Science of the Total Environment 813 (2022) 151882

Contents lists available at ScienceDirect

Science of the Total Environment

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

Development of a soil heavy metal estimation method based on a spectral index: Combining fractional-order derivative pretreatment and the absorption mechanism

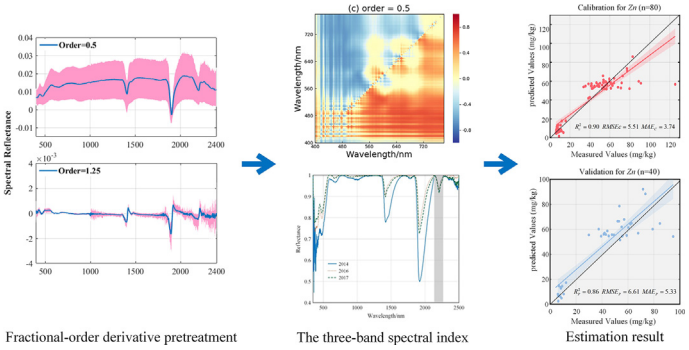
Lihan Chen [a], Jian Lai [b], Kun Tan [c,d,e,*], Xue Wang [c,d,e], Yu Chen [a], Jianwei Ding [f]

[a] Key Laboratory of Land Environment and Disaster Monitoring of MNR, China University of Mining and Technology, Xuzhou 221116, China
[b] Shanghai Institute of Satellite Engineering, Shanghai 200240, China
[c] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
[d] Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources, East China Normal University, Shanghai 200241, China
[e] School of Geographic Sciences, East China Normal University, Shanghai 200241, China
[f] The Second Surveying and Mapping Institute of Hebei, Shijiazhuang 050037, China

HIGHLIGHTS

- The potential of spectral index to estimate soil heavy metal was explored.
- Fractional-order derivate pretreatment has strong ability to amplify the feature of reflectance.
- The absorption mechanism of soil spectra was analyzed.
- The model of extreme learning machine shows a superior performance in soil heavy metal estimation.

GRAPHICAL ABSTRACT



Fractional-order derivative pretreatment | The three-band spectral index | Estimation result

ARTICLE INFO

Article history:
Received 31 August 2021
Received in revised form 26 October 2021
Accepted 18 November 2021
Available online 23 November 2021

Editor: Deyi Hou

Keywords:
Soil heavy metal
Fractional-order derivative pretreatment
Absorption mechanism
Spectral index

ABSTRACT

Visible and near-infrared (Vis–NIR) reflectance is an effective way to estimate soil heavy metal content. In this study, in order to magnify the spectral information of the soil heavy metals and solve the collinearity and redundancy of hyperspectral datasets, we aimed to explore the potential of the fractional-order derivative (FOD) spectral pretreatment method and the band combination algorithm in soil heavy metal estimation. A total of 120 soil samples were collected in Xuzhou city, Jiangsu province, China, and their heavy metal contents and spectra were measured. The FOD (intervals of 0.25, range of 0–2) and a new three-band spectral index which take into account the electronic transition of metal ions in the visible region and organic matter and clay minerals in the near-infrared region were utilized for the spectral pretreatment and the selection of characteristic bands, respectively. FOD with an order of 0.75 exhibited the best model performance for estimating *Cr* and *Zn*, yielding R_p^2 values of 0.74 and 0.81, respectively. As regards *Pb*, the highest estimation accuracy was achieved with the 0.5-order reflectance, yielding R_p^2 values of 0.56. The three-band spectral indices with the best performance were then combined for a better estimation. To improve the estimation accuracy and generalization, partial least squares (PLS), support vector machine (SVM), random forest (RF), ridge regression (RR), XGBoost and extreme learning machine (ELM) were used to estimate the heavy metals by incorporating

* Corresponding author at: School of Geographic Sciences, East China Normal University, Shanghai 200241, China.
E-mail address: tankuncu@gmail.com (K. Tan).

http://dx.doi.org/10.1016/j.scitotenv.2021.151882

multiple spectral indices, and it was found that ELM outperformed other counterparts (the highest $R_P^2 = 0.77$ for *Cr*, the highest $R_P^2 = 0.86$ for *Zn*, the highest $R_P^2 = 0.63$ for *Pb*). The main spectral absorption mechanisms and modes of heavy metals were also analyzed. This estimation method combining FOD and a three-band index will provide a reference to estimate soil heavy metals using Vis–NIR spectra over a large scale.

1. Introduction

Soil, as an important part of the terrestrial ecosystem, continuously circulates and transforms materials and energy with other components in the ecosystem. With the rapid development of industry and agriculture and the huge increase in population, environmental pollution, including soil heavy metal pollution, is an increasingly important problem worldwide (Salazar et al., 2012). The heavy metals in soil cannot be decomposed. Furthermore, some heavy metals can accumulate in the human body at harmful concentrations through the food chain, which is a serious hazard to human health (Li et al., 2012; Khan et al., 2008; Xia et al., 2020). For example, if a human ingests or inhales an excessive amount of a heavy metal, this can affect the normal function and growth of cells, and can cause a series of lesions in various organs of the body. Zinc (*Zn*) usually exists in a divalent form in soil, and is an essential micronutrient element for the growth and development of animals and plants. However, the excessive accumulation of *Zn* in soil can reduce soil microbial activity and inhibit crop growth, which results in the enrichment of *Zn* in crops (He et al., 2020). *Zn* can also further harm human health by polluting surface and groundwater bodies. The traditional methods for obtaining soil heavy metal concentration are mainly based on laboratory chemical analysis, which is labor-intensive, time-consuming, environmentally unfriendly, and requires a high level of expertise (Lassalle et al., 2020; Meng et al., 2020). Moreover, different analysis and detection methods are utilized for the various types of heavy metals, and the chemical analysis reagents can themselves be harmful to the environment. Therefore, it is difficult to achieve large-scale, dynamic, and rapid soil heavy metal content estimation. In recent years, the relationship between reflectance within the visible to near-infrared (Vis–NIR) region (380–2500 nm) and soil has been widely used as an inexpensive and rapid estimation tool for soil water content (Zhang et al., 2020a), organic carbon (Viscarra Rossel and Behrens, 2010), and heavy metals (Wu et al., 2007a; Tan et al., 2020a; Tan et al., 2020b).

When obtaining sample spectral information, hyperspectral data can be affected by the specific state of the sample (such as humidity and particle size), as well as the experimental conditions (such as instruments and operations), which can affect the spectral quality, to varying degrees. Due to the low content of heavy metals in soil, the spectral characteristics are not obvious. Therefore, it is necessary to preprocess the spectra to obtain spectra that can reflect the true properties of the sample and enhance the effective spectral information of heavy metals. The common preprocessing methods include Savitzky-Golay (SG) smoothing (Asadzadeh and de Souza Filho, 2016), derivative preprocessing, standard normal variate (SNV) preprocessing (Fearn et al., 2009), multiplicative scatter correction (MSC), and continuum removal (CR). The use of derivative spectra is an established technique in analytical chemistry for the elimination of background signals and for resolving overlapping spectral features. Derivative spectra also help to eliminate the differences in reflectance caused by the differences in particle size of the material. Nevertheless, a bottleneck of the conventional integer-order derivatives (i.e. the first- and second-order derivatives) is a lack of sensitivity to the gradual tilts or curvatures that may contain beneficial information regarding soil heavy metals (Hong et al., 2018). However, the fractional-order derivative (FOD), as an extension of integer-order derivatives, is of increasing importance in many fields (such as control systems, signal filtering, bioengineering, and image processing) (Lu and Jin, 2011; Tarasov, 2016; Baderia et al., 2015; Zhang, 2011; Hong, 2018), but few studies have explored its potential in soil heavy metal estimation through Vis–NIR spectroscopy. Weak overtones and combinations of these fundamental vibrations due to the stretching and bending of NH, OH, and CH groups dominate the NIR (700–2500 nm) region and electronic transitions dominate the Vis (400–700 nm) portion of the spectra (Mohamed et al., 2018). The overtones and combination modes make qualitative and quantitative interpretation in the Vis–NIR region more difficult (Rossel et al., 2006). Therefore, the FOD, with its strong ability to amplify the feature of reflectance and remove the mixed overlapping peaks, can be utilized to estimate soil heavy metals.

Due to the numerous bands of hyperspectral reflectance, it is necessary to select characteristic bands to simplify the model, shorten the running time, and improve the generalization ability of the model in soil heavy metal estimation. However, soil and soil spectra are rather complex phenomena, which prevents the straightforward prediction of reflectance properties by physical theories or models. Furthermore, the characteristic bands have usually been selected with statistical methods in most of the previous studies, and it is difficult to analyze how these bands influence the soil heavy metals (Tan et al., 2018; Yuan et al., 2020). Band combination algorithm (i.e. spectral index) is a explainable and effective way to analyze soil properties, and have been widely utilized in the development of hyperspectral techniques (Bao et al., 2017; Bartholomeus et al., 2008). Spectral indices can be obtained by mathematically transforming the reflectance values of two or more characteristic bands. Spectral indices can express the hyperspectral response characteristics of heavy metals from two-dimensional or even multi-dimensional spectral spaces, reducing the impact of other soil composition information on the estimation. They also have the ability to amplify the weak correlation between bands, reduce the complexity of the model, and remove redundant information variables. Therefore, using spectral indices to quantitatively express the interaction relationship between the characteristic bands of heavy metals can effectively improve the accuracy of heavy metal estimation models. Spectral indices of the optimal band combination algorithm, including different combinations (e.g. sum, difference, and ratio), have been utilized to explore the relationship between soil reflectance and soil components with hyperspectral data (Hong et al., 2020; Wang et al., 2018). Compared to two-band spectral indices, three-band spectral indices contain an additional band in a specific area through mathematical operations, which are more robust and stable for soil components estimation (Zhang et al., 2019). However, three-band spectral indices, as an improvement of two-band spectral indices, have been less studied. In this paper, we introduce a new three-band spectral index with band combination algorithm and compare the estimation ability with that of traditional two-band spectral index.

