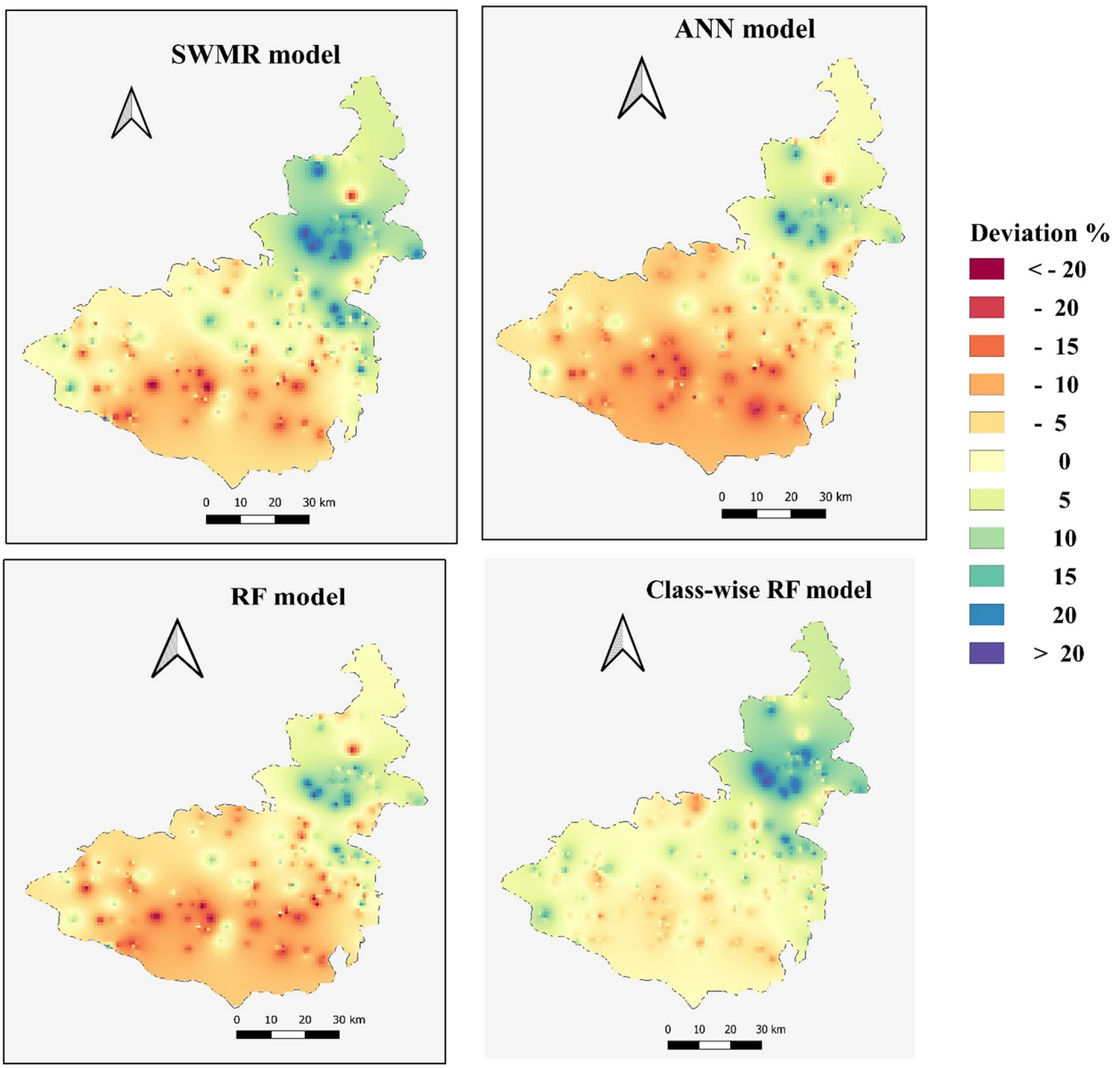

Fig. 8 Interpolated map of deviation percentage calculated between measured and estimated soil pH for SWMR, ANN, RF and class-wise RF models for Angul district

