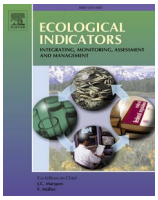
Ecological Indicators 158 (2024) 111437

Contents lists available at ScienceDirect

Ecological Indicators

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



Mapping vertical distribution of SOC and TN in reclaimed mine soils using point and imaging spectroscopy

Sihan Peng [a], Nisha Bao [a,*], Shijia Wang [b], Asa Gholizadeh [c], Mohammadmehdi Saberioon [d], Yi Peng [e]

[a] College of Resources and Civil Engineering, Northeastern University, Shenyang, China
[b] College of Arts, Humanities and Social Sciences, The University of Edinburgh, Edinburgh, the United Kingdom
[c] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Kamycka 129, Suchdol, Prague 16500, Czech Republic
[d] Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Section 1.4 Remote Sensing and Geoinformatics, Telegrafenberg, Potsdam 14473, Germany
[e] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China

ARTICLE INFO

Keywords:
Visible and near infrared reflectance spectroscopy
Soil horizon
Reclaimed soil
Coal mine area, Soil properties mapping

ABSTRACT

Soil organic carbon (SOC) and total nitrogen (TN) contents in different soil horizons are essential for vegetation growth and crucial indicators to evaluate soil quality in reclaimed mining areas. Compared with conventional wet chemistry methods, soil spectroscopy, including imaging spectroscopy, can be used as a cost and time-efficient soil analysis technique. However, there is a great challenge in combining laboratory point spectra and laboratory hyperspectral imagery for mapping vertical distribution of SOC and TN (0–100 cm) in reclaimed soils. This is primarily because mixing of spectral data from different sources and technologies to improve soil models is still in its infancy. The main objective of this study is to provide a generic workflow to efficiently evaluate and map reclaimed mine soils in different horizons using imaging spectroscopy and machine learning approaches. A total of 65 soil samples (0–100 cm) were collected from three reclaimed mining lands and one natural site in northern China. Both point soil spectral information and hyperspectral images (350–2500 nm) were obtained under laboratory condition. In order to enhance the relationship between soil quality indicators and spectral features, the stacked feature selection algorithms and three-bands spectral indices were proposed for further modelling. Three machine learning methods (partial least squares regression; PLSR, random forest; RF, and radial basis function model; RBF) based on the point spectra were applied to calibrate and map continuous vertical distribution of SOC and TN. According to the results, thirty spectral bands were identified as important spectral features for SOC and eighteen bands for TN. With feature spectral bands and optimized three-bands spectral indices, the RF model yielded the best predictions for both SOC ($R^2 = 0.97$, RMSE = 7.5 g kg^{-1}) and TN ($R^2 = 0.78$, RMSE = 0.33 g kg^{-1}). It was concluded that imaging spectroscopy can be used to quantify and map soil quality indicators for better monitoring ecological restoration process in reclaimed soil of mining site.

1. Introduction

Surface mining results in large areas of mined land with open pits and dumps causing serious environmental issues, such as plant elimination, soil removal, environmental pollution, and overburdening due to excavation, landform reshaping, and subsurface hydrologic regimes (Bi et al., 2018; Ussiri and Lal, 2005; Wang et al., 2014). Mine soils refer to Technosols, which formed during the mining process of "stripping-transporting-storing-transporting-covering" (Feng et al., 2019). This type of soil generally contains low organic matter and nutrients compared to soils in the natural environment (Liu et al., 2017). Therefore, land reclamation aims to improve soil quality by increasing the content of soil organic carbon (SOC), total nitrogen (TN), phosphorous (P), and exchangeable potassium (K) in Technosols, even reach its original status (Upadhyay et al., 2016).

According to mine reclamation regulations in China (Standard for quality control of land reclamation, 2013), the surface Technosols (0–50 cm) are required to be reclaimed according to different land use plans. On the other hand, Liu et al. (2017) pointed out that the vertical distributions of SOC and TN in the subsoil (60–100 cm) of reclaimed mine

* Corresponding author.
E-mail address: baonisha@mail.neu.edu.cn (N. Bao).

https://doi.org/10.1016/j.ecolind.2023.111437
Received 30 July 2023; Received in revised form 1 December 2023; Accepted 13 December 2023
Available online 21 December 2023

areas are highly dependent on different dumping methods. They also found that land use and vegetation cover could highly affect the changes in the continuous vertical distribution of SOC and TN contents up to 80 cm depth in the reclaimed mine site. In addition, Kumar et al. (2015) found that the vertical distribution of SOC and TN contents in soil profile was influenced by reclamation years. Thus, the SOC and TN distribution in different horizons could be considered as important indicators to show the efficacy of reclamation and even guide ecological restoration in future (Šourková et al., 2005). However, in order to systematically monitor vertical distribution of reclaimed soil in various mining locations, a large number of soil samples are required to obtain sufficient soil data and information.

During the last few decades, wet chemical analysis methods have widely been used to monitor mine soils around the world. However, field investigations with laboratory chemical analysis are time-consuming and costly. To overcome these challenges, numerous researchers attempted to use proximal sensors to estimate soil properties in a rapid, economical, and environmentally friendly way (Andrade et al., 2020; Ben-Dor and Banin, 1995; Gholizadeh et al., 2018). Several studies have shown that visible and near infrared (VIS-NIR; 350–2500 nm) reflectance spectroscopy could extract different soil information through various chemometric modelling approaches (Ben-Dor and Banin, 1995; Heller Pearlshtien and Ben-Dor, 2020; Tavakoli et al., 2023). This is because most of chemical, physical and mineralogical composition of soil information can be effectively characterized by soil spectral information. Meanwhile, different statistical methods have to be applied to efficiently calibrate and extract useful spectral information for targeted soil properties, because spectral information does not directly provide quantitative soil property values.

Moreover, some specific spectral characteristics and indices can be used to identify reclaimed soils in different landforms of dumping sites (Bao et al., 2017). Meanwhile, the SOC in mine soils can be estimated through spectroscopy techniques coupled with machine learning algorithms (Gholizadeh et al., 2020; Wei et al., 2020). To date, most studies mainly focused on topsoil (0–20 cm) using field and laboratory spectroscopy (Feng et al., 2019; Jiang et al., 2017; Wang et al., 2019). However, in case of the need to map and visualize the distribution of soil properties in reclaimed soils, particularly through the various horizons, imaging spectroscopic methods could be more efficient compared to point spectroscopy because imaging spectroscopy has the advantage of obtaining both spectral and spatial domain information (Ben-Dor et al., 2009). Many studies have also proved that spectral information from VIS-NIR range (400–2500 nm) could be particularly important for SOC and TN calibration process (Ben-Dor and Banin, 1995; Martin et al., 2002; Stenberg et al., 2005).

However, there were only very few studies applied imaging spectrocopy (400 nm-2500 nm) to spatially characterize the undisturbed soil cores in deeper soil horizons. Steffens et al. (2013) extracted 66 spectral information from hyperspectral imagery (400–990 nm) to predict SOC and TN contents in forest soils (0–30 cm). Sorenson et al. (2020) successfully estimate and map the continuous vertical distribution of SOC, TN, and clay content using imaging spectroscopy (1000–2500 nm) at the soil depth of 0–100 cm in forest soil of Canada. Moreover, Steffens et al. (2021) showed the great potential of using spectral images from two spectral ranges (410–970 nm and 970–2500 nm) to quantitatively map SOC from undisturbed soil cores (0–30 cm). They also found that imaging spectroscopy with its millions of spectra available in one image, overfitting can easily occur pretending a higher accuracy. Therefore, a robust and transfer predicting calibration model is of great importance to acquire spatial distribution of soil horizons from hyperspectral imagery.

To estimate and map SOC and TN using laboratory-based imaging spectroscopy various machine learning approaches have been investigated (Jordan and Mitchell, 2015). Recently, many researchers attempted to apply different machine learning methods for extracting soil spectral information in a more efficient way. Wu et al. (2018) showed that hyperspectral images with machine learning methods could efficiently map continuous vertical distribution (0–100 cm) of salt contents. Xu et al. (2021) compared various machine learning methods namely artificial neural networks (ANN), Cubist, gaussian process regression (GPR), Random Forest (RF), Support vector machine (SVM), extreme learning machine (ELM), k-nearest neighbour (KNN), multivariate adaptive regression splines (MARS), and extreme gradient boosting (XGBoost) to predict TN in paddy soils, fluvo-aquic soils, and black soils. They found that SVM yielded the best predictions with higher R^2 and lower RMSE values. Additionally, linear models such as partial least square regression (PLSR) maintain the capability to handle soil spectral information from a relatively homogeneous or a small geographical area. Meanwhile, the PLSR is the most common tool used to calibrate spectra and soil chemical components. The advantage of the PLSR algorithm is that it maximizes the covariance between variable X and reference Y and thus the resulting spectra are directly related to the soil characteristics (Wadoux et al., 2021). However, non-linear and non-monotonic methods could more efficiently deal with soil information from more complex environments (i.e., geological, mineralogical, or climatic) or large datasets such as images (de Santana et al., 2021; Jordan and Mitchell, 2015). The RF algorithm is typically known as a hierarchical nonparametric method and is often used to estimate complex nonlinear relationships between spectral data and soil properties (Douglas et al., 2018). In addition, radial basis function (RBF) neural network as a feed-forward neural network with only a 3-layer network structure attains a better function approximation ability and can approximate any non-linear function with an arbitrary accuracy (Emamgholizadeh et al., 2018). Li et al (2021) attempted to apply RBF to estimate soil salinity and showed RBF had the great potential to deal with soil information.

Due to the large number of bands, the VIS-NIR hyperspectral images (350–2500 nm) have high dimensional data. Thus, several feature selection methods have been proposed to be applied to the multivariate models in order to extract useful spectral information from a high dimension dataset. For instance, Xu et al. (2020) applied hyperspectral imagery with artificial neural networks (ANN), cubist regression tree (Cubist), gaussian process regression (GPR), and support vector machine (SVM) as well as the competitive adaptive reweighted sampling (CARS) feature selection method to map SOC in paddy soils. They found that feature selection methods could simplify the models and improve the computation efficiency. Araujo et al. (2001) proposed a successive projections algorithm (SPA) as a novel feature selection algorithm for multivariate calibration analysis, which could effectively decrease the complexity and collinearity of spectral data matrix by performing simple projection operations in a vector space. More than 10 years later, Peng et al. (2014) successfully applied SPA with SVM to improve SOC prediction accuracy based on the selected spectral bands.

Nevertheless, one of the major concerns about using machine learning models is that it is perceived as a black box with low interpretability. For instance, shapley additive explanation (SHAP) values could represent the contribution of covariates to the final model predictions (Lundberg and Lee, 2017). Until very recently, some researchers from soil science community started to apply SHAP values and interpret different geospatial models for a better understanding calibration process (Beucher et al., 2022; Haghi et al., 2021; Padarian et al., 2020).

