



# Inversion of soil properties in rare earth mining areas (southern Jiangxi, China) based on visible–near-infrared spectroscopy

Jiaxin Guo<sup>1,2</sup> · Xiaomin Zhao<sup>1,2</sup> · Xi Guo<sup>1,2</sup> · Qing Zhu<sup>3</sup> · Jie Luo<sup>1,2</sup> · Zhe Xu<sup>4</sup> · Liang Zhong<sup>1,2</sup> · Yingcong Ye<sup>1,2</sup>

Received: 24 January 2022 / Accepted: 22 May 2022 / Published online: 17 June 2022  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

**Purpose** Traditional measurement for soil properties is time-consuming and costly, while visible–near-infrared spectroscopy enables the rapid prediction of soil properties. In this study, visible–near-infrared spectroscopy was used to predict these four soil properties including OC (organic carbon) content, TN (total nitrogen) content, pH value, and clay content in rare earth mining areas based on different spectral transformation and calibration methods.

**Materials and methods** A total of 232 soil samples were collected from unexploited, in situ leaching, and heap leaching mining areas in southern Jiangxi Province, China. The chemical properties and reflectance spectra of air-dried samples were measured. Spectral transformations including first-order derivative (FOD), continuum removal (CR), and continuous wavelet transform (CWT) were selected to improve the prediction accuracy of the model. Partial least-squares regression (PLSR), the support vector machine (SVM), and extreme gradient boosting (XGBoost) were used to construct prediction models.

**Results and discussion** The highest prediction accuracies in terms of the coefficient of determination ( $R^2$ ), root mean square error (RMSE), and relative prediction deviation (RPD) were obtained using CWT spectra with XGBoost for organic carbon content ( $R^2=0.89$ , RMSE=0.24, RPIQ=4.67), total nitrogen content ( $R^2=0.86$ , RMSE=0.01, RPIQ=4.14), and pH value ( $R^2=0.73$ , RMSE=0.19, RPIQ=1.66). The best prediction result for clay content was obtained using CWT spectra with the SVM ( $R^2=0.67$ , RMSE=6.45, RPIQ=2.75).

**Conclusions** The CWT coupled with a non-linear model, such as XGBoost, is an effective method for the accurate prediction of soil properties in rare earth mining areas.

**Keywords** Visible–near-infrared spectroscopy · Rare earth mine · Soil organic carbon · Extreme gradient boosting

---

Responsible editor: Xiuping Jia

---

✉ Xiaomin Zhao  
zhaoxm889@126.com

<sup>1</sup> Academy of Land Resource and Environment, Jiangxi Agricultural University, Nanchang 330045, China

<sup>2</sup> Key Laboratory of Poyang Lake Watershed Agricultural Resources and Ecology of Jiangxi Province, Nanchang 330045, China

<sup>3</sup> School of Design, Shanghai Jiao Tong University, Shanghai 200240, China

<sup>4</sup> State Key Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

## 1 Introduction

Rare earths, also known as industrial vitamins, are vital strategic resources widely used in many fields, such as the military, petrochemicals, and textile (Yang et al. 2013). Locations with ion-absorption rare earth deposits, such as Ganzhou, Jiangxi Province, China, are commonly characterized by a warm and humid climate, and low undulating hilly landforms. Owing to disorderly activity and outdated technology in the earlier stages of mining, substantial abandoned tailings are present in rare earth mining areas, resulting in serious eco-environmental problems that need to be solved urgently. In particular, the soil properties in rare earth mining areas are seriously affected by the leaching process (Yang et al. 2013).

The major soil type of rare earth mining area is red soil in Ganzhou. With a hot and rainy conditions, the process of

desilicification and allitization during the formation of red soil results in a high content of iron and aluminum oxide in the soil. And clay minerals composed of halloysite and kaolin are formed to provide a good environment for the accumulation of rare earth elements (Li and Zhou 2020). There are three typical types of rare earth mining areas namely unexploited, in situ leaching, and heap leaching mining areas. The unexploited mining area is less affected by human activity and has high vegetation coverage. The in situ leaching and heap leaching areas extract rare earth elements by chemical methods, but the degree of impact on the environment is different. As for the in situ leaching mining area, overall damage to the mountain is minor, with greater damage to the soil and vegetation due to mining mainly occurring around the injection wells and collection ditch. The heap leaching mining area is mainly covered by rare earth tailings, and the soil shows severe desertification with only a few *Pinus massoniana* plants. Collecting soil samples in these three types of mining areas can conduct a more comprehensive study of the soil in rare earth mining areas in order to deal with the eco-environmental problems.

As a sink for atmospheric carbon dioxide, soil plays an important role in achieving global carbon neutrality (Paustian et al. 2016). The impact of human activities on changes in soil properties is a long-term and complicated process (Gu et al. 2021). The extraction process of rare earth elements by using  $(\text{NH}_4)_2\text{SO}_4$  negatively affects soil properties, the most important of which are organic carbon (OC) content, total nitrogen (TN) content, pH value, and clay content. Soil OC and TN influence soil functions related to water and nutrient retention, while also providing nutrients for plant growth. When using  $(\text{NH}_4)_2\text{SO}_4$ ,  $\text{NH}_4^+$  replaces  $\text{Ca}^{2+}$  adsorbed on soil colloids, which destroys the soil aggregate structure and causes a subsequent loss of soil OC and TN. Soil pH is related to the growth and development of animals and plants, with the low pH values in rare earth mining areas found to affect soil health. The  $\text{NH}_4^+$  replaces  $\text{H}^+$  on soil colloids, which will increase  $\text{H}^+$  in soil, leading to soil acidification and compaction (Guo et al. 2010). Soil clay plays an essential role in soil that affect many soil properties and process (Song et al. 2021); it also acts as “glue” to hold soil particles together (Bronick and Lal 2005). Rare earth elements are adsorbed and enriched by clay minerals in ionic form. The usage of  $(\text{NH}_4)_2\text{SO}_4$  results in the replacement of rare earth ions on clay minerals with  $\text{H}^+$  and  $\text{NH}_4^+$  and the destruction of soil binding agent (clay). Therefore, these four soil properties in rare earth mining areas should be monitored in a timely manner to provide support data for soil erosion control and ecological restoration.

Traditionally, soil properties are measured using physical and chemical methods in the laboratory (Greenberg et al. 2020). Although accurate results can be obtained, traditional soil testing methods require substantial labor, materials, and

financial resources. Furthermore, these methods have limitations in large-scale monitoring owing to spatial variability in soil properties. The emergence of visible–near-infrared spectroscopy provides a powerful tool for the rapid monitoring of soil properties (Chen et al. 1989), based on acquiring soil spectral data using a ground spectroradiometer. Studies have shown the potential of soil spectra to predict soil properties. For example, Kovačević et al. (2010) successfully used the Gaussian kernel with the support vector machine (SVM) to predict soil pH values. Zhang et al. (2019) predicted the soil TN content using feature bands and the SVM method. Ji et al. (2019) combined data from four soil spectral sensors to predict soil organic matter (SOM) and pH, and the concentrations of soil ions, including phosphorus, potassium, and calcium. Tsakiridis et al. (2020) found that the convolutional neural network method performed well in predicting the soil clay, silt, and sand contents, pH value, cation exchange capacity (CEC), and OC,  $\text{CaCO}_3$ , and N contents in the LUCAS topsoil database. The mechanisms for predicting soil properties based on visible–near-infrared spectroscopy depend on different spectral interactions of the main soil chromophores (Vohland et al. 2011).

Extracting useful information from the original soil spectrum is difficult owing to spectral overlaps occurring in the visible and near-infrared range (Stenberg et al., 2010; Chen et al. 2020). Therefore, spectral transformation methods are used for spectral pre-processing to reduce the influence of environmental noise and enhance the useful information. The first-order derivative (FOD) can remove interference from linear or nearly linear background noise to improve analysis accuracy (Ben-Dor et al. 1997). The continuum removal (CR) generally magnifies the absorption and reflection characteristics in spectra, and the spectra are normalized to a consistent spectrum background, which is beneficial to identify feature bands (Clark and Roush 1984; Tziolas et al. 2020). As an effective signal processing method, the continuous wavelet transform (CWT) is able to decompose the original spectrum into multi-scale wavelet coefficients through operations such as scaling and translation. This decomposition process can enhance certain information in the spectrum, including the location and nature of high-frequency features (narrow absorption features, spikes, and noise), or the size and shape of continuous features on a large scale (Vohland et al. 2016).

Soil spectral analysis commonly uses linear and non-linear calibration methods. As a linear multivariate regression method, partial least-squares regression (PLSR) is superior to other regression methods, such as stepwise multiple regression (SMLR) and principal component regression (PCR), in processing multi-dimensional collinearity data (Conforti et al. 2015; Shi et al. 2013). However, soil is formed under the effects of multiple factors, such as parent-material, topography, and climate. Owing to the complex

composition of soil, the relationship between soil spectra and soil properties might not be a simple linear relationship (Vohland et al. 2011). Non-linear methods might outperform linear methods in dealing with such issues. For example, the SVM is a useful tool for solving non-linear problems with multi-dimensional data and small sample sizes (Nawar et al. 2016). Based on soil spectroscopy, the SVM has been successfully used to predict soil properties. Nawar and Mouazen (2017) found that the SVM and multivariate adaptive regression splines (MARS) outperformed PLSR and achieved similar accuracy for predicting soil TN, total carbon (TC), and water content at different geographical scales. Furthermore, extreme gradient boosting (XGBoost) based on the gradient descent algorithm can solve classification, regression, and sorting issues (Chen and Guestrin 2016), but is rarely used to predict soil properties based on spectroscopy. Wei et al. (2019) found that XGBoost performed well in estimating the arsenic (As) content in soil, suggesting that this method has potential to predict other soil properties.