Fig. 9 Proportion of Balangir and Angul soil pH data estimated by four prediction models partitioned into 11 classes of percent deviation ranging from < -20 to $> 20\%$

Discussion

Soil pH Prediction Models

The soil data of Angul and Balangir districts collected under the Bhoochetana project indicated that the majority soils are acidic. As documented by Mishra in his review regarding the Soils of Orissa, the predominant soils of Angul and Balangir of Orissa are Alfisols (Mishra, 2007). Even in this study, most of the soils of the study area were classified as acidic (Fig. 2).

The generally used vegetation indices NDVI, NMSI1 and NMSI2 on an average for the districts are 0.3, − 0.35, 0.02 indicating scanty or no vegetation with very little moisture in the soils of the study area during the image acquisition time. The model efficiency depends on the use of the optimum number of variables with less multi-collinearity; as a huge number of multi-collinear, dependent variables increase the standard error of the predictions. Therefore, using the VIF method the multi-collinear variables were removed and used for model development consequently. SWMR method was found useful in variable selection. The factors that were selected by the SWMR model soil pH prediction are B2, B11, Brightness index, SI2, SI5, T11 and VH/VV indicated that the Blue, Red, Red Edge and SWIR regions of the electromagnetic spectrum were affected by changes in the soil pH. Similar results have been reported in an article by (Lee et al., 2003) which emphasizes the importance of the visible region, red edge and short wave infra-red spectral reflectance in estimating soil pH of Alfisols. The exact reason for the response of these bands cannot be ascertained as soil pH is influenced by many factors such as parent material, climate, topography, soil water content, organic matter content, land management and many others (Neina, 2019; Pahlavan-Rad & Akbarimoghaddam, 2018; Zhang et al., 2018). Similar findings have been reported by (Bai et al., 2016) in which Landsat-8 OLI (Operational Land Imager) satellite data are used to estimate soil pH. This study also found that the model for prediction was based on blue (0.45–0.51 µm) and SWIR (1.57–1.65 µm) bands with 30 m spatial resolution which has also been reported by (Bannari et al., 2016).

From the results (Table 4), it is quite evident that the RF model performance was better than other models, i.e., SWMR and ANN. Although, RF showed an R^2 value of 0.8 for calibration dataset, indicating a higher performance model for predicting soil pH, for validation and test dataset the R^2 drastically reduced implying that the model cannot be applied for prediction with a new dataset.

The better performance of class-wise RF models over single-class models can be attributed to different spectral characteristics of different soil pH groups. Every soil character has a unique spectral signature, and any changes in the soil's physical and chemical properties also alter its spectral signature. Therefore, one model for all the classes will not be sufficient to provide reliable soil pH estimated using satellite data proxies. The outperformance of random forest regression over methods of regression for estimating soil characteristics using spatial and satellite data has earlier been reported by (Ließ et al., 2012; Yang et al., 2016). Generally, the random forests regression have given more reliable soil pH estimates than linear and neural network regression; as random forests have unique characteristics such as (1) it incorporates the interaction between predictors, (2) it is based on ensemble learning theory, which allows it to learn both simple and complex problems; (3) random forest does not require much fine-tuning of its hyper-parameters as compared to deep learning techniques (ANN). However, ANN requires more number of dependent variables and huge dataset for developing several hidden layers which in turn provide final estimates (Ahmad et al., 2017; Gopal & Bhargavi, 2019; Mekonnen et al., 2019). As we have only provided less than 15 dependent variables to the model, the ANN model performance was hindered.

