An Improved Vis-NIR Estimation Model of Soil Organic Matter Through the Artificial Samples Enhanced Calibration Set

Xibo Xu, Yunhao Chen, Xiujuan Dai, Tianjie Lei, Sijia Wang, and Kangning Li

***Abstract*—A suitable calibration sample set is extremely important to acquire an accurate spectral-based model for estimating soil organic matter (SOM). However, an unrepresentative calibration sample set was frequently collected due to the inappropriate samplings pattern caused by problematic transportation logistics and complex geographic conditions, which resulted in fairly poor generalization and low accuracy of the spectroscopic model. Thus, we hypothesized that a soil sample dataset equivalent to natural soil samples could be prepared under controlled laboratory conditions, and increase the accuracy of spectroscopic estimation of SOM content by use of a coverage assessment method that added laboratory-simulated near-natural samples to the natural samples set in order to enhance the representative sample size and variability of the calibration set. The results showed that the near-natural samples enhanced (NSE) calibration set contained 42 natural soil samples and 28 near-natural soil samples. This set exhibited sufficient coverage and better information integrity within estimators space than the initial calibration set that included 43 natural soil samples. Random forest model based on the NSE calibration set ($R^2 = 0.90$; RPIQ = 4.17) more accurately estimated SOM content than the spectral-based model built with the initial calibration set ($R^2 = 0.73$; RPIQ = 2.32); the SOM chemical compositions (e.g., lipids, polysaccharides, and lignin) and their relative abundance from the laboratory-simulated near-natural soil samples were basically consistent with those of natural soil samples. The inclusion of near-natural soil samples in the calibration set improved the SOM spectral-based estimation model, and was observed to be a practical method. Our results provided a calibration set enhancement strategy that effectively supports spectroscopic estimation model of SOM contents in the case of pattern-biased field samplings at the local scale.**

***Index Terms*—Artificial samples enhanced calibration set, estimation model, proximally sensed Vis-NIR spectroscopy, sampling pattern bias, soil organic matter (SOM).**

Manuscript received 8 December 2022; revised 15 March 2023; accepted 9 May 2023. Date of publication 12 May 2023; date of current version 24 May 2023. This work was supported by the Ningxia Agriculture and the Animal Husbandry Department East–West Cooperation Project. *(Corresponding authors: Yunhao Chen; Tianjie Lei.)*

Xibo Xu, Yunhao Chen, and Sijia Wang are with the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China (e-mail: xu_xibo@126.com; cyh@bnu.edu.cn; sijia.wang@mail.bnu.edu.cn).

Xiujuan Dai and Kangning Li are with the Beijing Key Laboratory of Environmental Remote Sensing and Digital City, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China (e-mail: 2394992508@qq.com; 201631190001@mail.bnu.edu.cn).

Tianjie Lei is with the Institute of Environment and Sustainable Development in Agriculture, Chinese Academy of Agricultural Sciences, Beijing 100081, China (e-mail: leitianjie@caas.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2023.3275745, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2023.3275745

I. Introduction

SOIL organic matter (SOM) is a soil fertility indicator and an indispensable part of the earth's carbon pool [1], [2], [3], [4], [5]. For quantitative analysis of SOM contents in the soil, spectroscopy techniques have attracted the attention of many scholars in recent decades. Compared with conventional chemical analysis in the laboratory, quantitative spectroscopic techniques can be used to efficiently and inexpensively determine SOM contents without sample destruction [6], [7].

The quantitative spectroscopic analysis of SOM contents between spectral reflectance and the actual SOM contents in soil samples can be acquired from a multivariate statistical model, such as random forest (RF), partial least squares regression, and support vector machine [8], [9], [10]. A suitable calibration set was one of the vital parts of spectral-based model construction [11], [12]. Theoretically, an effective multivariate statistical model could be calibrated by a representative sample set that covers all possible sources of the variability of the target area soils faced during the estimation [13], [14], [15]. Furthermore, several approaches (e.g., Kennard-Stone, D-optimal procedure, and auxiliary information method) were developed to select the representative samples from the field samplings dataset to build a calibration set, and in order for improving the robustness and generalization of spectral-based model [11], [16]. However, the calibration samples selection strategy is ineffective when an insufficient number of representative samples were collected due to the inappropriate sampling patterns. Obviously, field sample collection is affected by historical data, expert-based experience, and geographic conditions [17], [18]. An insufficient number of representative samples and an inappropriate spatial pattern of sampling may result when the local environment is harsh and transportation is not accessible. As a result, the constructed spectroscopic models are especially similar, resulting in local specificity at the loss of generalization capacity that may damage the spectral-based estimation performance [19], [20]. Therefore, building an effective spectral-based estimation model under the condition of field sampling bias deserves to be explored.

For this purpose, we hypothesized that near-natural soil samples with different known SOM contents could be prepared under a controlled experimental environment, and such samples would be approximately equal to natural soil samples within the research area. These samples would then be integrated with natural soil samples to enhance the representative samples size and

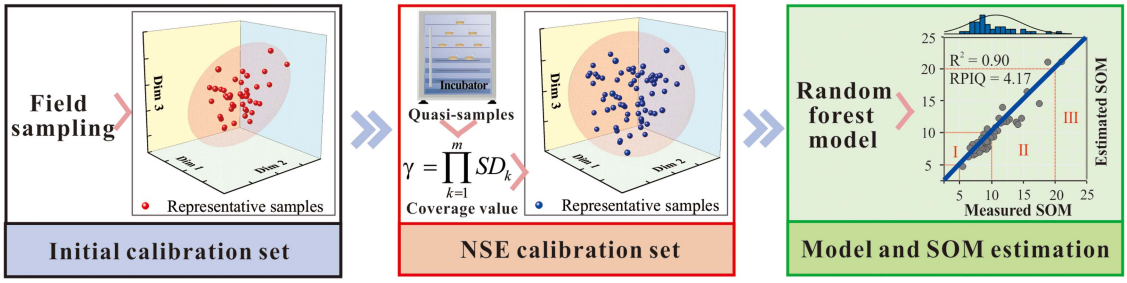

Fig. 1. Flowchart for the NSE calibration set and the construction of a spectral-based estimation model of the SOM contents in the Yinbei area, China.

variability of the calibration set by use of a coverage assessment method [i.e., near-natural samples enhanced (NSE) calibration set] in order to improve the robustness of the spectral-based estimation model at the local scale. In addition, machine learning has been proven to be an excellent method for modeling soil properties through spectra data [21]. Machine learning is a data-driven programming model, and as such, the quality of the calibration sample set is decisive in acquiring an accurate spectral-based model for SOM estimation. An important learning prerequisite is that the calibration set must contain samples from a distribution that is equal to that which the machine learning model is expected to predict [22]. Nevertheless, this assumption is commonly violated due to bias in the field sampling pattern. Hence, optimizing the sample data distribution and information integrity of the calibration set through near-natural soil samples would help improve a machine learning-based estimation model of SOM content.

The near-natural samples method is regarded as an easy and effective experiment for supporting quantitative spectroscopic analysis of soil attributes [18], [23]. Farifteh et al. [24], Linsinger et al. [25], Zou et al. [20], and Wang et al. [26] created near-natural samples to characterize the spectral signals of the key attributes in soils, and developed a spectroscopic technique to recognize the content status of soil chemical components. The near-natural soil samples with various known SOM contents created under controlled laboratory conditions exhibit the standard behaviors of SOM absorption features that are in agreement with the natural soils, and therefore can be help to compensate for sampling pattern shortages under field conditions. Additionally, RF (a popular machine learning paradigm) has been widely used in the quantitative spectroscopic analysis, which was characterized with the low-computational expenses, easy-to-implementation, and outstanding performance in spectral-based estimation of SOM contents [27]. RF combined with the NSE calibration set has shown great potential for improving quantitative analysis of SOM under field conditions.

The objectives of this study were to

1) collect the natural soil samples for building the initial calibration set;
2) prepare the near-natural soil samples under controlled laboratory conditions to enhance the initial calibration set by using the coverage assessment method; and
3) integrate the NSE calibration set with the RF algorithm to construct a spectral-based estimation model of SOM contents under field conditions.