According to several studies about the mechanism for the estimation of heavy metal concentrations in soil by Vis–NIR reflectance spectroscopy, there are two aspects playing important roles in the estimation: (i) the absorption of soil over the Vis–NIR spectral region (350–2500 nm) is primarily associated with Fe-oxides, clay minerals, water, and organic matter, as a consequence of the vibrational energy transitions of these dominant molecular bonds (Shi et al., 2014). Metal cations (M^{2+}) adsorbed onto hydroxylated surface sites (*R*OH, in which *R* can be *Al*, *Fe*, *Mn*, *Si*, etc. upon mineral surfaces) are generally described as follows: $ROH + M^{2+} = RO\text{-}M^{+} + H^{+}$. Consequently, an increase in the metal cations results in a decrease in *R*OH and an increase in *R*O (e.g. FeO) on the surfaces of clay and oxide minerals (Choe et al., 2008). Moreover, heavy metals can be bound to soil organic matter due to the metal complexation resulting from the overtones and combination bands of the NH, CH, and CO groups (Wu et al., 2019). (ii) Some transition metals, such as *Ni*, *Cr*, and *Co*, have an unfilled d shell. The energy levels of d-orbitals split when the atom of a transition element is located in a crystal field. Electromagnetic energy is absorbed when an electron moves from a lower level into a higher one (Wu et al., 2007a), which causes the absorption features of heavy metals in spectral reflectance.

However, to the best of our knowledge, few studies have explored the potential of soil heavy metal estimation directly with characteristic spectral indices. Therefore, the objectives of the present study were: (1) to analyze the influence of FOD on soil spectra; (2) to introduce a new three-band spectral index with the band combination algorithm and investigate the optimal band combination with different fractional orders; (3) to study the absorption mechanism and characteristic bands of soil heavy metals; and (4) to investigate the estimation results for heavy metals using both a single spectral index and multiple spectral indices.

2. Study area and materials preparation

2.1. Study area and experimental design

The study area is the remote sensing experimental site (34°13′N, 117°08′E) near the north gate of Nanhu Campus of the China University of Mining and Technology, Xuzhou city, Jiangsu province, China (Fig. 1). The soil type in the study area is cinnamon soil. The climate of Xuzhou is a warm temperate semi-humid monsoon climate, with four distinct seasons and abundant sunshine. The experimental site has average annual temperature and precipitation of 14 °C and 847 mm, respectively. The precipitation of the rainy season accounts for 56% of the whole year.

This experiment focused on analyzing the impact of heavy metals on soil, so the small experiment field was selected to ensure that soil properties including soil spectra were mainly affected by heavy metal concentration added. Four plots were studied in the experimental site, three of which were artificially enriched with *Cr*, *Pb* and *Zn*, and the last plot had nothing added. The plots were approximately 12.0 × 11.8 m^2. The sampling points at the control plot and *Cr* addition plot adopted the plum blossom pattern. There were two closed dry wells located at the *Pb* addition plot and *Zn* addition plot, respectively. In order to reduce the influence of closed dry wells, the sampling points at these two plots were arranged in a chessboard pattern and in a S shaped path, respectively. A plexiglass column with a diameter of 20 cm and a height of 20 cm (bottom sealed) was vertically inserted into each soil point. Meanwhile, the upper end of the plexiglass column was set flush with the surface soil. The heavy metal compounds concentrations added were based on the second level standard (level II) of the national environmental quality standards for soils (GB15618-1995) (Agency, 1995), in which level II is used to the threshold values for protecting human health. In October 2013, 30.8 g of chromium (III) chloride hexahydrate, 8.3 g of lead acetate trihydrate and 16.5 g of zinc sulfate heptahydrate in nylon bags (size: 10 × 15 cm^2; density: 100 mesh; material: polyethylene) were added to the plexiglass column bottoms of the plots. After this date, no further heavy metals were artificially added. In October of 2013, 2015, and 2016, 13–15 winter wheat seeds of the same type were sown in each plexiglass column. In July of 2014, 2016, and 2017, after the wheat had matured, about 1 kg of surface soil was collected from each plexiglass column. In total, 40 samples were collected each year and 120 samples were collected over the three years. The soil samples were sealed, marked, and brought back to the laboratory.

2.2. Soil sample analysis and reflectance measurements

In the laboratory, the sundries in the soil samples, such as stones, leaves, and roots, were removed. The soil samples were then dried, ground, and passed through a 100-mesh nylon sieve. After drying, grinding, and sieving, each sample was divided into two subsamples. One sample was used for the spectral measurements and the other was analyzed for the heavy metal concentrations.

The reflectance was measured by an Analytical Spectral Devices (ASD) spectrometer that covers the Vis–NIR spectral region (350–2500 nm). This was conducted in a dark room to avoid any interference by stray

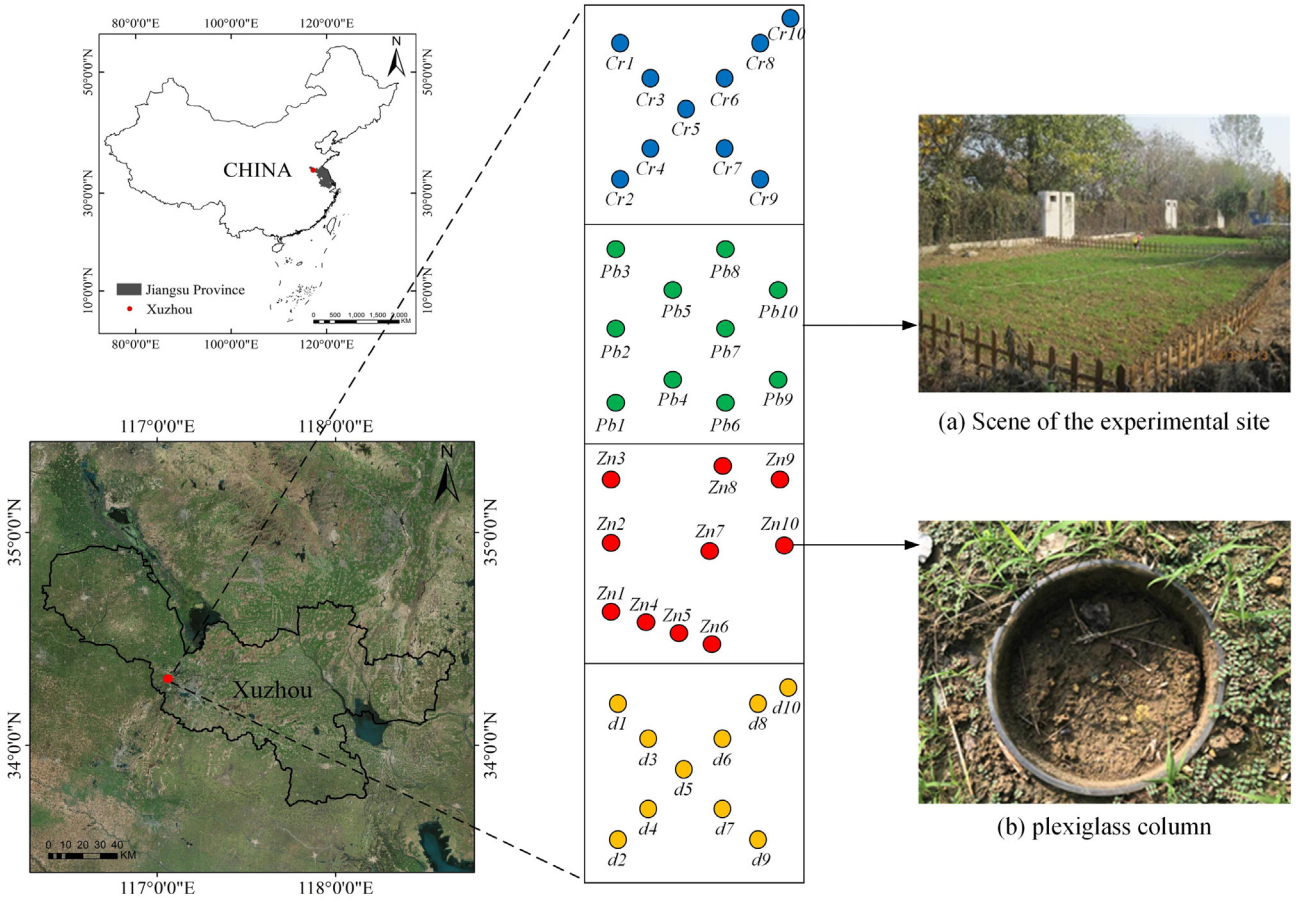


Fig. 1. Location of the study area and sampling points.

light. The sampling interval of the spectrometer in the 350–1000 nm range is 1.4 nm, and in the range of 1000–2500 nm, the sampling interval is 2 nm. The resampling interval in the full-band range is 1 nm. The sensor probe was positioned perpendicular to the sample surface at a distance of 20 cm. For each sample, 10 spectral measurements were taken, the anomalous spectra were removed, and the results were averaged to present the spectral characteristic of the sample. Because of the low signal-to-noise levels near 350 nm and 2500 nm, only the 400–2400 nm wavelength range was used. To reduce the noise, the spectra were smoothed using an SG smoothing algorithm.

The heavy metal of the soil samples was detected by inductively coupled plasma-mass spectrometry (ICP-MS). Soil samples were added to a precleaned digestion flask. A solution of HNO_3 and HCL with a ratio of 1:1 was then poured onto the samples. The soil samples were then heated using an oven, before the samples were removed and left to cool down to room temperature. After cooling down, the samples were diluted using pure deionized water and were placed on a hot plate until they evaporated to nearly a dry state. The samples were then left to cool down and were diluted again using deionized water. Finally, after filtering, the heavy metal concentrations of the soil samples were measured by ICP-MS. A basic statistical analysis of the soil heavy metal contents in three years is provided in Table 1. The CV can reflect the degree of dispersion of the data. This table shows that the three heavy metals have the similar CV.

In three years, the soil pH values of the experimental sites were 7.83, 7.79 and 7.61. The soil pH was higher than 7.5.

3. Methods

3.1. Fractional-order derivative (FOD)

The FOD extends the concept of integer-order derivatives, and is a field devoted to the study of the properties and applications of arbitrary-order derivatives. During the last few decades, the FOD has increasingly attracted the attention of researchers in many different fields. There are three main types of FOD algorithms: Riemann-Liouville (R-L), Grünwald-Letnikov (G-L), and Caputo (Benkhettou et al., 2015). The definition of G-L is relatively simple, and was applied in our research (Hong et al., 2018).