In the case of the RF model, the coefficient of determination and RMSE for calibration dataset was found to indicate a good model but a look at R^2 and RMSE for validation and test datasets showed that it is similar to SWMR and ANN models. When examined, the misclassification of single class models to identify the correct soil pH class using the prediction models; it is found that the models failed to identify the alkaline soils correctly in many instances leading to poor accuracy of 3.1%, 5.3% and 9.5% for SWMR, ANN and RF models, respectively. The highest accuracy of classification is calculated for acidic soils with an accuracy percentage of 88%, 89% and 91.5%, respectively, for SWMR, ANN and RF models (Table 5). The overall classification accuracy was affected by higher misclassifications in the alkaline group of soils. The lower percentage of accuracy can be attributed to the less number of soil samples of alkaline soils that affect the training set and ultimately the model performance. The soil pH predicted by RF-Acidic, RF-Neutral and RF-Alkaline models have been consolidated and compared with other single class models to verify the performance of class-wise RF models. It is obvious and understandable that the accuracy of classification will be more than 90% as we are already providing the class details to the models. But R^2 and RMSE are the measures that indicate the higher model performance of class-wise RF models with the highest R^2 and lowest RMSE.

Table 5 Coefficient of determination (R^2) and RMSE for RF-Acidic, RF-Neutral and RF-Alkaline models

Datasets	R^2			RMSE		
	Acidic	Neutral	Alkaline	Acidic	Neutral	Alkaline
Calibration	0.86	0.79	0.66	0.27	0.18	0.11
Validation	0.60	0.44	0.33	0.38	0.27	0.14
Test	0.41	0.25	–	0.54	0.29	–

Class-wise RF Models

The better performance of the class-wise RF models can be attributed to the multiple decision trees. Comparatively less performance of RF-neutral and RF-alkaline models is basically due to the less number of data points compared to RF-acidic, as (Millard & Richardson, 2015) mentioned model performance depends on the quality and quantity of the training dataset. Error percentage of more than 15% for all the models is observed toward the northern part of the district which can be due to the presence of haze or a thin layer of cirrus clouds in the satellite image. Any model and in particular the RF models can be tuned with good training data. More number of training samples helps the model to understand the behavior of the data to classify the data into various classes. The out performance of random forest instead is that it combines the predictions of many decision trees into a single model. The logic is that a single even made up of many mediocre models will still be better than one good model. A random forest can reduce the high variance from a flexible model like a decision tree by combining many trees into one ensemble model.

Millard and Richardson (Millard & Richardson, 2015) tried to examine the relationship between the size of training data and model performance; they found that in addition to being as large as possible, the training data sets used in RF classification should also be randomly distributed.

The alkaline soils mostly influence the reflectance in visible and NIR regions whereas acidic and neutral soils influence the SWIR and TIR regions of the electromagnetic spectrum. For RF and RF acidic models, B11, SI6, T11 and B5 contributed up to 40–50% (Fig. 10). As the majority of the soils in the study area are acidic, the variable contributions for the RF model and RF-acidic model are almost similar. For the RF-alkaline model, the major contribution was observed from T11 and VV bands. Similarly for RF-neutral model, the Sentinel-2 spectral bands B2, B4, B5, B8 and B11 contributed more than 40% for the model generation (Fig. 10). However, for acidic soils, the model failed to provide the right estimates for locations with soil pH less than 5. Use of soil and vegetation indices to estimate soil pH with better accuracy than interpolation method has been reported by several researchers (Bai et al., 2016; Chang & Islam, 2000; Malley et al., 1999; Merry & Janik, 2001; Roelofsen et al., 2015; Zhang et al., 2018) as interpolation is just a statistical method of estimating the soil pH without any other soil information. Remote sensing data to estimate soil pH also gives an idea of spectral characteristics of the location which also alters with time, climate, vegetation, soil condition, etc. So, the use of remote sensing data can give a better picture of the soil properties of the given location better than interpolation. These models have been applied to Balangir and Angul districts of Orissa to estimate the soil pH areas whose soil pH is not known which is presented in Fig. 11.