II. Materials and Methods

A. Key Workflow

The workflows for the analysis processes are summarized in Fig. 1. Initially, natural soil samples and those spectra data were collected to build an initial calibration set. Then, the 120 near-natural soil samples were produced under controlled laboratory conditions and taken to enhance the initial calibration set based on the coverage assessment method. Finally, the NSE calibration set was employed to construct the RF model for SOM estimation under field conditions.

B. Study Area and Sample Collection

The study area was located in the Yinbei area of western China [see Fig. 2(a)], having a population of 1.22 million and covering a 1000 km^2 area. The Yinbei area is characterized by a temperate continental climate, and the average temperature and average annual precipitation are 9.69 °C and 187 mm, respectively. Anthropogenic-alluvial soil is the dominant soil type and widely covers the Yinbei area [28]. Additionally, cultivated land accounts for more than 60% of the total area, and the western part of the area is in the Helan Mountains. Due to low precipitation and low temperature in the dry season, agricultural activities are conducted during the wet season (from late April to early October) every year, and the bare soil was exposed to remote sensors during the dry season [see Fig. 2(c)].

With consideration for soil type, geological condition, and land use status, sites for 43 natural soil samples and two background soil sample were predetermined using ArcGIS 10.2 software. The 45 soil samples (0–20 cm) were collected in April 2018. Five soil subsamples were taken at each sample location within a square with a side length of 10 m, mixed into a representative sample (1 kg) in a sealed bag, and sent to the laboratory for chemical analysis. The actual locations of all samples were documented using on a global position system [see Fig. 2(b)]. Additionally, 10 kg of background soil was collected for subsequent production of the near-natural soil samples, and the above-mentioned procedures were reimplemented.

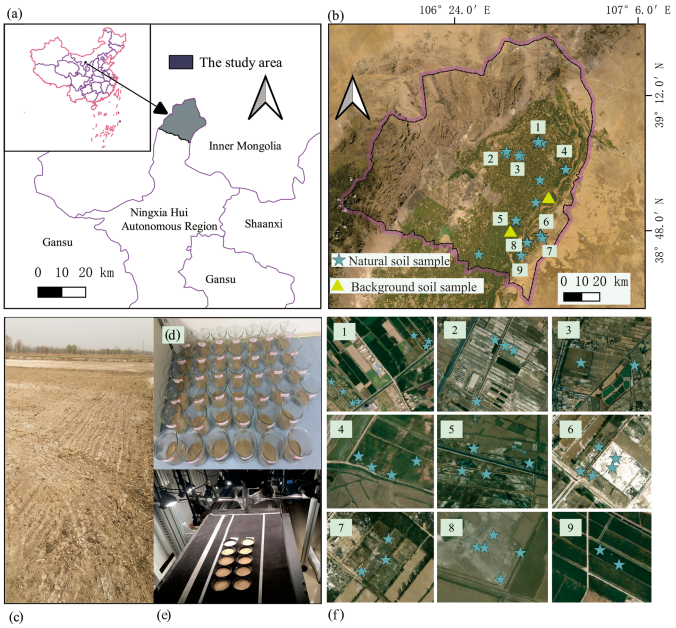

Fig. 2. Sampling sites and laboratory treatments for soil samples obtained in Yinbei, China. (a) Geographic location of the Yinbei area in China. (b) Actual locations of natural soil samples and the background soil sample. (c) Soil surface in the Yinbei area. (d) Created near-natural soil samples. (e) Spectral soil sample measurement process in the laboratory. (f) Detailed images of soil sample locations shown in (b).

C. Proximal Vis-NIR Sensing Measurement and Chemical Analysis

Both field spectra of natural soil samples and laboratory spectra of the near-natural soil samples were measured with a Spectra Vista Corporation (SVC) field-portable spectroradiometer (HR-1024i). The SVC HR-1024i covered a wavelength range from 350 to 2500 nm, with 1024 spectral bands and a spectrometer sampling interval of 1 nm. Field Vis-NIR spectra were measured under cloudless (clear sky) weather conditions. At each sampling site, the stones and roots on the soil surface were simply removed (i.e., soil surface smoothed) to eliminate any shading effects on soil reflectance. A white reference panel was initially used to calibrate the spectrometer, followed by fiberoptic cable scanning of the soils and recording of the spectral reflectance five times at each sampling site. In order to minimize any changes in radiance values, the white reference panel was used for recalibrating when the sampling locations were different [16].

In the laboratory, the soil samples were air-dried (25 °C) and sieved through a 1 mm mesh screen. Then, six 50-W halogen lamps were applied for the light source, with an incident angle of 25°, and the soil was placed in a container with a diameter of 6 cm and a depth of approximately 2 cm [see Fig. 2(e)]. Finally, the SVC HR-1024i spectrometer was used to scan the soil samples five times, and the spectral data were recorded in a computer.

Any abnormal spectral curves from either laboratory-obtained spectra or field-obtained spectra were deleted. The measured spectral reflectance of each sample was the mean value of the remaining spectral data. The Savitzky–Golay method is a common and easy technique used to remove random noise from spectral reflectance data, and has been successfully applied to smooth both laboratory-obtained spectral reflectance data and field-obtained spectral reflectance data of natural soil samples [29]. A noticeable noise-induced spectral range (350–399 nm) having to do with the influence of the device and environmental conditions was removed, and the remaining 946 spectral bands were included in the subsequent calculation [10], [30]. After the pretreatments of the natural soil samples described above were completed, the SOM contents were determined using the potassium dichromate oxidation-external heating method in the laboratory.

D. Building an NSE Calibration Set

1) Laboratory-Preparation of the Near-Natural Soil Samples and Measurement of Those Spectra: Producing the near-natural soil samples and measuring those Vis-NIR spectra under controlled laboratory conditions were the crucial steps for calibration set enhancement. The background soil used to prepare the near-natural samples was initially collected from the same soil types and locations as the natural soil samples in the research area to ensure consistency of the soil background minerals. During the preparation process (Fig. S1), the collected background soils were air-dried (25 °C) and sieved through a 1 mm mesh screen, and the SOM was removed from the background soils using H_2O_2 oxidation technology. Then, the soil samples with SOM removed were divided into 50 g soil samples for a total of 121 soil samples and stored in glass beakers [see Fig. 2(d)]. A single sample was analyzed for actual SOM content, and the remaining 120 background soil samples were the basic materials for building the near-natural soil samples. Based on the actual SOM contents of the background soil and the technical specifications of soil-forming factors [31], [32], a production strategy for the near-natural soil samples with different SOM contents was created.

Specifically, different weight of standard organic fertilizers (NY884-2012: SOM content, 45%; average) (AMPRC, 2012) [33] was added to the background soil (with the SOM removed), the temperature (10 °C), humidity (10%), precipitation (20 ml), and ploughing work (that were consistent with historical data in the study area) simulated in a constant temperature humidity chamber, and finally, 120 near-natural soil samples with an expected content gradient of 0.5 g·kg^{-1} of different SOM content would be prepared

$$L = M \times \frac{\rho - \mu}{\mu - 0.216} \tag{1}$$

where L is the weight (g) of the added organic fertilizer (note that the organic fertilizers added in the study were consistent with those applied in agricultural production activities in the study area); M and μ represent the initial background soil weight (g) and target SOM contents of near-natural soil samples (g·kg^{-1}), respectively; and ρ is the measured SOM content in the background soil, with the value set to 3 g·kg^{-1} in this study.

In the near-natural soil sample preparation process, soil respiration and soil weight measuring errors may cause a deviation between expected SOM contents and actual SOM contents;

thereby, the actual SOM contents of the 120 near-natural soil samples were determined based on the potassium dichromate oxidation-external heating method. Additionally, a pyrolysis-gas chromatograph/mass spectrometer (7890B-5977A, Agilent Corporation, USA) was used to analyze SOM chemical composition and relative abundance in both the natural soil samples and the near-natural soil samples to ensure the consistency of both samples [34], [35].

Correspondingly, spectral reflectance values (i.e., near-natural soil sample spectra) were measured in the laboratory based on the procedures described in Section II-C. With the aim of identifying the SOM response bands for the predictor variables in the subsequent calculation, the above-mentioned SOM removal technology was reimplemented to remove the SOM from the 120 near-natural soil samples, and the final state of the reflectance spectroscopy values for near-natural soil samples was recorded.

