Heliyon 9 (2023) e19782

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Inversion of soil heavy metals in metal tailings area based on different spectral transformation and modeling methods

Nannan Yang [a], Ling Han [a,b,*], Ming Liu [a]

[a] School of Land Engineering, Chang'an University, Xi'an, 710054, China
[b] Shaanxi Key Laboratory of Land Consolidation, Chang'an University, Xi'an, 710054, China

ARTICLE INFO

Keywords:
Soil heavy metals
Hyperspectral
Spectral transformation
Multispectral simulation
Model inversion

ABSTRACT

The exploitation of mineral resources has seriously polluted the environment around mines, notably in terms of heavy metal contamination of tailings pond soil. Hyperspectral remote sensing, as opposed to conventional on-site sampling and laboratory analysis, offers a potent tool for effective monitoring the content of soil heavy metals. Therefore, we investigated the inversion models of heavy metal content in metal tailings area based on measured hyperspectral and multispectral data. Hyperspectral and its transformation, as well as the simulated Landsat8-OLI multispectral were used for model inversion respectively. Stepwise Multiple Linear Regression (SMLR), Partial Least Squares Regression (PLSR) and Back Propagation Neuron Network (BPNN) were established to study the spectral inversion of eight heavy metals (Cu, Cd, Cr, Ni, Pb, Zn, As, and Hg). The direct inversion models were established on the basis of correlation analysis and the adjust coefficient of determination (*Adjust_*R^2) and Root Mean Square Error (*RMSE*) were used for model evaluation. Then the best combination of spectral transformation and inversion model were explored. The model inversion results suggested that: (1) Hyperspectral transformation can generally improve the model accuracy, especially the second derivative spectral, based on which the training *Adjust_*R^2 of Hg SMLR and PLSR models are as high as 0.795 and 0.802. (2) The BP neural network inversion based on the denoised hyperspectrum demonstrate that both the training and testing *Adjust_*R^2 of Cd, Ni and Hg models are all greater than 0.5, indicating good applicability in practical extrapolation. (3) Both the training and testing *Adjust_*R^2 of Cu and Hg PLSR models based on simulated *R*_Landsat8-OLI multispectral are greater than 0.5, and Hg has lower *RMSE* and lager *Adjust_*R^2 with training and testing *Adjust_*R^2 values of 0.833 and 0.553 respectively. (4) Multispectral remote sensing detection and mapping of Hg contamination were realized by the optimal simulation model of Hg. Hence, it is feasible to simulate the multispectral with hyperspectral data for investigating heavy metal contamination.

1. Introduction

Soil heavy metal pollution is a significant issue that threatens both the sustainable development of social economy and the safety of ecological environment [1]. As a major energy consumer, China has continuously expanded the scale of mineral resources exploitation in recent years, which has caused serious negative effects on the soil environment, especially the soil heavy metal contamination [2]. Heavy metals are deposited in the soil of the mining area due to waste water, solid waste, and dust from mining activity. These heavy

* Corresponding author. School of Land Engineering, Chang'an University, Xi'an 710054, China.
E-mail address: hanlingmail@163.com (L. Han).

https://doi.org/10.1016/j.heliyon.2023.e19782
Received 5 May 2023; Received in revised form 30 August 2023; Accepted 31 August 2023
Available online 7 September 2023