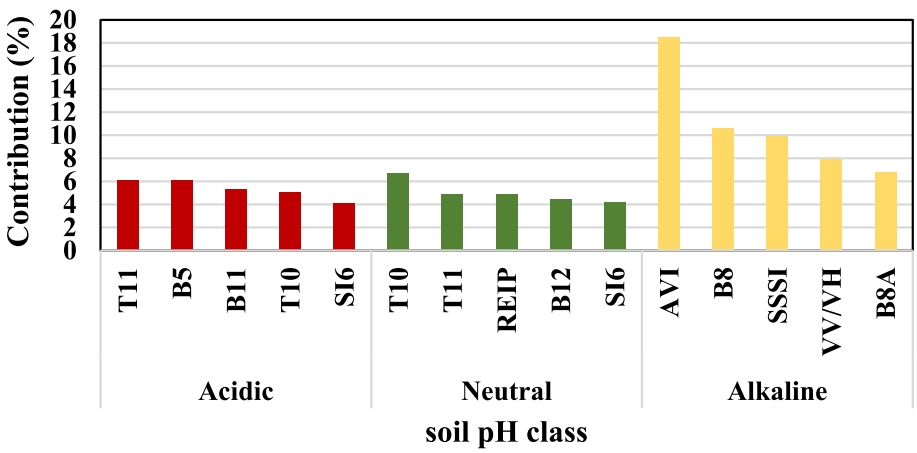

Fig. 10 Percent contribution of five important spectral bands and indices for RF-Acidic, RF-Neutral and RF-Alkaline models

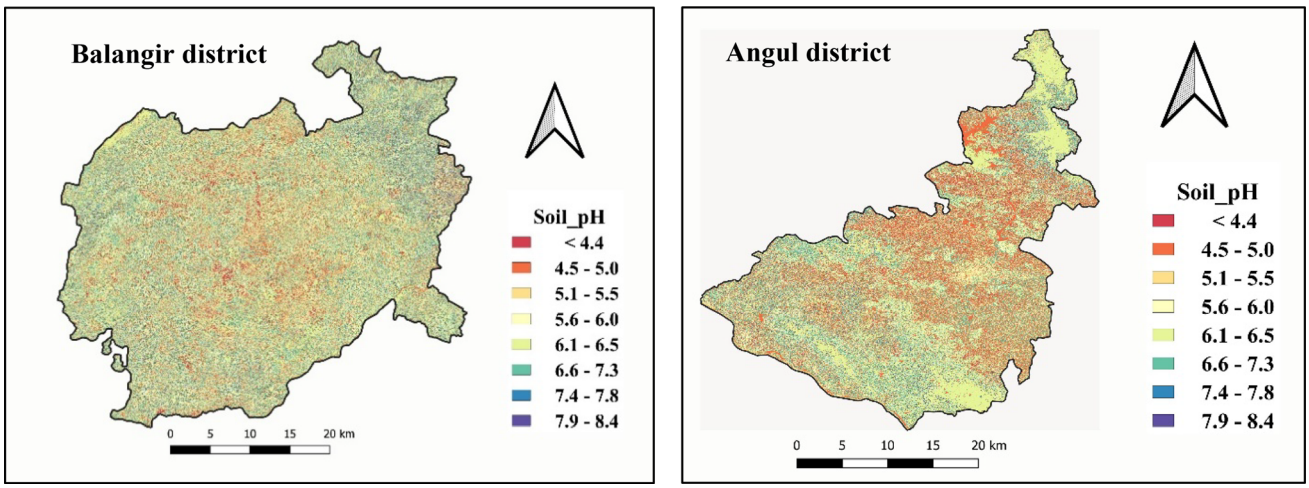

Fig. 11 Maps of class-wise RF model predicted soil pH for Balangir and Angul districts

Conclusions

In this research, it was observed that the satellite data with high spatial, spectral and temporal resolutions can estimate soil pH with fairly good accuracy. Among the three statistical models developed, the random forest model performed better than other models. The RF model misclassified the alkaline group of soils due to which the overall accuracy was affected. As every soil type or every soil pH class has its spectral signature, therefore models were developed for each pH class. The R^2 and RMSE of class-wise random forest models were far better than an all-inclusive RF model.