2) Application of the External Parameter Orthogonalization for Spectra Pretreatment: The spectra pretreatment aimed to narrow the spectral differences between field-obtained spectra of natural soil samples and near-natural soil sample spectra, and thereby improve the application ability of near-natural soil sample spectra in calibration set optimization. A noticeable spectral difference (i.e., arising from soil moisture, particle size, and random effects) can be observed between field-obtained spectra of natural soil samples and near-natural soil sample spectra [14]. As a result, computational complexity increases, and accuracy is impaired in the calibration set enhancement.

An external parameter orthogonalization approach was developed by Minasny et al. [36], and successfully used to decompose moisture-influenced signals from the field spectra data by comparing the differences between field-obtained spectra and laboratory-obtained spectra of natural soil samples. Thus, in this study, an external parameter orthogonalization was employed to analyze the spectral differences between field-obtained spectra of natural soil samples and near-natural soil sample spectra for supporting the calibration set enhancement. The field-obtained spectra data was initially transformed into a weight coefficient matrix. Then, the weight correction matrix was extracted by analyzing the differences between spectral transition matrix and near-natural soil sample spectra, and a matrix inverse was implemented to generate the pretreated spectra [36], [37], [38].

3) Using the Coverage Assessment Method for Enhancing the Initial Calibration Set: Adding the various representative samples from the near-natural soil sample spectral data to the natural soil samples dataset (i.e., initial calibration set) enables the identification of all possible sources of variability of the target site soils included in the calibration set, which can promote the generalization and accuracy of the recalibration model [39].

The coverage (*COV*) value is an index that can assess the coverage distribution and information integrity of the sample dataset within estimator space [40], [41], [42]. Normally, a sample set with a high coverage value indicates uniform distribution and excellent information integrity in the estimator space, which can promote the generalization and accuracy of the recalibration model. The coverage assessment method was implemented for the calibration set construction and enhancement in this stage. In step 1, the *COV* value was calculated [(2) and (3)] for each sample from the natural soil samples dataset (initial calibration set, $n = 43$), and the samples were removed from the initial calibration set as a sample with the replicated coverage value. In step 2, m near-natural soil samples were randomly selected and added to the initial calibration set to update the coverage value. Note that each near-natural sample should be selected at least once. Changes in coverage values with different sample sets were assessed. Finally, the near-natural samples with positive changes in coverage values were highlighted and inserted into the initial calibration set to acquire an NSE calibration set

$$COV = \prod_{k=1}^{m} \sqrt{\frac{\sum_{j=1}^{t} (M_k - h_{kj})^2}{t}} \tag{2}$$

$$M_k = \frac{\sum_{j=1}^{t} h_{kj}}{t} \tag{3}$$

where m and t are the number of spectral bands and samples, respectively; h_{kj} represents the spectral reflectance value of the kth band of the jth sample ($j = 1, 2, 3, \ldots, t$); M_k is the mean value of the spectral reflectance value of the kth band.

E. Predictor Variable Selection and RF Model

Changes in the spectral reflectance values in the preparation process of near-natural soil samples (i.e., background soil samples with SOM removed to near-natural soil samples) were used to identify the SOM response bands for the predictor variables in spectroscopic models. The change intensity value (t) for the spectral reflectance between SOM removal soils and near-natural soil samples accounted for the SOM content changes (Fig. S1). A high t value suggests bands in which the spectral response for SOM is significant. The t calculation was

$$t = \frac{1}{2n} \sqrt{\sum_{i=1}^{i=n} \left(\frac{B_{\mathrm{ki}} - b_{\mathrm{ki}}}{b_{\mathrm{ki}}} \right)^2 - \sum_{i=1}^{i=n} \left(\frac{C_{\mathrm{ki}} - B_{\mathrm{ki}}}{B_{\mathrm{ki}}} \right)^2} \tag{4}$$

where B_{ki} represents the reflectance value of the kth spectral band of the ith near-natural soil spectral sample [see Fig. 3(b)]. b_{ki} and C_{ki} are reflectance values of the kth bands of the ith spectral sample from soil samples with SOM removed (Fig. S1).

Furthermore, a spectral-based SOM estimation model was constructed using the RF algorithm [43]. RF is an ensemble method of multiple decision trees in which the unbiased estimation result was chosen from various tree structures by voting as the created features of each decision tree were assessed [44]. In this process, the number of trees and the predictor variables in each split of the trees were defined as 800 and 3, respectively; Model accuracy was evaluated by using a 10-fold cross-validation method.

Generally, the coefficient of determination (R^2) and ratio of performance to interquartile distance (RPIQ) are the key parameters for evaluating model accuracy and stability [26], [45], [46]. R^2 indicates the level to which the target variables are fully explained by the predictor variables. RPIQ is the ratio of the quartile ranges (the difference value between the third and first quartiles) to the root mean square error of the validation set.

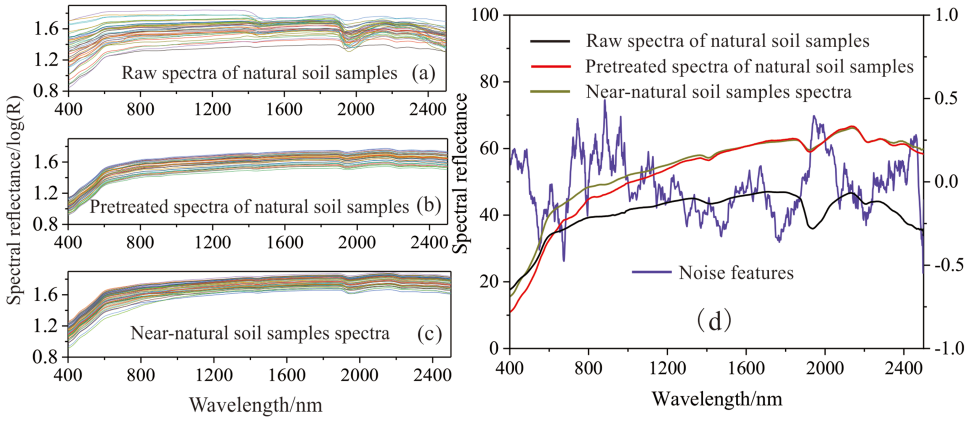

Fig. 3. Characteristics of the multisource spectra of samples from the Yinbei area, China. (a) is the raw spectra of natural soil samples; (b) is the pretreated spectra of natural soil samples; (c) is near-natural soil sample spectra; (d) indicates average spectra of the different soil samples.

TABLE I
STATISTICAL CHARACTERISTICS OF SOM CONTENTS IN BOTH NATURAL SOIL SAMPLES AND NEAR-NATURAL SOIL SAMPLES

Sample set	Range/ g·kg^{-1}	Mean/ g·kg^{-1}	Median /g·kg^{-1}	Interquartile range	S.D.[1]
Natural soil samples (n = 43)	6.09–21.00	10.17	9.26	7.93	3.35
Near-natural soil samples (n = 120)	5.20–99.20	36.91	32.20	22.03	19.06

[1]S.D. is the standard deviation.

A reliable estimation model is generally characterized by an RPIQ > 4.05, while RPIQ values between 3.37 and 4.05 indicate that the model provides good accuracy. An RPIQ value between 2.70 and 3.37 suggests that the model gives an approximate estimation accuracy. When RPIQ values ranged from 2.02 to 2.70, the model is considered to be fair in estimating SOM contents of the natural soil samples [47]. A good accuracy model is generally indicated by larger R^2 and RPIQ values.

All calculations in Section II were performed using Python 3.8.