Currently, most predicting models have been based on spectral data extracted from hyperspectral imagery of dry samples with constant conditions. However, the hyperspectral imageries usually are obtained from fresh samples, resulting in various uncertainties due to different measurement conditions (e.g., roughness, moisture, roots, et). Those uncertainties can largely affect final predictions, when models apply to images. Therefore, some soil scientists tried to develop models from laboratory spectra (350–2500 nm) and directly transfer models to hyperspectral images for predicting and mapping different soil properties. For instance, Zhang et al. (2022) also established a workflow to transfer models from laboratory spectra to satellite imagery (i.e., GF-5)

for some heavy metals (Zn, Ni, and Cu) mapping. The results revealed that transferring the model could be a reliable approach for mapping Zn, Ni, and Cu in coal mine soils.

So far, the imaging spectroscopy has not yet been used to characterize SOC and TN in the reclaimed soils (0–100 cm) and to monitor reclamation process in mining sites. Moreover, to the best of our knowledge, transfer of a model based on laboratory point spectroscopy on dry soil samples to laboratory hyperspectral imagery on fresh soil samples for mapping vertical distribution of SOC and TN (0–100 cm) in reclaimed soils of a mining area has never been conducted. Therefore, there is still a great challenge in establishing a robust model based on point spectra and then applying to hyperspectral imagery for mapping deeper soil horizons. In order to fill this knowledge gap and facilitate the assessment of reclaimed soils in deeper horizons through the use of spectroscopy technology, we aimed to develop a generic workflow to evaluate and map reclaimed soils in different horizons using spectral information from both point and imaging spectroscopy through various machine learning approaches. This work focused on three reclaimed sites of coal mine area in the northern part of China. The other two objectives of this study were 1) to improve SOC and TN prediction using machine learning methods based on the calculated three-bands spectral indices and the selected feature bands and 2) to map and differentiate the distribution of SOC and TN contents through the horizons (0–100 cm).

2. Materials and methods

2.1. Study area

The study area is located in northeastern China, with a remarkable continental climate. The annual average temperature is below 0℃. The precipitation varies from 450 to 550 mm and is concentrated in July and August. The elevation of the study area ranges from 550 m to 800 m (Yang et al., 2014). Three reclaimed study sites and one natural site were selected for the current study, as shown in Fig. 1 (based on the standard map released by the Ministry of Natural Resources of the People's Republic of China [No. GS (2019)1822]). The reclaimed sites belong to a typical fragile ecological area of grassland with poor ecosystem resistance in which, soils are predominantly chestnut soil with a 20–50 cm humus layer and some low-lying areas having meadow soils. The detailed information on the three study sites and one natural site is listed below:

1) Reclaimed site I: Bao Ri Xi Le surface coal mine (119°51′ E, 49°22′ N) with the area of about 42 km^2, where intensive mining and artificial reclamation activities started since 2004 and 2009, respectively. The soil pH ranges between 6.8 and 7.7 and the main plant species are *icegrass, sheepgrass, elymus dahuricus* (major reclamation plant), and *leather oats*.
2) Reclaimed site II: Yimin surface coal mine (YSCM, 119°42′ E, 48°37′ N) with the area of about 10 km^2, where intensive mining and natural recovery were started in 1985 and 1999, respectively. The soils are deficient in phosphorus, rich in potassium, and moderate in nitrogen with the pH values of 8.0–9.1. The plants grown on the surface are *goatgrass, mosswort, artemisia, dandelion, solanum,* and *ashwort.*
3) Reclaimed site III: Dalai Nur coal mine, which is a postmining site (DNCM, 117°45′ E, 49°27′ N) with a mining life of more than 100 years and the mining activities were closed in 2012. The waste dump covers an area of 13 km^2. Soil salinization after pasture degradation has already occurred in some areas. Soils are deficient in nitrogen and phosphorus and pH ranges from 5.5 to 7.5. The main plants that

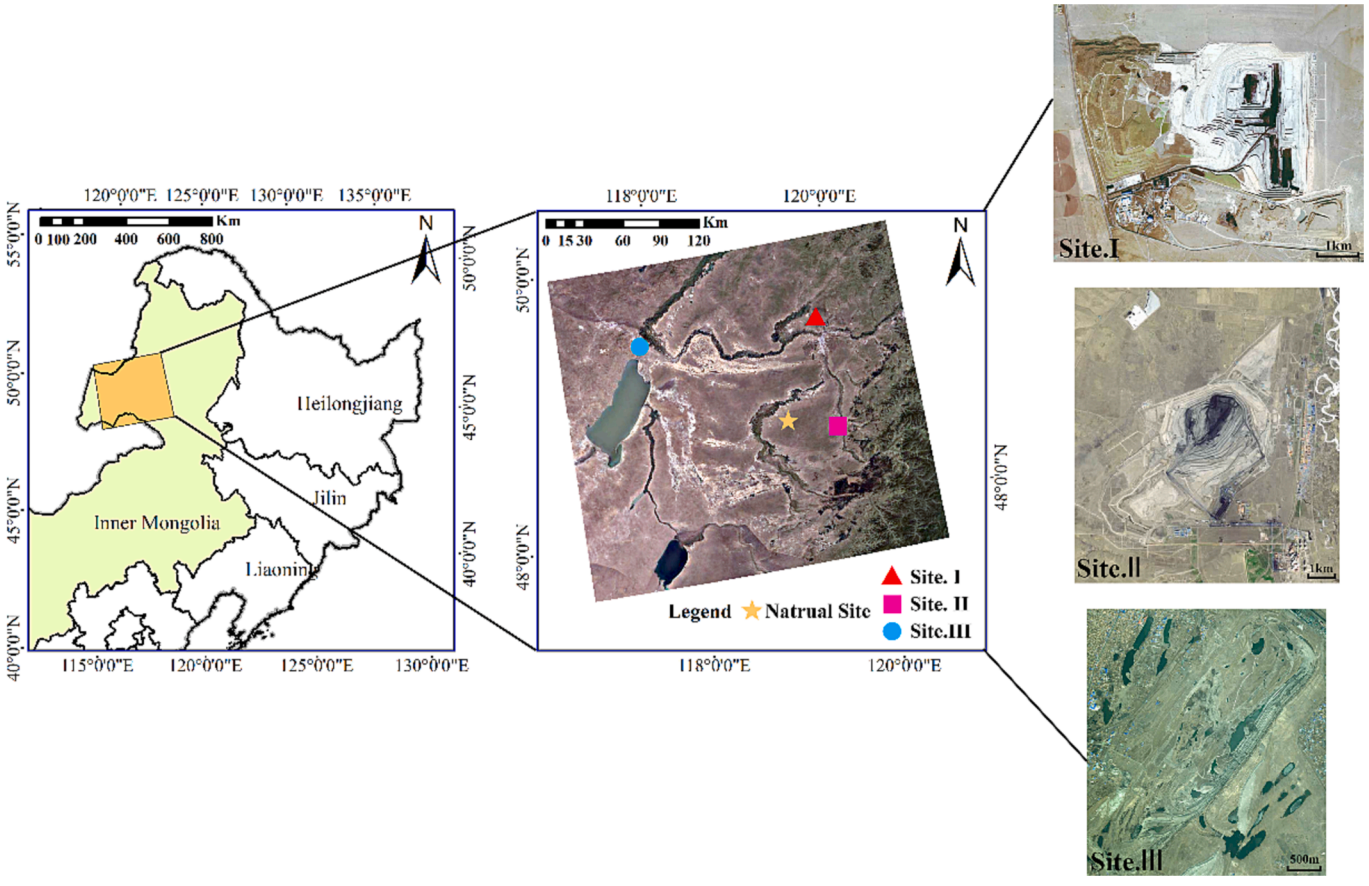


Fig. 1. Study area and location of three open mines.

recover naturally on the surface are *ashwort, cryptomeria, echinocereus, goatgrass, tussock grass,* and *commissures.*

4) Natural site: an undisturbed area is located in a typical meadow grassland near three reclaimed study sites. The pH values are in the range of 8.0–9.1 and the main vegetation covers can be classified as *Potentilla, Carex, Stipa,* and *Oxytropis*.

2.2. Soil sampling and data sources

2.2.1. Soil sampling and chemical analyses

The fieldwork was conducted in August 2017. All soil cores were collected using a closed sampler with an internal diameter of 10 cm.

Site I: four completed soil cores (0–100 cm) from four different reclaimed years (2011, 2012, 2013, and 2014) were collected.

Site II: two completed soil cores (0–100 cm) from two different reclaimed years (2005 and 2009) were collected.

Site III: two completed soil cores (0–100 cm) from two different reclaimed years (2000 and 2005) were collected.

The collected soil cores were kept as fresh samples in individual plastic tubes. We kept the cores in the fridge with 2–5 degree after sampling. Due to the appearance of the rock in deeper horizons, it was difficult to take all soil cores from the depth 0–100 cm in reclaimed mining fields. Therefore, three uncompleted soil cores (one core from 0 to 30 cm and two cores from 0 to 50 cm) only for point spectra and wet chemistry analysis were collected, which resulted in 11 point-spectra with the associated SOC and TN values. Meanwhile, nine completed soil cores (0–100 cm) covered the variation of reclamation mode and time information were taken to obtain the point spectra, hyperspectral imagery, and wet chemistry analysis. It provided nine hyperspectral images and 54 point-spectra with their associated SOC and TN values. Finally, a total of 65 spectra according to different depths (0–10, 10–20, 20–30, 30–50, 50–70, and 70–100 cm) with the associated soil properties were used for the calibration process, and nine hyperspectral images were used for mapping purposes. Additionally, these 65 soil samples for point spectral and wet chemistry analysis were collected from all soil cores.

For the chemical analysis of the soil samples, every fresh sample was weighed approximately 200 g. All samples were then air-dried, crushed, and sieved through a 2 mm mesh for SOC and TN analysis using a dry combustion (Vitti et al., 2016) and an elemental analyzer (LECO CHN, LECO Laboratory Equipment Corporation, MI, USA) in a commercial laboratory with certification of proficiency by International Organization for Standardization ISO 17025:2005.

2.2.2. Hyperspectral imaging and point spectra acquisition and processing

The undisturbed fresh soil samples from soil cores were directly used for hyperspectral imageries collection. We first removed upper part of the plastic tube to expose entire soil profile for image scanning. Before scanning, we used knife to carefully smooth the surface of the soil profile, stones larger than 1 cm were removed from the surface before scanning. Therefore, image spectral information was obtained directly from the soil surface without any effects from plastic tube. All hyperspectral imageries were captured using the GaiaSorter hyperspectral imaging system with two combined sensors developed by Zolix Co., Ltd (Beijing, China). The first sensor covers spectral range from 400 to 1000 nm with 2.8 nm spectral resolution, giving 176 bands, while the second sensor covers spectral range from 1000 to 2500 nm with 10 nm spectral resolution, giving 272 bands. Four 200 W bromine tungsten lamps provided a constant light source during the data acquisition. The exposure time was 12 ms, the scanning speed was 2 cm/s, and the distance between the samples and detector was 30 cm. The raw spectra were consisted of 448 spectral bands. The hyperspectral images were acquired and processed by SpecVIEW software (SpecView Ltd., Uckfield, UK) and saved in RAW format. All hyperspectral images were processed and analyzed using ENVI v5.3 software (Exelis Visual Information Solutions, Boulder, CO, USA). To eliminate noise at the edges of each spectrum, eight bands were removed from each spectrum at 2450–2500 nm. Finally, each imagery consisted of 440 spectral bands in total and were used for mapping the vertical distribution of SOC and TN contents. In order to reduce noises from light scattering, all hyperspectral images were pre-treated by the Savizky-Golay technique (Savitzky and Golay, 1964) with a window size of 11.