Generally, the first derivative of function $f(x)$ is defined as:

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \tag{1}$$

where h is the increment of the independent variable x. The second derivative of function $f(x)$ can then be defined as:

$$f''(x) = \lim_{h\to 0} \frac{f(x+2h)-2f(x+h)+f(x)}{h^2} \tag{2}$$

If the integer order is increased to a higher order (v) and simultaneously extended to a non-integer order, we can obtain the v-order fractional derivative formula in the interval of [a, b] (G-L):

$$d^v f(x) = \lim_{h\to 0} \frac{1}{h^v} \sum_{m=0}^{[(b-a)/h]} (-1)^m \frac{\Gamma(v+1)}{m!\Gamma(v-m+1)} f(x-mh) \tag{3}$$

Table 1
Statistical results for the soil heavy metal contents in three years.

Heavy metal	Max (mg/kg)	Min (mg/kg)	Mean (mg/kg)	Std. (mg/kg)	CV. (%)
Cr	125.50	5.90	40.70	26.31	64.64
Zn	78.73	6.30	30.82	17.08	55.42
Pb	44.82	1.60	12.74	7.90	62.01

Table 2
Band absorption of the soil components and their overtones and combinations in the Vis–NIR reflectance region (ν represents stretching vibration and δ represents bending vibration).

Band (nm)	Soil component	Mode
404	Hematite (Viscarra Rossel and Behrens, 2010)	Electronic transition
409	Ferrihydrite, goethite (Scheinost, 1998)	
427, 434	Goethite (Viscarra Rossel and Behrens, 2010; Scheinost, 1998)	
444	Hematite (Viscarra Rossel and Behrens, 2010)	
470	Fe^{3+}, ferric oxide (Wickersheim and Lefever, 1962)	
480	Goethite (Sherman and Waite, 1985)	
490	Goethite, hematite (Viscarra Rossel and Behrens, 2010; Sherman and Waite, 1985)	
484–499	Ferrihydrite (Scheinost, 1998)	
510, 529, 531	Hematite (Viscarra Rossel and Behrens, 2010; Sherman and Waite, 1985)	
620	Goethite, hematite (Stenberg et al., 2010)	
570–700	Organic matter (Galvão and Vitorello, 1998)	ν
716	Ferrihydrite (Stenberg et al., 2010)	Electronic transition
751	Amine (NH) (Clark, 1999; Clark et al., 1990a)	ν
825	CH (Clark, 1999; Clark et al., 1990a)	ν
1400	Molecular H_2O, OH, AlOH, or MgOH (Hunt, 1977)	ν
1900	OH, molecular H_2O (Srasra et al., 1994)	ν
2200	ALOH and OH, organic matter (Clark et al., 1990b)	$\nu+\delta$

where h is the step length, which is set to 1, and [(b − a)/h] is the integer part of (b − a)/h. The Gamma function is characterized by:

$$\Gamma(z) = \int_0^\infty \exp(-u)u^{z-1}du = (z-1)! \tag{4}$$

Eq. (3) can then be converted to:

$$\frac{d^v f(x)}{dx^v} \approx f(x) + (-v)f(x-1) + \frac{(-v)(-v+1)}{2} f(x-2) + \cdots + \frac{\Gamma(-v+1)}{m!\Gamma(-v+m+1)} f(x-m) \tag{5}$$

In this study, v was allowed to vary from 0 to 2 (incremented by 0.25 at each step).

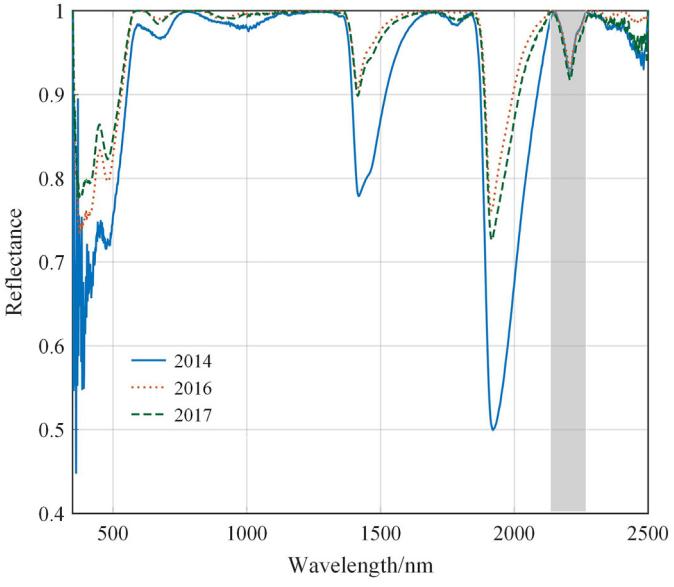


Fig. 2. Average spectra for 2014, 2016, and 2017 after continuum removal.

3.2. Band combination strategy

3.2.1. Absorption mechanism analysis

In the visible range, the main process by which molecules absorb energy is electronic transitions in atoms from the ground to higher energy states. In the NIR region, the absorption characteristics of the soil spectra are mainly generated by the stretching vibration and bending vibration of the functional groups. Table 2 presents a summary of the important fundamental absorptions in the Vis–NIR region and the occurrence of their overtones and combinations, which can be used to assist the interpretation (Viscarra Rossel and Behrens, 2010; Knadel et al., 2013; Nayak and Singh, 2007).

3.2.2. Band combination algorithm

In recent years, many researchers have devoted themselves to estimating the soil components with spectral indices based on the correlation coefficient (Zhang et al., 2019; Wang et al., 2018). The two-band spectral index is beneficial for visualizing the external response and internal meaning of spectra (Zhang et al., 2020b; Wang et al., 2019). The ratio index can also amplify the weak correlation between bands, as shown in Eq. (6).

$$\mathrm{RSI} = \frac{R_{\lambda 1}}{R_{\lambda 2}} \tag{6}$$

where $R_{\lambda 1}$ and $R_{\lambda 2}$ represent spectral bands λ_1 and λ_2, respectively, in the range of 400–780 nm.

In addition, we attempt to propose a three-band spectral index to improve the estimation accuracy of the spectral index and enhance the anti-interference ability. When selecting the characteristic bands, we adopted the following strategies. First of all, the soil samples spectra for the three years were processed by continuum removal (CR), as shown in Fig. 2, where it can be seen that there are several obvious absorption bands, namely, 480, 730, 1400, 1900 and 2200 nm. Except for the deepest water absorption bands of 1400 and 1900 nm, the 2200 nm band was the clearest. Meanwhile, in the NIR range, the absorption characteristics of soil spectra are mainly due to the stretching vibration and bending vibration of the functional groups of organic matter and clay minerals (Minasny et al., 2011; Wang et al., 2020). The band of 2200 nm has been proved to be the feature band which is associated with organic matter and clay minerals in the NIR range (Viscarra Rossel and Behrens, 2010; Choe et al., 2008; Chittleborough et al., 2011). It has demonstrated the feasibility of estimating heavy metal concentrations with hyperspectral imageries from spectral absorption features parameters of 2200 nm, so 2200 nm can also be used in remote sensing data estimation (Choe et al., 2008). The band of 2200 nm was selected as one band of the three-band spectral index.

A lot of studies have demonstrated that, in the Vis range, the soil absorption characteristics are primarily on account of the electronic transition of metal ions from the ground to higher energy states. Therefore, the other two bands were selected in the Vis range by correlation coefficients. In this study, we established a new three-band spectral index (TSI), which had a specific sensitive third band. The spectral index is shown as:

$$\mathrm{TSI} = \frac{R_{\lambda 1}}{R_{\lambda 2} + R_{\lambda 3}} \tag{7}$$

where $R_{\lambda 1}$ and $R_{\lambda 2}$ represent spectra of bands λ_1 and λ_2, respectively, in the range of 400–780 nm. $R_{\lambda 3}$ is the spectrum of 2200 nm.