III. RESULTS AND ANALYSIS

A. Statistical Summary of SOM Contents

The summary statistics for SOM contents in both the natural soil samples and laboratory-prepared near-natural soil samples from the Yinbei area of China are provided in Table I, and the created near-natural soil samples and their actual SOM contents are listed in Table S1. As shown in Table I, the mean SOM value for the 43 natural soil samples was 10.17 g·kg^{-1}, ranging from 6.09 to 21.00 g·kg^{-1}. According to the Chinese classification criteria of soil nutrient materials [48], the SOM content level was the lowest (10–20 g·kg^{-1} and 6–10 g·kg^{-1}). Furthermore, the SOM contents of the near-natural soil samples ranged from 5.20 to 99.20 g·kg^{-1}, and were widely distributed in all content intervals. The mean SOM content was 36.91 g·kg^{-1}. The SOM values for the 43 natural soil samples suggested low distribution variation with a standard deviation of 3.35, while the standard deviation value in the sample set of near-natural soil was 19.06. The clear variations in SOM values in the sample set had a positive effect on model convergence and the ability to decrease errors in the estimated values in the model calibration [49]. With regard to the actual SOM contents of near-natural soil samples (Table S1), wide SOM content ranges and large sampling sizes contributed to the high median and average SOM values of the near-natural soil samples set compared with the natural soil samples set. All possible sources of the variability of the target area soils can be provided with the acquired near-natural samples set, thus effectively enhancing the initial calibration set. Importantly, SOM's absorption feature identification and differences analysis of spectral curves can also benefit from such near-natural soil samples sets.

B. Characteristics of Multisource Proximally Sensed Spectra

The characteristics of the spectral curves, including the raw spectra and the pretreated spectra of the 43 natural soil samples and 120 near-natural soil sample spectra, are displayed in Fig. 3. Logarithmic transformation of spectral reflectance is an effective method for highlighting the absorption features of spectra. This transformation also provides support for analyzing the differences between different soil samples among the multiple spectra sources [20]. Specifically, all logarithmic spectral curves showed smooth curves. The pretreated spectra were observed to have the same trend as the spectra of the near-natural samples, indicating their spectral differences have been minimized. The raw spectral curves were pretreated by the external parameter orthogonalization, and suggested regularity in contrast to the dispersion and irregularity of the raw spectral data. The spectral curves of the near-natural soil samples also illustrated regular and uniform changes. Furthermore, the spectral reflectance curves indicated rapidly increasing trends in the range of 400–600 nm. Significant water absorption features at approximately 1900 nm are clearly shown in the spectral curves.

Clear differences in the spectral absorption valleys and depths resulted from the various chemical components in the soils, and provide the basis for SOM estimation under given environmental

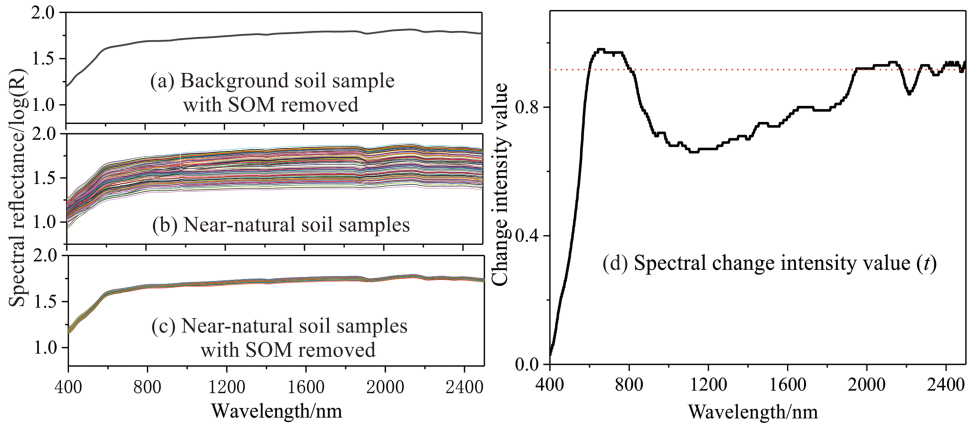

Fig. 4. Results for the spectral change intensity value and the spectral characteristics of (a) the collected background soils from which SOM has been removed. (b) Created near-natural soil samples with different known SOM contents. (c) Near-natural soil samples from which SOM has been removed. (d) Spectral change intensity value (t).

conditions even though the negative effects on the spectral curves due to the environment and uncertain factors were also significant. Therefore, analysis of the SOM spectral features and noise sources should be conducted. The raw spectral curve, pretreated spectral curve, and near-natural soil spectral curve are shown in Fig. 3(d). The spectral reflectance values of the raw spectra ranging from 400 to 2500 nm increased as the spectra pretreatment was conducted, indicating that the moisture effects from the spectral data were minimized. Additionally, the performance differences in the spectral pretreatments used for removal of spectral differences in various wavelength ranges were significant, such as in the 400–1400 nm and 1400–2500 nm ranges. Thus, source identification of the external noise may be useful for explaining this phenomenon. The noise features (i.e., spectral differences features) and their change curves are also illustrated in Fig. 3(d). Spectral ranges of 500–650 nm, 700–950 nm, 1700–1850 nm, and 1900–2000 nm showed obvious noise features. Specifically, spectral noise in the range of 700–950 nm may be generated by vegetation residue, and the vibration of –OH in the soil moisture was indicated by spectral absorption at approximately 1900 nm [50]. Spectral noise in the wavelength range of 500–650 nm may be derived from mixed environmental factors, such as organic matter and iron [51]. Furthermore, the materials responsible for generating the spectral noise in the wavelength range of 1700–1850 nm cannot be accurately identified, and artificial operations (e.g., drying, grinding, and sieving) may be the cause of the noise generation. Additionally, few noise features occurred close to the SOM-related material bands, increasing the difficulty of noise source identification. Note that the spectral reflectance features of soil salt could affect the reflectance values of SOM, as the spectral reflectance values in the visible wavelength range improved by 10%–50% [52]. Wang et al. [53] and Xu et al. [28] have indicated that the study area suffers from salinization, and that the soil is characterized by high salt content. However, salt information was not considered in the design of the external parameter orthogonalization correction method, and this oversight may have contributed to the correction bias of the field-obtained spectra in the wavelength range of 400–1400 nm. Overall, the raw spectral data pretreated by the external parameter orthogonalization method were close to the near-natural soil spectra data under laboratory conditions, suggesting that the spectral differences were minimized.

C. Identified Spectral Bands for Predictor Variables

The calculation results for the spectral change intensity value (t) are shown in Fig. 4, in which the change intensity in soil spectral reflectance between soils with SOM removed and the near-natural soil samples with various SOM contents is illustrated. Based on prior experience with the SOM response bands and environment-induced spectral noise described in the literature [19], [54], the spectral bands were selected as the input variables for building a spectroscopic quantitative model when the t values were greater than 0.91. Subsequently, a total of 265 spectral bands used as input variables were determined, consisting of five wavelength ranges: 600–800 nm, 2040–2180 nm, 2270–2320 nm, and 2390–2465 nm.

Using near-natural soil sample spectral data can support the identification of the SOM response bands for predictor variables. The spectral wavelengths in the range of 600–800 nm were associated with the chromophores and the humic acid in the organic matter. The vibrations of N–H and C–O structure shows spectral features at approximately 2100 nm, and is considered to have a close relationship with SOM. Our results agree with findings previously reported in the literature [9], [55]. Additionally, the SOM-related bands for C–H bonds were also indicated by the spectral response at approximately 2300 nm [56], [57].

D. Near-Natural Soil Samples Enhanced (NSE) Calibration Set

The NSE calibration set enabled the sample set to cover the spectral signals (i.e., representative samples) not captured in the initial calibration dataset; thus, the model's generalization capacity was increased as the bias was corrected [58], [59]. The first two principal component (PC) scores of spectra data for the initial calibration set (that included 43 natural soil samples) and

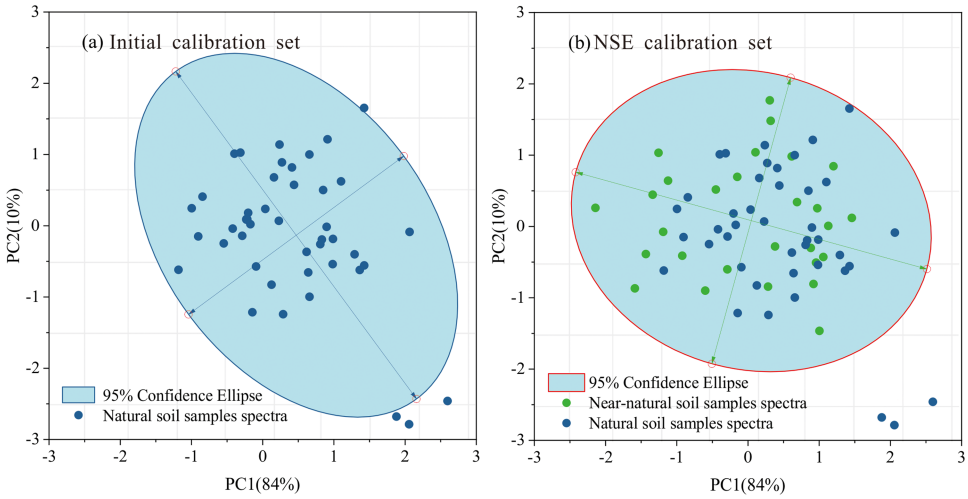

Fig. 5. PC scores within the projection space of the (a) initial calibration sample set and (b) NSE calibration set.