metals are subsequently enriched by soil bio-enrichment, which threatens human health and the food chain [3]. Therefore, studying on soil heavy metal pollution in mining areas is crucial to enhancing both the sustainable use of land and the ecological environment protection. Conventional laboratory testing is labor-intensive, time-consuming, and challenging to acquire accurate distribution of soil contaminants across a continuous range. Due to its effectiveness, low cost, and extensive potential for dynamic monitoring, hyperspectral remote sensing technology has emerged as the industry standard for analyzing heavy metal contamination [4–6]. The heavy metal concentration of mining soil has been predicted by soil hyperspectral analysis [7–9]. However, the spectrum characteristic absorption bands are not readily apparent and it is challenging to directly invert the heavy metal content in the mining soil due to the trace amounts of heavy metals in the mining soil. Consequently, current research mainly focus on improving model algorithms, selecting spectral features, transforming hyperspectral to achieve the indirect prediction of soil heavy metals [10–12]. Derivation, logarithm, continuum elimination, and wavelet transformation are common spectral transformation forms, from which the spectral derivative transformation can effectively highlight the response characteristics of soil heavy metals [13,14]. Typically, univariate and conventional linear regression cannot satisfy the prediction needs, whereas models coupling with feature selection algorithms, such as partial least squares regression (PLSR), neural network (BP, RBF), support vector machine (SVM), and random forest (RF), which thoroughly consider multivariate responses, have been widely used to predict soil heavy metal concentrations [5,15,16]. Multispectral remote sensing images, on the other hand, rich in spectral information, simple to obtain, inexpensive, and have a wide range of spatial distribution, which make it better suited for investigating small mining areas that actually require spatial analysis of soil heavy metals [17]. The multispectral inversion approach, which creates a link between the heavy metal content of the sampling point and the reflectance of the multispectral to find the best fitting model, thus providing a reliable basis for soil environmental monitoring [18]. There are, however, few studies on regional-scale soil heavy metal remote sensing inversion combined with remote sensing images due to the difficulty in obtaining hyperspectral remote sensing images and the number limitation of multispectral image bands [19]. Even worse, the influence of topography and land use types on soil model information have not been completely taken into account in previous studies [20,21]. Additionally, different soil types have different compositions, structures, and textures, which affect how light is absorbed, scattered, and transmitted as well as how heavy metals are adsorbted, moved, and released from the soil [22].