Recently, owing to the emergence of data mining and deep-learning methods, a growing number of studies have focused on the prediction of soil properties using large spectral libraries (Tsakiridis et al. 2020; Zhong et al. 2021). Although a model with high overall accuracy can be obtained, its applicability might not be high when the prediction is downscaled to a single soil type in a specific area, owing to the large coverage scale of the soil samples. Therefore, the aims of this study were (1) to verify the feasibility of using visible–near-infrared spectroscopy to predict the OC content, TN content, pH value, and clay content of soil in rare earth mining areas; and (2) to select the optimal spectral transformation method (FOD, CR, and CWT) coupled with different calibration methods (PLSR, SVM, and XGBoost) for predicting soil properties based on visible–near-infrared spectroscopy.

## 2 Materials and methods

### 2.1 Study area and sampling

The study area, located in Longnan, Dingnan, and Xinfeng Counties, Jiangxi Province, China, was rich in ionic rare earth ores. This area has a mid-subtropical humid climate, with a mean annual rainfall of 1500–1600 mm and a mean annual temperature of 19.0 °C. The elevation is in the range of 200–400 m, and the major landform is low hills, with forestland as the dominant land-use type. The soil type in the study area is mainly red soil (Alumi-Ferric Alisols) and the clay minerals mainly consist of kaolinite.

Nine typical mining sites representing unexploited, in situ leaching, and heap leaching mining areas (three sites each) were selected in the study area (Fig. 1). Considering

the danger of sampling in the mountains, soil samples of the unexploited mining area were mainly collected on both sides of the mountain road (> 5 m distance from the road). In the in situ leaching mining area, soil sampling was mainly conducted near the injection wells. Soil sampling points in the heap leaching mining areas were evenly distributed.

A total of 232 topsoil samples (depth, 0–20 cm) were collected using a five-point sampling method in June and July 2020. The center position of each sample was recorded using a handheld global positioning system (Fig. 1). After removing plant roots and stones by hand, all soil samples were kept in resealable bags and labeled. The samples were then air-dried, ground, and passed through a 10-mesh sieve (2 mm). Each sample was equally divided into three parts for the determination of soil properties, measurement of soil spectra, and further use.

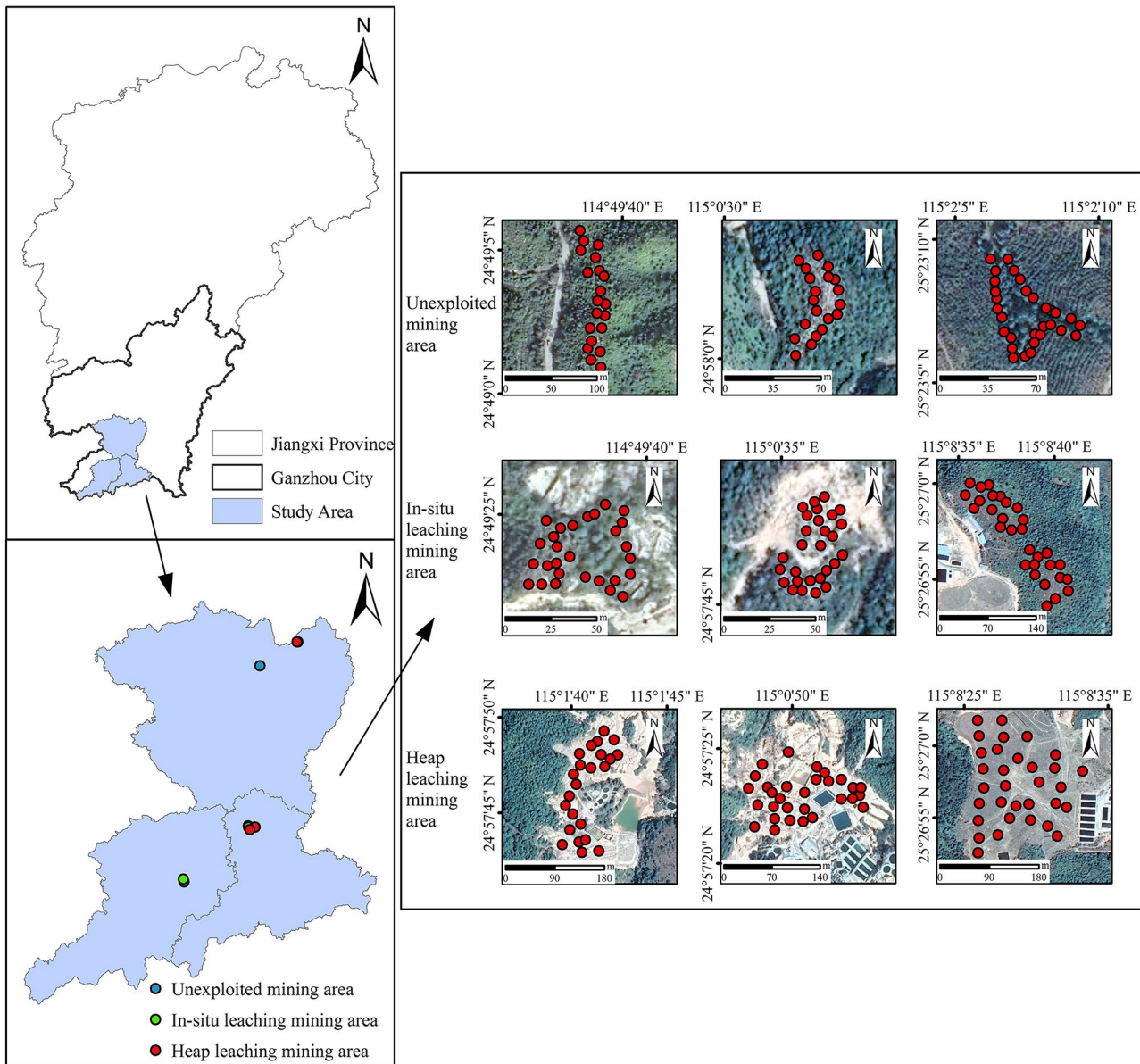
### 2.2 Chemical analysis

The OC and TN contents of soil samples passed through a 100-mesh sieve (0.149 mm) were determined using a Vario MACRO cube elemental analyzer (Elementar, Hanau, Germany) based on the combustion–oxidation method (Wang et al. 2020). As the measured soil pH values were strongly acidic, the samples contained nearly no inorganic carbon, and the OC content was considered to be equal to the TC content. The pH values of soil samples (2 mm) were measured using a potentiometric method with a water–soil ratio of 2.5:1 (v/w) (Kovačević et al. 2010). The clay content of the soil samples (2 mm) was measured using the pipette method (Kilmer and Alexander 1949).

### 2.3 Spectral measurement and pre-processing

An ASD FieldSpec 4 spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO, USA) was used to obtain the spectral reflectance of soil samples in the range of 350–2500 nm. The spectral sampling resolutions of the instrument were 3 nm (at 700 nm) and 10 nm (at 1400 and 2100 nm). The spectra were resampled to 2-nm intervals and 2151 bands were exported for each spectrum. Spectral measurements were conducted in a dark room to reduce interference from external light sources. Soil samples were placed in a black sample container kept in the slot at the top of the MugLite instrument (Analytical Spectral Devices Inc.) and measured with the built-in light source. A white Spectralon panel (Analytical Spectral Devices Inc.) was used to calibrate the instrument every 10 min. Each soil sample was scanned five times and the mean of the spectra was used as the final spectrum for each sample.

The splice correction function in ViewSpecPro v6.0 (Analytical Spectral Devices Inc.) was used to eliminate the effects of breakpoints generated by the instrument when



**Fig. 1** Locations of the study area and soil sampling points in Jiangxi Province, China

measuring soil spectra. The sections at 350–399 nm and 2451–2500 nm, which were considerably affected by the instrument and environmental noise during the measurement, were removed. The Savitzky–Golay filter was then used to smooth the spectrum and remove noise caused by the instrument and environment, while maintaining the original spectral characteristics (Savitzky and Golay 1964).

## 2.4 Spectral transformation methods

The first-order derivative (FOD), continuum removal (CR), and continuous wavelet transform (CWT) were selected to compare with the original reflectance.

The Mexican hat (Torrence and Compo 1998) was selected as the mother wavelet function for CWT, and transformed it into a set of wavelet coefficients on different scales (Mallat 1989). The decomposition scales were set at  $2^1$ ,  $2^2$ ,  $2^3$ , ..., and  $2^{10}$  to prevent data redundancy (Cheng et al. 2011).

## 2.5 Calibration methods

### 2.5.1 Partial least-squares regression

PLSR is a linear multivariate regression method that projects the independent (X) and dependent (Y) variables into

a new space and identifies the relationship between them to construct a prediction model (Viscarra Rossel and Behrens 2010; Wold et al. 2001). PLSR is able to extract the main information from multiple independent variables by reducing the dimensionality and effect of multicollinearity in the independent variables. The correlation between independent and dependent variables is also considered, with the dependent variables predicted through several latent variables extracted from multiple independent variables. This method is suitable for situations where the number of samples is less than the number of independent variables (Kuang et al. 2015). In this study, the number of latent variables in the PLSR model was determined by ten-fold cross-validation, and model construction was implemented in The Unscrambler X v10.4 (CAMO, Oslo, Norway).

### 2.5.2 Support vector machine

The SVM is a non-linear model in machine learning that projects the input data into a feature plane and finds an optimal plane that can minimize the distance from all samples to the plane (Wang et al. 2019). To reduce the complexity of the calculation and prevent dimensional disaster, the kernel function is introduced, which can solve high-dimension problems by calculating them under low dimensions (Smola and Schölkopf 2004). In this study, the radial basis function (RBF) kernel was used to construct the SVM model. Two main parameters,  $C$  (cost parameter) and  $\gamma$  (kernel parameter), must be optimized in the construction of the SVM model (Hong et al. 2018; Dong et al. 2021). Therefore, a grid search with ten-fold cross-validation was used to optimize the  $C$  and  $\gamma$  values. The SVM model was constructed using Matlab R2017b (MathWorks Inc.).