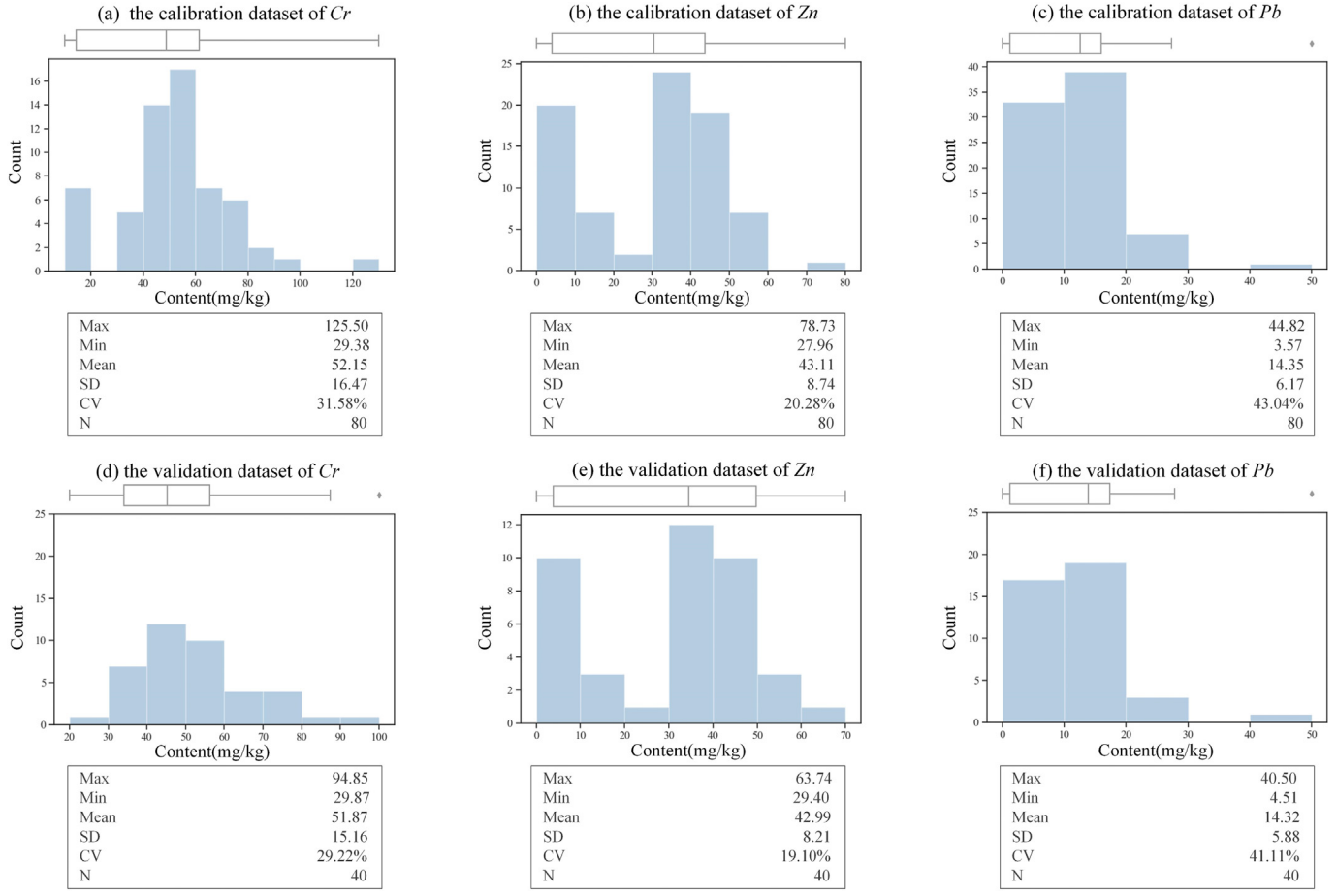


Fig. 3. Histograms, box-plots, and statistical data of heavy metals: (a) the calibration dataset of *Cr*; (b) the calibration dataset of *Zn*; (c) the calibration dataset of *Pb*; (d) the validation dataset of *Cr*; (e) the validation dataset of *Zn*; (f) the validation dataset of *Pb*. Min: minimum, Max: maximum, SD: standard deviation, CV: coefficient of variation, N: the number of soil samples.

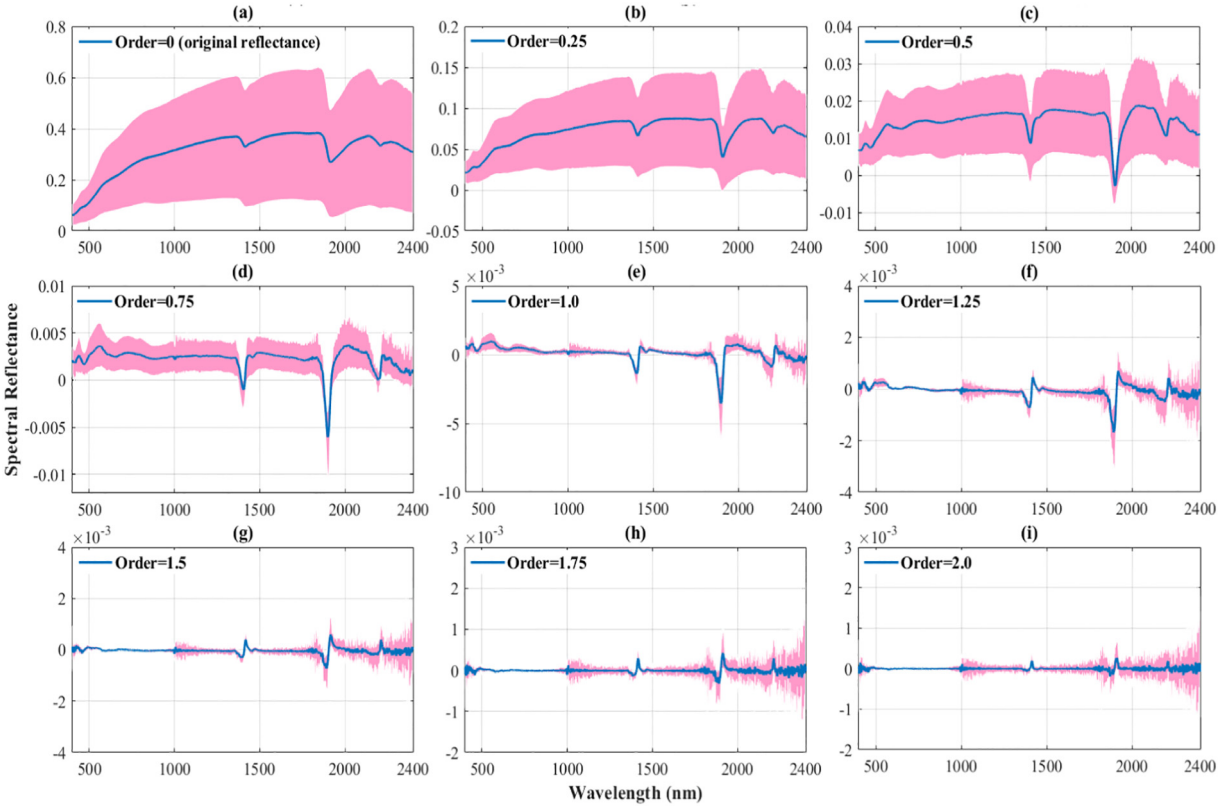


Fig. 4. FOD spectra (0 to 2, with an increment of 0.25 per step). The pink areas represent the whole scope of the spectra. The blue lines represent the mean spectra.

3.3. Modeling and model evaluation methods

3.3.1. Partial least squares (PLS)

PLS uses principal component analysis to condense multiple X and multiple Y into components (X corresponds to the principal component U, Y corresponds to the principal component V) (Leone et al., 2012), and then with the help of the canonical correlation principle, the relationship between X/U and Y/V can be analyzed. Combined with the principle of multiple linear regression analysis, the relationship between X and Y is studied through analyzing the relationship between X and V.

3.3.2. Support vector machine (SVM)

The SVM model is a kernel-based method that was proposed by Vapnik (1999). It is a nonlinear modeling method based on statistical learning theory. SVM can use support vectors in the training samples to design an optimal decision boundary. It can handle both linear and nonlinear problems, and can solve regression modeling problems. SVM performs well in dealing with high-dimensional and small sample data.

3.3.3. Ridge regression (RR)

RR (Mcdonald, 2010) is a biased estimation model for collinear data analysis. It improves the singularity of the coefficient matrix of the normal equations, which least square method unable to dispose when estimating the regression coefficients. At the cost of the partial precision of the least squares regression equation, a regression equation with a strong tolerance to ill-conditioned data is obtained, which can better solve the problem of the collinearity of hyperspectral data.

3.3.4. Random forest (RF)

RF is a kind of regression set algorithm that was proposed by Breiman (2004). The essence of the RF algorithm is an improvement of the decision tree (DT) algorithm. RF can handle a large number of input variables. By the bootstrap resampling technique, random sampling is repeated K times to generate a fixed number of subset training samples from all the samples (where K is the number of trees in the forest). Meanwhile, for each sample, only a fixed number of sub-attributes are selected. Each randomly selected subsample with its corresponding sub-attributes can then be used to generate a regression tree, and all the trees make up the forest. Finally, the results are obtained according to the scores of the class voting from all the trees. RF is a good way to estimate missing data, and it can maintain the accuracy in the case of missing data. These properties of RF make it suitable for processing hyperspectral data.

3.3.5. XGBoost

XGBoost (Zheng et al., 2017) is an improvement of the gradient boosting algorithm. It realizes the generation of weak learners by optimizing the structured loss function (the loss function with the regular term is added, which can reduce the risk of overfitting). XGBoost directly uses the first and second derivative values of the loss function, and greatly improves the performance of the algorithm through techniques such as presorting and weighted quantiles.

3.3.6. Extreme learning machine (ELM)

ELM (Chen and Wu, 2017) is a novel training algorithm for single-hidden-layer feedforward networks (SLFNs), which only needs setting the number of hidden layer nodes. Since it functions without extra adjustion of the input weight of the network and the offset of hidden elements, ELM maintains faster training speeds No parameters need to be manually tuned, and the output layer weights are estimated by the least-squares method.

3.3.7. Model evaluation method

The model evaluation was conducted with five determinant indicators: the coefficient of determination (R^2), the root-mean-square error (*RMSE*),

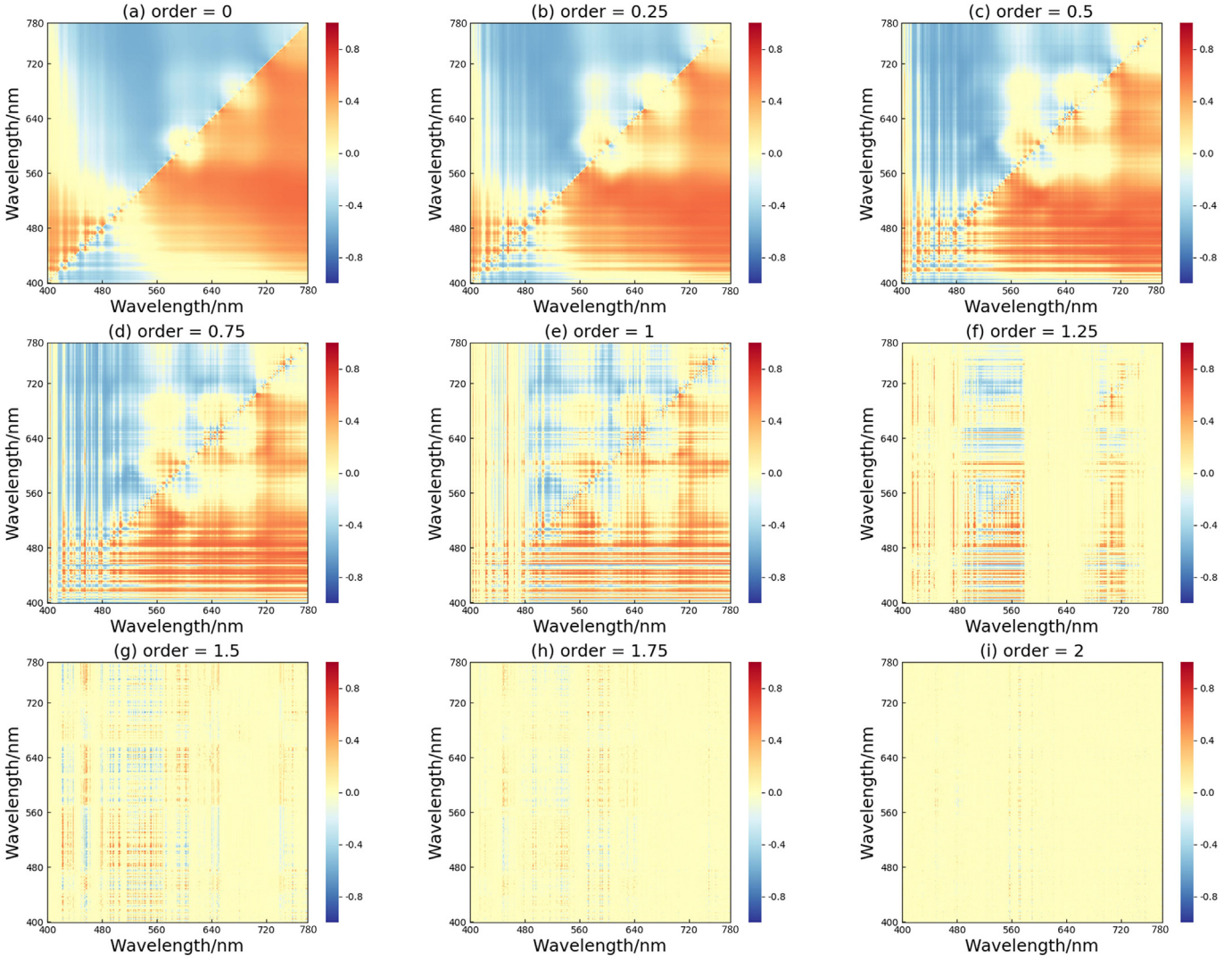


Fig. 5. Correlation between *Cr* and the two-band spectral index.