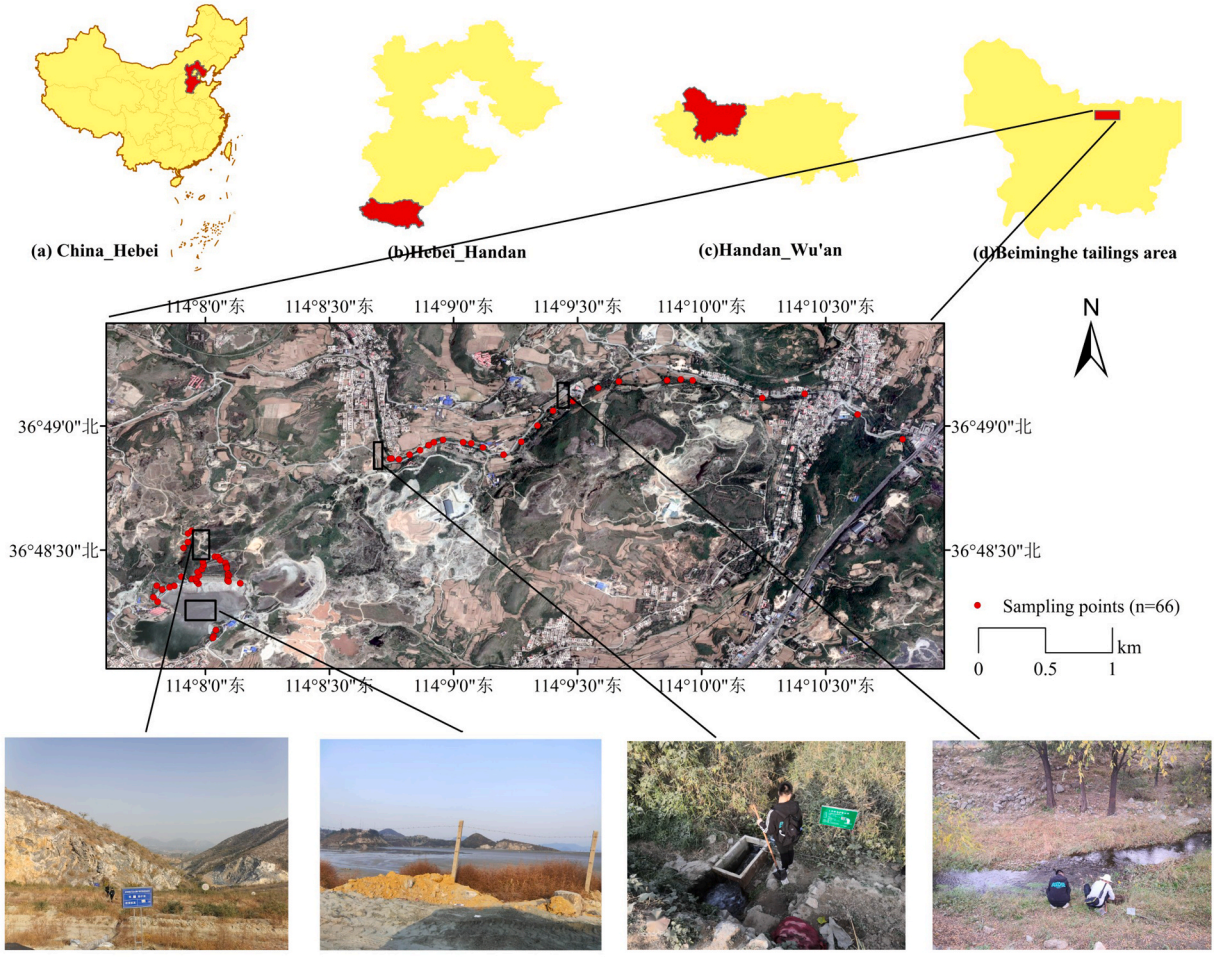

Fig. 1. General situation of the Beiminghe tailings area and sampling sites distribution.

Consequently, it is challenging to connect contamination levels with hyper or multispectral responses.

The terrain in Beiminghe metal tailings area is complex and undulant, and the mines are densely distributed. The regional ecological has been severely damaged by long-term continuous mining, and the heavy metal content in the soil has far exceeded the standard, which seriously threatens the life and health of local residents [23]. Due to the limits of the available study on soil heavy metal pollution in mining sites, the Beiminghe tailings area have been chosen as our investigation target. The heavy metal spectral response characteristics were identified through the correlation analysis of heavy metal content and transformation spectrum, and the direct inversion model of Cu, Cd, Cr, Ni, Pb, Zn, As, and Hg have been established, which primarily composed of three parts: (1) SMLR and PLSR inversion models based on the characteristic bands of measured soil hyperspectral and its transformation; (2) BPNN inversion models based on denoised spectral *R*; (3) SMLR and PLSR inversion models based on the spectral of simulated Landsat8-OLI multispectral and its associated vegetation index. The study investigated the optimal combination of spectral transformation and inversion models for Cu, Cd, Cr, Ni, Pb, Zn, As, and Hg. Then the models of each kind soil heavy metals based on simulated Landsat8-OLI spectral characters were evaluated, and based on the optimal model and Landsat8 images, the corresponding soil heavy metal multispectral pollution mapping was realized. It is envisioned that these initiatives will be able to offer technical support and reference for local soil heavy metal monitoring and soil environmental management.

2. Materials and methods

2.1. Study area

Beiminghe metal tailings area (113°45′–114°22′ E and 36°28′–37°01′ N) is located in the middle section of the eastern foothills of Taihang Mountains, in Wu'an City, Hebei Province (Fig. 1). The average annual temperature is 11 °C–13.5 °C and the precipitation variations in seasonal. This metal tailings area is approximately 20.95 Km^2 with ravines and rugged terrain, and the difference in height is 93 m. The ore body in this area belongs to the contact metasomatism iron deposit of Ordovician limestone and Yanshanian diorite, and the metallic minerals in the ore are mainly magnetite and pyrite. It has a significant impact on the local economy and ecosystem as a contemporary large-scale mine that integrates mining and dressing. The mining area has been continuously mined for 20 years, which has brought significant economic benefits to the local community. However, tailings and wastewater have been discharged casually, which not only takes up farmland but also seriously pollutes the environment. Harmful substances in the tailings and ore are blown by the wind, and nearby villagers suffered from skin diseases. More seriously, the piled-up tailings dumped into the riverchannel pose a serious threat to the river's flood discharge. For these reasons, Beiminghe tailings area is an important soil environment research area. Therefore, high accuracy inversion models are necessary to established to promptly and effectively monitor soil heavy metal contamination, so as to facilitate the prevention and management of local soil heavy metal pollution.