The salient features of this study are.

1. Use of open-source satellite data, multiple sensors; their spectral and soil, and vegetation indices developed from them.
2. Processing of the satellite data in an open-source, high-performance Google Earth Engine (GEE) platform.
3. Use of simple linear regression as well as deep learning (ANN) and machine learning (RF) statistical techniques to develop soil pH, estimation models.
4. Availability of extensive, well-distributed, and reliable village level measured soil pH data of Angul and Balangir districts of Odisha state.

All these features enabled us to develop class-wise RF soil pH estimation models which can give soil pH estimation.

Acknowledgements The authors want to acknowledge the grants from the Department of Agriculture, Odisha state to undertake *Bhoochetana* project by ICRISAT. We are also grateful to all the participating of farmers, departmental staff, NGOs and University students of University of Agriculture and Technology, Odisha.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77–89.

Bai, L., Wang, C., Zang, S., Zhang, Y., Hao, Q., & Wu, Y. (2016). Remote sensing of soil alkalinity and salinity in the Wuyu'er-Shuangyang River Basin, Northeast China. *Remote Sensing, 8*(2), 163.

Banerjee, K., Panda, S., Bandyopadhyay, J., & Jain, M. K. (2014). Forest canopy density mapping using advance geospatial technique. *International Journal of Innovative Science, Engineering & Technology, 1*(7), 358–363.

Bannari, A., Guédon, A., & El-Ghmari, A. (2016). Mapping slight and moderate saline soils in irrigated agricultural land using advanced land imager sensor (EO-1) data and semi-empirical models. *Communications in Soil Science and Plant Analysis, 47*(16), 1883–1906.

Breaux, H. J. (1967). On stepwise multiple linear regression. Army Ballistic Research Lab Abredeen Proving Ground MD.

Breitenbach, M., Nielsen, R., & Grudic, G. (2003). Probabilistic Random Forests: Predicting Data Point Specific Misclassification Probabilities. *CU-CS-954–03*.

Buerge, I. J., Bächli, A., Kasteel, R., Portmann, R., López-Cabeza, R., Schwab, L. F., & Poiger, T. (2019). Behavior of the chiral herbicide imazamox in soils: PH-dependent, enantioselective degradation, formation and degradation of several chiral metabolites. *Environmental Science & Technology, 53*(10), 5725–5732.

Byrne, J. M., & Yang, M. (2016). Spatial variability of soil magnetic susceptibility, organic carbon and total nitrogen from farmland in northern China. *CATENA, 145*, 92–98.

Chang, D.-H., & Islam, S. (2000). Estimation of soil physical properties using remote sensing and artificial neural network. *Remote Sensing of Environment, 74*(3), 534–544.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

De Sousa, L. M., Poggio, L., Batjes, N. H., Heuvelink, G. B., Kempen, B., Riberio, E., & Rossiter, D. (2020). SoilGrids 2.0: Producing quality-assessed soil information for the globe. *Soil Discussions, 2020*, 1–37.

Douaoui, A., Hartani, T., & Lakehal, M. (2006). La salinisation dans la plaine du Bas-Cheliff: acquis et perspectives. Presented at the Economies d'eau en Systèmes IRrigués au Maghreb. Deuxième atelier régional du projet Sirma.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment, 120*, 25–36.

Eisele, A., Chabrillat, S., Hecker, C., Hewson, R., Lau, I. C., Rogass, C., et al. (2015). Advantages using the thermal infrared (TIR) to detect and quantify semi-arid soil properties. *Remote Sensing of Environment, 163*, 296–311.

Eli-Chukwu, N. C. (2019). Applications of artificial intelligence in agriculture: A review. *Engineering, Technology & Applied Science Research, 9*(4), 4377–4383.

Elshorbagy, A., & Parasuraman, K. (2008). On the relevance of using artificial neural networks for estimating soil moisture content. *Journal of Hydrology, 362*(1–2), 1–18.