### 2.5.3 Extreme gradient boosting

XGBoost (Chen and Guestrin 2016) is a scalable end-to-end tree boosting system algorithm inspired by the gradient enhancement algorithm (Friedman 2001). XGBoost not only uses the first derivative of the loss function, but also performs a second-order Taylor expansion of the loss function by accounting for the second derivative information. Consequently, the model converges quickly and its operating efficiency is improved (Friedman et al. 2000). In this study, the root mean square error (RMSE) was used as the loss function to evaluate the optimal objective function. A regular term was added to the calculation process of the model objective function, which can improve the generalization ability to prevent over-fitting of the prediction model. When encountering a situation with a large amount of data, a multi-threaded parallel method was used to improve the computational efficiency (Wei et al. 2019).

For a given dataset of  $n$  samples and  $m$  independent variables, the objective function of XGBoost can be defined as follows:

$$\text{Obj}(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{t=1}^T \Omega(f_t) \quad (1)$$

where  $l$  is the loss function.  $y_i$  and  $\hat{y}_i$  are the measured and predicted value of the number  $i$  sample, respectively.  $f_t$  is the number  $t$  tree.  $\sum_{t=1}^T \Omega(f_t)$  is the sum complexity of  $t$  trees, which is used as a regular term in the objective function.

Expanding the objective function according to Taylor's formula, the second-order Taylor expression of the loss function after  $t$  iterations can be obtained, which can be approximately expressed as follows:

$$L^{(t)} = \sum_{i=1}^k [l(y_i, \hat{y}^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_t) \quad (2)$$

where  $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$  and  $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)})$  are the first-order and second-order partial derivatives of the loss function, respectively.

In the XGBoost model, grid search was used to optimize hyperparameters, and the hyperparameters were tuned as follows: the maximum depth of the tree was set to 7; learning rate was set to 0.1 to control length of each iteration step; and the number of trees (n-estimators) was set to 80. The XGBoost model was constructed using the xgboost package in Python v3.7 (<https://www.python.org/>).

## 2.6 Model accuracy evaluation

The accuracy of the constructed prediction models was evaluated using the coefficient of determination ( $R^2$ ), RMSE, and ratio of performance to inter-quartile distance (RPIQ). Larger  $R^2$  or RPIQ values and a smaller RMSE value indicated better prediction accuracy. The prediction performance of models was divided into four categories according to the RPIQ values, as follows: excellent ( $\text{RPIQ} \geq 4.05$ ); good ( $3.37 \leq \text{RPIQ} < 4.05$ ); approximately quantitative ( $2.70 \leq \text{RPIQ} < 3.37$ ); distinguishing between high and low values ( $2.02 \leq \text{RPIQ} < 2.70$ ); and unsuccessful ( $\text{RPIQ} \leq 2.02$ ) (Saeys et al. 2005; Ludwig et al. 2017).

## 2.7 Statistical analysis

Statistical analysis was conducted using IBM SPSS Statistics v22.0 (IBM Corp., Armonk, NY, USA) and Microsoft Excel v2019 (Microsoft Corp., Redmond, WA, USA). Graphical drawing was performed using ArcGIS v10.5 (ESRI Inc., Redlands, CA, USA) and OriginPro v2021 (OriginLab Corp., Northampton, MA, USA).

The FOD, CR, and CWT spectral transformations were performed using OriginPro v2021 (OriginLab Corp., Northampton, MA, USA), ENVI Classic v5.5 (Harris Geospatial Inc., Bloomfield, CO, USA), and Matlab R2017b (MathWorks Inc., Natick, MA, USA), respectively.

### 3 Results

#### 3.1 Descriptive statistics of soil properties

The obtained 232 soil samples were divided into two parts by the Kennard–Stone method (Kennard and Stone 1969), namely, the calibration dataset ( $N=174$ ) and the validation dataset ( $N=58$ ). Descriptive statistics of the measured soil OC content, TN content, pH value, and clay content for the whole, calibration, and validation datasets are summarized in Table 1. It could be observed that the OC content, TN content, pH value, and clay content of the whole dataset ranged from 0.04 to 2.08%, 0.01 to 0.18%, 3.90 to 5.84%, and 3.82 to 47.30%, respectively. The CV (coefficient of variation) of OC content, TN content, and clay content were higher than 35%, whereas the CV of pH value was below 15%, which meant that the pH value of the study area had little variability according to Wilding (1985).

The mean, SD (standard deviation), and CV of the whole, and calibration and validation datasets were similar. And Levene's test (Levene 1960) was conducted to prove the reliability of the method used to split datasets. The  $p$ -values of the four soil properties from Levene's test were 0.011, 0.298, 0.666, and 0.759 (significance level,  $\alpha = 0.01$ ), respectively, indicating that the calibration and validation datasets had equal variances and could represent the whole dataset.

#### 3.2 Soil spectra and transformations

Figure 2a shows the original (OR) spectrum after averaging the obtained 232 soil spectra. The soil spectrum increased rapidly in the visible range, showing a steep slope. Slight changes were observed from 800 to 1300 nm and from 1500 to 1800 nm, while the spectrum exhibited a slow downward trend in the range of 2200–2450 nm.

Figures 2b–d show the soil spectra after three different transformations. When transformed using the FOD, the spectrum mainly fluctuated near 0, and bands, such as those at 500, 1400, 1900, and 2200 nm, became more distinct. After CR transformation, absorption valleys were observed in the spectrum at approximately 500, 900, 1400, 1900, and 2200 nm. The range of wavelet coefficients obtained by the CWT increased with increasing scale. The wavelet coefficients at scales 1–6 had a small variation range, fluctuating between  $-1$  and  $1$ . The wavelet coefficients at scales 7–10 exhibited a remarkable expansion of the value range. At scale 8, two wavelet coefficient peaks appeared near 800 and 2200 nm, respectively, while at scales 9 and 10, the curves were smooth with a convex center at approximately 1400 nm.

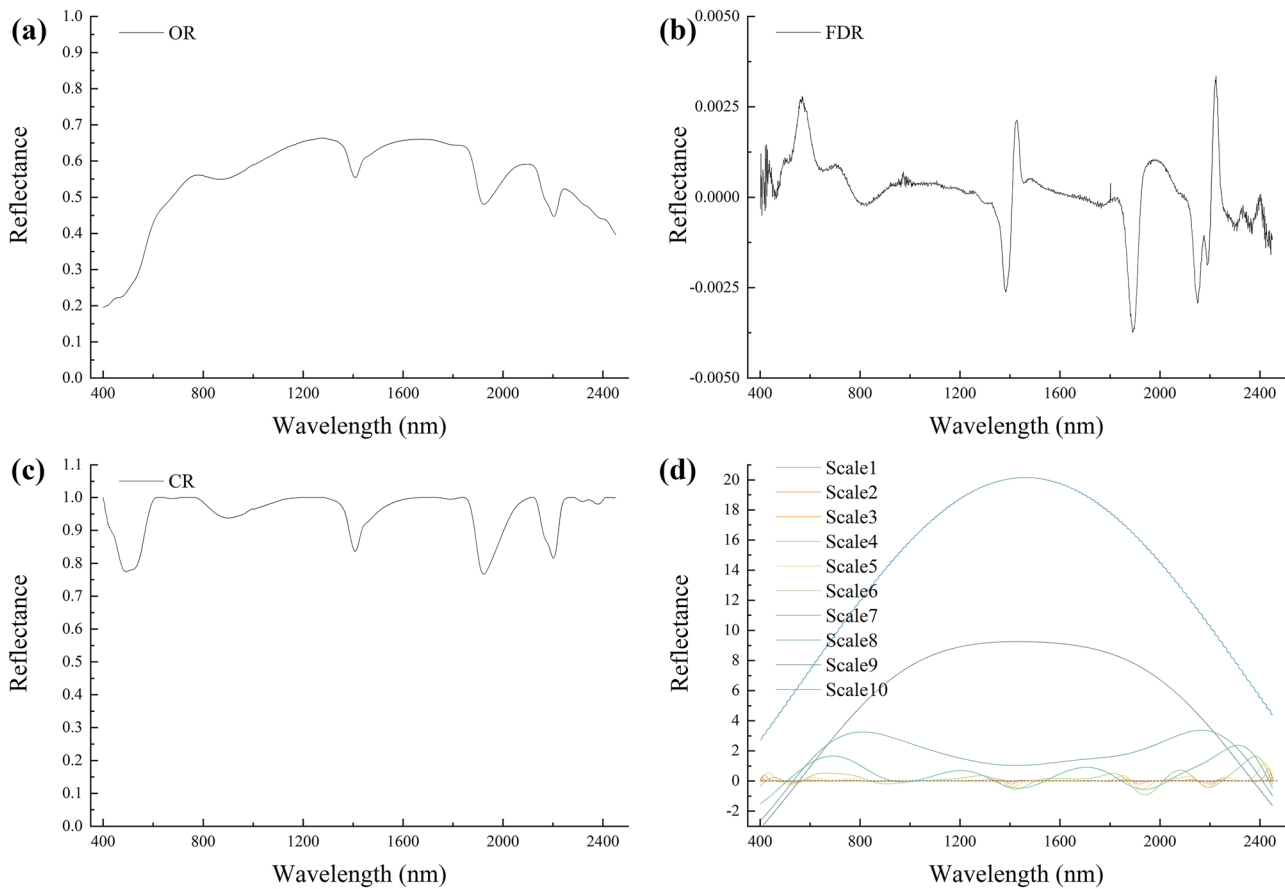
#### 3.3 Correlation between soil properties and spectra

Pearson correlation analysis was used to measure the relationship between soil properties (OC, TN, pH, and clay) and soil spectra (OR, FOD, CR, and CWT). Bands in the range of 600–800 and 2000–2400 nm were highly correlated with the OC content (Fig. 3). The highest correlation with the OC content was observed at 793 nm for FOD spectra, and at 2110, 1290, and 1945 nm for OR, CR, and CWT1 spectra, respectively. The OR spectra were negatively correlated