2.2. Data acquisition

2.2.1. Soil sample collection and preparation

By viewing the Google's high-definition satellite images and prospecting on the site, a strip sampling belt was set up along the tailing accumulation step dam and its sewage discharge direction. The field surveys we conducted from mid to late October 2020, and 66 soil samples were collected (Fig. 1). What needs to be noted is that the tailings wastewater and residue in the drainage channel section have little impact on the surrounding soil environment due to the isolation effect of the gutter, so the sampling points are only laid in the tailing accumulation step dam at the top of the channel and the downstream discharge section. 1∼1.5 kg of topsoil (0–20 cm in depth) was excavated at each sampling point with weeds and larger gravels removed. And GPS positioning was performed simultaneously with the sampling. Soil properties, land use types, and vegetation cover conditions around sampling points were recorded in detail. And when necessary, the hand-held GPS camera was used to take pictures of the sampling site to record environmental information. Finally, these samples were packed into labeled fresh-keeping bags and brought back to the laboratory. The soil were naturally air-dried by spreading them in a ventilated and dry place for grinding. Then, the soil was screened via a 200-mesh sieve to further determine heavy metal content and soil spectrum.

2.2.2. Chemical detection of soil heavy metal content

The soil specimens were treated by the microwave digestion system with an appropriate acid solution (HNO3-HCl-HF-HC104). Then the content of Cu, Cd, Cr, Ni, Pb, Zn, and As were determined by inductively coupled plasma atomic emission spectrometry (ICP-AES). Due to low boiling point and volatile nature of Hg, the results of conventional digestion measurements will be low. Therefore, the DMA-80 automatic mercury analyzer was used to directly determine the Hg content of solid soil samples. This method does not require sample pretreatment and has no reagent contamination [24]. The detection limits of all elements are less than 0.05 mg kg^{-1}, and the quality was controlled by blank samples and spiked recovered samples.

2.2.3. Soil spectra measurement

The soil spectrum was measured by the FiledSpec 4 portable ground object spectrometer produced by the ASD company of America. The spectrometer probe consists of a 512-element array PDA detector and two independent InGaAs detectors and detectable wavelength range is 350–2500 nm. The spectrometer is easy to operate with powerful software package. It is suitable for various applications in remote sensing measurement, crop monitoring, forest research, industrial lighting measurement, oceanographic research and mineral exploration [25]. Spectral measurements were performed in a simulated darkroom with interior spectral measurement

conditions are as follows: 1000 W halogen lamp was the only light source, 30° between the illumination and vertical directions, and 30 cm distance between light source and soil. The soil sample was placed in a black and non-metallic container with a diameter of 10 cm and a depth of approximately 1 cm. Spectral reflectance data was gathered by irradiating the surface of the material vertically at a distance of 15 cm with the instrument's high-density reflection probe. The equipment was preheated for 30 min prior to data collection, and the whiteboard was optimized before the experiment and every 3 samples were optimized during the experiment. With the spectrum range of 350~2500 nm and the spectral resolution of 1 nm, 12 spectral curves were gathered for each soil sample. The obviously incorrect spectral curve was identified and removed using the software View Spec Pro, and then the average of spectral was calculated as the actual spectrum of the soil sample.

2.3. Spectral smoothing and transformation

Due to the influence of instrument internal structure, human errors, and external environment factors, pre-processing of spectra is required to reduce the interference from illumination and background noise. This allows for the decomposition of mixed spectral features, enhancing spectral sensitivity, and thereby improving the accuracy of the spectral model [13,14]. In this study, the soil reflectance spectra were initially modified using the splice correction tool of the View SpecPro software [26]. Additionally, Savitzky-Golay smoothing was employed to reduce noise and enhance the smoothness of the spectra, efficiently preserving the original characteristics of the data, ensuring the accuracy and reliability of the model [27]. As the content of heavy metals in soil occurs at a microscopic level, their individual reflectance spectra may not distinctly highlight the response of heavy metals. Spectral mathematical transformations, on the other hand, typically increase the signal-to-noise ratio, thereby enhancing the target spectral information and highlighting soil spectral features. Therefore, the spectrum after splicing correction as well as smoothed and denoised was employed as the original reflection spectrum *R*, and eight spectral transformation indices such as the first derivative and second derivative of the spectrum, the inverse-log of the spectrum, the first derivative of the inverse-log of the spectrum, the second derivative of the inverse-log of the spectrum, the square root of the spectrum, and the continuum removal of the spectrum, which were represented by *FD(R)*, *SD(R)*, *LR*, *FD (LR)*, *SD (LR)*, *Sqrt (R)*, and *CR*, respectively, have been obtained respectively.