the mean absolute error (*MAE*), the residual prediction deviation (*RPD*), and the ratio of prediction performance to interquartile range (*RPIQ*), which are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{9}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{10}$$

Table 3
The maximum absolute correlation coefficients (MACC) between heavy metals and the two-band spectral index.

Element	Order								
	0	0.25	0.5	0.75	1	1.25	1.5	1.75	2
Cr	0.696	0.716	0.727	0.734	0.638	0.628	0.611	0.549	0.458
Zn	0.773	0.751	0.805	0.783	0.675	0.630	0.657	0.628	0.495
Pb	0.609	0.611	0.627	0.650	0.596	0.554	0.578	0.575	0.458

$$RPD = \frac{SD}{RMSE_P} \tag{11}$$

$$RPIQ = \frac{IQ}{RMSE_P} \tag{12}$$

where y_i is the measured value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average of the measured value, N is the number of samples, SD is the standard deviation of the validation set, and IQ is the interquartile distance of the validation set ($IQ = Q_3 - Q_1$). R_C^2, $RMSE_C$, and MAE_C represent the calibration dataset evaluation, in order to evaluate the fitting ability of the model. Meanwhile, R_P^2, $RMSE_P$, MAE_P, RPD, and $RPIQ$ represent the validation dataset evaluation, to evaluate the generalization ability, respectively. Overall, higher R_P^2, RPD, and $RPIQ$ values combined with lower $RMSE_P$ and MAE_P values are considered as good predictions, respectively.

4. Results and discussion

4.1. The content and spectra analysis

The heavy metal contents of the 120 soil samples were ranked in ascending order (from the lowest to the highest). These samples were further divided into 40 stratums. In each stratum, the middle sample and the

remaining two samples were assigned to validation and calibration sets, respectively. Finally, there were 80 samples in the calibration set, and the remaining 40 samples formed the validation set. The statistical data for the calibration dataset and validation dataset are shown in Fig. 3.

It can be seen from Fig. 3 that the mean, standard deviation (SD), and coefficient of variation (CV) of the two datasets are relatively close, that is, the data partition can be considered to be reasonable.

The FOD spectra with different orders are shown in Fig. 4.

From the original reflectance (order = 0) in Fig. 4, the reflectance shows a steep slope at 400–700 nm. After that, the rate of increase slows down. There are three major water absorption bands (at approximately 1400 nm, 1900 nm, and 2200 nm), especially the 2200 nm absorption feature, which is also characteristic of clays and soil organic matter (Zhang et al., 2019). When the FOD increases from 0 to 0.5, the reflection peak of water becomes sharper, and some small reflection peaks gradually became prominent. When the FOD increases from 0.5 to 1, the reflection peaks and absorption valleys become sharper. There are two reflection peaks at 420 nm and 560 nm (related to goethite and iron oxide, respectively), and two absorption valleys at 480 nm (associated with goethite and the spectral absorption feature of iron/manganese oxide) and 2200 nm (OH stretching vibration and ALOH bending vibration directly related to organic matter). When the FOD increases from 1 to 1.5, the differences between the spectra are small (almost zero), but the noise increases with the further increase in derivative order. The reflection peaks at 1400 nm, 1900 nm, and 2200 nm gradually become sharp positive and negative peaks. When the FOD increases from 1.5 to 2, the soil reflectance appears stable in the region of −0.002 to 0.002, which indicates that the baseline offsets and overlapping peaks are gradually removed (Zhang et al., 2019). The reflectance intensity gradually stabilizes, but numerous tiny peaks begin to appear and grow.

4.2. Spectral indices estimation

4.2.1. Modeling using a single two-band spectral index

The correlation coefficients between the two-band spectral index and the heavy metal contents using different FOD orders are shown in Fig. 5, Figs. S1 and S2. The horizontal and vertical axes denote the spectral bands. The maximum absolute correlation coefficients (i.e. the maximum between the positive maximum correlation coefficient and the negative maximum correlation coefficient) are listed in Table 3.

In the range of 0 to 2 fractional order, the maximum absolute correlation coefficient show first increasing and then decreasing, which reaches the maximum value (MACC = 0.734) when the order is 0.75 for *Cr*. The maximum absolute correlation coefficient presents fluctuating in the whole range and reaches the maximum value (MACC = 0.805) when the order is 0.5 for *Zn*. Similar to the case of *Cr*, it reaches the maximum value (MACC = 0.650) when the order is 0.75 for *Pb*. Then, the characteristic bands with the maximum absolute correlation coefficient for different fractional orders are used to estimate the contents of heavy metals. The characteristic bands and estimation results are listed in Table 4.

Table 4 shows that the best R_P^2 of *Cr*, *Zn* and *Pb* are 0.61 (order = 0.25), 0.65 (order = 0.25) and 0.47 (order = 0.5). Meanwhile, the highest values combined with lowest $RMSE_P$ and lower MAE_P values are good predictions to these three heavy metals. The estimation results perform better with 0- to 0.75-order than with 1- to 2-order for these three heavy metals, which states noise signals generate gradually and performances become unstable with the derivative order further increasing. These characteristic bands are related to the presence of hematite, goethite, organic matter and ferrihydrite.

4.2.2. Modeling using a single three-band spectral index

The correlation coefficients between the three-band spectral index and the heavy metals contents with different FOD orders are shown in Fig. 6, Figs. S3 and S4. The horizontal and vertical axes denote the spectral bands. The maximum absolute correlation coefficients (i.e. the maximum

Table 4
The characteristic bands and estimation results for heavy metals with the two-band spectral index by linear regression.

Element	Order	Characteristic bands	R_c^2	$RMSE_c$	MAE_c	R_P^2	$RMSE_P$	MAE_P	RPD	RPIQ
Cr	0	526,779	0.71	13.18	10.01	0.55	19.33	13.16	1.01	1.70
	0.25	504,725	0.64	14.71	11.35	**0.61**	17.95	12.26	1.26	2.26
	0.5	486,723	0.67	14.13	10.10	0.54	19.54	13.38	1.05	1.62
	0.75	705,706	0.78	11.33	8.34	0.59	18.39	11.85	1.13	1.95
	1	607,455	0.42	18.82	13.14	0.36	20.91	14.76	0.66	0.89
	1.25	510,423	0.13	23.06	19.29	0.48	25.71	20.43	0.19	0.30
	1.5	504,506	0.44	18.54	14.83	0.43	23.21	17.23	0.49	0.55
	1.75	706,574	0.57	16.18	12.18	0.38	23.33	15.60	0.99	1.04
	2	702,573	0.32	20.28	15.57	0.21	25.71	19.84	0.48	0.43
Zn	0	519,777	0.53	11.22	8.64	0.58	12.19	9.31	0.85	1.14
	0.25	504,723	0.58	10.64	8.53	**0.65**	10.72	8.69	1.27	2.08
	0.5	486,657	0.70	8.89	7.15	0.62	11.21	8.40	1.33	1.66
	0.75	444,570	0.69	9.01	6.67	0.62	11.51	8.62	0.99	1.03
	1	606,455	0.20	14.71	11.57	0.38	15.18	12.43	0.40	0.26
	1.25	453,552	0.32	13.52	10.45	0.29	15.36	11.89	0.46	0. 41
	1.5	504,506	0.26	14.07	10.89	0.48	13.98	11.11	0.57	0.80
	1.75	706,572	0.54	11.03	8.57	0.13	33.48	15.16	1.06	0.39
	2	706,572	0.31	13.64	10.04	0.16	23.46	14.49	1.00	0.32
Pb	0	422,406	0.44	5.76	4.20	0.43	6.11	4.28	0.75	0.89
	0.25	457,456	0.44	5.80	4.02	0.38	6.36	4.61	0.67	0.77
	0.5	736,504	0.45	5.71	4.08	**0.47**	5.97	4.14	0.74	1.08
	0.75	404,582	0.46	5.67	3.87	0.38	6.34	4.47	0.76	0.92
	1	707,618	0.20	6.93	2.16	0.16	7.41	5.41	0.40	0.61
	1.25	484,476	0.27	6.63	5.06	0.33	6.45	4.76	0.60	0.82
	1.5	579,552	0.36	6.17	4.53	0.31	6.65	4.68	0.72	0.78
	1.75	518,449	0.19	6.96	5.45	0.33	6.87	5.57	0.47	0.50
	2	533,449	0.01	7.71	6.22	0.13	7.96	6.33	0.04	0.03

Note: R_C^2, $RMSE_C$, and MAE_C represent the coefficient of determination, the root-mean-square error and the mean absolute error of the calibration dataset, respectively. R_P^2, $RMSE_P$, MAE_P, *RPD*, and *RPIQ* represent the coefficient of determination, the root-mean-square error, the mean absolute error, the residual prediction deviation and the ratio of prediction performance to interquartile range of the validation dataset, respectively. The unit of $RMSE_C$, MAE_C, $RMSE_P$ and MAE_P is mg/kg. The boldfaces represent the order and R_p^2 with the best perfomance.