**Table 1** Descriptive statistics of soil properties in the study area

Soil property	Dataset	Sample number	Min	Max	Mean	SD	CV	Skewness	Kurtosis
OC	Whole	232	0.04%	2.08%	0.66%	0.57%	86.58%	0.69	-0.69
	Calibration	174	0.04%	2.08%	0.67%	0.54%	81.44%	0.62	-0.69
	Validation	58	0.05%	2.03%	0.65%	0.66%	101.80%	0.81	-0.80
TN	Whole	232	0.01%	0.18%	0.06%	0.04%	63.90%	0.73	-0.27
	Calibration	174	0.01%	0.18%	0.06%	0.04%	64.44%	0.75	-0.23
	Validation	58	0.01%	0.14%	0.06%	0.04%	62.05%	0.59	-0.73
pH	Whole	232	3.90	5.84	4.55	0.37	8.07%	0.95	0.78
	Calibration	174	3.90	5.84	4.56	0.37	8.11%	0.89	0.69
	Validation	58	3.97	5.65	4.53	0.36	7.99%	1.17	1.34
Clay	Whole	232	3.82%	47.30%	20.08%	11.26%	56.17%	0.52	-1.05
	Calibration	174	3.82%	47.30%	20.56%	11.29%	54.90%	0.45	-1.09
	Validation	58	5.34%	42.12%	18.63%	11.17%	60.05%	0.79	-0.77

OC organic carbon, TN total nitrogen, SD standard deviation, CV coefficient of variance



**Fig. 2** a Soil spectral curves of original reflectance (OR) and (b–d) transformed reflectance

with the OC content over the full wavelength range, with almost all bands passing the significance test. The correlation coefficients between FOD, CR, or CWT spectra and the OC content fluctuated between positive and negative values (Fig. 3a). The highest correlation between soil spectra and the TN content was found at 793 nm in the FOD spectra, with the correlation coefficient being higher than the corresponding coefficient between soil spectra and the OC content. However, the number of bands that passed the significance test was slightly smaller for the TN content than for the OC content (Fig. 3b).

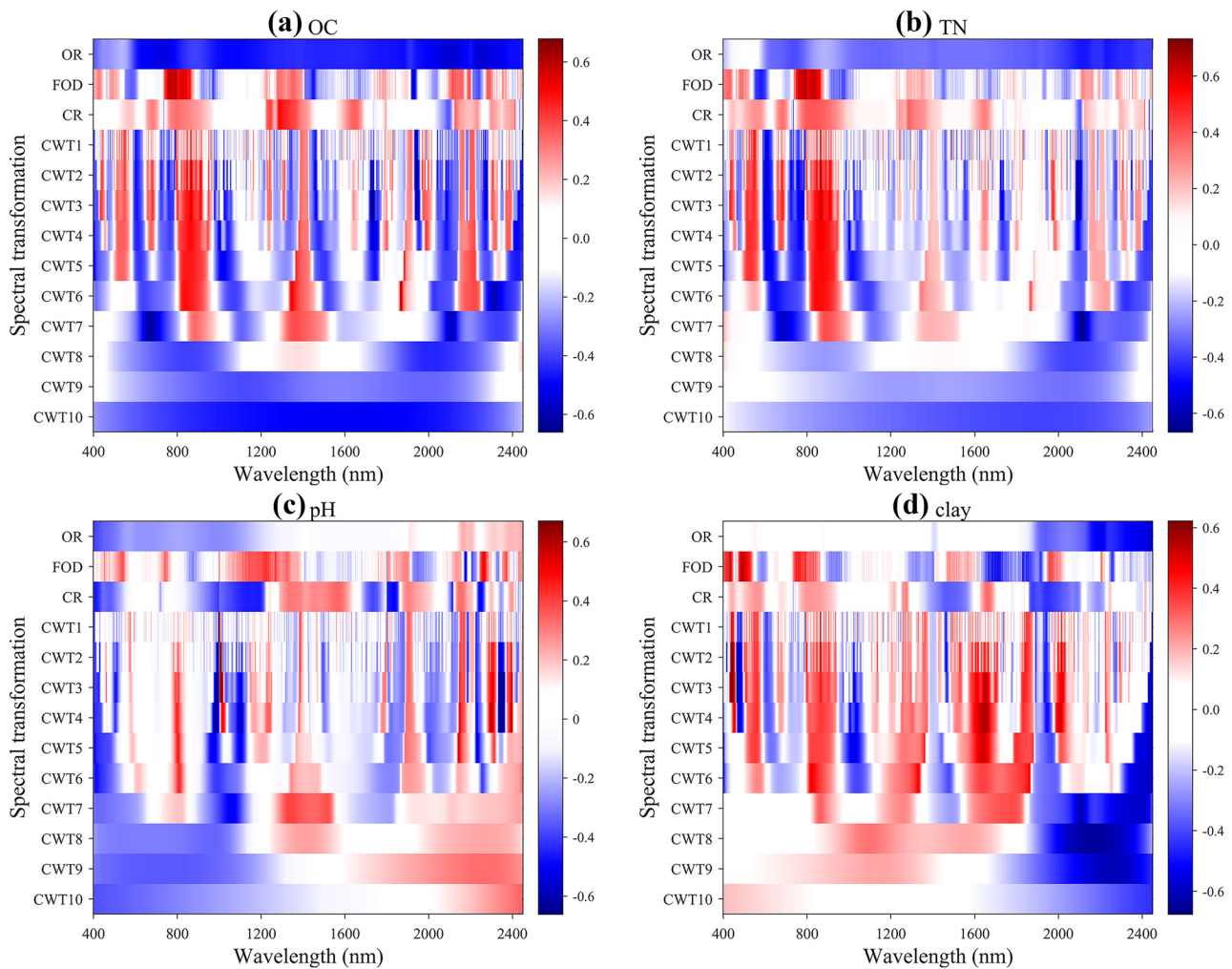
The soil pH value showed a relatively low correlation with the OR, FOD, and CR spectra. The CWT spectra showed a high correlation with the pH value, mainly at approximately 400, 1000, and 2400 nm at low decomposing scales, with the highest correlation observed at 1011 nm in the CWT2 spectra (Fig. 3c). The clay content of soil showed a lower correlation with the OR spectra compared with the OC content. Bands in transformed spectra in the range of 400–500, 1600–1700, and 2000–2450 nm showed a relatively high correlation with the clay content, with the highest correlation observed at 1672 nm in the FOD spectra.

For the CWT spectra, bands from 400 to 500 nm at scales 1–4 appeared as narrow features, while bands from 2000 to 2450 nm at scales 7–9 appeared as broad features (Fig. 3d).

### 3.4 Prediction of soil properties

For calibration dataset, the prediction accuracy of the XGBoost method for soil properties were higher than the PLSR and SVM method. The best results of the OC content, TN content, pH value, and clay content for the calibration dataset were obtained using XGBoost based on CWT spectra (OC:  $R^2=0.99$ , RMSE=0.05; TN:  $R^2=0.99$ , RMSE=0.01; pH:  $R^2=0.97$ , RMSE=0.07; clay:  $R^2=0.99$ , RMSE=1.54; Table 2).

To evaluate the influence of different spectral transformation and calibration methods on the prediction of soil properties, RPIQ value was added to calculate for the validation dataset. For the OC content, nearly all models constructed based on FOD and CWT spectra outperformed the models based on OR spectra (Table 3 and Fig. 4). In particular, the best results with CWT spectra were obtained using SVM ( $R^2=0.88$ , RMSE=0.26, RPIQ=4.37) and XGBoost



**Fig. 3** Correlations between soil properties (**a** OC, **b** TN, **c** pH, and **d** clay) and spectra. Pearson correlation coefficients that pass the  $p=0.01$  significance test are shown

( $R^2=0.89$ , RMSE=0.24, RPIQ=4.67). Models based on CR spectra yielded the worst result when coupled with SVM ( $R^2=0.62$ , RMSE=0.45, RPIQ=2.44). Compared with the OC content, the prediction accuracy for the TN content was lower, with the best result obtained using CWT spectra with XGBoost ( $R^2=0.86$ , RMSE=0.01, RPIQ=4.14). For each calibration method, the results obtained with OR, FOD, and CWT spectra had higher accuracy than those obtained with CR spectra (Table 3 and Fig. 5). The model based on CWT spectra performed well with the SVM and XGBoost methods. For OR and FOD spectra, the models' accuracies obtained using SVM and XGBoost methods were similar, and lower than those based on CWT spectra. The best prediction result for soil pH value was obtained using CWT spectra with XGBoost ( $R^2=0.73$ , RMSE=0.19, RPIQ=1.95). However, each prediction model had an RPIQ value below 2.02, indicating that the model could not successfully predict soil pH value (Table 3 and Fig. 6). For clay

content, the best result was obtained with CWT spectra and the SVM method ( $R^2=0.67$ , RMSE=6.45, RPIQ=3.12). The models constructed with original or transformed spectra and different calibration methods were mostly able to approximately quantitative the clay content (Table 3 and Fig. 7).