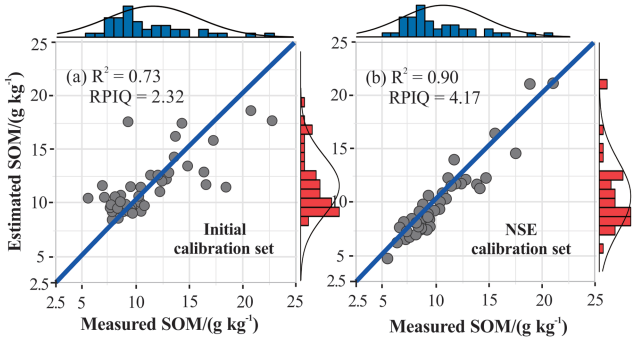

Fig. 6. Scatter plot of measured SOM values versus estimated SOM values obtained using (a) initial calibration set and (b) NSE calibration set; R^2 and RPIQ indicates coefficient of determination and RPIQ distance, respectively.

the NSE calibration set (that included 42 natural soil samples and 28 near-natural soil samples) are illustrated in Fig. 5. A total of 94% of the variation was accounted for by the first two PCs. This result indicated that the NSE calibration set was characterized by vast spectral diversity and sufficient coverage compared with the sparse distribution of projection scores within the PC space on the initial calibration set. Additionally, some PC spaces on the initial calibration set were deficient, such as projection space ranges of −2 to −1.1 of PC2, and this deficiency may damage the subsequent calibration analysis of the spectral-based estimation model. The NSE calibration set provided a sufficient estimator space with good coverage, abundant spectral signals, and information integrity, and effectively supported spectral-based model calibration.

E. Performance of Models Built With Initial Calibration Set and NSE Calibration Set

As shown in Fig. 6(a), the model built with the initial calibration set exhibited poor performance ($R^2 = 0.73$; RPIQ = 2.32) for SOM estimation. SOM estimates were significantly different from the measured SOM contents for the natural soil samples. In addition, the NSE calibration sample set was also employed to model estimates of SOM, as indicated by the R^2 and RPIQ values of 0.90 and 4.17, respectively [see Fig. 6(b)]. The RPIQ value increased from 2.32 to 4.17 when the enhanced calibration set was involved in the SOM spectral estimation, suggesting the model estimation capability was satisfactory.

IV. Discussion

A. Feasibility of Calibration Set Enhancement Via Near-Natural Soil Samples

The SOM chemical compositions and their relative abundance from the near-natural soil samples were basically consistent with those of natural soil samples. SOM is normally developed based on mineral background materials exposed to natural weathering and anthropologic activities [60], [61]. Thus, during the production process of near-natural soil samples, the soil background materials are initially collected from the same locations as the natural soil samples in the research area to eliminate inconsistencies in spectral information generated by the mineralogy components of the soil. Then, the soil-forming environment (temperature and precipitation) and human activities (fertilizer application and plowing work) are simulated under controlled laboratory conditions. Near-natural soil samples with different known SOM contents are created and can be representative of the case research area. Finally, the laboratory-measured spectra of near-natural soil samples (i.e., near-natural soil sample spectra) are collected.

Additionally, the SOM chemical compositions of soil samples were obtained by the pyrolysis-gas chromatograph/mass spectrometer [34], [35], indicating that lipids and polysaccharides were important SOM components in both natural soil samples and near-natural soil samples with relative abundances of 30% and 17% as well as 25% and 19%, respectively. The relative abundances of lignin, n-bearing, and nonlignin aromatics in SOM were approximately 10% (see Fig. 7).

Laboratory prepared near-natural soil samples with different known SOM contents were roughly equal to the natural soil

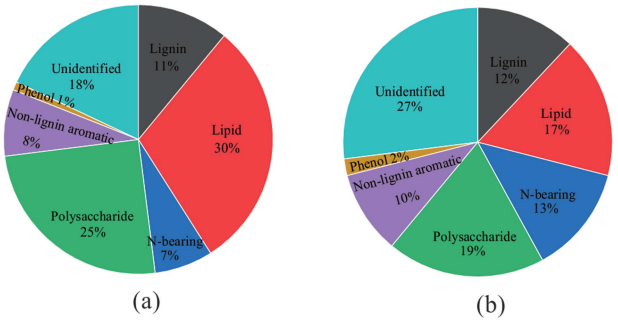

Fig. 7. Relative abundances of major compositions in SOM acquired from a pyrolysis-gas chromatograph/mass spectrometer. (a) Natural soil samples. (b) Near-natural soil samples.

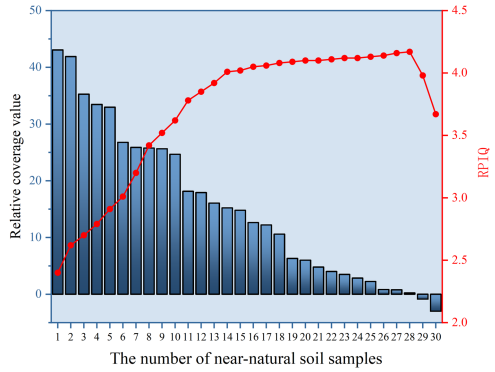

Fig. 8. NSE calibration set with various numbers of near-natural samples and the calibrated model performance.

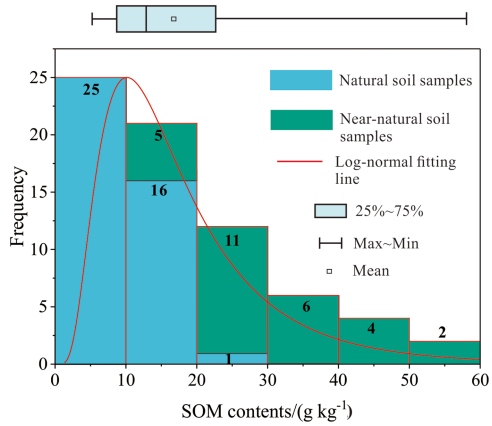

Fig. 9. Frequency distribution of SOM contents of the samples from the NSE calibration set in Yinbei, China.

samples, and can be used to compensate for the shortages of the insufficient number of representative soil samples in calibration set. Furthermore, plant secretions and secondary metabolites are also essential factors contributing to SOM generation, and vegetation information needs to be considered when further eliminating inconsistencies between natural and near-natural soil samples [62].

B. Applicability of Calibration Set Enhancement Strategy

Narrowing the spectral differences between natural and near-natural samples is an essential issue in implementing calibration set enhancement; the coverage assessment of the calibration sample set can be affected by spectral differences. Laboratory-prepared near-natural soil samples were initially reported by Farifteh et al. [24] and Zhou et al. [18]. These samples were primarily used to determine the spectral absorption features of the soil components because the near-natural soil samples were not affected by the heterogeneous constituents of the soils. Furthermore, Wang et al. [26] attempted to enlarge the field sample size with samples prepared under controlled laboratory conditions. Still, a model built with mixed samples did not show satisfactory performance and was even lower than the performance of the initial model, suggesting that the spectral difference between field-obtained spectra of natural soil samples and near-natural soil sample spectra may damage the generalization and robustness of the model. In addition, Zou et al. [20] suggested that spectral differences could be removed using a laboratory-field spectral transformation method (e.g., external parameter orthogonalization, direct standardization, and piecewise direct standardization method) in order to increase the accuracy of laboratory-field spectral-integrated estimation. Spectral difference removal is crucial for calibration set enhancement via laboratory-field spectra-data integration.