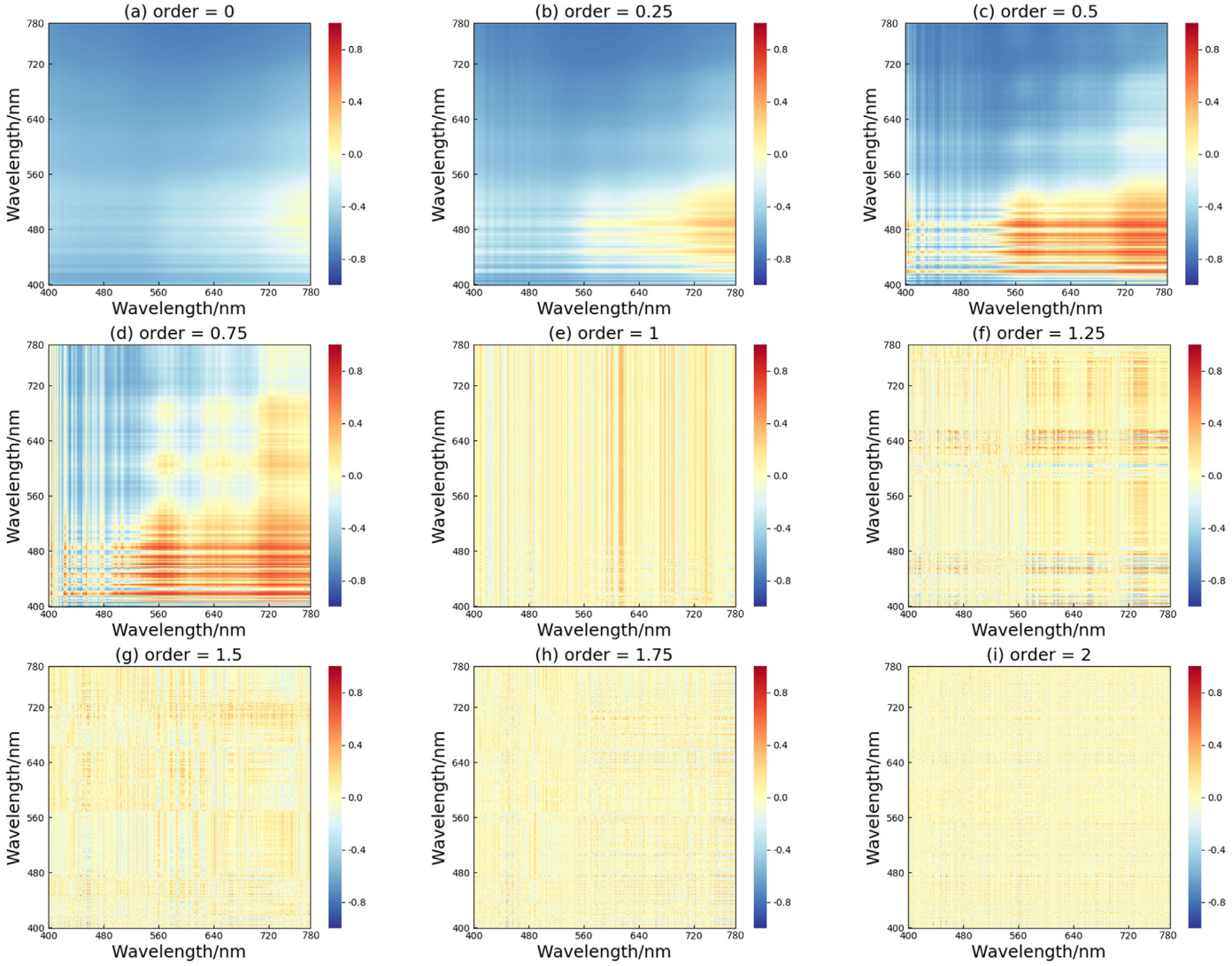


Fig. 6. Correlation between *Cr* and the three-band spectral index.

between the positive maximum correlation coefficient and the negative maximum correlation coefficient) are also listed in Table 5.

The maximum absolute correlation coefficient of *Cr* increases as the fractional order increases from 0 to 0.5. It then reaches a maximum value (MACC = 0.792) when the order is 0.5, before beginning to descend. The maximum absolute correlation coefficient reaches the minimum value (MACC = 0.416) when the order is 1, and it fluctuates with the order from 1 to 2. In the range of 0 to 1 fractional order, there is a trend of first increasing and then decreasing for the correlation coefficient of *Zn*, which reaches the maximum value (MACC = 0.835) when the order is 0.25. The maximum absolute correlation coefficient reaches the minimum value (MACC = 0.460) when the order is 1, and it also goes up and then down when the order is from 1 to 2. Similar to the case of *Cr*, it reaches the maximum value (MACC = 0.741) when the order is 0.5 and the minimum value (MACC = 0.485) when the order is 1 for *Pb*. The band combinations with the maximum correlation coefficients are defined as λ_1 and λ_2, respectively. These two bands and the band of 2200 nm are then combined as the three-band spectral index to establish heavy metal concentration by linear regression. The characteristic bands and estimation results are listed in Table 6.

Table 5
The maximum absolute correlation coefficients (MACC) between heavy metals and the three-band spectral index.

Element	Order								
	0	0.25	0.5	0.75	1	1.25	1.5	1.75	2
Cr	0.767	0.787	0.792	0.769	0.416	0.646	0.644	0.535	0.538
Zn	0.810	0.835	0.828	0.832	0.460	0.712	0.717	0.600	0.553
Pb	0.678	0.713	0.741	0.651	0.485	0.577	0.611	0.501	0.588

The models obtain different estimation performances, depending on the combination bands and the fractional orders. For *Cr*, the estimation accuracy improves with R_P^2 increasing, $RMSE_P$ and MAE_P decreasing when the order is from 0 to 0.75, and the best result is obtained using 0.75-order reflectance (R_P^2 = 0.74, $RMSE_P$ = 13.17, and RPD = 1.78), indicating that the model with 0.75-order reflectance is stable and predictive. It is clear that the estimation performance decreases since the 1-order reflectance. When the order is 1.25, although R_P^2 reaches 0.52, R_c^2 is only 0.18, which demonstrates that the estimation ability is poor. For the 1.5-order reflectance, R_c^2 and R_P^2 are 0.62 and 0.55, respectively, and the estimation result for the 1.5-order reflectance is better than the other estimation results when the order is from 1 to 2. The accuracy from 0- to 0.75-order reflectance is superior to that of 1- to 2-order reflectance, because the spectra introduce external noise and may be subject to intense peak deformations when the order is from 1 to 2. The FOD with 0.25-order, 0.5-order, and 0.75-order reflectance can reveal the masked soil heavy metal information with a better estimation performance than the original spectra and the 1- and 2-order derivatives. This result shows the capability of FOD in

Table 6
The characteristic bands and estimation results for heavy metals with the three-band spectral index by linear regression.

Element	Order	Characteristic bands	R_c^2	$RMSE_c$	MAE_c	R_P^2	$RMSE_P$	MAE_P	RPD	RPIQ
Cr	0	779,607	0.54	18.07	13.43	0.65	15.77	12.36	1.02	1.48
	0.25	779,604	0.61	16.75	12.38	0.69	14.79	11.90	1.18	2.10
	0.5	778,525	0.62	12.28	12.30	0.69	14.45	11.25	1.31	2.43
	0.75	448,563	0.74	13.59	9.81	**0.74**	13.17	10.15	1.78	2.55
	1	406,615	0.06	25.76	20.84	0.16	24.09	20.04	0.17	0.05
	1.25	706,458	0.18	24.09	18.33	0.52	19.61	15.42	0.51	0.47
	1.5	706,547	0.62	15.18	10.78	0.55	19.69	13.36	0.93	1.65
	1.75	408,447	0.02	24.58	21.18	0.02	29.14	23.88	0.33	0.07
	2	410,447	0.02	26.34	21.74	0.11	24.61	20.63	0.14	0.07
Zn	0	779,607	0.65	9.61	8.10	0.68	10.84	8.20	1.02	1.48
	0.25	779,604	0.69	9.06	7.63	0.74	9.99	7.78	1.17	2.08
	0.5	776,525	0.69	9.53	7.51	0.73	8.66	7.20	1.68	2.43
	0.75	400,432	0.70	9.49	6.94	**0.81**	7.12	5.67	2.15	3.94
	1	400,483	0.14	15.23	12.48	0.10	18.62	14.94	0.63	0.18
	1.25	641,585	0.25	14.82	11.11	0.33	13.65	10.89	0.60	0.74
	1.5	706,458	0.44	12.87	9.44	0.41	12.75	10.47	0.84	0.67
	1.75	409,457	0.04	16.81	14.41	0.02	16.49	14.28	0.10	0.05
	2	641,431	0.07	16.56	14.13	0.01	17.26	14.67	0.20	0.11
Pb	0	779,608	0.46	5.68	3.87	0.46	6.03	3.90	0.75	1.17
	0.25	779,595	0.54	5.22	3.47	0.46	6.02	4.11	1.01	1.37
	0.5	745,533	0.55	5.17	3.51	**0.56**	5.41	3.47	0.93	1.76
	0.75	400,515	0.41	5.95	4.05	0.46	5.98	3.92	0.78	0.83
	1	407,443	0.15	7.13	5.74	0.05	11.83	7.17	0.91	0.17
	1.25	457,470	0.07	7.48	5.99	0.14	7.66	5.86	0.18	0.17
	1.5	706,458	0.30	6.48	4.65	0.40	6.60	4.64	0.45	0.50
	1.75	407,467	0.24	6.74	5.00	0.25	6.98	5.12	0.64	0.51
	2	747,447	0.08	7.44	5.97	0.07	7.75	6.12	0.31	0.14

Note: R_C^2, $RMSE_C$, and MAE_C represent the coefficient of determination, the root-mean-square error and the mean absolute error of the calibration dataset, respectively. R_P^2, $RMSE_P$, MAE_P, *RPD*, and *RPIQ* represent the coefficient of determination, the root-mean-square error, the mean absolute error, the residual prediction deviation and the ratio of prediction performance to interquartile range of the validation dataset, respectively. The unit of $RMSE_C$, MAE_C, $RMSE_P$ and MAE_P is mg/kg. The boldfaces represent the order and R_p^2 with the best perfomance.

processing reflectance spectra. The wavelengths at 406(408,410), 447, 604 (607), and 779(776) nm are selected multiple times, and are thus particularly important for the spectral estimation of *Cr*. These bands may be linked to the presence of ferrihydrite, goethite, hematite, ferric oxide, and CH. The estimation accuracy generally improves and then decreases from 0- to 1-order reflectance, and the change trend of 1- to 2-order is similar to that of 0- to 1-order of *Zn*. The estimation model with 0.75-order reflectance is superior to the other orders in the estimation accuracy. Based on this model, the R_P^2 is 0.81, the $RMSE_P$ is 7.12, and the *RPD* is 2.15. The model performs the worst when the order is 2. The wavelengths at 400, 431 (432), 457(458), and 779(776) nm are selected multiple times, and are thus particularly important for the spectral estimation of *Zn*. These bands may be linked to the presence of hematite, goethite, hematite, and CH. With regards to *Pb*, there are the same phenomena. The best result is obtained using 0.5-order reflectance (R_P^2 = 0.56, $RMSE_P$ = 5.41, and *RPD* = 0.93). The wavelengths at 407, 457(458), 467(470), 745(747) and 779 nm are selected multiple times. These bands may be linked to the presence of ferrihydrite, goethite, hematite, ferric oxide, NH and CH.