## 4 Discussion

### 4.1 Features of soil spectra and transformations

For the original soil spectra (OR), five absorption features were observed at approximately 400–600, 900, 1400, 1900, and 2200 nm, respectively. The absorption feature at approximately 400–600 nm was associated with humus and iron (Palacios-Orueta and Ustin 1998; Stoner and Baumgardner 1981). The broad absorption band at



**Table 2** Comparison of the prediction accuracy for soil properties using different spectral transformation and calibration methods based on the calibration dataset ( $N=174$ )

Soil property	ST	PLSR		SVM		XGBoost	
		$R^2$	RMSE	$R^2$	RMSE	$R^2$	RMSE
OC	OR	0.82	0.23	0.59	0.37	0.97	0.10
	FOD	0.91	0.16	0.93	0.16	0.99	0.06
	CR	0.68	0.31	0.74	0.30	0.99	0.06
	CWT	0.77	0.26	0.98	0.08	0.99	0.05
TN	OR	0.75	0.02	0.43	0.03	0.95	0.01
	FOD	0.85	0.02	0.83	0.02	0.99	0.01
	CR	0.65	0.02	0.66	0.03	0.93	0.01
	CWT	0.75	0.02	0.95	0.01	0.99	0.01
pH	OR	0.68	0.21	0.58	0.25	0.86	0.14
	FOD	0.57	0.30	0.73	0.30	0.96	0.08
	CR	0.29	0.33	0.75	0.31	0.94	0.10
	CWT	0.63	0.22	0.94	0.10	0.97	0.07
Clay	OR	0.73	5.90	0.51	7.98	0.96	2.51
	FOD	0.75	5.62	0.80	5.24	0.93	3.22
	CR	0.64	6.76	0.57	7.47	0.96	2.58
	CWT	0.70	6.17	0.93	3.12	0.99	1.54

ST spectral transformation, OR original reflectance, FOD first-order derivative, CR continuum removal, CWT continuous wavelet transform, PLSR partial least-squares regression, SVM support vector machine; XGBoost extreme gradient boosting,  $R^2$  determination of coefficient, RMSE root mean square error

approximately 900 nm was primarily attributed to ferric ion (Stoner and Baumgardner 1981). The absorption valleys at 1400, 1900, and 2200 nm were associated with O–H groups (Viscarra Rossel and Behrens 2010; Whiting et al. 2004).

Specifically, the valley at 1400 nm was attributed to the O–H stretching of water and clay minerals. The absorption at 1900 nm was dominated by hygroscopic water and lattice water retained in the air-dried soil samples. The absorption

**Table 3** Comparison of the prediction accuracy for soil properties using different spectral transformation and calibration methods based on the validation dataset ( $N=58$ )

Soil property	ST	PLSR			SVM			XGBoost		
		$R^2$	RMSE	RPIQ	$R^2$	RMSE	RPIQ	$R^2$	RMSE	RPIQ
OC	OR	0.80	0.33	3.35 <sup>c</sup>	0.69	0.43	2.56 <sup>d</sup>	0.76	0.35	3.17 <sup>c</sup>
	FOD	0.84	0.28	4.00 <sup>b</sup>	0.88	0.28	4.02 <sup>b</sup>	0.77	0.32	3.45 <sup>b</sup>
	CR	0.69	0.39	2.85 <sup>c</sup>	0.62	0.45	2.44 <sup>d</sup>	0.69	0.37	2.98 <sup>c</sup>
	CWT	0.77	0.35	3.17 <sup>c</sup>	0.88	0.26	4.37 <sup>a</sup>	0.89	0.24	4.67 <sup>a</sup>
TN	OR	0.67	0.02	2.76 <sup>c</sup>	0.66	0.02	2.59 <sup>d</sup>	0.67	0.02	2.81 <sup>c</sup>
	FOD	0.75	0.02	3.24 <sup>c</sup>	0.77	0.02	3.21 <sup>c</sup>	0.71	0.02	2.96 <sup>c</sup>
	CR	0.53	0.03	2.34 <sup>d</sup>	0.55	0.03	2.32 <sup>d</sup>	0.51	0.03	2.28 <sup>d</sup>
	CWT	0.67	0.02	2.77 <sup>c</sup>	0.84	0.02	3.91 <sup>b</sup>	0.86	0.01	4.14 <sup>a</sup>
pH	OR	0.50	0.25	1.46 <sup>e</sup>	0.44	0.28	1.34 <sup>e</sup>	0.64	0.22	1.72 <sup>e</sup>
	FOD	0.35	0.30	1.23 <sup>e</sup>	0.34	0.31	1.18 <sup>e</sup>	0.67	0.21	1.78 <sup>e</sup>
	CR	0.09	0.35	1.07 <sup>e</sup>	0.31	0.32	1.14 <sup>e</sup>	0.52	0.25	1.48 <sup>e</sup>
	CWT	0.41	0.28	1.34 <sup>e</sup>	0.71	0.20	1.90 <sup>e</sup>	0.73	0.19	1.95 <sup>e</sup>
Clay	OR	0.60	7.03	2.87 <sup>c</sup>	0.50	7.85	2.57 <sup>d</sup>	0.51	7.99	2.52 <sup>d</sup>
	FOD	0.61	6.96	2.89 <sup>c</sup>	0.62	6.89	2.92 <sup>c</sup>	0.52	7.85	2.56 <sup>d</sup>
	CR	0.61	6.97	2.89 <sup>c</sup>	0.49	8.19	2.46 <sup>d</sup>	0.37	8.88	2.27 <sup>d</sup>
	CWT	0.59	7.08	2.85 <sup>c</sup>	0.67	6.45	3.12 <sup>c</sup>	0.63	6.78	2.97 <sup>c</sup>

RPIQ ratio of performance to inter-quartile distance

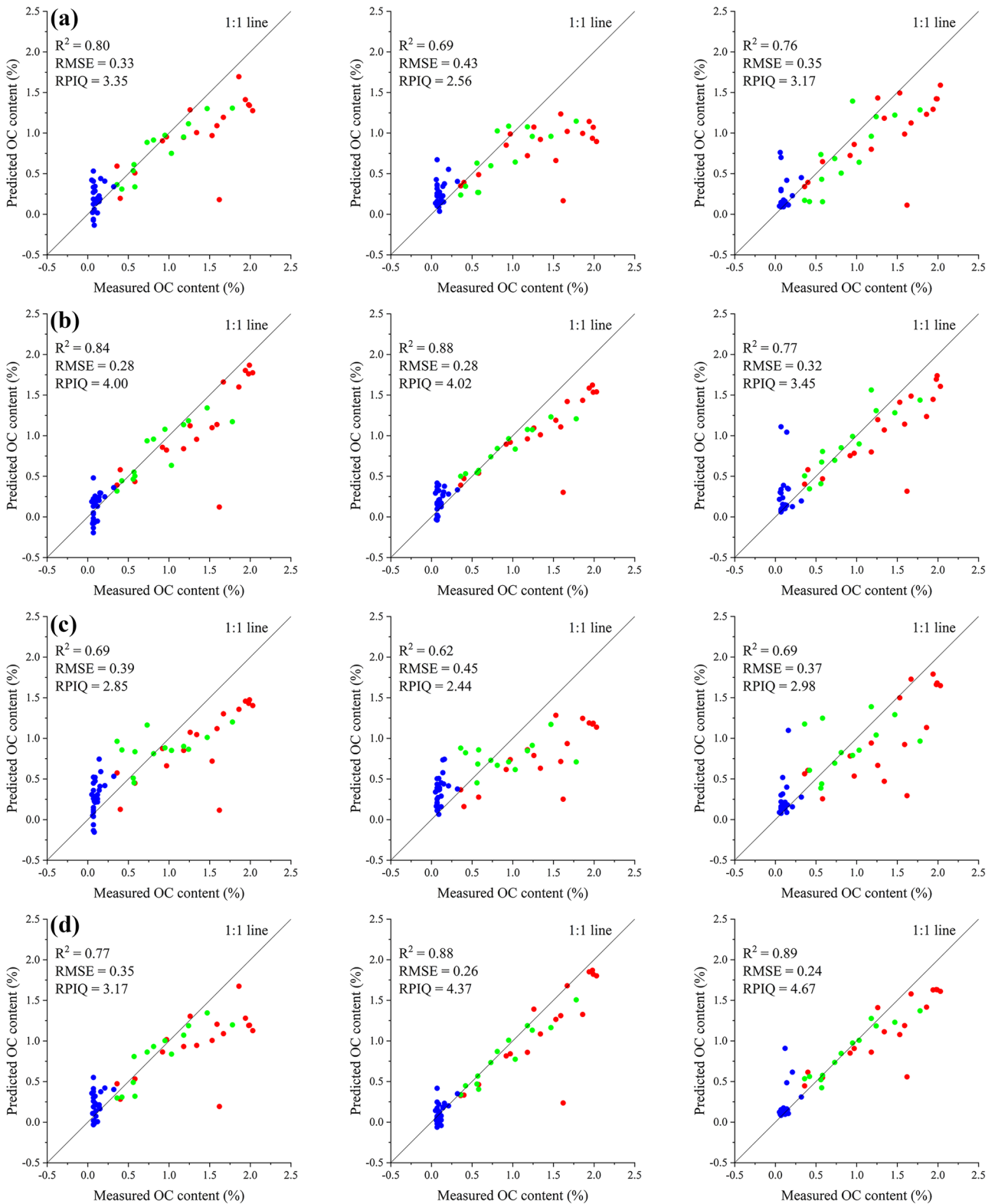
<sup>a</sup>excellent ( $RPIQ \geq 4.05$ )

<sup>b</sup>good ( $3.37 \leq RPIQ < 4.05$ )

<sup>c</sup>approximately quantitative ( $2.70 \leq RPIQ < 3.37$ )