For collecting point spectral information, soil samples were air-dried, crushed, and sieved through a 2 mm mesh. The point spectral information of dried samples was obtained with an SVC-HR-1024 spectroradiometer (Spectra Vista Corporation, NY, USA). The spectral range was between 350 and 2500 nm with a spectral interval of 1.5 nm, giving 973 data points. The light source was a 50 W halogen bulb at a 45° zenith angle. A Spectralon^{TM} white plate (Lab-sphere, NH, USA) was used every 30 min as the standard reference. Each sample was repeatedly scanned 10 times over the central area of the sample, and the average value was taken as the reference spectrum. The collected point spectra were resampled through a cubic spline interpolation to match the spectral resolution of the hyperspectral imagery (440 bands). Finally, the point spectra were used to develop SOC and TN prediction models. In order to enhance the relevant peaks of the spectra and reduce the influence of particle size, different preprocessing approaches were applied. First, the Savizky-Golay technique was followed by a window size of 11 then baseline corrections were applied.

2.3. Modelling methods

2.3.1. Feature selection and spectral indices calculation

In order to improve the computation efficiency and accuracy of the final predictions, two feature selection methods were conducted.

The Pearson correlation coefficient (PCC) was firstly applied to resolve high-dimension issues through identifying the specific spectral regions that can be related to the SOC and TN (He et al., 2015; Wang et al., 2019). A subset of the selected bands could be generated through the PCC feature selection to specifically target SOC and TN ($p < 0.001$). However, the selected feature bands after PCC still had the situation of strong collinearity. We then applied the SPA feature selection method to solve feature redundancy-related issues (Araújo et al., 2001) on the results of the PCC- feature band selection. The SPA is a projection-based wrapper approach that accounts correlation between variables by minimizing the vector space collinearity (Galvão et al., 2001). The SPA algorithm has earlier been used for band screening of various spectral information (Peng et al., 2014).

Band combination algorithms have commonly been used to generate different spectral indices through various mathematical transformation methods to enhance final predictions of different soil properties (Wang et al., 2018; Zhang et al., 2020a). We applied three-bands spectral indices of the optimal band combination algorithm (Tian et al., 2011) to explore more potential spectral response relationships between the soil properties and feature bands. The following equations were proposed for the three-bands spectral indices' calculations:

$$TBI1 = (R_i - R_j)/(R_j - R_k) \quad (1)$$

$$TBI2 = (R_i + R_j)/R_k \quad (2)$$

$$TBI3 = (R_i - R_j)/(R_j + R_k) \quad (3)$$

$$TBI4 = R_i/(R_j \times R_k) \quad (4)$$

$$TBI5 = (R_i - R_j)/[(R_i - R_j) - (R_j - R_k)] \quad (5)$$

where, TBI is three-band index, and R_i, R_j and R_k are reflectance values of different selected feature bands (for SOC, $i, j, k = 1$ to 30; for TN, $i, j, k = 1$ to 18). Each calculation was based on equation from (1) to (5) individually and calculated band by band, which gave five initial TBI values that linked to each three bands combination. The correlation

coefficient between soil properties and each initial TBI result were compared, the optimal spectral index was selected according to the highest correlation coefficient.

2.3.2. Model development and accuracy assessment

In order to select representative soil samples to cover all soil variations for calibration purposes, 50 soil samples (~75 %) were selected, as the training set, by set partitioning based on joint X-Y distance (SPXY) method, which has been originated from the Kennard–Stone (KS) algorithm (Kennard and Stone, 1969). The SPXY first selects two samples with the farthest Euclidean distance, then selects samples with the maximum and minimum distances until the number of samples reaches the specified number (Galvao et al., 2005). The rest of the 15 samples (~25 %) were used for the independent validation purpose (testing set).

The calibration models were developed based on the point spectra and soil properties (SOC and TN contents), then further applied to map the hyperspectral images. Three the most common algorithms (PLSR, RF and RBF) which cover both linear and non-linear situations were used to calibrate spectra and soil properties in this study. The PLSR (Geladi and Kowalski, 1986) and RF (Breiman, 2001) have intensively been used for modelling practices in soil science during the last 30 years (Atzberger et al., 2010; Bartholomeus et al., 2008; Ogrič et al., 2019; S. Wang et al., 2018). The RBF neural network is a type of leading network with an intermediate layer (Ghorbani et al., 2013) along with a stimulus function for the neuron and a radial function with a specific center and width (Kalra et al., 2005). Therefore, the RBF neural network is expected to outperform traditional multilayer perceptions and has been applied in predicting some soil properties (Girosi and Poggio, 1990; Li et al., 2021).

Therefore, three commonly used machine learning approaches for chemometric modelling were used to calibrate point spectra as variable with the associated soil properties as response variable (SOC and TN contents). Each approach (PLSR, RF, and RBF) was individually applied to two different spectral datasets: 1) full spectral dataset with 440 bands and 2) the selected feature bands + five calculated three-bands spectral indices. All calibration models were developed with leave-one-out cross-validation. For the PLSR model, the optimal numbers of factors were determined by minimizing the prediction error of validation. The models were subsequently applied and tested on independent validation set. Finally, a total of 6 models were developed for each studied soil property (SOC and TN), and the independent validation results from different models were compared. In this study, the PLSR calculations were implemented using the "pls" package in R version 4.2.1. The RF regression model was built using the "TreeBagger" function in MATLAB software and the number of trees was set to 1000, as recommended by Wang et al. (2018). Finally, the RBF programs were operated using MATLAB software's "newrbe" function.

To evaluate the final performance of the models, a coefficient of determination (R^2), root mean square error (RMSE), Lin's concordance correlation coefficient (LCCC), and bias were utilized. R^2 is a measure of the proportion of explained variance in data. The larger the value of R^2 in the range of 0 to 1, the more stable the model and the better its fit. The deviation between the predicted and observed values can be detected using the RMSE and bias. Therefore, the smaller the RMSE and bias of the model, the more accurate its prediction ability is. LCCC determines the extent to which the 1:1 regression line deviates from the uniform slope based on accuracy and bias measurements (Lin, 1989). Large values of LCCC are indicators of a good predictive model. In addition, the statistical differences between the mean values of soil SOC and TN in each layer and the sample points after inversion were analyzed.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

$$LCCC = \frac{2rs_{\hat{y}}s_y}{s_{\hat{y}}^2 + s_y^2 + (\bar{\hat{y}} - \bar{y})^2} \tag{8}$$

$$bias = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{9}$$

where, y_i and $\hat{y}_i$ are the observed and predicted values at the time i, n is the number of total samples, $\bar{y}$ and $\bar{\hat{y}}$ are the mean of the observed and predicted values, r is the Pearson correlation coefficient between the observed and predicted values, and s_y and $s_{\hat{y}}$ are the standard deviations of observed and predicted values.

In this study, SHAP analysis (Lundberg et al., 2018) was used to understand input variables' effects and interpret machine learning models' performance. The SHAPs are constructed to clarify the importance of each input variable by comparing the contribution to the model according to the presence or absence of each variable (Haghi et al., 2021). Fig. 2 indicates a concise description of the procedure used in this study.

3. Results and discussion

3.1. Exploratory data analysis

3.1.1. SOC and TN descriptive statistics

Fig. 3 shows the distributions of SOC and TN from three reclaimed sites and one natural site. The number in the sample label represents the year when the reclamation of the sample collection area began. At the natural site, the concentration of SOC and TN gradually decreased with increasing the depth. At the reclamation site I, the SOC and TN contents in topsoil (0–20 cm) were much higher than soils below 20 cm depth during the years 2012 and 2013. Indeed, the reclamation activities of artificial revegetation quickly increased SOC and TN contents in topsoil through litter accumulation (Liu et al., 2017). However, subsoils in site I (2014) had higher SOC and TN contents than the topsoil.

It can also be seen that at the reclamation site II, the contents of SOC and TN in 2005, greatly increased with increasing depth (Fig. 3). This changing trend was contrary to what we have shown in the natural site. Meanwhile, in 2005 and 2009, the SOC contents in the site II were higher than those of the natural site at all depths. Though, TN contents in 0–10 cm were lower than the soils in the natural site. Various dumping strategies could explain this unusual trend of SOC and TN vertical distribution.

Regarding the reclamation site III, similar patterns of SOC content were found in the years 2000 and 2005 at the depths below 20 cm (Fig. 3). Moreover, in 2000, the changes of SOC and TN contents in topsoil (0–20 cm) were similar to those in natural site. This is mainly because longer reclamation years occurred in this site during the year 2000. In addition, soils with the highest SOC and TN were located between the depths 30 and 50 cm. No regular SOC and TN trends through the horizons were found in these two sites (site I and II).

3.1.2. Soil spectra characteristic

Based on different SOC (5.22–84.91 g kg^{-1}) and TN (0.33–4.16 g kg^{-1}) contents, selected seven representative laboratory spectra were shown in Fig. 4 to demonstrated how SOC and TN contents could affect spectral information. It can be seen that the reflectance intensity was gradually decreased by increasing the SOC and TN contents. Soil samples with the highest SOC content (84.91 g kg^{-1}) and TN content (4.16 g kg^{-1}) showed the lowest reflectance values (Fig. 4). Considering these two spectra, we detected the first peak around 665 nm, which were associated with soil samples with low SOC (5.22 g kg^{-1}) and TN (0.50 g kg^{-1}) contents. This was mainly due to iron-containing minerals (Scheinost, 1998). The absorption features that appeared near the wavelengths of 1400, 1900, and 2200 nm in the spectra of all samples, caused by the combined vibration of water bound in the interlayer lattices as a hydrated cation and water adsorbed on the particle surface

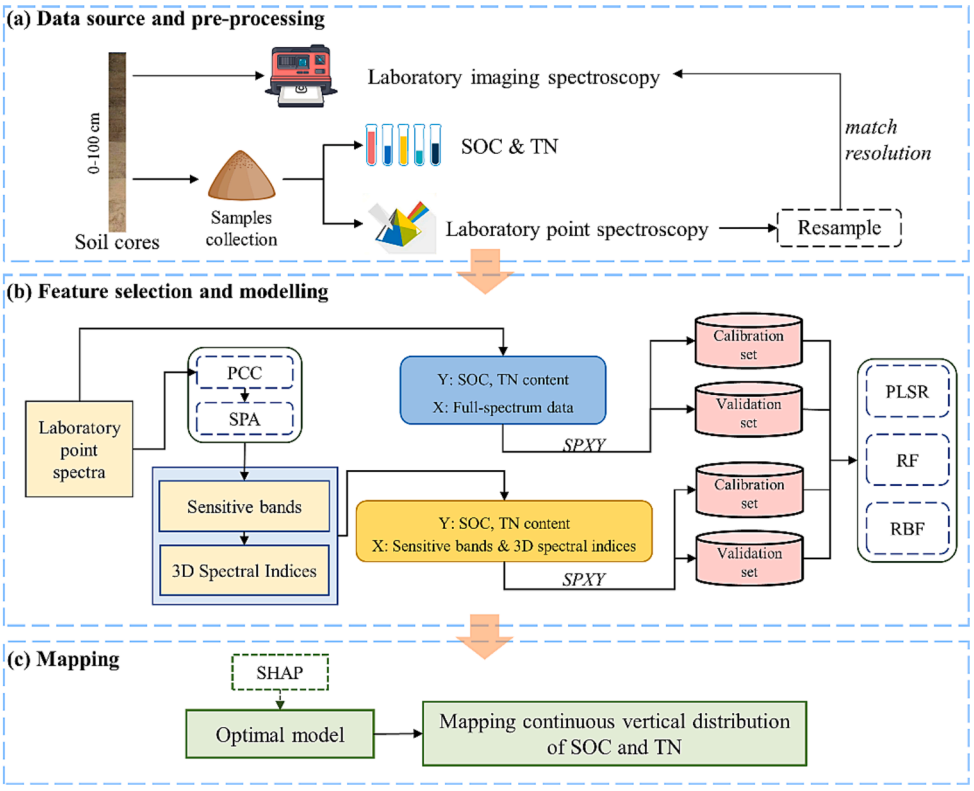


Fig. 2. Flowchart of the procedure applied in this study.