Fichter, W. (1984). Reduction of root-mean-square error in faceted space antennas. *AIAA Journal, 22*(11), 1679–1684.

Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PloS one, 12*(1), e0170478.

Gao, B.-C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment, 58*(3), 257–266.

Gascon, F., Cadau, E., Colin, O., Hoersch, B., Isola, C., Fernández, B. L., & Martimort, P. (2014). Copernicus Sentinel-2 mission: products, algorithms and Cal/Val (Vol. 9218, p. 92181E). Presented at the Earth observing systems XIX, International Society for Optics and Photonics.

Gopal, P. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. *Computers and Electronics in Agriculture, 165*, 104968.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18–27.

Gorelick, N. (2013). Google earth engine (Vol. 15, p. 11997). Presented at the EGU General Assembly Conference Abstracts.

Grishin, I., & Timirgaleeva, R. (2020). Remote sensing: The method of GIS application for monitoring the state of soils (Vol. 175, p. 06009). Presented at the E3S Web of Conferences, EDP Sciences.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One, 12*(2), e0169748.

Kah, M., Beulke, S., & Brown, C. D. (2007). Factors influencing degradation of pesticides in soil. *Journal of Agricultural and Food Chemistry, 55*(11), 4487–4492.

Kartalopoulos, S. V., & Kartakapoulos, S. V. (1997). *Understanding neural networks and fuzzy logic: basic concepts and applications*. Wiley-IEEE Press.

Khan, N. M., Rastoskuev, V. V., Shalina, E. V., & Sato, Y. (2001). Mapping salt-affected soils using remote sensing indicators-a simple approach with the use of GIS IDRISI.

Lee, W., Sanchez, J., Mylavarapu, R., & Choe, J. (2003). Estimating chemical properties of Florida soils using spectral reflectance. *Transactions of the ASAE, 46*(5), 1443.

Li, J., & Mocko, M. (2020). Machine learning for a citizen data scientist: an experience with JMP.

Li, S., & Chen, X. (2014). A new bare-soil index for rapid mapping developing areas using landsat 8 data. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 40*(4), 139.

Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. *Geoderma, 170*, 70–79.

Liu, K., He, Y., Xu, S., Hu, L., Luo, K., Liu, X., et al. (2018). Mechanism of the effect of pH and biochar on the phytotoxicity of the weak acid herbicides imazethapyr and 2, 4-D in soil to rice (Oryza sativa) and estimation by chemical methods. *Ecotoxicology and Environmental Safety, 161*, 602–609.

Loveland, T. R., & Irons, J. R. (2016). Landsat 8: The plans, the reality, and the legacy. *Remote Sensing of Environment, 185*, 1–6.

Malley, D. F., Yesmin, L., Wray, D., & Edwards, S. (1999). Application of near-infrared spectroscopy in analysis of soil mineral nutrients. *Communications in Soil Science and Plant Analysis, 30*(7–8), 999–1012.

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma, 117*(1–2), 3–52.

Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S. (2019). Machine learning techniques in wireless sensor network based precision agriculture. *Journal of the Electrochemical Society, 167*(3), 037522.

Merry, R., & Janik, L. (2001). Mid infrared spectroscopy for rapid and cheap analysis of soils. Presented at the Proceedings of the 10th Australian agronomy conference, Australian society of agronomy.

Millard, K., & Richardson, M. (2015). On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing, 7*(7), 8489–8515.

Minasny, B., McBratney, A., Malone, B., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in agronomy, 118*, 1–47.

Mishra, A. (2007). A review on genesis and taxonomic classification of soils of Orissa. *Orissa Review, 63*(6), 53–56.

Neina, D. (2019). The role of soil pH in plant nutrition and soil remediation. *Applied and Environmental Soil Science, 2019*, 1–9.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97*(2), 307.

Pahlavan-Rad, M. R., & Akbarimoghaddam, A. (2018). Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). *CATENA, 160*, 275–281.