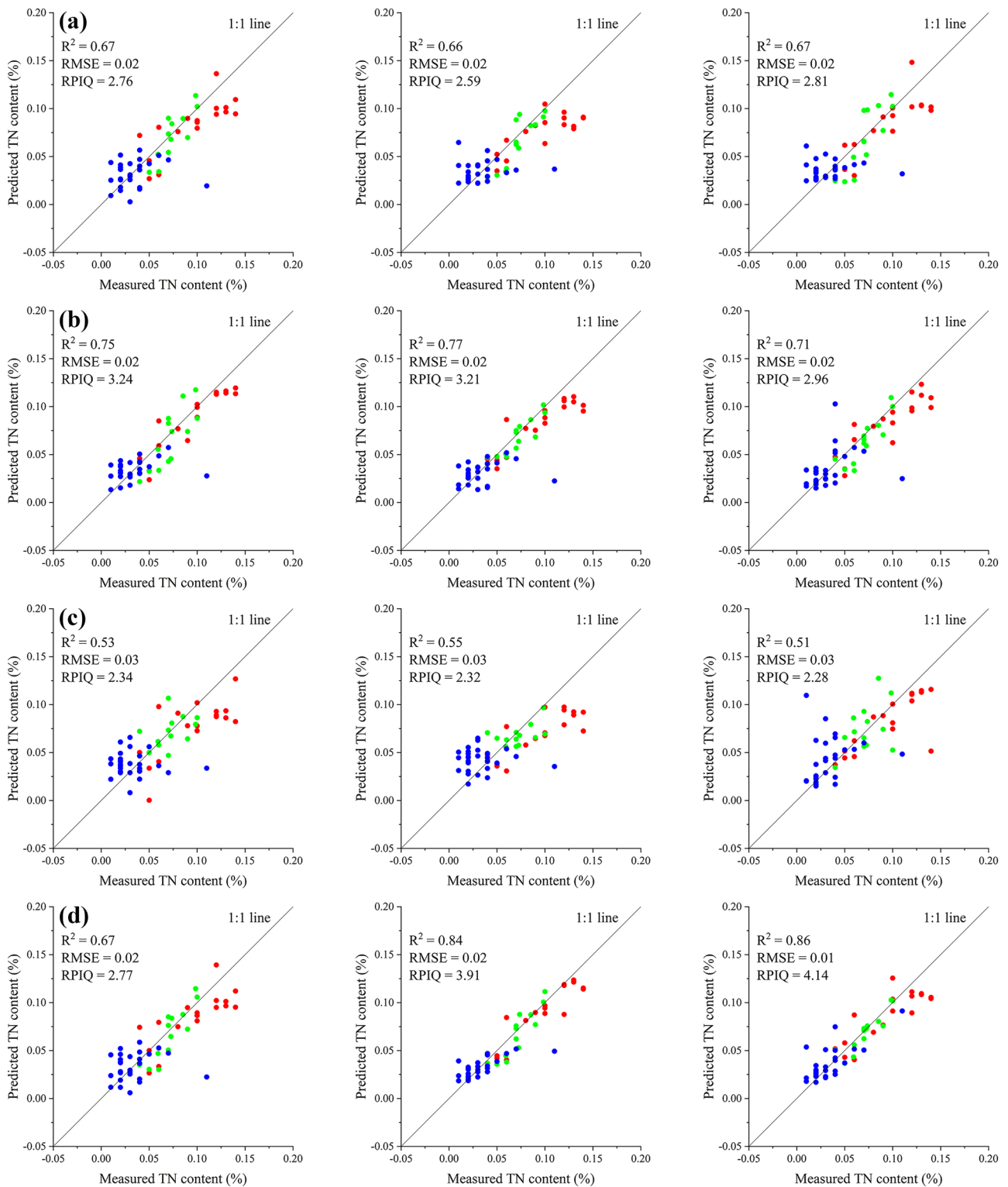
<sup>d</sup>distinguishing between high and low values ( $2.02 \leq RPIQ < 2.70$ )

<sup>e</sup>unsuccessful ( $RPIQ \leq 2.02$ )

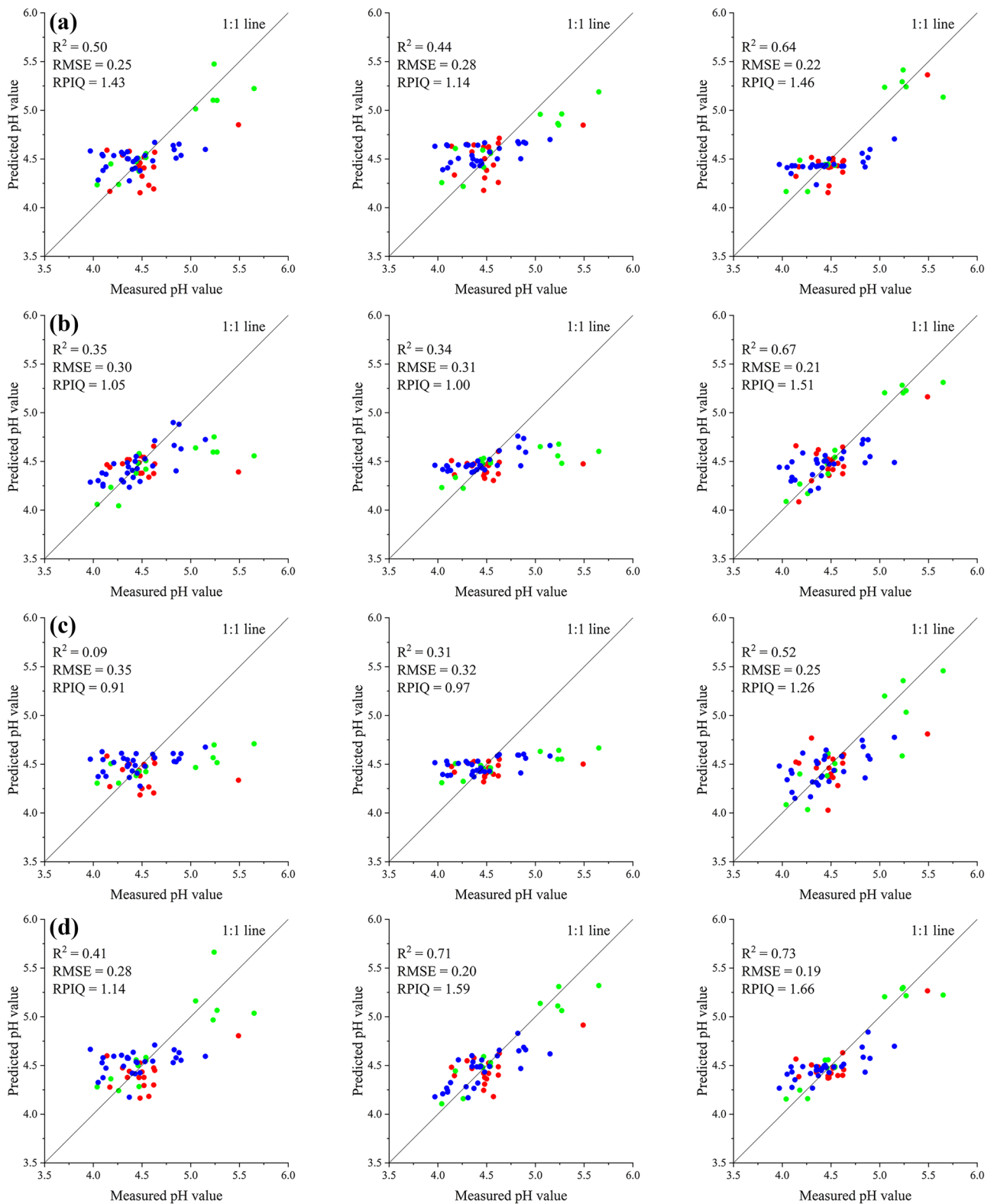


**Fig. 4** Scatter plot of OC content models based on **a** OR, **b** FOD, **c** CR, and **d** CWT spectra with partial least-squares regression (PLSR, left), support vector machine (SVM, middle), and extreme gradient boosting (XGBoost, right) methods using the validation dataset. The

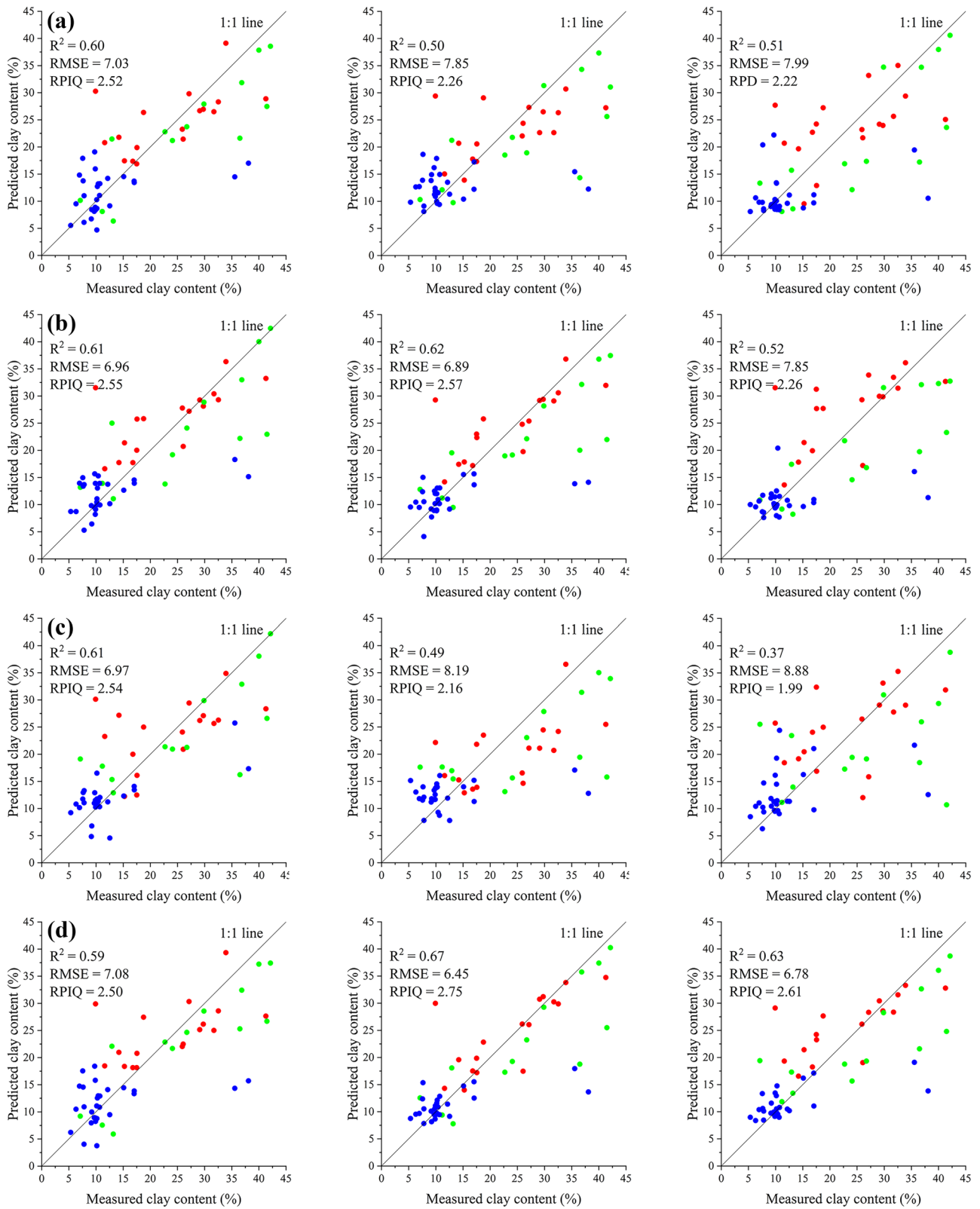
soil samples from unexploited, in situ leaching, and heap leaching mining areas are in red, green, and blue colors, respectively (the same below)