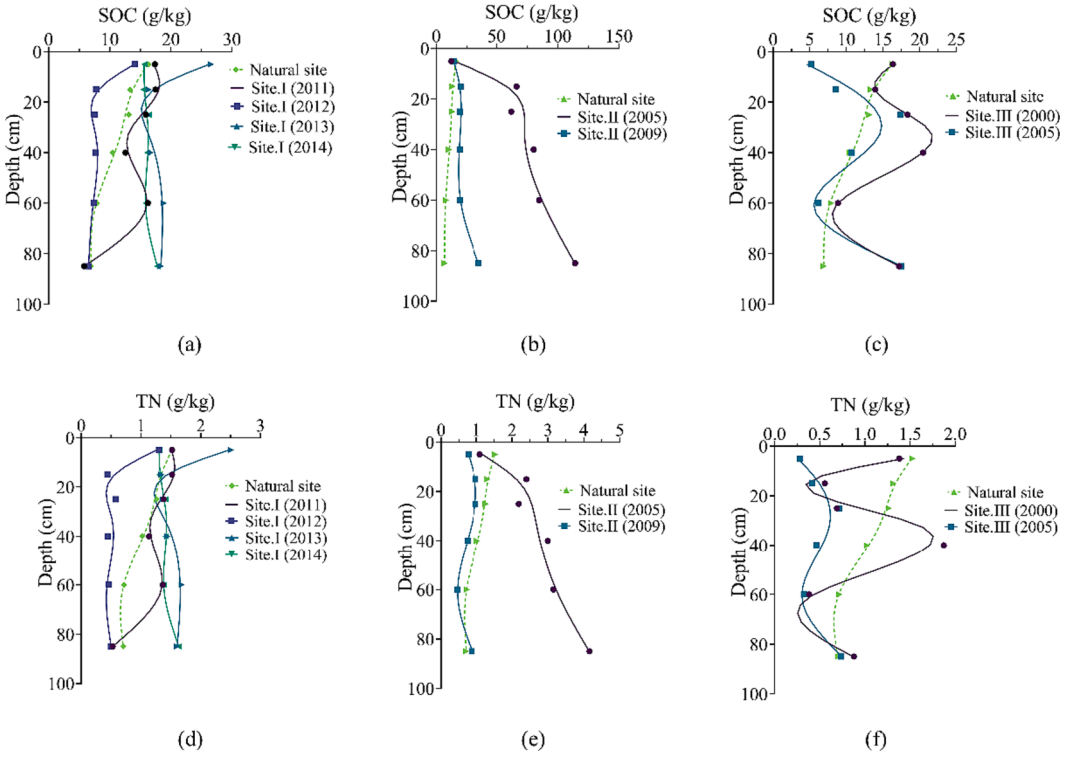


Fig. 3. Characteristics spline of the three reclaimed sites (a, b, and c) for SOC and (d, e, and f) TN. The numbers represent the reclamation years, for example, site I (2005) indicates reclamation was done in 2005.

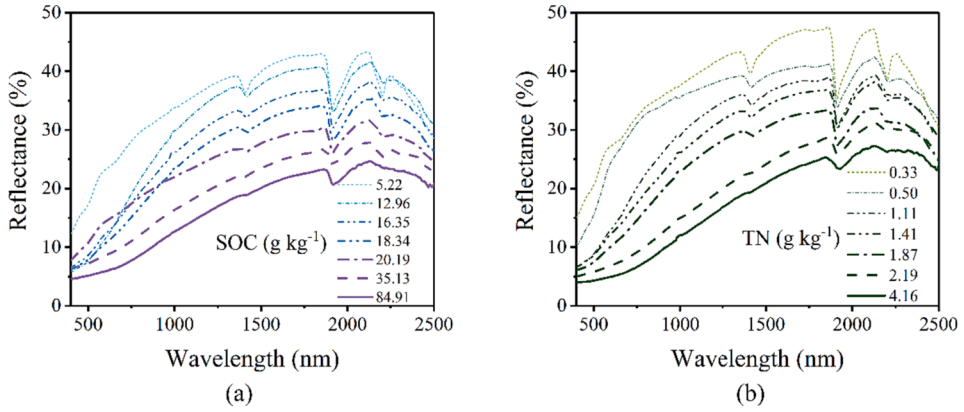


Fig. 4. Reflectance after preprocessing of partial soil samples with different contents of (a) SOC and (b) TN.

(Stenberg et al., 2010), while the spectral feature at 2300 nm can be related to organic matter (Araújo et al., 2014).

3.2. Feature selection

3.2.1. Feature bands selection

The correlation between the soil properties (SOC and TN) of 65 samples and corresponding reflectance values from each band were calculated and displayed in Fig. 5. For SOC, this study showed that the correlation coefficient increased dramatically from the wavelength 350 to 1400 nm and then decreased slightly until 2450 nm. The highest correlation value (0.76) was recorded at 1400 nm. For TN, the correlation coefficient ranged from 0.36 to 0.70 for the wavelengths 2450 and 400 nm, respectively. Although, the correlation generally indicated a gradual decrease from 350 to 2450 nm (Fig. 5).

The SPA feature selection was also performed and the results were shown in Fig. 6. The best combination of feature bands was based on the lowest root mean square error of cross-validation ($RMSE_{CV}$) from multiple linear regression. For SOC, a total of 30 bands were selected as feature bands, in which, 29 bands were from 1100 to 1500 nm and one band was from 1890 nm. Several studies have also highlighted the importance of the NIR spectral region for determining SOC contents (Sarkhot et al., 2011; Vohland and Emmerling, 2011), which is in correspondence to this study. The reason is that the overtones and combinations of fundamental vibrations related to soil composition and stretching and bending of N–H, O–H, and C–H groups are concentrated in the NIR region (Thissen et al., 2004; Viscarra Rossel et al., 2006). A total of 18 bands were chosen and considered as feature bands for the TN evaluation, which 7 bands were located in the visible (VIS) region and 11 bands were in the NIR region. Our results agreed with previous findings, Jiang et al. (2017) reported important bands (420, 470, 520, 680, 1000, 2200 and 2300 nm) for TN estimation. Sorenson et al. (2018) also reported similar result for SOC (1093, 2312, 2374, 1874, 2355, 1917, 1930, 2493, 2262, 1867 nm) and TN (2043, 1010, 985, 991, 1484, 2024, 2011, 2062, 2087, 1490 nm) estimation. Nevertheless, our study has not found important feature bands after 2000 nm for both SOC and TN estimation. This probably because of the reclaimed soils generally contain more SOC than soils from nature sites. Also, due to NIR region is characterized by broad, superimposed, and weak vibrational modes, which gives the broad and overlapping bands (Stenberg et al., 2010). Consequently, in this case, the majority of spectral features occurred after 2000 nm that are mainly associated with soil minerals, not strongly linked to SOC and TN.

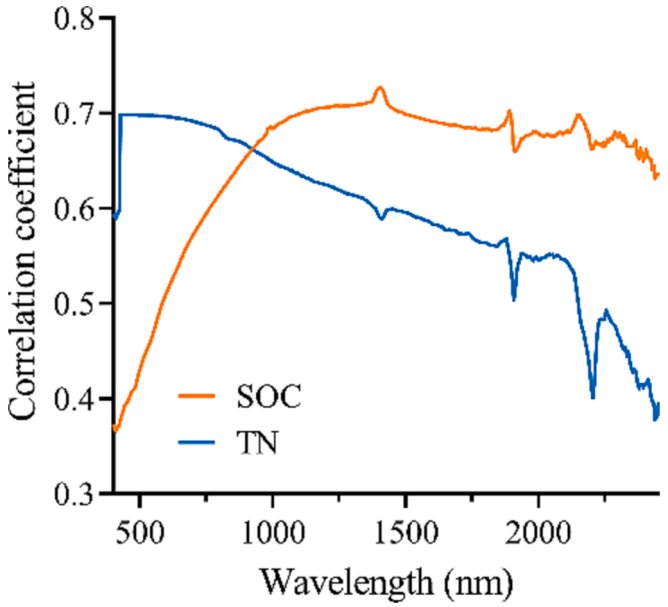


Fig. 5. The PCC between SOC and TN contents, and reflectance (65 samples).

3.2.2. Three-bands spectral indices

The chosen optimal three-bands spectral indices with the highest coefficient values were summarized in Table.1. It can be seen that for spectral indices related to SOC content, both TBI4 and TBI5 had the highest correlation coefficient of 0.86. However, for TN content, the spectral indices of TBI1 produced the highest correlation coefficient of 0.84. Compared to raw spectral information, the spectral indices showed stronger correlation with SOC and TN contents. This is most likely because the mathematical calculation could fully consider the interaction information between the bands, minimize the influence of irrelevant bands, and enhance the relationship between soil composition and spectral information (Zhang et al., 2021, 2020b; Zhu et al., 2020).

3.3. Comparison of the developed models

3.3.1. Performance of different prediction models for SOC and TN

Table 2 shows the general statistics of SOC and TN of the 50 calibration samples and 15 validation samples were used in this study. Samples from three study sites represent a highly heterogeneous dataset. Samples for SOC calibration ranged from 5.22 to 84.91 g kg^{-1}, for SOC validation ranged from 13.94 to 80.40 g kg^{-1}. Samples for TN calibration ranged from 0.28 to 4.16 g kg^{-1}, for TN validation ranged from 0.56 to 3.15 g kg^{-1}.