Parastatidis, D., Mitraka, Z., Chrysoulakis, N., & Abrams, M. (2017). Online global land surface temperature estimation from Landsat. *Remote Sensing, 9*(12), 1208.

Potin, P., Bargellini, P., Laur, H., Rosich, B., & Schmuck, S. (2012). Sentinel-1 mission operations concept (pp. 1745–1748). Presented at the 2012 IEEE International Geoscience and Remote Sensing Symposium, IEEE.

Ranjbar, F., & Jalali, M. (2016). The combination of geostatistics and geochemical simulation for the site-specific management of soil salinity and sodicity. *Computers and Electronics in Agriculture, 121*, 301–312.

Rikimaru, A., Roy, P., & Miyatake, S. (2002). Tropical forest cover density mapping. *Tropical Ecology, 43*(1), 39–47.

Rodrigo-Comino, J., López-Vicente, M., Kumar, V., Rodríguez-Seijo, A., Valkó, O., Rojas, C., et al. (2020). Soil science challenges in

a new era: A transdisciplinary overview of relevant topics. *Air, Soil and Water Research, 13*, 1178622120977491.

Roelofsen, H. D., van Bodegom, P. M., Kooistra, L., van Amerongen, J. J., & Witte, J.-P.M. (2015). An evaluation of remote sensing derived soil pH and average spring groundwater table for ecological assessments. *International Journal of Applied Earth Observation and Geoinformation, 43*, 149–159.

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., et al. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment, 145*, 154–172.

Sall, J., Stephens, M. L., Lehman, A., & Loring, S. (2017). *JMP start statistics: A guide to statistics and data analysis using JMP*. Sas Institute.

Spadotto, C. A., & Hornsby, A. G. (2003). Organic compounds in the environment: Soil sorption of acidic pesticides: modeling pH effects. *Embrapa Meio Ambiente-Artigo em periódico indexado (ALICE)*.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., & 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.

Todd, S. W., & Hoffer, R. M. (1998). Responses of spectral indices to variations in vegetation cover and soil background. *Photogrammetric Engineering and Remote Sensing, 64*, 915–922.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., et al. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment, 120*, 9–24.

Tucker, C. J., Elgin, J., Jr., McMurtrey Iii, J., & Fan, C. (1979). Monitoring corn and soybean crop development with hand-held radiometer spectral data. *Remote Sensing of Environment, 8*(3), 237–248.

Vogelmann, J., Rock, B., & Moss, D. (1993). Red edge spectral measurements from sugar maple leaves. *TitleREMOTE SENSING, 14*(8), 1563–1575.

von Tucher, S., Hörndl, D., & Schmidhalter, U. (2018). Interaction of soil pH and phosphorus efficacy: Long-term effects of P fertilizer and lime applications on wheat, barley, and sugar beet. *Ambio, 47*(1), 41–49.

Wang, X.-X., Liu, S., Zhang, S., Li, H., Maimaitiaili, B., Feng, G., & Rengel, Z. (2018). Localized ammonium and phosphorus fertilization can improve cotton lint yield by decreasing rhizosphere soil pH and salinity. *Field Crops Research, 217*, 75–81.

Wani, S. P., Chander, G., Bhattacharyya, T., & Patil, M. (2016). Soil health mapping and direct benefit: Transfer of fertilizer subsidy, research report IDC-6.

Wilson, H. F., Satchithanantham, S., Moulin, A. P., & Glenn, A. J. (2016). Soil phosphorus spatial variability due to landform, tillage, and input management: A case study of small watersheds in southwestern Manitoba. *Geoderma, 280*, 14–21.

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. *Ecological Indicators, 60*, 870–878.

Zhang, Y., Sui, B., Shen, H., & Wang, Z. (2018). Estimating temporal changes in soil pH in the black soil region of Northeast China using remote sensing. *Computers and Electronics in Agriculture, 154*, 204–212.