**Fig. 5** Scatter plot of TN content models based on **a** OR, **b** FOD, **c** CR, and **d** CWT spectra with PLSR (left), SVM (middle), and XGBoost (right) methods using the validation dataset



**Fig. 6** Scatter plot of pH value models based on **a** OR, **b** FOD, **c** CR, and **d** CWT spectra with PLSR (left), SVM (middle), and XGBoost (right) methods using the validation dataset



**Fig. 7** Scatter plot of clay content models based on **a** OR, **b** FOD, **c** CR, and **d** CWT spectra with PLSR (left), SVM (middle), and XGBoost (right) methods using the validation dataset

at 2200 nm was attributed to Al–OH bending and stretching in clay minerals (Bishop et al. 1994).

Many studies have explored the relationship between soil properties and spectra. Dalal and Henry (1986) reported that soil OC and TN contents have the same feature bands at approximately 1870 and 2052 nm. Shi et al. (2013) found that bands at 1450, 1850, 2250, 2330, and 2430 nm were essential for predicting the soil TN content. Jiang et al. (2017) observed 400–800, 1900, and 2000–2350 nm as essential band ranges for predicting soil OC and TN contents. The range of bands highly correlated with soil OC and TN contents was fairly similar, mainly concentrated at 600–800 and 2000–2400 nm. However, there is no consensus on whether the TN content in soil is predicted by its correlation with the OC content or based on its own spectral features (Jiang et al. 2017; Zhang et al. 2019). Despite this, the TN content clearly has a close relationship with the OC content of soil, because most nitrogen in the topsoil is organic and generally accounts for one-tenth of the OC carbon content in soil (Stenberg et al. 2010). In the present study, the OC and TN contents showed a significant correlation ( $r=0.84$ ,  $p<0.01$ ), and their correlation coefficients with OR spectra were similar, despite the slight variation in coefficient values (Fig. 3a and b).

The mechanisms used to predict soil pH based on visible–near-infrared spectroscopy are mainly related to organic materials, iron oxides, and clay minerals (Viscarra Rossel and Behrens 2010). Vašát et al. (2014) found that, in the range of 350–2500 nm, no band was highly correlated with soil pH, while bands at 400, 800, 1400, 1850, and 2300 nm recognized by PLSR might be correlated with soil color, OC content, and clay minerals. The results of our study also showed that bands at 400, 1000, and 2300 nm were highly correlated with soil pH, in partial agreement with previous studies. Furthermore, Peng et al. (2014) attributed bands at 410–572, 1400, 1900, 2200, and 2300 nm to iron, water, O–H stretching, aluminum, and magnesium in clay minerals. Nawar et al. (2016) observed that bands at 1900, 2000, and 2200 nm in CR spectra showed strong correlations with the clay content of soil. Similarly, the results of our study showed absorption features attributed to clay content at approximately 400–500 and 2000–2450 nm.

## 4.2 Comparison of prediction models

Three calibration methods (PLSR, SVM, and XGBoost) were used to compare the prediction accuracy of models constructed based on OR, FOD, CR, and CWT spectra for soil properties in the study area.

For spectral transformations, compared with OR, FOD and CWT had a better improvement in prediction accuracy, while the accuracy of CR decreased. The effectiveness of CR transformation has been reported by Vašát et al. (2014).

However, in the present study, CR was the worst spectral transformation method for predicting soil properties. For CR spectra, the most prominent objects in the data normalization process were feature peaks and troughs. Other detailed information in the data might not be displayed well, leading to reduced prediction accuracy. FOD transformation could remove baseline drift and enhance absorption features, thereby improving prediction accuracy (Hong et al. 2018). CWT was the optimal spectral transformation method for predicting the soil properties in the present study. Decomposing the spectrum at multiple scales provided more possibilities for predicting soil properties and capturing useful information hidden in the spectrum. As the decomposing scale increases, the width of the adsorption features correlated with soil properties also increases.

For calibration methods, SVM and XGBoost outperformed PLSR in predicting soil properties. Nawar et al. (2016) and Yang et al. (2019) all proved that non-linear models are superior for predicting soil properties based on visible–near-infrared spectroscopy. The linear regression method, PLSR, might not be able to integrate a large amount of information and extract effective information, while the non-linear methods, SVM and XGBoost, could solve this problem well (Zhang et al. 2019). Previously, Viscarra Rossel and Behrens (2010) found that the prediction accuracy of the non-linear model was positively correlated with the number of variables used for modeling. When there were more variables, the model could extract more features.

The best prediction accuracy for OC and TN were obtained using CWT spectra with XGBoost, and the RPIQ values were both higher than 4.05, indicating excellent results. Furthermore, although XGBoost outperformed the PLSR and SVM methods in predicting pH value, all of the constructed models were unsuccessfully to predict it ( $RPIQ < 2.02$ ). For prediction of the clay content, most constructed models could approximately quantify it.

The OC has broad absorption bands in the visible range, while the TN has almost the same feature bands. TN in soil is closely related to organic matter, most of the N is organic and stored in nitrogen-containing compounds (Stenberg et al. 2010). The clay is negatively correlated with soil spectral and is mainly affected at 1400, 1900, and 2200 nm (Peng et al. 2014). However, the pH has no direct spectral response in the visible–near-infrared range, the previous studies reported the prediction mechanisms for pH value might be due to other soil properties (Viscarra Rossel and Behrens 2010; Vašát et al. 2014). Furthermore, the variation of the sample source is also an important factor affecting the prediction accuracy of the models (Stenberg et al. 2010). The greater the variation of the soil sample dataset, the better the prediction accuracy. The CV of pH value was below 10%, which might influence the prediction accuracy.

## 5 Conclusions

In this study, three spectral transformation methods (FOD, CR, and CWT) were used to predict soil OC content, TN content, pH value, and clay content in rare earth mining areas based on visible–near-infrared spectroscopy. The accuracies of the prediction models constructed using different calibration methods (PLSR, SVM, and XGBoost) were compared. Spectral transformations based on FOD and CWT were useful for predicting the soil properties. Overall, the models based on CWT spectral transformation coupled with XGBoost calibration outperformed other models in predicting the OC content, TN content, and pH value. However, the optimal model for clay content estimation was CWT spectra coupled with SVM. Soil spectra used in this study were measured in the laboratory, resulting in less disturbance from the external environment compared with field spectra. In future research, we will explore the potential of field and satellite spectra in predicting large-area soil properties.