The performance of different prediction models for SOC and TN is shown in Table 3. Based on the feature bands and spectral indices derived from point spectra, the RF approach produced the best prediction accuracy for both SOC ($R^2 = 0.97$, RMSE = 7.5 g kg^{-1}, LCCC = 0.84, bias = 3.70 g kg^{-1}) and TN ($R^2 = 0.78$, RMSE = 0.33 g kg^{-1}, LCCC =

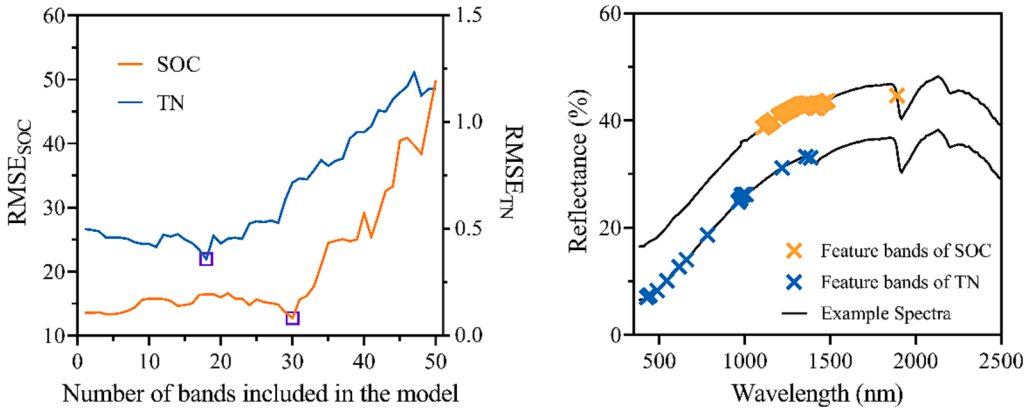


Fig. 6. The operation processes using SPA (left), selected feature bands (right), 30 bands for SOC and 18 bands for TN.

Table 1
Pearson correlation coefficient of the band combinations of the optimal spectral indices with SOC and TN contents.

Index	SOC		TN	
	Bands combination	|r|	Bands combination	|r|
TBI1	(1324.2 nm-1425.0 nm)/ (1425.0 nm-1441.8 nm)	0.74	(1217.7 nm-543.6 nm)/ (543.6 nm-488.4 nm)	0.84
TBI2	(1486.6 nm + 1105.7 nm)/ 1167.3 nm	0.77	(1385.8 nm + 617.0 nm)/ 781.4 nm	0.80
TBI3	(1486.6 nm-1324.2 nm)/ (1486.6 nm + 1105.7 nm)	0.75	(1357.8 nm-1217.7 nm)/ (1357.8 nm + 617.0 nm)	0.79
TBI4	1486.6 nm/(1105.7 nm × 1105.7 nm)	0.86	1385.8 nm/(659.4 nm × 659.4 nm)	0.83
TBI5	(1486.6 nm-1402.6 nm)/ [(1486.6 nm-1402.6 nm) - (1402.6 nm-1324.2 nm)]	0.86	(426.9 nm-961.8 nm)/ [(426.9 nm-961.8 nm) - (961.8 nm-1217.7 nm)]	0.81

Table 2
Statistical descriptions of the SOC and TN contents (g kg^{-1}).

Parameters	SOC		TN	
	Calibration set	Validation set	Calibration set	Validation set
Sample number	50	15	50	15
Minimum	5.22	13.94	0.28	0.56
Maximum	84.91	80.40	4.16	3.15
Mean	17.56	26.20	1.14	1.67
Median	16.10	18.83	1.04	1.63

0.74, bias = 0.19 g kg^{-1}) compared to the rest of methods (PLSR and RBF). This is consistent with the results of the previous studies (Hong et al., 2019; Tan et al., 2020), which demonstrated that RF could achieve stable and satisfactory predictions. However, RBF could not perform as well as it was expected for SOC prediction. The reason might be that the RBF approach could work more effectively and extract enough features from full spectral information in dealing with non-linearities in high dimension data (Kalra et al., 2005).

Regarding the RF technique, comparing the models developed from the full spectra feature bands, and spectral indices (Table 3), it can be noticed that the models from the feature bands, and spectral indices gave better results by decreasing 15 % of RMSE value in the TN predictions. Because through the feature selection process, the invalid information and noise were removed (Xiaobo et al., 2010), and the correlation between TN and reflectance values increased. Based on the feature bands and spectral indices, SOC prediction had minor improvement due to the high correlation between SOC and reflectance values. Nevertheless, it would save much computing time when applied to a large dataset.

For the PLSR models extracted from the feature bands and spectral indices, it was found that the RMSE value significantly decreased by 38 % for SOC prediction and 27 % for TN compared to the PLSR models based on the full spectra. Moreover, the LCCC value increased from 0.36 to 0.75 for SOC and from 0.30 to 0.66 for TN. The accuracy of the models developed by PLSR with feature bands and spectral indices datasets was comparable to or better than the RBF-based models. While, previous studies have demonstrated that PLSR as a linear method does not perform well, when the non-linearity appears in the dataset (Peng et al., 2014; Xu et al., 2021, 2020).These results indicated that feature selection could efficiently enhance the PLSR model's predictability and extract linear information from high dimensional data.

3.3.2. *Model interpretability using SHAP*

In order to reveal the importance of each input variable, the SHAP analysis provides a detailed interpretation (Mai et al., 2022). Hence, in Fig. 7, we highlighted the best SOC and TN predicting results from the RF approach with the associated SHAP values. We found that compared to the feature bands, most of the optimal spectral indices provided relatively remarkable contribution to both SOC and TN contents prediction. For SOC modelling process, all spectral indices contributed the most. Meanwhile, apart from the spectral indices (TBI1, TBI3, and TBI4), the wavelengths 433.7 and 1385.8 nm also ranked in the top 5

Table 3
Performance of the SOC and TN models based on full-spectra (440 bands), and feature bands + optimal spectral indices (SOC: 30 bands + 5 indices, TN: 18 bands + 5 indices).

Input datasets	ML approach	SOC				TN			
		R^2	RMSE	LCCC	bias	R^2	RMSE	LCCC	bias
Full-spectra	PLSR	0.70	14.8	0.36	5.03	0.57	0.51	0.30	0.33
	RF	**0.94**	**7.5**	**0.86**	**3.92**	**0.68**	**0.39**	**0.67**	**0.22**
	RBF	0.89	9.5	0.77	5.24	0.65	0.41	0.61	0.23
Feature bands + five optimal spectral indices	PLSR	0.95	9.2	0.75	4.22	0.75	0.37	0.66	0.25
	RF	**0.97**	**7.5**	**0.84**	**3.70**	**0.78**	**0.33**	**0.74**	**0.19**
	RBF	0.95	10.0	0.69	4.53	0.73	0.40	0.64	0.27

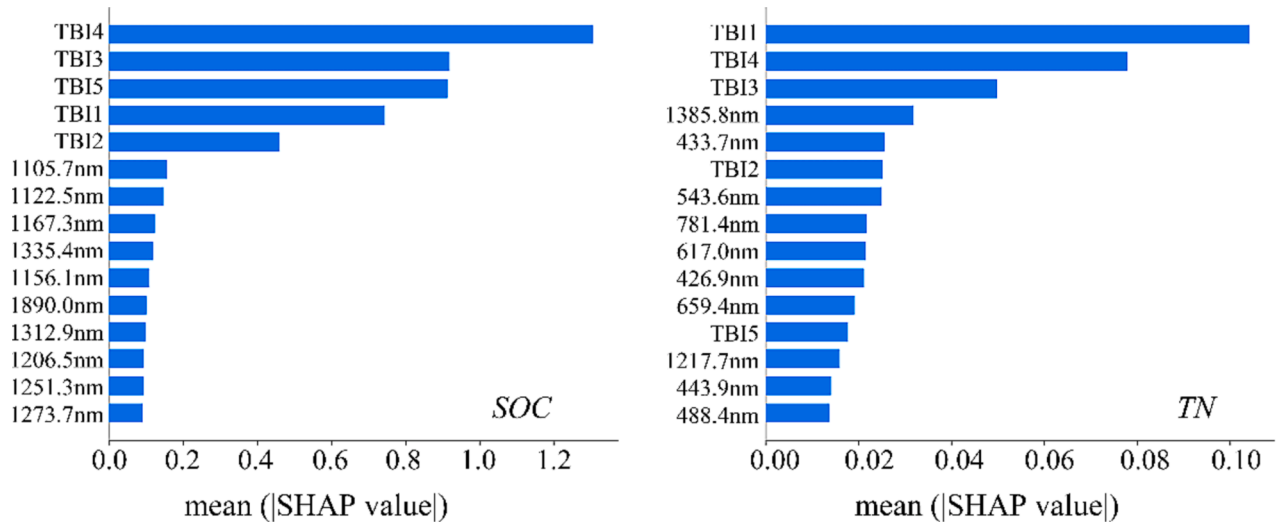


Fig. 7. The important rankings of optimal featured set of SOC and TN.

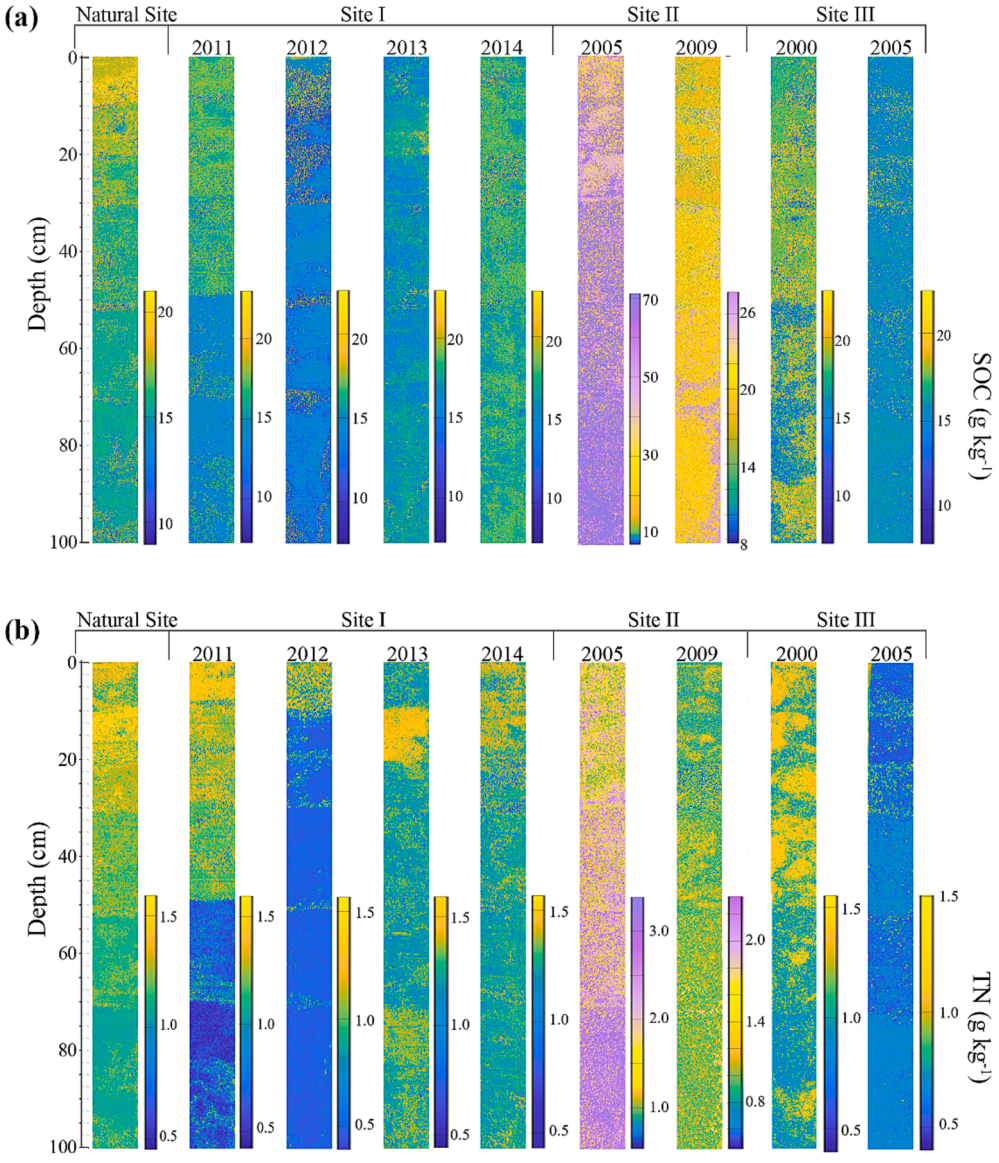


Fig. 8. Soil maps for the vertical distribution of SOC (a) and TN (b) content (g kg^{-1}).

predictors of TN content. We also found that TBI1, TBI3, and TBI4 had the highest SHAP values for SOC evaluation too. However, the TBI5 could not contribute to the TN modelling process as much as the other examined indices; which could be linked to the spectral bands for the TBI5 calculation.