Tables 4 and 6 show that the best R_P^2, $RMSE_P$ and MAE_P of three heavy metals using two-band index are worse than using three-band index. The best R_P^2 of *Cr*, *Zn* and *Pb* are 0.74 (order = 0.75), 0.81 (order = 0.75) and 0.56 (order = 0.5) using three-band index. This result proved that using three-band index form could promote the sensitivity and estimation ability to the heavy metals.

Table 7
The estimation results with multiple spectral indices for heavy metals.

Element	Model	R_c^2	$RMSE_c$	MAE_c	R_P^2	$RMSE_P$	MAE_P	RPD	RPIQ
Cr	PLS	0.65	15.62	26.82	0.72	13.75	25.35	1.38	2.58
	SVM	0.64	19.88	15.34	0.72	18.95	15.66	0.44	0.82
	RR	0.64	19.72	15.45	0.73	18.96	15.84	0.44	0.82
	RF	0.75	13.23	8.03	0.77	13.72	9.89	1.89	3.72
	XGBoost	0.97	4.58	3.09	0.69	13.83	8.67	1.72	3.73
	ELM	0.79	12.21	7.48	**0.77**	13.40	10.24	1.88	4.09
Zn	PLS	0.78	7.63	17.63	0.73	9.50	17.59	1.41	2.54
	SVM	0.72	9.27	7.35	0.66	11.64	8.25	1.02	1.51
	RR	0.79	7.77	18.10	0.81	7.03	18.27	2.31	4.21
	RF	0.81	7.61	5.46	0.82	7.07	5.42	2.09	4.42
	XGBoost	0.98	1.75	1.21	0.77	7.98	6.21	2.01	4.15
	ELM	0.90	5.51	3.74	**0.86**	6.61	5.33	2.50	4.89
Pb	PLS	0.57	5.07	7.51	0.59	5.21	7.23	1.00	1.90
	SVM	0.54	5.39	3.40	0.55	5.70	3.50	0.80	1.28
	RR	0.48	5.61	3.77	0.51	5.81	3.74	0.77	1.04
	RF	0.63	10.25	8.25	0.52	11.57	7.92	1.23	1.89
	XGBoost	0.96	1.57	1.15	0.51	12.84	8.16	2.11	3.24
	ELM	0.62	9.89	7.57	**0.63**	10.19	7.75	1.28	2.60

Note: R_C^2, $RMSE_C$, and MAE_C represent the coefficient of determination, the root-mean-square error and the mean absolute error of the calibration dataset, respectively. R_P^2, $RMSE_P$, MAE_P, *RPD*, and *RPIQ* represent the coefficient of determination, the root-mean-square error, the mean absolute error, the residual prediction deviation and the ratio of prediction performance to interquartile range of the validation dataset, respectively. The unit of $RMSE_C$, MAE_C, $RMSE_P$ and MAE_P is mg/kg. The boldfaces represent the order and R_p^2 with the best perfomance.

4.2.3. Modeling using multiple spectral indices

In order to improve the fitting and generalization ability, the regression models of PLS, SVM, RF, RR, XGBoost, and ELM were used to conduct a comprehensive analysis of the spectral indices with a high accuracy. The models obtain different degrees of accuracy depending on the spectral estimation models applied and the FODs used. For these three heavy metals, the original spectra, 0.25-order, 0.5-order, and 0.75-order spectral indices are better than those with other FOD intervals in terms of improving the spectral response to heavy metals and mining potential information, which provides a potential to establish a more robust heavy metals estimation model. Therefore, these spectral indices are combined for the further multiple estimation. The estimation results are listed in Table 7.

For *Cr* and *Zn*, it can be seen that the estimation accuracy is improved by combining the multiple spectral indices with RF and ELM. In contrast, the estimation performance of PLS and SVM is not as good as that of linear regression by one spectral index. For *Cr*, the other evaluation indicators of the ELM model are better than those of the RF model, while the R_P^2 values of these two methods are the same. Therefore, the ELM model outperforms

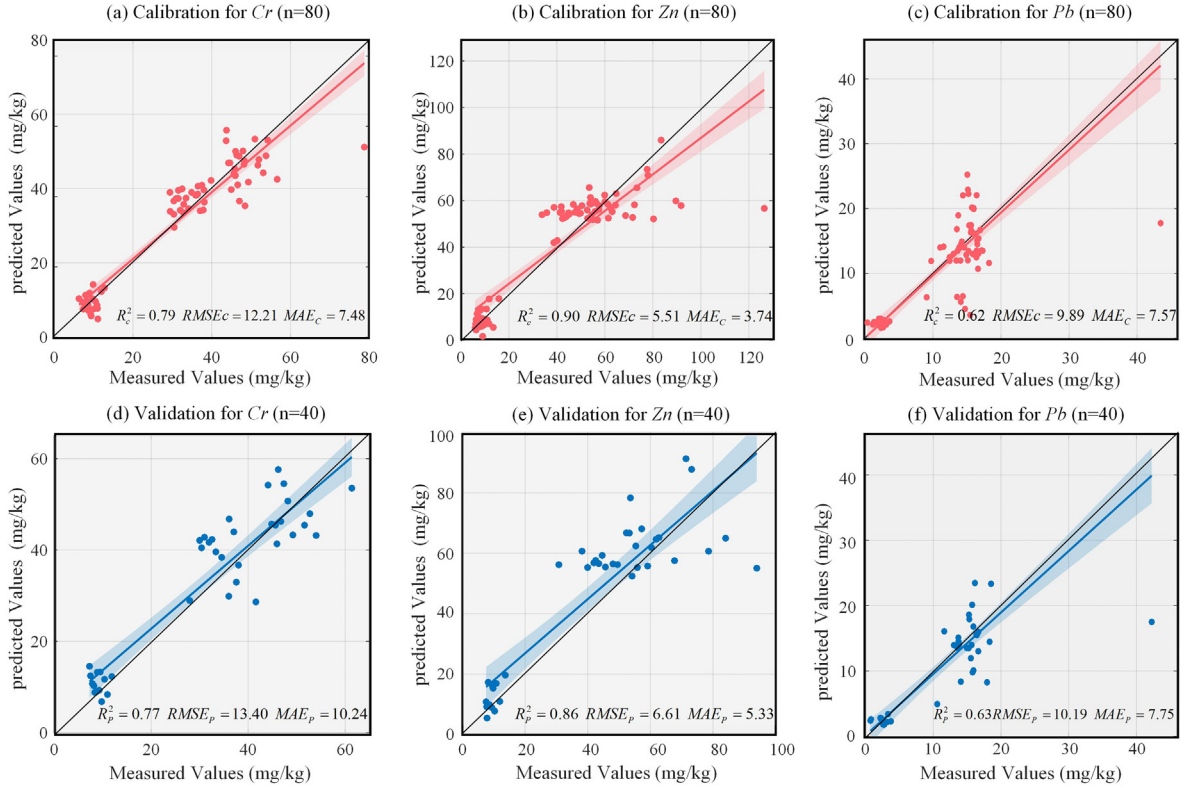


Fig. 7. Scatter plots of ELM model in the calibration and validation datasets: (a) Calibration for *Cr*; (b) Calibration for *Zn*; (c) Calibration for *Pb* (d) Validation for *Cr*; (e) Validation for *Zn*; (f) Validation for *Pb*.

other counterparts in estimating *Cr*. For *Zn*, the ELM model performs better than the other three models, with the R_P^2, *RPD*, and *RPIQ* values being the highest, while the $RMSE_P$ and MAE_P are the lowest. The estimation results of ELM are better than other methods for *Pb*. To these three heavy metals, the R_C^2 of XGBoost exceeds 0.95, which appears in the accuracy evaluation of the training set, and the testing set accuracy is lower. In contrast, ELM has the best performance on the testing set compared with other court parts, which reflects its strong generalization performance. Therefore, ELM is stable and robust under the limited training set and can be competed for the practical applications. The scatter plots of ELM model are shown in Fig. 7.

The measured-predicted points are distributed well around the 1:1 line for *Cr* and *Zn*, which indicates that the model has a stable performance. And the accuracy for the validation set is lower than that for the calibration set. As regard *Pb*, it performs better in low values but distributes dispersedly in high values, which is also the reason for the higher *RMSE*, but with acceptable limits. The results show that the use of the most appropriate FOD order and band combination algorithm can be used to estimate the three heavy metals effectively.

In the study of soil heavy metals estimation, R^2 of the PLS model from Vis/NIR spectroscopy for the monitoring toxic elements of *Pb* in the agricultural soils of the Changjiang River reaches 0.68 (Song et al., 2012). Sun et al. (2017) applied GA-PLS to predict *Pb*, and their result was slightly lower ($R^2 = 0.44$) with spectral angle. Based on soil spectra mechanism, Wu et al. estimated *Zn* using spectra absorption features, yielding R^2 values of 0.56 (Wu et al., 2007b). It could be observed that the ELM model yielded superior model performance comparing with the above research.