The size of the NSE calibration set and the calibrated model performances were influenced by the relative coverage value of the near-natural soil sample in coverage assessment. Initially, a sample did not affect the coverage value of the sample set and could be removed from the initial calibration set ($n = 43$). As shown in Fig. 8, near-natural soil samples (≤ 28) with positive relative coverage values could enhance the coverage and information integrity within the estimators space of the calibration set, thereby improving the model generalization. In contrast, the near-natural soil samples (sample 29 and sample 30) characterized with the negative relative coverage value impaired the performance of the spectral-based model based on the NSE calibration set. Thus the calibration set enhancement strategy may be ineffective when the near-natural soil samples posed no effect or negative effects on the coverage values of the natural soil sample set. In other words, an expert-based sampling design and a sufficient number of representative samples are also vital for improving the SOM spectral-based estimation at the local scale.

The NSE calibration set was employed to improve the spectral-based model by changing the size of the representative samples and those distribution patterns over the estimator space. The frequency distribution of SOM contents of soil samples from the optimized calibration set is shown in Fig. 9. Results showed that the SOM values of 43 natural soil samples in the initial calibration set were mainly distributed in the ranges of 0–10 g·kg^{-1} and 10–20 g·kg^{-1}, while the samples in the NSE calibration set (that included 42 natural soil samples and 28 near-natural soil samples) exhibited a log-normal distribution ranging from 0 to 60 g·kg^{-1}. The calibration sample set was distributed over a wide SOM content range that could cover the SOM variation of future

predicted samples. Theoretically, natural soil derived from parent materials should follow a normal distribution. However, soil data obtained from the NSE calibration set showed a log-normal distribution. In fact, the soil system of the research area has experienced high-intensity human activities (especially agricultural practices) for a long time, and the initial data distribution has been changed and commonly exhibits a log-normal distribution. Various studies have reported these articles [61], [63], [64], and [65]. The log-normal distributed calibration set was close to the SOM frequency distribution of natural soil data, and thus satisfies the data distribution prerequisites of a machine learning model, and strengthens the model's prediction ability for SOM contents under field conditions [22]. Furthermore, Lucà et al. [12] suggested that the minimum number of soil samples for a machine learning model (e.g., support vector machine) to be sufficiently trained is 72. The sample size of the initial calibration set ($n = 43$) is smaller than this threshold, which may be a reason for the poor performance of the initial calibrated model. Note that the enhanced calibration set in this study consisted of 70 representative samples, which is basically consistent with the results of Lucà et al. [12].

In addition, a widely-used representative sample selection strategy (Kennard-Stone method) [10] was used for calibration set construction in this study. The calibrated model's performance ($R^2 = 0.75$; RPIQ = 2.71) was better than the spectral-based model built with the initial calibration set ($R^2 = 0.73$; RPIQ = 2.32), while it was lower than the spectral-based model ($R^2 = 0.90$; RPIQ = 4.17) calibrated with the NSE calibration set. The effect of an inappropriate sampling pattern on the initial calibration set cannot be eliminated by a sample selection method. The calibration set enhancement strategy based on the near-natural soil samples shows the obvious advantage of building an effective calibration set under the conditions of sampling pattern bias.

In general, near-natural soil samples involved in calibration set enhancement to build an improved spectral-based model for SOM estimation is a practical method. However, an important prerequisite is to eliminate the spectral difference between natural and near-natural samples because the spectral difference would interfere with the coverage assessment of the sample set, damaging the calibration set enhancement. Additionally, near-natural soil sample preparation may be more expensive than field sampling costs in the short term, but the constructed near-natural sample dataset can support the calibration set enhancement for a long time into the future. This strategy still seems to be an efficient and low-cost method for SOM spectral-based estimation at the field scale.

C. Uncertainty Analysis

The established national soil spectral database for calibration set enhancement may be an alternative solution. Rossel et al. [39], Seidel et al. [66], and Zhao et al. [67] integrated the field-obtained spectra of natural soil samples with the spectra from the national soil spectroscopic database (e.g., Chinese Soil Spectroscopic Database), and the calibration model produced high accuracy for key soil component estimation. However, some issues with calibration set enhancement based on the national soil spectroscopic database need to be resolved. First, the acquisition standards of natural soil spectra data should be consistent with those of the national soil spectroscopic database and open to the public [6]. Second, a national soil spectroscopic database should be convenient for users to access. However, this is very difficult to accomplish in less developed regions, such as Asia and Africa (i.e., locations where national soil spectral databases are still under construction). Additionally, databases should also cover all soil types within the regional scope. Otherwise, the complexity of calibration set enhancement will be increased and unreliable estimates will be generated [16], [68].

The strategy for calibration set enhancement needs further optimization. A spiking algorithm is an approach that can be used to support the calibration set enhancement for spectral-based estimation of SOM contents. Li et al. [69] and Ji et al. [70] successfully spiked the local sampling dataset into the Chinese Soil Spectroscopic Database data for content estimation of several soil parameters at the local scale. However, if the spectral characteristics of the soil spectral library data are similar to those of the local spectra data of natural soil samples, the spiking method would show fairly poor accuracy [59]. In fact, the enhancement strategy of augmenting representative samples in the calibration set was devoted to explaining the spatial variability, and relied on the spatial variability of the site studied, compared with the spiking method, covering the variability of samples from the different locations in different soil types and landscapes [7]. Furthermore, SOM is not a unique factor affecting soil spectral reflectance, and the coverage assessment and spectral-based estimation model may be affected by various components of soils [16], [55]. In the future, a matrix effect correction should be considered in constructing the spectral-based estimation model. Additionally, as a result of the autocorrelations in the observed samples, the data cross-validation approach method may generate overly optimistic validation results in spectral-based SOM estimates, and thus not be recognized as a robust validation method. An unbiased validation strategy based on the design-obtained sampling method or adding additional independent samples may be an ideal method for validating model accuracy [71], [72]. Although the issue described above is obviously beyond the scope of our current research, it deserves further study.

V. Conclusion

In this study, an NSE calibration set strategy was proposed for improving the spectral-based estimation model of SOM contents for use under conditions when the field sampling patterns are biased. Results showed that the NSE calibration set (that included 42 natural soil samples and 28 near-natural soil samples) posed sufficient coverage and better information integrity within the estimators space than the initial calibration set (that included 43 natural soil samples), and this calibration set would be beneficial to the machine learning-based model calibration. The RF model based on the NSE calibration set produced a satisfactory result ($R^2 = 0.90$; RPIQ = 4.17) in SOM estimation, in contrast

to the fairly poor accuracy ($R^2 = 0.73$; RPIQ = 2.32) of the spectroscopy model when the initial calibration set calibrated it. The SOM chemical compositions (e.g., lipids, polysaccharides, and lignin) and their relative abundance from the laboratory-simulated near-natural soil samples were basically consistent with those of natural soil samples. The use of near-natural soil samples and a coverage assessment method for calibration set enhancement was proven to be a practical method. The results presented in this study provide an efficient strategy for building a calibration set that can be applied to a pattern-biased field sampling dataset to improve the spectral estimation of SOM contents under field conditions at the local scale.

REFERENCES

[1] T. Wu, J. Luo, W. Dong, Y. Sun, L. Xia, and X. Zhang, "Geo-object-based soil organic matter mapping using machine learning algorithms with multi-source geo-spatial data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1091–1106, Apr. 2019.

[2] A. D. Bayer, M. Bachmann, D. Rogge, A. Müller, and H. Kaufmann, "Combining field and imaging spectroscopy to map soil organic carbon in a semiarid environment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 3997–4010, Sep. 2016.

[3] H. Eswaran, E. Van Den Berg, and P. Reich, "Organic carbon in soils of the world," *Soil Sci. Soc. Amer. J.*, vol. 57, no. 1, pp. 192–194, 1993.

[4] X. Li, J. Ding, J. Liu, X. Ge, and J. Zhang, "Digital mapping of soil organic carbon using sentinel series data: A case study of the Ebinur lake watershed in Xinjiang," *Remote Sens.,* vol. 13, no. 4, 2021, Art. no. 769.