3.4. Mapping vertical distribution of SOC and TN contents

The RF model with the highest accuracy based on point spectra data was finally applied to the hyperspectral imagery for the mapping purpose. The vertical distribution of SOC and TN (0–100 cm) in different soil cores were illustrated in Fig. 8a and b, respectively. It can be seen that the topsoil from the natural site had a clear decreasing trend for both SOC and TN content with increasing of the depth, which is proved by the results in Fig. 3 too. The SOC contents start to be stabilized (approx.: 16 g kg^{-1}) from the depth 20 cm, while TN contents stabilized (approx.: 1.2 g kg^{-1}) after 30 cm. We have also noticed that the litter layer of natural grassland resulted in higher SOC and TN contents in the first 20 cm than in the deeper horizons, according to Jobbágy and Jackson (2000) and Li et al. (2015).

At the site I (2011), we found a clear difference between top 50 cm and the lower part of soil core for both SOC and TN contents (Fig. 8). This could be explained by reclamation history and vegetation covers. The reclamation project at this site was conducted six years before the sampling campaign of the current study and most probably the increasing of SOC and TN contents could be positively correlated with 6 years of revegetation (Burke et al., 1990). The dominant specie is *Elymus dahuricus Turcz*, which could accelerate the SOC accumulation in the topsoil (Li et al., 2016). For the site I (2012), which was located on the dumping site with slope gradient of approximately 15–20°, the SOC and TN contents were lower in general than other sites. The topsoil (0–30 cm) was more heterogeneous than soils from the deeper horizon. This was most likely due to the effect of landform and vegetation type. The surface vegetation type of *Artemisia sieversiana Ehrhart ex Willd* was selected to be planted to prevent soil erosion and loss instead of improving soil properties (Wang et al., 2011). Conversely, the site I (2013 and 2014) were located in a relatively flat area. Therefore, they presented higher SOC contents compared to the site I (2012). The main vegetation cover in this area was *Melilotus officinalis*, which might be difficult to significantly influence SOC changes in such a short period (3–4 years) (Moyes and Bowling, 2016). On the other hand, we found that this type of vegetation had some azotification effects (Wolf et al., 2004), which resulted in higher TN contents accumulation in topsoil (0–20 cm).

For the site II, the reclamation project was conducted ten years before sampling campaign in 2017. However, the vertical distribution of SOC and TN was not as expected as the natural site. One reason might be a disordered dumping technique during the mining activities. As shown in Fig. 8, at the site II (2005), the soils with high SOC and TN contents were found in the deeper depths (70–100 cm). We assume that surface natural soils were not kept and stored at the beginning of the mining activities as requested by the guideline (Shaughnessy et al., 2022), but were directly dumped into an open-pit. In addition, this site was reclaimed without any man-made revegetation as the vegetation in these two sites was generated and recovered through natural seeding. The vegetation coverage in the natural recovery area was far lower than in the artificial vegetation coverage area (such as site I). This was because artificial planting significantly accelerated soil redevelopment process in mining lands (Singh and Singh, 2006).

The mining activities in the reclaimed site III were kept active for over 100 years and were then closed in 2012. This site was reclaimed under a 100 % natural process. At the site III (2000), the reclaimed project was carried out over 17 years before the sampling for this study. Based on the results in Fig. 8, it can be found that the SOC and TN contents in the first 50 cm were relatively higher than in the deeper horizons. It was also seen that the TN content of the topsoil was similar to the soils from the natural site (Fig. 8). The reclamation years significantly influence the SOC changes in this site after a long term period (Srivastava et al., 1989). In addition, we have also observed high SOC and TN contents after the depth 90 cm in the site III (2000); the disordered dumping way could also explain this. Nevertheless, in the site III (2005), no clear changes in the SOC and TN contents through the soil cores were observed and the site was remained low in concentration compared to the other sites even after ten years of reclamation. We believe that the quality of the mine soil will not be improved even after over ten years, if no planting and topsoil covering project is executed. Hence, it is important to monitor soil reclamation progressively in mining areas.

Generally, our results indicated that dumping strategies could significantly influence the distribution of SOC and TN through all horizons and ultimately affect the soil quality after reclamation in the mining areas, which has also been confirmed by Cao et al. (2015) and Feng et al. (2019).

In soil science, soil profiles are commonly classified as O-A-B-C horizons (Fig. 9). The O and A horizons are normally rich in case of the SOC and TN contents to support the growth of surface vegetation (Gholizadeh et al., 2022; Wang et al., 2022). Nevertheless, this study revealed the reality and consequences of disordered dumping way in targeted three reclamation sites. For example, if topsoil is not be kept and dumped in order of O-C-B-A mode, the O and A horizons might wrongly be moved to the bottom as C-B-A-O or disorder pattern that is highlighted in Fig. 9. Meanwhile, the B or C horizons with relatively low nutrients and high stone content are transferred to the upper layers, which cannot provide sufficient nutrients and water for vegetation growth. Latinopoulos (1981) emphasized that inappropriate dumping and reclamation measures can largely affect the establishment of a sustainable soil ecosystem after intensive mining activities. This study also demonstrated no significant improvement in SOC and TN contents even after over ten years of reclamation by disordered horizons. Therefore, the sequential dumping strategy is curial to reconstructing soil horizons for supporting vegetation growth and establishing a sustainable soil ecosystem (Feng et al., 2019). Meanwhile, VIS-NIR imaging spectroscopy can reveal soil quality changes in different horizons and has appeared to be a promising technique for soil quality assessment of reclaimed sites in mining areas.

4. Conclusions

This study proposed a workflow that combines point and imaging spectroscopy to estimate and map vertical distribution (0–100 cm) of SOC and TN contents in different reclaimed sites of a mining area in China in order to evaluate the quality of soil health in reclaimed mine sites.

We concluded that: 1) in total of 30 feature bands were selected and 5 spectral indices were calculated for SOC modelling, and 18 feature bands were selected and 5 spectral indices were calculated for TN modelling. 2) the model derived from the RF machine learning method with optimized bands and spectral indices yielded the best predictions for both SOC ($R^2 = 0.97$, RMSE = 7.5 g kg^{-1}, LCCC = 0.84, bias = 3.70 g kg^{-1}) and TN ($R^2 = 0.78$, RMSE = 0.33 g kg^{-1}, LCCC = 0.74, bias = 0.19 g kg^{-1}). 3) based on the SHAP values, all spectral indices contributed the most to SOC modelling process, while the spectral indices (TBI1, TBI3, and TBI4) and feature bands (433.7 nm and 1385.8 nm) made significant contributions to TN prediction. 4) based on the hyperspectral imagery technique, the vertical distributions of SOC and TN contents in the natural site were decreased gradually from the topsoil to the deeper layers. However, in the reclaimed mine soils, the changes in the vertical distribution of both SOC and TN were highly depended on the various dumping, reclaiming strategies and reclaiming years. The results suggested that mapping vertical distribution of SOC and TN could better indicate how dumping technique affects the reclaimed soil layer in the mining activities.

In future studies, imaging spectroscopy can be used to monitor other

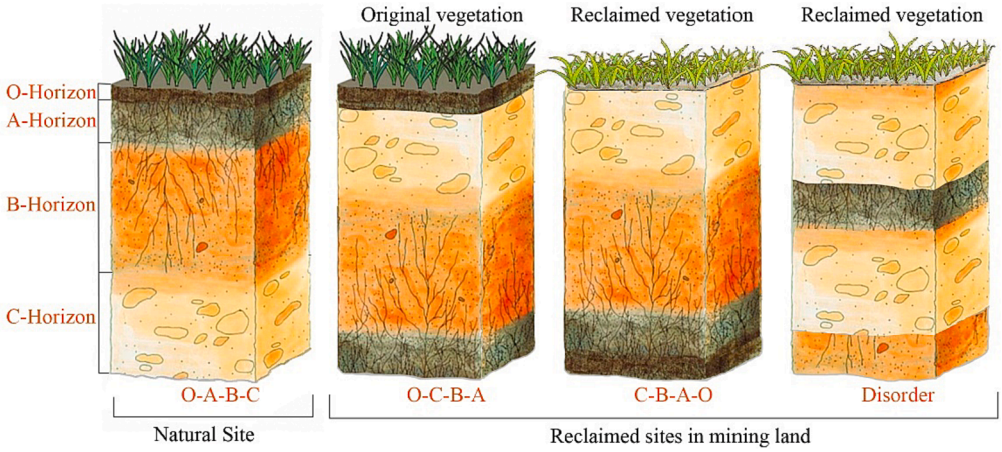


Fig. 9. Open-pit mining activities and different landfill methods in targeted three reclamation sites.

soil properties, such as soil moisture content and particle size, which are closely related to reclamation processes. In addition, the laboratory spectral-based model needs to be further verified in more reclaimed sites, and soil maps can be validated independently or map uncertainly assessment can be conducted. Successful implementation of these ideas will significantly reduce the cost of sampling and chemical analysis related to the monitoring of mine areas' reclamation.

CRediT authorship contribution statement

Sihan Peng: Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Nisha Bao:** Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Shijia Wang:** Data curation. **Asa Gholizadeh:** Writing – review & editing. **Mohammadmehdi Saberioon:** Writing – review & editing. **Yi Peng:** Writing – review & editing.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

This study was supported by the National Key R&D Program of China (2022YFC2903903), National Natural Science Foundation of China (No. 52074063), Fundamental Research Funds for the Central Universities (No. 2023GFZD03), and 2021 Liaoning Provincial People's Livelihood Science and Technology Project (No. 2021JH2/10200035). We sincerely thank all our colleagues and editors for their dedicated efforts and efficient work on this study.

References

Andrade, R., Silva, S.H.G., Faria, W.M., Poggere, G.C., Barbosa, J.Z., Guilherme, L.R.G., Curi, N., 2020. Proximal sensing applied to soil texture prediction and mapping in Brazil. Geoderma Reg. 23, e00321.