5. Conclusions

In this study, the potential of the FOD pretreatment method with different orders and the new three-band spectral index based on the band combination algorithm to estimate soil heavy metal was investigated. The main conclusions can be summarized as follows. The FOD can attenuate the baseline drift and separate the overlapping peaks while detecting more detailed spectral characteristics, but it does introduce intense peak deformations and spectral noise with further increases in the derivative order. For the three heavy metals, the highest estimation accuracy is achieved with the 0.75-order reflectance (i.e. the highest $R_P^2 = 0.74$ for *Cr*, the highest $R_P^2 = 0.81$ for *Zn*) and 0.5-order reflectance (i.e. the highest $R_P^2 = 0.56$ for *Pb*). In addition, the ELM model combining several spectral indices with better performance can improve the estimation results (i.e. the highest $R_P^2 = 0.77$ for *Cr*, the highest $R_P^2 = 0.86$ for *Zn*, the highest $R_P^2 = 0.63$ for *Pb*). In this paper, the experiment was conducted in an experiment field, and it has been proved that the comprehensive use of the FOD and the band combination algorithm can obtain fine spectral features, low data redundancy, and a high prediction accuracy, which will provide an important reference to estimate soil heavy metals using Vis–NIR spectra over a large scale.

CRediT authorship contribution statement

Lihan Chen: Conceptualization, Methodology, Data curation, Validation, Writing – original draft. **Jian Lai:** Formal analysis, Writing – original draft. **Kun Tan:** Conceptualization, Methodology, Writing – original draft. **Xue Wang:** Formal analysis, Writing – review & editing. **Yu Chen:** Writing – review & editing. **Jianwei Ding:** 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.

Acknowledgements

The authors would like to thank Prof. Jihong Dong with China University of Mining and Technology for providing the soil samples. This research

was supported in part by National Natural Science Foundation of China (No. 41871337 and No. 42171335).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.151882.

References

Agency, N.E.P., 1995. Environmental Quality Standard for Soils (GB15618-1995). National Environmental Protection Agency, Beijing, pp. 1–5.

Asadzadeh, S., de Souza Filho, C.R., 2016. A review on spectral processing methods for geological remote sensing. Int. J. Appl. Earth Obs. Geoinf. 47, 69–90.

Baderia, K., Kumar, A., Singh, G.K., 2015. Hybrid method for designing digital FIR filters based on fractional derivative constraints. ISA Trans. 58, 493–508.

Bao, N., et al., 2017. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. Geoderma 288 (00167061), 47–55.

Bartholomeus, H., et al., 2008. Spectral reflectance based indices for soil organic carbon quantification. Geoderma 145 (1), 28–36.

Benkhettou, N., da Cruz, A., Torres, D.F.M., 2015. A fractional calculus on arbitrary time scales: fractional differentiation and fractional integration. Signal Process. 107, 230–237.

Breiman, L., 2004. Random forests. Mach. Learn. 45, 5–32.

Chen, Y., Wu, W., 2017. Mapping mineral prospectivity using an extreme learning machine regression. Ore Geol. Rev. 80, 200–213.

Chittleborough, D., et al., 2011. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. Ecol. Indic. 11 (1), 123–131.

Choe, E., et al., 2008. Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: a case study of the Rodalquilar mining area, SE Spain. Remote Sens. Environ. 112 (7), 3222–3233.

Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy. Remote Sens. Earth Sci. Man. Remote Sens. 3, 3–58.

Clark, R.N., et al., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. Solid Earth 95, 12653–12680.

Clark, R.N., et al., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. 95, 12653–12680.

Fearn, T., et al., 2009. On the geometry of SNV and MSC. Chemom. Intell. Lab. Syst. 96 (1), 22–26.

Galvão, L.S., Vitorello, Í., 1998. Role of organic matter in obliterating the effects of iron on spectral reflectance and colour of brazilian tropical soils. Int. J. Remote Sens. 19, 1969–1979.

He, C., et al., 2020. Phytoremediation of soil heavy metals (Cd and Zn) by castor seedlings: tolerance, accumulation and subcellular distribution. Chemosphere 252, 126471.

Hong, Y., 2018. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. Geoderma 337, 758–769.

Hong, Y., et al., 2018. Combining fractional order derivative and spectral variable selection for organic matter estimation of homogeneous soil samples by VIS–NIR spectroscopy. Remote Sens. 10 (3), 479.

Hong, Y.S., et al., 2020. Exploring the potential of airborne hyperspectral image for estimating topsoil organic carbon: effects of fractional-order derivative and optimal band combination algorithm. Geoderma 365, 114228.

Hunt, G.R., 1977. Spectral signatures of particulate minerals in the visible and near infrared. Geophysics 42, 501–513.

Khan, S., et al., 2008. Health risks of heavy metals in contaminated soils and food crops irrigated with wastewater in Beijing, China. Environ. Pollut. 152 (3), 686–692.

Knadel, M., et al., 2013. Visible-near infrared spectra as a proxy for topsoil texture and glacial boundaries. Soil Sci. Soc. Am. J. 77 (2), 568–579.

Lassalle, G., et al., 2020. Monitoring oil contamination in vegetated areas with optical remote sensing: a comprehensive review. J. Hazard. Mater. 393, 122427.

Leone, A.P., et al., 2012. Prediction of soil properties with PLSR and Vis-NIR spectroscopy: application to Mediterranean soils from southern Italy. Curr. Anal. Chem. 8 (2), 283–299.

Li, Q.S., et al., 2012. Health risk of heavy metals in food crops grown on reclaimed tidal flat soil in the Pearl River Estuary, China. J. Hazard. Mater. 227, 148–154.

Lu, D., Jin, W.M., 2011. Fully phase color image encryption based on joint fractional fourier transform correlator and phase retrieval algorithm. Chin. Opt. Lett. 9 (2), 021002.

Mcdonald, G.C., 2010. Ridge regression. Wiley Interdiscip. Rev. Comput. Stat. 1 (1), 93–100.

Meng, X., et al., 2020. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. Int. J. Appl. Earth Obs. Geoinf. 89, 102111.

Minasny, B., et al., 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma 167, 118–124.

Mohamed, E., et al., 2018. Application of near-infrared reflectance for quantitative assessment of soil properties. Egypt. J. Remote Sens. Space Sci. 21 (1), 1–14.

Nayak, P.S., Singh, B.K., 2007. Instrumental characterization of clay by XRF, XRD and FTIR. Bull. Mater. Sci. 30 (3), 235–238.

Rossel, R.A.V., et al., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1), 59–75.

Salazar, M.J., et al., 2012. Effects of heavy metal concentrations (Cd, zn and Pb) in agricultural soils near different emission sources on quality, accumulation and food safety in soybean [Glycine max (L.) Merrill]. J. Hazard. Mater. 233, 244–253.

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. Clay Clay Miner. 46 (5), 528–536.

Sherman, D.M., Waite, T.D., 1985. Electronic spectra of Fe3+ oxides and oxide hydroxides in the near IR to near UV. Am. Mineral. 70 (11), 1262–1269.

Shi, T., et al., 2014. Visible and near-infrared reflectance spectroscopy-an alternative for monitoring soil contamination by heavy metals. J. Hazard. Mater. 265, 166–176.

Song, Y.X., et al., 2012. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. Appl. Clay Sci. 64, 75–83.

Srasra, E., Bergaya, F., Fripiat, J.J., 1994. Infrared spectroscopy study of tetrahedral and octahedral substitutions in an interstratified illite-smectite clay. Clay Clay Miner. 42, 237–241.

Stenberg, B., et al., 2010. Visible and near infrared spectroscopy in soil science. Adv. Agron. 107, 163–215.

Sun, W.C., et al., 2017. Exploring the potential of spectral classification in estimation of soil contaminant elements. Remote Sens. 9 (6), 632.

Tan, K., et al., 2018. An improved estimation model for soil heavy metal(loid) concentration retrieval in mining areas using reflectance spectroscopy. J. Soils Sediments 18 (5), 2008–2022.

Tan, K., et al., 2020. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. J. Hazard. Mater. 382, 120987.

Tan, K., et al., 2020. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. J. Hazard. Mater. 401, 123288.

Tarasov, V.E., 2016. On chain rule for fractional derivatives. Commun. Nonlinear Sci. Numer. Simul. 30, 1–4.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10 (5), 988–999.

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2), 46–54.

Wang, X.P., et al., 2018. 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.

Wang, J., et al., 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.

Wang, H., et al., 2020. Study of the retrieval and adsorption mechanism of soil heavy metals based on spectral absorption characteristics. Spectrosc. Spectr. Anal. 40 (1), 316–323.

Wickersheim, K.A., Lefever, R.A., 1962. Absorption spectra of ferric iron-containing oxides. J. Chem. Phys. 36 (3), 844–850.

Wu, Y., et al., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71 (3), 918–926.

Wu, Y., et al., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71, 918–926.

Wu, Y., et al., 2019. Hyperspectral characteristics of soil organic matter and inversion methods. Journal of Shanghai Jiaotong University 37 (4), 37–44.

Xia, Y., et al., 2020. Enhanced phosphorus availability and heavy metal removal by chlorination during sewage sludge pyrolysis. J. Hazard. Mater. 382, 121110.

Yuan, Z.-R., et al., 2020. Hyperspectral inversion and analysis of heavy metal arsenic content in farmland soil based on optimizing CARS combined with PSO-SVM algorithm. Spectrosc. Spectr. Anal. 40 (2), 567–573.

Zhang, J., 2011. Performance analysis of fractional Fourier transform optical imaging based on fractional Fourier-domain filtering. Acta Opt. Sin. 31 (11) p. 1111003-1-1111003-8.

Zhang, Z., et al., 2019. Prediction of soil organic matter in northwestern China using fractional- order derivative spectroscopy and modified normalized difference indices. Catena 185, 172–187.

Zhang, Y., et al., 2020. Retrieval of soil moisture content based on a modified hapke photometric model: a novel method applied to laboratory hyperspectral and Sentinel-2 MSI data. Remote Sens. 12 (14), 2239.

Zhang, Z.P., et al., 2020. Prediction of soil organic matter in northwestern China using fractional-order derivative spectroscopy and modified normalized difference indices. Catena 185, 104257.

Zheng, H., Yuan, J., Chen, L., 2017. Short-term load forecasting using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation. Energies 10 (8), 1168.