**Funding** This work was supported by the National Key R&D Program of China (Grant No. 2020YFD1100603-02).

## Declarations

**Conflict of interest** The authors declare no competing interests.

## References

- 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* 61:1–15. [https://doi.org/10.1016/S0034-4257\(96\)00120-4](https://doi.org/10.1016/S0034-4257(96)00120-4)
- Bishop JL, Pieters CM, Edwards JO (1994) Infrared spectroscopic analyses on the nature of water in montmorillonite. *Clay Clay Miner* 42:702–716. <https://doi.org/10.1346/ccmn.1994.0420606>
- Bronick CJ, Lal R (2005) Soil structure and management: a review. *Geoderma* 124:3–22. <https://doi.org/10.1016/j.geoderma.2004.03.005>
- Chen CT, Landgrebe DA, Szilagyi A, Henderson TL, Baumgardner MF (1989) Spectral band selection for classification of soil organic matter content. *Soil Sci Soc Am J* 53:1778–1784. <https://doi.org/10.2136/sssaj1989.03615995005300060028x>
- Chen T, Guestrin C (Ed.) (2016) Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM
- Cheng T, Rivard B, Sánchez-Azofeifa A (2011) Spectroscopic determination of leaf water content using continuous wavelet analysis. *Remote Sens Environ* 115:659–670. <https://doi.org/10.1016/j.rse.2010.11.001>
- Chen Y, Li YQ, Wang XY, Wang JL, Gong XW, Niu YY, Liu J (2020) Estimating soil organic carbon density in Northern China's agro-pastoral ecotone using vis-NIR spectroscopy. *J Soils Sediments* 20:3698–3711. <https://doi.org/10.1007/s11368-021-02977-0>
- Clark RN, Roush TL (1984) Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. *J Geophys Res-Sol Ea* 89:6329–6340. <https://doi.org/10.1029/JB089iB07p06329>
- Conforti M, Castrignanò A, Robustelli G, Scarciglia F, Stelluti M, Buttafuoco G (2015) Laboratory-based Vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. *CATENA* 124:60–67. <https://doi.org/10.1016/j.catena.2014.09.004>
- Dalal RC, Henry RJ (1986) Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. *Soil Sci Soc Am J* 50:120–123. <https://doi.org/10.2136/sssaj1986.03615995005000010023x>
- Dong ZY, Wang N, Liu JB, Xie JC, Han JC (2021) Combination of machine learning and VIRS for predicting soil organic matter. *J Soils Sediments* 21:2578–2588. <https://doi.org/10.1007/s11368-021-02977-0>
- Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. *Ann Stat* 28:337–374. <https://doi.org/10.1214/aos/1016218223>
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29:1189–1232. <https://doi.org/10.2307/2699986>
- Greenberg I, Linsler D, Vohland M, Ludwig B (2020) Robustness of visible near-infrared and mid-infrared spectroscopic models to changes in the quantity and quality of crop residues in soil. *Soil Sci Soc Am J* 84:963–977. <https://doi.org/10.1002/saj2.20067>
- Gu BJ, Chen DL, Yang Y, Vitousek P, Zhu YG (2021) Soil-food-environment-health nexus for sustainable development. *Research*. 2021:9804807. <https://doi.org/10.34133/2021/9804807>
- Guo JH, Liu XJ, Zhang Y, Shen JL, Han WX, Zhang WF, Christie P, Goulding KWT, Vitousek PM, Zhang FS (2010) Significant acidification in major Chinese croplands. *Science* 327:1008–1010. <https://doi.org/10.1126/science.1182570>
- Hong YS, Chen YY, Yu L, Liu YF, Liu YL, Zhang Y, Liu Y, Cheng H (2018) Combining fractional order derivative and spectral variable selection for organic matter estimation of homogeneous soil samples by VIS-NIR spectroscopy. *Remote Sens-Basel* 10:479. <https://doi.org/10.3965/10.3390/rs10030479>
- Ji WJ, Adamchuk VI, Chen SC, Mat Su AS, Ismail A, Gan QI, Shi Z, Biswas A (2019) Simultaneous measurement of multiple soil properties through proximal sensor data fusion: a case study. *Geoderma* 341:111–128. <https://doi.org/10.1016/j.geoderma.2019.01.006>
- Jiang QH, Li QX, Wang XG, Wu Y, Yang XL, Liu F (2017) Estimation of soil organic carbon and total nitrogen in different soil layers using VNIR spectroscopy: effects of spiking on model applicability. *Geoderma* 293:54–63. <https://doi.org/10.1016/j.geoderma.2017.01.030>
- Kennard RW, Stone LA (1969) Computer aided design of experiments. *Technometrics* 11:137–148. <https://doi.org/10.1080/00401706.1969.10490666>
- Kilmer VJ, Alexander LT (1949) Methods of making mechanical analysis of soils. *Soil Sci* 68:15–24. <https://doi.org/10.1097/00010694-194907000-00003>
- Kovačević M, Bajat B, Gajić B (2010) Soil type classification and estimation of soil properties using support vector machines. *Geoderma* 154:340–347. <https://doi.org/10.1016/j.geoderma.2009.11.005>
- Kuang BY, Tekin Y, Mouazen AM (2015) Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil Till Res* 146:243–252. <https://doi.org/10.1016/j.still.2014.11.002>
- Levene H (1960) Robust tests for equality of variances. Stanford University Press
- Li MYH, Zhou MF (2020) The role of clay minerals in formation of the regolith-hosted heavy rare earth element deposits. *Am Miner* 105:92–108. <https://doi.org/10.2138/am-2020-7061>
- Ludwig B, Vormstein S, Niebuhr J, Heinze S, Marschner B, Vohland M (2017) Estimation accuracies of near infrared spectroscopy for

- general soil properties and enzyme activities for two forest sites along three transects. *Geoderma* 288:37–46. <https://doi.org/10.1016/j.geoderma.2016.10.022>
- Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE T Pattern Anal* 11:674–693. <https://doi.org/10.1109/34.192463>
- Nawar S, Mouazen AM (2017) Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *CATENA* 151:118–129. <https://doi.org/10.1016/j.catena.2016.12.014>
- Nawar S, Buddenbaum H, Hill J, Kozak J, Mouazen AM (2016) Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil Till Res* 155:510–522. <https://doi.org/10.1016/j.still.2015.07.021>
- Palacios-Orueta A, Ustin SL (1998) Remote sensing of soil properties in the Santa Monica Mountains I. spectral analysis. *Remote Sens Environ* 65:170–183. [https://doi.org/10.1016/S0034-4257\(98\)00024-8](https://doi.org/10.1016/S0034-4257(98)00024-8)
- Paustian K, Lehmann J, Ogle S, Reay D, Robertson GP, Smith P, Smith P (2016) Climate-smart soils. *Nature* 532:49–57. <https://doi.org/10.1038/nature17174>
- Peng Y, Knadel M, Gislum R, Schelde K, Thomsen A, Greve MH (2014) Quantification of SOC and clay content using visible near-infrared reflectance-mid-infrared reflectance spectroscopy with Jack-Knifing partial least squares regression. *Soil Sci* 179:325–332. <https://doi.org/10.1097/SS.0000000000000074>
- Saeyns W, Mouazen AM, Ramon H (2005) Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. *Biosys Eng* 91:393–402. <https://doi.org/10.1016/j.biosystemseng.2005.05.001>
- Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* 36:1627–1639. <https://doi.org/10.1021/ac60214a047>
- Shi TZ, Cui LJ, Wang JJ, Fei T, Chen YY, Wu GF (2013) Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. *Plant Soil* 366:363–375. <https://doi.org/10.1007/s11104-012-1436-8>
- Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14:199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Song X, Chen C, Arthur E, Tuller M, Zhou H, Ren TS (2021) Effects of increasing water activity on the relationship between water vapor sorption and clay content. *Soil Sci Soc Am J* 85:520–525. <https://doi.org/10.1002/saj2.20236>
- Stenberg B, Viscarra Rossel RA, Mouazen AM, Wetterlind J (2010) Visible and near infrared spectroscopy in soil science. *Adv Agron* 107:163–215. [https://doi.org/10.1016/S0065-2113\(10\)07005-7](https://doi.org/10.1016/S0065-2113(10)07005-7)
- Stoner ER, Baumgardner MF (1981) Characteristic variations in reflectance of surface soils. *Soil Sci Soc Am J* 45:1161–1165. <https://doi.org/10.2136/sssaj1981.03615995004500060031x>
- Torrence C, Compo GP (1998) A practical guide to wavelet analysis. *B Am Meteorol Soc* 79:61–78. [https://doi.org/10.1175/1520-0477\(1998\)079%3c0061:APGTWA%3e2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079%3c0061:APGTWA%3e2.0.CO;2)
- Tsakiridis NL, Keramaris KD, Theocharis JB, Zalidis GC (2020) Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* 367:114208. <https://doi.org/10.1016/j.geoderma.2020.114208>
- Tziolas N, Tsakiridis N, Ogen Y, Kalopesa E, Ben-Dor E, Theocharis J, Zalidis G (2020) An integrated methodology using open soil spectral libraries and Earth Observation data for soil organic carbon estimations in support of soil-related SDGs. *Remote Sens Environ* 244:111793. <https://doi.org/10.1016/j.rse.2020.111793>
- Vašát R, Kodešová R, Borůvka L, Klement A, Jakšík O, Gholizadeh A (2014) Consideration of peak parameters derived from continuum-removed spectra to predict extractable nutrients in soils with visible and near-infrared diffuse reflectance spectroscopy (VNIR-DRS). *Geoderma* 232–234:208–218. <https://doi.org/10.1016/j.geoderma.2014.05.012>
- Viscarra Rossel RA, Behrens T (2010) Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158:46–54. <https://doi.org/10.1016/j.geoderma.2009.12.025>
- Vohland M, Besold J, Hill J, Fründ H (2011) Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma* 166:198–205. <https://doi.org/10.1016/j.geoderma.2011.08.001>
- Vohland M, Ludwig M, Harbich M, Emmerling C, Thiele-Bruhn S (2016) Using variable selection and wavelets to exploit the full potential of visible-near infrared spectra for predicting soil properties. *J Near Infrared Spec*. 24:255–269. <https://doi.org/10.1255/jnirs.1233>
- Wang SJ, Chen YH, Wang MG (2019) Performance comparison of machine learning algorithms for estimating the soil salinity of salt-affected soil using field spectral data. *Remote Sens-Basel* 11:2605. <https://doi.org/10.3390/rs11222605>
- Wang LX, Pang XY, Li N, Qi KB, Huang JS, Yin CY (2020) Effects of vegetation type, fine and coarse roots on soil microbial communities and enzyme activities in eastern Tibetan plateau. *CATENA* 194:104694. <https://doi.org/10.1016/j.catena.2020.104694>
- Wei LF, Yuan ZR, Yu M, Huang C, Cao LQ (2019) Estimation of arsenic content in soil based on laboratory and field reflectance spectroscopy. *Sensors* 19:3904. <https://doi.org/10.3390/s19183904>
- Whiting ML, Li L, Ustin SL (2004) Predicting water content using Gaussian model on soil spectra. *Remote Sens Environ* 89:535–552. <https://doi.org/10.1016/j.rse.2003.11.009>
- Wilding LP (1985) Spatial variability: Its documentation, accommodation and implication to soil surveys. In: Nielsen DR, Bouma J (eds) *Soil Spatial Variability*. Pudoc, Wageningen, pp 166–187
- Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemometr Intell Lab* 58:109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
- Yang MH, Xu DY, Chen SC, Li HY, Shi Z (2019) Evaluation of machine learning approaches to predict soil organic matter and pH using Vis-NIR spectra. *Sensors* 19:263. <https://doi.org/10.3390/s19020263>
- Yang XJ, Lin A, Li X, Wu Y, Zhou W, Chen Z (2013) China's ion-adsorption rare earth resources, mining consequences and preservation. *Environ Dev* 8:131–136. <https://doi.org/10.1016/j.envdev.2013.03.006>
- Zhang Y, Li MZ, Zheng LH, Qin QM, Lee WS (2019) Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. *Geoderma* 333:23–34. <https://doi.org/10.1016/j.geoderma.2018.07.004>
- Zhong L, Guo X, Xu Z, Ding M (2021) Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. *Geoderma* 402:115366. <https://doi.org/10.1016/j.geoderma.2021.115366>