[5] Z. Zhang et al., "Strategies for the efficient estimation of soil organic matter in salt-affected soils through Vis-NIR spectroscopy: Optimal band combination algorithm and spectral degradation," *Geoderma,* vol. 382, 2021, Art. no. 114729.

[6] E. B. Dor, C. Ong, and I. C. Lau, "Reflectance measurements of soils in the laboratory: Standards and protocols," *Geoderma,* vol. 245, pp. 112–124, 2015.

[7] S. Nawar and A. M. Mouazen, "Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy," *Comput. Electron. Agriculture*, vol. 151, pp. 469–477, 2018.

[8] J. P. Ackerson, C. Morgan, and Y. Ge, "Penetrometer-mounted Vis-NIR spectroscopy: Application of EPO-PLS to in situ VisNIR spectra," *Geoderma*, vol. 286, pp. 131–138, 2017.

[9] R. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma,* vol. 158, no. 1-2, pp. 46–54, 2010.

[10] Z. Zhang, J. Ding, J. Wang, and X. Ge, "Prediction of soil organic matter in northwestern China using fractional-order derivative spectroscopy and modified normalized difference indices," *Catena*, vol. 185, 2020, Art. no. 104257.

[11] Y. Liu et al., "The influence of spectral pretreatment on the selection of representative calibration samples for soil organic matter estimation using Vis-NIR reflectance spectroscopy," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 450.

[12] F. Lucà, M. Conforti, A. Castrignanò, G. Matteucci, and G. Buttafuoco, "Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy," *Geoderma,* vol. 288, pp. 175–183, 2017.

[13] P. Berzaghi, J. S. Shenk, and M. O. Westerhaus, "LOCAL prediction with near infrared multi-product databases," *J. Near Infrared Spectrosc.*, vol. 8, no. 1, pp. 1–9, 2000.

[14] C. Guerrero, R. Zornoza, I. Gómez, and J. Mataix-Beneyto, "Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy," *Geoderma,* vol. 158, no. 1-2, pp. 66–77, 2010.

[15] P. T. Von Hippel, "New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples," *Struct. Equation Model., Multidisciplinary J.*, vol. 23, no. 3, pp. 422–437, 2016.

[16] Y. Bao et al., "Vis-SWIR spectral prediction model for soil organic matter with different grouping strategies," *Catena,* vol. 195, 2020, Art. no. 104703.

[17] A. M.-C. Wadoux, B. Minasny, and A. B. McBratney, "Machine learning for digital soil mapping: Applications, challenges and suggested solutions," *Earth-Sci. Rev.*, vol. 210, 2020, Art. no. 103359.

[18] M. Zhou, B. Zou, Y. TU, and J. Xia, "Hyperspectral modeling of Pb content in mining area based on spectral feature band extracted from near standard soil samples," *Spectrosc. Spectral Anal.*, vol. 40, no. 7, pp. 2182–2187, 2020.

[19] R. V. Rossel et al., "A global spectral library to characterize the world's soil," *Earth-Sci. Rev.,* vol. 155, pp. 198–230, 2016.

[20] B. Zou, X. Jiang, H. Feng, Y. Tu, and C. Tao, "Multisource spectral-integrated estimation of cadmium concentrations in soil using a direct standardization and spiking algorithm," *Sci. Total Environ.,* vol. 701, 2020, Art. no. 134890.

[21] H. Meyer, C. Reudenbach, S. Wöllauer, and T. Nauss, "Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction," *Ecological Modelling*, vol. 411, 2019, Art. no. 108815.

[22] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, Art. no. 114.

[23] Y. Jiang, L. Shi, A.-L. Guang, Z. Mu, H. Zhan, and Y. Wu, "Contamination levels and human health risk assessment of toxic heavy metals in street dust in an industrial city in Northwest China," *Environ. Geochem. Health*, vol. 40, no. 5, pp. 2007–2020, 2018.

[24] J. Farifteh, F. Van der Meer, M. Van der Meijde, and C. Atzberger, "Spectral characteristics of salt-affected soils: A laboratory experiment," *Geoderma,* vol. 145, no. 3-4, pp. 196–206, 2008.

[25] T. P. J. Linsinger, J. Pauwels, A. M. H. van der Veen, H. Schimmel, and A. Lamberty, "Homogeneity and stability of reference materials," *Accreditation Qual. Assurance*, vol. 6, no. 1, pp. 20–25, 2001.

[26] Y. Wang, C. Tao, and B. Zou, "A transfer learning approach utilizing combined artificial samples for improved robustness of model to estimate heavy metal contamination in soil," *IEEE Access*, vol. 8, pp. 176960–176972, 2020.

[27] J. Wang, J. Ding, A. Abulimiti, and L. Cai, "Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS–NIR) spectroscopy, Ebinur Lake Wetland, Northwest China," *PeerJ,* vol. 6, 2018, Art. no. e4703.

[28] X. Xu, Y. Chen, M. Wang, S. Wang, K. Li, and Y. Li, "Improving estimates of soil salt content by using two-date image spectral changes in Yinbei, China," *Remote Sens.*, vol. 13, no. 20, 2021, Art. no. 4165.

[29] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[30] A. Gholizadeh, D. Žižala, M. Saberioon, and L. Borůvka, "Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging," *Remote Sens. Environ.,* vol. 218, pp. 89–103, 2018.

[31] M. Kibblewhite, K. Ritz, and M. Swift, "Soil health in agricultural systems," *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 363, no. 1492, pp. 685–701, 2008.

[32] A. M. C. Wadoux et al., "Ten challenges for the future of pedometrics," *Geoderma,* vol. 401, 2021, Art. no. 115155.

[33] Agricultural Ministry of the People's Republic of China (AMPRC), Bio-organic fertilizer implementation standards (New York, NY, USA 884—2012), Beijing, 2012.

[34] S. Derenne and K. Quenea, "Analytical pyrolysis as a tool to probe soil organic matter," *J. Anal. Appl. Pyrolysis,* vol. 111, pp. 108–120, 2015.

[35] G. S. Santana, H. Knicker, F. J. González-Vila, J. A. González-Pérez, and D. P. Dick, "The impact of exotic forest plantations on the chemical composition of soil organic matter in Southern Brazil as assessed by Py–GC/MS and lipid extracts study," *Geoderma Regional*, vol. 4, pp. 11–19, 2015.

[36] B. Minasny et al., "Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon," *Geoderma,* vol. 167, pp. 118–124, 2011.

[37] S. Mirzaei et al., "Minimising the effect of moisture on soil property prediction accuracy using external parameter orthogonalization," *Soil Tillage Res.*, vol. 215, 2022, Art. no. 105225.

[38] X. Zhao and B. Ye, "Selection of effective singular values using difference spectrum and its application to fault diagnosis of headstock," *Mech. Syst. Signal Process.,* vol. 25, no. 5, pp. 1617–1631, 2011.

[39] R. V. Rossel, S. R. Cattle, A. Ortega, and Y. Fouad, "In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy," *Geoderma,* vol. 150, no. 3-4, pp. 253–266, 2009.