Araújo, M.C.U., Saldanha, T.C.B., Galvão, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemom. Intell. Lab. Syst. 57, 65–73. https://doi.org/10.1016/S0169-7439(01)00119-8.

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. Eur. J. Soil Sci. 65, 718–729. https://doi.org/10.1111/ejss.12165.

Atzberger, C., Guérif, M., Baret, F., Werner, W., 2010. Comparative analysis of three chemometric techniques for the spectroradiometric assessment of canopy chlorophyll content in winter wheat. Comput. Electron. Agric. 73, 165–173. https://doi.org/10.1016/j.compag.2010.05.006.

Bao, N., Wu, L., Ye, B., Yang, K., Zhou, W., 2017. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. Geoderma 288, 47–55. https://doi.org/10.1016/j.geoderma.2016.10.033.

Bartholomeus, H.M., Schaepman, M.E., Kooistra, L., Stevens, A., Hoogmoed, W.B., Spaargaren, O.S.P., 2008. Spectral reflectance based indices for soil organic carbon quantification. Geoderma 145, 28–36. https://doi.org/10.1016/j.geoderma.2008.01.010.

Ben-Dor, E., Banin, A., 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Sci. Soc. Am. J. 59, 364–372. https://doi.org/10.2136/sssaj1995.03615995005900020014x.

Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Taylor, G.R., Hill, J., Whiting, M.L., Sommer, S., 2009. Using Imaging Spectroscopy to study soil properties. Remote Sens. Environ. 113, S38–S55. https://doi.org/10.1016/j.rse.2008.09.019.

Beucher, A., Rasmussen, C.B., Moeslund, T.B., Greve, M.H., 2022. Interpretation of convolutional neural networks for acid sulfate soil classification. Front. Environ. Sci. 9, 809995 https://doi.org/10.3389/fenvs.2021.809995.

Bi, Y., Zhang, Y., Zou, H., 2018. Plant growth and their root development after inoculation of arbuscular mycorrhizal fungi in coal mine subsided areas. Int. J. Coal Sci. Technol. 5, 47–53. https://doi.org/10.1007/s40789-018-0201-x.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Burke, I.C., Schimel, D.S., Yonker, C.M., Parton, W.J., Joyce, L.A., Lauenroth, W.K., 1990. Regional modeling of grassland biogeochemistry using GIS. Landsc. Ecol. 4, 45–54. https://doi.org/10.1007/BF02573950.

Cao, Y., Wang, J., Bai, Z., Zhou, W., Zhao, Z., Ding, X., Li, Y., 2015. Differentiation and mechanisms on physical properties of reconstructed soils on open-cast mine dump of loess area. Environ. Earth Sci. 74, 6367–6380. https://doi.org/10.1007/s12665-015-4607-0.

de Santana, F.B., Otani, S.K., de Souza, A.M., Poppi, R.J., 2021. Comparison of PLS and SVM models for soil organic matter and particle size using vis-NIR spectral libraries. Geoderma Reg. 27, e00436.

Douglas, R.K., Nawar, S., Alamar, M.C., Mouazen, A.M., Coulon, F., 2018. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. Sci. Total Environ. 616–617, 147–155. https://doi.org/10.1016/j.scitotenv.2017.10.323.

Emamgholizadeh, S., Esmaeilbeiki, F., Babak, M., Zarehaghi, D., Maroufpoor, E., Rezaei, H., 2018. Estimation of the organic carbon content by the pattern recognition method. Commun. Soil Sci. Plant Anal. 49, 2143–2154. https://doi.org/10.1080/00103624.2018.1499750.

Feng, Y., Wang, J., Bai, Z., Reading, L., 2019. Effects of surface coal mining and land reclamation on soil properties: a review. Earth-Sci. Rev. 191, 12–25. https://doi.org/10.1016/j.earscirev.2019.02.015.

Galvao, R., Araujo, M., Jose, G., Pontes, M., Silva, E., Saldanha, T., 2005. A method for calibration and validation subset partitioning. Talanta 67, 736–740. https://doi.org/10.1016/j.talanta.2005.03.025.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.

Gholizadeh, A., Saberioon, M., Ben-Dor, E., Borůvka, L., 2018. Monitoring of selected soil contaminants using proximal and remote sensing techniques: background, state-of-the-art and future perspectives. Crit. Rev. Environ. Sci. Technol. 48, 243–278. https://doi.org/10.1080/10643389.2018.1447717.

Gholizadeh, A., Saberioon, M., Ben-Dor, E., Viscarra Rossel, R.A., Borůvka, L., 2020. Modelling potentially toxic elements in forest soils with vis–NIR spectra and learning

algorithms. Environ. Pollut. 267, 115574 https://doi.org/10.1016/j.envpol.2020.115574.
Gholizadeh, A., Saberioon, M., Pouladi, N., Ben-Dor, E., 2022. Quantification and depth distribution analysis of carbon to nitrogen ratio in forest soils using reflectance spectroscopy. Int. Soil Water Conserv. Res. S2095633922000454 https://doi.org/10.1016/j.iswcr.2022.06.004.
Ghorbani, M.A., Khatibi, R., Hosseini, B., Bilgili, M., 2013. Relative importance of parameters affecting wind speed prediction using artificial neural networks. Theor. Appl. Climatol. 114, 107–114. https://doi.org/10.1007/s00704-012-0821-9.
Girosi, F., Poggio, T., 1990. Networks and the best approximation property. Biol. Cybern. 63, 169–176. https://doi.org/10.1007/BF00195855.
Haghi, R.K., Pérez-Fernández, E., Robertson, A.H.J., 2021. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: a comparison of near infrared and mid-infrared spectroscopy. Geoderma 396, 115071. https://doi.org/10.1016/j.geoderma.2021.115071.
He, Y.-L., Geng, Z.-Q., Zhu, Q.-X., 2015. Data driven soft sensor development for complex chemical processes using extreme learning machine. Chem. Eng. Res. Des. 102, 1–11. https://doi.org/10.1016/j.cherd.2015.06.009.
Heller Pearlshtien, D., Ben-Dor, E., 2020. Effect of organic matter content on the spectral signature of iron oxides across the VIS–NIR spectral region in artificial mixtures: an example from a red soil from israel. Remote Sens. 12, 1960. https://doi.org/10.3390/rs12121960.
Hong, Y., Shen, R., Cheng, H., Chen, Y., Zhang, Y., Liu, Y., Zhou, M., Yu, L., Liu, Y.i., Liu, Y., 2019. Estimating lead and zinc concentrations in peri-urban agricultural soils through reflectance spectroscopy: effects of fractional-order derivative and random forest. Sci. Total Environ. 651, 1969–1982. https://doi.org/10.1016/j.scitotenv.2018.09.391.
Jiang, Q., Li, Q., Wang, X., Wu, Y., Yang, X., 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.
Jobbágy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10, 423–436. https://doi.org/10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349, 255–260. https://doi.org/10.1126/science.aaa8415.
Kalra, R., Deo, M.C., Kumar, R., Agarwal, V.K., 2005. RBF network for spatial mapping of wave heights. Mar. Struct. 18, 289–300. https://doi.org/10.1016/j.marstruc.2005.09.003.
Kawakami Harrop Galvão, R., Fernanda Pimentel, M., Cesar Ugulino Araujo, M., Yoneyama, T., Visani, V., 2001. Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry. Anal. Chim. Acta 443, 107–115. https://doi.org/10.1016/S0003-2670(01)01182-5.
Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148. https://doi.org/10.1080/00401706.1969.10490666.
Kumar, S., Maiti, S.K., Chaudhuri, S., 2015. Soil development in 2–21 years old coalmine reclaimed spoil with trees: A case study from Sonepur-Bazari opencast project, Raniganj Coalfield. India. Ecol. Eng. 84, 311–324. https://doi.org/10.1016/j.ecoleng.2015.09.043.
Latinopoulos, P., 1981. The effect of infiltration distribution on artificial recharge schemes. J. Hydrol. 49, 279–285. https://doi.org/10.1016/0022-1694(81)90218-3.
Li, Y., Li, Q., Guo, D., Liang, S., Wang, Y., 2016. Ecological stoichiometry homeostasis of Leymus chinensis in degraded grassland in western Jilin Province, NE China. Ecol. Eng. 90, 387–391. https://doi.org/10.1016/j.ecoleng.2016.01.079.
Li, W., Liu, J., Bao, N., Mao, X., Mao, Y., Fu, Y., Cao, W., Huang, J., Zhao, Z., 2021. Salinity monitoring at saline sites with visible–near-infrared spectral data. Minerals 11, 1086. https://doi.org/10.3390/min11101086.
Li, S., Shi, Z., Chen, S., Ji, W., Zhou, L., Yu, W., Webster, R., 2015. In situ measurements of organic carbon in soil profiles using vis-NIR spectroscopy on the qinghai-tibet plateau. Environ. Sci. Technol. 49, 4980–4987. https://doi.org/10.1021/es504272x.
Lin, L.-I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255. https://doi.org/10.2307/2532051.
Liu, X., Bai, Z., Zhou, W., Cao, Y., Zhang, G., 2017. Changes in soil properties in the soil profile after mining and reclamation in an opencast coal mine on the Loess Plateau. China. Ecol. Eng. 98, 228–239. https://doi.org/10.1016/j.ecoleng.2016.10.078.
Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent individualized feature attribution for tree ensembles. ArXiv Prepr. 1802, 03888. https://doi.org/10.48550/ARXIV.1802.03888.
Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Presented at the Nips, arXiv.
Mai, J., Lu, T., Xu, P., Lian, Z., Li, M., Lu, W., 2022. Predicting the maximum absorption wavelength of azo dyes using an interpretable machine learning strategy. Dyes Pigments 206, 110647. https://doi.org/10.1016/j.dyepig.2022.110647.
Martin, P.D., Malley, D.F., Manning, G., Fuller, L., 2002. Determination of soil organic carbon and nitrogen at the field level using near-infrared spectroscopy. Can. J. Soil Sci. 82, 413–422. https://doi.org/10.4141/S01-054.
Moyes, A.B., Bowling, D.R., 2016. Plant community composition and phenological stage drive soil carbon cycling along a tree-meadow ecotone. Plant Soil 401, 231–242. https://doi.org/10.1007/s11104-015-2750-8.
Ogrič, M., Knadel, M., Kristiansen, S.M., Peng, Y., De Jonge, L.W., Adhikari, K., Greve, M. H., 2019. Soil organic carbon predictions in subarctic greenland by visible–near infrared spectroscopy. Arct. Antarct. Alp. Res. 51, 490–505. https://doi.org/10.1080/15230430.2019.1679939.
Padarian, J., McBratney, A.B., Minasny, B., 2020. Game theory interpretation of digital soil mapping convolutional neural networks. SOIL 6, 389–397. https://doi.org/10.5194/soil-6-389-2020.
Peng, X., Shi, T., Song, A., Chen, Y., Gao, W., 2014. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. Remote Sens. 6, 2699–2717. https://doi.org/10.3390/rs6042699.
Sarkhot, D.V., Grunwald, S., Ge, Y., Morgan, C.L.S., 2011. Comparison and detection of total and available soil carbon fractions using visible/near infrared diffuse reflectance spectroscopy. Geoderma 164, 22–32. https://doi.org/10.1016/j.geoderma.2011.05.006.
Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047.
Scheinost, A.C., 1998. Use and limitations of second-derivative diffuse reflectance spectroscopy in the visible to near-infrared range to identify and quantify fe oxide minerals in soils. Clays Clay Miner. 46, 528–536. https://doi.org/10.1346/CCMN.1998.0460506.
Shaughnessy, B.E., Dhar, A., Naeth, M.A., 2022. Natural recovery of vegetation on reclamation stockpiles after 26 to 34 years. Écoscience 29, 55–67. https://doi.org/10.1080/11956860.2021.1943931.
Singh, A.N., Singh, J.S., 2006. Experiments on ecological restoration of coal mine spoil using native trees in a dry tropical environment, india: a synthesis. New for. 31, 25–39. https://doi.org/10.1007/s11056-004-6795-4.
Sorenson, P.T., Quideau, S.A., Rivard, B., 2018. High resolution measurement of soil organic carbon and total nitrogen with laboratory imaging spectroscopy. Geoderma 315, 170–177. https://doi.org/10.1016/j.geoderma.2017.11.032.
Sorenson, P.T., Quideau, S.A., Rivard, B., Dyck, M., 2020. Distribution mapping of soil profile carbon and nitrogen with laboratory imaging spectroscopy. Geoderma 359, 113982. https://doi.org/10.1016/j.geoderma.2019.113982.
Šourková, M., Frouz, J., Šantrůčková, H., 2005. Accumulation of carbon, nitrogen and phosphorus during soil formation on alder spoil heaps after brown-coal mining, near Sokolov (Czech Republic). Geoderma 124, 203–214. https://doi.org/10.1016/j.geoderma.2004.05.001.
Srivastava, S.C., Jha, A.K., Singh, J.S., 1989. Changes with time in soil biomass c, n and p of mine spoils in a dry tropical environment. Can. J. Soil Sci. 69, 849–855. https://doi.org/10.4141/cjss89-085.
Standard for quality control of land reclamation, 2013. , Land Management Industry Standard of the People's Republic of China.
Steffens, M., Buddenbaum, H., 2013. Laboratory imaging spectroscopy of a stagnic Luvisol profile — high resolution soil characterisation, classification and mapping of elemental concentrations. Geoderma 195–196, 122–132. https://doi.org/10.1016/j.geoderma.2012.11.011.
Steffens, M., Zeh, L., Rogge, D.M., Buddenbaum, H., 2021. Quantitative mapping and spectroscopic characterization of particulate organic matter fractions in soil profiles with imaging VisNIR spectroscopy. Sci. Rep. 11, 16725. https://doi.org/10.1038/s41598-021-95298-8.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and Near Infrared Spectroscopy in Soil Science, in: Advances in Agronomy. Elsevier, pp. 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.
Stenberg, B., Jonsson, A., Börjesson, T., 2005. Use of near infrared reflectance spectroscopy to predict nitrogen uptake by winter wheat within fields with high variability in organic matter. Plant Soil 269, 251–258. https://doi.org/10.1007/s11104-004-0556-1.
Tan, K., Wang, H., Chen, L., Du, Q., Du, P., Pan, C., 2020. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. J. Hazard. Mater. 382, 120987 https://doi.org/10.1016/j.jhazmat.2019.120987.
Tavakoli, H., Correa, J., Sabetizade, M., Vogel, S., 2023. Predicting key soil properties from Vis-NIR spectra by applying dual-wavelength indices transformations and stacking machine learning approaches. Soil Tillage Res. 229, 105684 https://doi.org/10.1016/j.still.2023.105684.
Thissen, U., Pepers, M., Üstün, B., Melssen, W.J., Buydens, L.M.C., 2004. Comparing support vector machines to PLS for spectral regression applications. Chemom. Intell. Lab. Syst. 73, 169–179. https://doi.org/10.1016/j.chemolab.2004.01.002.
Tian, Y.C., Yao, X., Yang, J., Cao, W.X., Hannaway, D.B., Zhu, Y., 2011. Assessing newly developed and published vegetation indices for estimating rice leaf nitrogen concentration with ground- and space-based hyperspectral reflectance. Field Crops Res. 120, 299–310. https://doi.org/10.1016/j.fcr.2010.11.002.
Upadhyay, N., Verma, S., Pratap Singh, A., Devi, S., Vishwakarma, K., Kumar, N., Pandey, A., Dubey, K., Mishra, R., Kumar Tripathi, D., Rani, R., Sharma, S., 2016. Soil ecophysiological and microbiological indices of soil health: a study of coal mining site in sonbhadra, uttar Pradesh. J. Soil Sci. Plant Nutr. 0–0 https://doi.org/10.4067/S0718-95162016005000056.
Ussiri, D.A.N., Lal, R., 2005. Carbon Sequestration in Reclaimed Minesoils. Crit. Rev. Plant Sci. 24, 151–165. https://doi.org/10.1080/07352680591002147.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.
Vitti, C., Stellacci, A.M., Leogrande, R., Mastrangelo, M., Cazzato, E., Ventrella, D., 2016. Assessment of organic carbon in soils: a comparison between the Springer-Klee wet digestion and the dry combustion methods in Mediterranean soils (Southern Italy). CATENA 137, 113–119. https://doi.org/10.1016/j.catena.2015.09.001.
Vohland, M., Emmerling, C., 2011. Determination of total soil organic C and hot water-extractable C from VIS-NIR soil reflectance with partial least squares regression and