[40] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, 2017.
[41] M. Li, L. Zhen, and X. Yao, "How to read many-objective solution sets in parallel coordinates," *IEEE Comput. Intell. Mag.*, vol. 12, no. 4, pp. 88–100, 2017.
[42] L. X. Liu G. and Z. Han, "Selection of building energy consumption prediction machine learning algorithms and parameter setting based on quality of samples," *J. Chongqing Univ.,* vol. 45, no. 5, pp. 79–95, 2022.
[43] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[44] W. Chen et al., "Estimating PM2. 5 with high-resolution 1-km AOD data and an improved machine learning model over Shenzhen, China," *Sci. Total Environ.,* vol. 746, 2020, Art. no. 141093.
[45] J. Farifteh, F. Van der Meer, C. Atzberger, and E. Carranza, "Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN)," *Remote Sens. Environ.,* vol. 110, no. 1, pp. 59–78, 2007.
[46] X. Meng et al., "Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data," *Int. J. Appl. Earth Observ. Geoinf.,* vol. 89, 2020, Art. no. 102111.
[47] V. Bellon-Maurel, E. Fernandez-Ahumada, B. Palagos, J.-M. Roger, and A. McBratney, "Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy," *TrAC Trends Anal. Chem.*, vol. 29, no. 9, pp. 1073–1081, 2010.
[48] National Soil Census Office (NSCO), *Soil Census Technology in China*. Beijing, China: China Agriculture Press, 1992.
[49] B. Kuang and A. Mouazen, "Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms," *Eur. J. Soil Sci.*, vol. 62, no. 4, pp. 629–636, 2011.
[50] Y. Ge, C. L. Morgan, and J. P. Ackerson, "VisNIR spectra of dried ground soils predict properties of soils scanned moist and intact," *Geoderma*, vol. 221, pp. 61–69, 2014.
[51] J. Peng, X. Li, Q. Zhou, S. Zhou, J. Wenjun, and W. Jiaqiang, "Influence of iron oxide on the spectral characteristics of organic matter," *J. Remote Sens.,* vol. 17, no. 06, pp. 1396–1412, 2013.
[52] J. Peng, Q. Zhou, Y. Zhang, and H. Xiang, "Effect of soil organic matter on spectral characteristics of soil," *Acta Pedologica Sinica*, vol. 50, no. 03, pp. 517–524, 2013.
[53] S. Wang, Y. Chen, M. Wang, Y. Zhao, and J. Li, "SPA-Based methods for the quantitative estimation of the soil salt content in saline-alkali land from field spectroscopy data: A case study from the yellow river irrigation regions," *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 967.
[54] J. Wang, X. Hu, T. Shi, L. He, W. Hu, and G. Wu, "Assessing toxic metal chromium in the soil in coal mining areas via proximal sensing: Prerequisites for land rehabilitation and sustainable development," *Geoderma,* vol. 405, 2022, Art. no. 115399.
[55] D. Ou et al., "Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction," *Geoderma,* vol. 385, 2021, Art. no. 114875.
[56] Y. Hong et al., "Prediction of soil organic matter by VIS–NIR spectroscopy using normalized soil moisture index as a proxy of soil moisture," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 28.
[57] R. Viscarra Rossel, E. Bui, P. De Caritat, and N. McKenzie, "Mapping iron oxides and the color of Australian soil using visible–near-infrared reflectance spectra," *J. Geophysical Res., Earth Surf.*, vol. 115, 2010, Art. no. F4.
[58] A. L. Guy, S. D. Siciliano, and E. G. Lamb, "Spiking regional vis-NIR calibration models with local samples to predict soil organic carbon in two high arctic polar deserts using a vis-NIR probe," *Can. J. Soil Sci.*, vol. 95, no. 3, pp. 237–249, 2015.
[59] W. Ng, B. Minasny, E. Jones, and A. McBratney, "To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library," *Geoderma*, vol. 406, 2022, Art. no. 115501.
[60] A. B. McBratney, M. M. Santos, and B. Minasny, "On digital soil mapping," *Geoderma*, vol. 117, no. 1-2, pp. 3–52, 2003.
[61] J. Lv, Y. Liu, Z. Zhang, and J. Dai, "Factorial kriging and stepwise regression approach to identify environmental factors influencing spatial multi-scale variability of heavy metals in soils," *J. Hazard Mater.*, vol. 261, pp. 387–397, Oct. 15, 2013.
[62] S. Nardi, G. Concheri, D. Pizzeghello, A. Sturaro, R. Rella, and G. Parvoli, "Soil organic matter mobilization by root exudates," *Chemosphere*, vol. 41, no. 5, pp. 653–658, 2000.
[63] D. McGrath, C. Zhang, and O. T. Carton, "Geostatistical analyses and hazard assessment on soil lead in Silvermines area, Ireland," *Environ. Pollut.*, vol. 127, no. 2, pp. 239–248, 2004.
[64] X. Xu, M. Ren, J. Cao, Q. Wu, P. Liu, and J. Lv, "Spectroscopic diagnosis of zinc contaminated soils based on competitive adaptive reweighted sampling algorithm and an improved support vector machine," *Spectrosc. Lett.*, vol. 53, no. 2, pp. 86–99, 2020.
[65] C. Zhang, Y. Tang, X. Xu, and G. Kiely, "Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland," *Appl. Geochem.,* vol. 26, no. 7, pp. 1239–1248, 2011.
[66] M. Seidel, C. Hutengs, B. Ludwig, S. Thiele-Bruhn, and M. Vohland, "Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations," *Geoderma*, vol. 354, 2019, Art. no. 113856.
[67] D. Zhao, M. Arshad, J. Wang, and J. Triantafilis, "Soil exchangeable cations estimation using Vis-NIR spectroscopy in different depths: Effects of multiple calibration models and spiking," *Comput. Electron. Agriculture*, vol. 182, 2021, Art. no. 105990.
[68] X. Wang et al., "The minimum level for soil allocation using topsoil reflectance spectra: Genus or species?," *Catena*, vol. 174, pp. 36–47, 2019.
[69] H. Li, Y. Li, M. Yang, S. Chen, and Z. Shi, "Strategies for efficient estimation of soil organic content at the local scale based on a national spectral database," *Land Degradation Develop.*, vol. 33, no. 10, pp. 1649–1661, 2022.
[70] W. Ji, R. Viscarra Rossel, and Z. Shi, "Accounting for the effects of water and the environment on proximally sensed vis–NIR soil spectra and their calibrations," *Eur. J. Soil Sci.*, vol. 66, no. 3, pp. 555–565, 2015.
[71] D. Brus, B. Kempen, and G. Heuvelink, "Sampling for validation of digital soil maps," *Eur. J. Soil Sci.*, vol. 62, no. 3, pp. 394–407, 2011.
[72] A. M. C. Wadoux, G. B. Heuvelink, S. De Bruin, and D. J. Brus, "Spatial cross-validation is not the right way to evaluate map accuracy," *Ecological Modelling*, vol. 457, 2021, Art. no. 109692.

Xibo Xu received the B.S. degree in geographical science from the School of Tourism and Resource Environment, Zaozhuang University, Zaozhuang, China, in 2016. He is currently working toward the Ph.D. degree in cartography and geography information system at the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China.

His research interests include quantitative spectroscopic analysis in soils and risk assessment of soil heavy metals.

Yunhao Chen received the B.S. and M.S. degrees in resource management from Anhui University of Science and Technology, Huainan, China, in 1994 and 1997, respectively, and the Ph.D. degree in geodetic engineering from the China University of Mining and Technology, Beijing, China, in 1999.

From 2000 to 2001, he was a Postdoctoral Researcher with Beijing Normal University, Beijing, China. Since 2001, he has been with the Faculty of Geographical Science, Beijing Normal University, where he is currently a Professor with the State Key Laboratory of Remote Sensing Science. His research interests include thermal remote sensing of urban environment and applications of remote sensing in ecology.

Xiujuan Dai received the B.S. degree in geographic information system from the School of Geosciences and Surveying Engineering, China University of Mining and Technology, Beijing, China, in 2020. She is currently working toward the M.S. degree in cartography and geography information system at the Beijing Key Laboratory of Environmental Remote Sensing and Digital City, Faculty of Geographical Science, Beijing Normal University, Beijing.

Her research interests include soil hyperspectral analysis and urban thermal environment.

Tianjie Lei received the B.S. degree in land resource management from Henan University of Technology, in 2008, the M.S. and Ph.D. degrees in geographical science from Beijing Normal University, Beijing, China, in 2011 and 2015, respectively.

He is mainly engaged in the research of monitoring and evaluation of natural disasters, and ecological security based-on air-space-ground big data.

Sijia Wang received the B.S. degree in geographic information system from the School of Geosciences and Surveying Engineering, China University of Mining and Technology, Beijing, China, in 2016. She is currently working toward the Ph.D. degree in cartography and geography information system at the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing.

Her research interests include soil salinity mapping.

Kangning Li received the B.S. degree in geographic information system from the College of Urban and Environmental Science, Northwest University, Xi'an, China, in 2016. She is currently working toward the Ph.D. degree with the Faculty of Geographical Science, Beijing Normal University, Beijing, China.

Her research interests include thermal remote sensing of urban environment.