spectral feature selection techniques. Eur. J. Soil Sci. 62, 598–606. https://doi.org/10.1111/j.1365-2389.2011.01369.x.
Wadoux, A.M.J.-C., Malone, B., Minasny, B., Fajardo, M., McBratney, A.B., 2021. Soil Spectral Inference with R: Analysing Digital Soil Spectra using the R Programming Environment, Progress in Soil Science. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-64896-1.
Wang, S., Cao, Y., Pietrzykowski, M., Zhou, W., Bai, Z., 2022. Research on the influence of vegetation restoration in loess open-pit coal mines of China: influencing factors and mechanism. Ecol. Eng. 177, 106549 https://doi.org/10.1016/j.ecoleng.2022.106549.
Wang, J., Jiao, Z., Bai, Z., 2014. Changes in carbon sink value based on RS and GIS in the Heidaigou opencast coal mine. Environ. Earth Sci. 71, 863–871. https://doi.org/10.1007/s12665-013-2488-7.
Wang, J., Ding, J., Yu, D., Ma, X., Zhang, Z., Ge, X., Teng, D., Li, X., Liang, J., Lizaga, I., Chen, X., Yuan, L., Guo, Y., 2019. Capability of Sentinel-2 MSI data for monitoring and mapping of soil salinity in dry and wet seasons in the Ebinur Lake region, Xinjiang, China. Geoderma 353, 172–187. https://doi.org/10.1016/j.geoderma.2019.06.040.
Wang, N., Jiao, J.-Y., Jia, Y.-F., Wang, D.-L., 2011. Seed persistence in the soil on eroded slopes in the hilly-gullied Loess Plateau region. China. Seed Sci. Res. 21, 295–304. https://doi.org/10.1017/S0960258511000195.
Wang, X., Zhang, F., Ding, J., Kung, H., Latif, A., Johnson, V.C., 2018b. Estimation of soil salt content (SSC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR), Northwest China, based on a Bootstrap-BP neural network model and optimal spectral indices. Sci. Total Environ. 615, 918–930. https://doi.org/10.1016/j.scitotenv.2017.10.025.
Wang, S., Zhuang, Q., Jia, S., Jin, X., Wang, Q., 2018a. Spatial variations of soil organic carbon stocks in a coastal hilly area of China. Geoderma 314, 8–19. https://doi.org/10.1016/j.geoderma.2017.10.052.
Wei, L., Yuan, Z., Wang, Z., Zhao, L., Zhang, Y., Lu, X., Cao, L., 2020. Hyperspectral inversion of soil organic matter content based on a combined spectral index model. Sensors 20, 2777. https://doi.org/10.3390/s20102777.
Wolf, J.J., Beatty, S.W., Seastedt, T.R., 2004. Soil characteristics of rocky mountain national park grasslands invaded by melilotus officinalis and M. Alba: montane grassland soil invaded by melilotus. J. Biogeogr. 31, 415–424. https://doi.org/10.1046/j.0305-0270.2003.00983.x.
Wu, S., Wang, C., Liu, Y., Li, Y., Liu, J., Xu, A., Pan, K., Li, Y., Pan, X., 2018. Mapping the salt content in soil profiles using Vis-NIR hyperspectral imaging. Soil Sci. Soc. Am. J. 82, 1259–1269. https://doi.org/10.2136/sssaj2018.02.0074.
Xiaobo, Z., Jiewen, Z., Povey, M.J.W., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667, 14–32. https://doi.org/10.1016/j.aca.2010.03.048.
Xu, S., Wang, M., Shi, X., 2020. Hyperspectral imaging for high-resolution mapping of soil carbon fractions in intact paddy soil profiles with multivariate techniques and variable selection. Geoderma 370, 114358. https://doi.org/10.1016/j.geoderma.2020.114358.
Xu, S., Wang, M., Shi, X., Yu, Q., Zhang, Z., 2021. Integrating hyperspectral imaging with machine learning techniques for the high-resolution mapping of soil nitrogen fractions in soil profiles. Sci. Total Environ. 754, 142135 https://doi.org/10.1016/j.scitotenv.2020.142135.
Yang, T., Li, P., Wu, X., Hou, X., Liu, P., Yao, G., 2014. Assessment of vulnerability to climate change in the Inner Mongolia steppe at a county scale from 1980 to 2009. Rangel. J. 36, 545. https://doi.org/10.1071/RJ14011.
Zhang, Z., Ding, J., Wang, J., Ge, X., 2020a. Prediction of soil organic matter in northwestern China using fractional-order derivative spectroscopy and modified normalized difference indices. CATENA 185, 104257. https://doi.org/10.1016/j.catena.2019.104257.
Zhang, Z., Ding, J., Zhu, C., Wang, J., 2020b. Combination of efficient signal pre-processing and optimal band combination algorithm to predict soil organic matter through visible and near-infrared spectra. Spectrochim. Acta. a. Mol. Biomol. Spectrosc. 240, 118553 https://doi.org/10.1016/j.saa.2020.118553.
Zhang, Z., Ding, J., Zhu, C., Wang, J., Ma, G., Ge, X., Li, Z., Han, L., 2021. 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 382, 114729. https://doi.org/10.1016/j.geoderma.2020.114729.
Zhang, B., Guo, B., Zou, B., Wei, W., Lei, Y., Li, T., 2022. Retrieving soil heavy metals concentrations based on GaoFen-5 hyperspectral satellite image at an opencast coal mine, Inner Mongolia. China. Environ. Pollut. 300, 118981 https://doi.org/10.1016/j.envpol.2022.118981.
Zhu, C., Zhang, Z., Wang, H., Wang, J., Yang, S., 2020. Assessing soil organic matter content in a coal mining area through spectral variables of different numbers of dimensions. Sensors 20, 1795. https://doi.org/10.3390/s20061795.