2.4. Landsat8-OLI image data acquisition and preprocessing

The Landsat 8-OLI remote sensing images with no cloud on October 29, 2020 were selected to integrate with the field measured data. The spectral response function of the Landsat8-OLI sensor was obtained from the ENVI software [28], according to which the soil spectrum was resampled, and the spectral reflectance consistent with each band of the sensor were acquired. The DEM data are ASTER GDEM provided by the Chinese Academy of Sciences Data Center, with a resolution of 30 m. The original Landsat 8 OLI Level 1T data product has already undergone systematic radiometric and geometric correction, but terrain correction was not applied. Therefore, in this study, the original Landsat 8 OLI imagery was first subjected to radiometric calibration and atmospheric correction. Subsequently, DEM data were utilized for terrain correction. NDVI were calculated by the pixel dichotomy model, and then the Modified Normalized Difference Water Index (MNDWI), the Difference Vegetation Index (DVI), the Enhanced Vegetation Index (EVI), the Clay Mineral Ratio (CMR), as well as characteristic components like Greenness, Brightness, and Wetness, were then calculated to reflect soil characteristics. The formulaes and references of NDVI, MNDWI, DVI, EVI, CMR, Greenness, Brightness, and Wetness were listed in Table 1. And B2, B3, B4, B5, B6 and B7 mentioned in the calculations in Table 1 correspond to the simulated bands of Blue, Green, Red, NIR, SWIR1, and SWIR2 of the Landsat 8-OLI image. Therefore, the simulated bands and indices were collectively refer to the simulated multispectral spectrum *R*_Landsat 8-OLI.

2.5. Model inversion and verification methods

2.5.1. Model approach

The accuracy of the soil heavy metal spectral inversion models is influenced by the soil composition, the rationality of the sampling data, and the selection of the inversion methodology [35,36]. It is difficult to estimate heavy metal elements in soil by direct physical model. The direct inversion method based on the correlation between heavy metals and soil hyperspectral is primarily applicable to regions with severe heavy metal pollution [37]. The Stepwise Multiple Linear Regression (SMLR) is a method that variables are selected into the regression equation based on *F* statistics, and the variance contribution values of all variables are considered [11,38].

Table 1
Spectral index definition of the Landsat 8-OLI image.

Index	Definition
MNDWI	$(B3-B6)/(B3+B6)$ [29]
DVI	$B5-B4$ [30]
CMR	$B6/B7$ [31]
EVI	$2.5\times(B5-B4)/(B5+6\times B4-7.5\times B2+1)$ [32]
NDVI	$(B5-B4)/(B5+B4)$ [33]
Greenness	$-0.294\times B2-0.243\times B3-0.5424\times B4+0.7276\times B5+0.0713\times B6-0.1608\times B7$ [34]
Brightness	$0.3029\times B2+0.2786\times B3+0.4733\times B4+0.5599\times B5+0.508\times B6+0.1782\times B7$ [34]
Wetness	$0.1511\times B2+0.1973\times B3+0.3283\times B4+0.3407\times B5-0.7117\times B6-0.4559\times B7$ [34]

Partial least squares regression (PLSR) integrates the advantages of principal component analysis, canonical correlation analysis and linear regression analysis. It is typically employed for regression modeling in the case of the number of samples to be fewer than the number of variables. It is widely used in the field of hyperspectral inversion since it can eliminate the influence of multiple correlations [39–42]. The Backpropagation neural network (BPNN), as a multi-layer feedforward neural network trained through error back-propagation, fundamentally utilizes the sum of squared network errors as the objective function. It employs the gradient descent method for training, thereby determining the objective function at its minimum error state [43]. Due to its sophisticated pattern classification capability and powerful multidimensional feature mapping ability, the BPNN has been widely applied in various fields [44]. Therefore, in this study SMLR, PLSR and BPNN were employed to establish predictive models for soil heavy metal content.

The empirical formula (Eq. (1)) was adopted to determine the number of hidden layer nodes in BP neural network, which ranges from 5 to 14 [45]. The number of hidden layer nodes was finally determined through repeated experiments, ensuring that the number of hidden layer nodes was less than the number of training samples [46]. *Tansig* and *purelin* were used as the transfer function of the hidden layer and the output layer respectively, while the *trainlm* based on Levenberg-Marquardt method was used for training which was according to numerical optimization theory.

$$N = \sqrt{n+m} + a \tag{1}$$

where N is the number of hidden layer nodes, n is the number of input nodes, which is equal to the number of independent variables, that means $n = 15$. m is the number of output nodes, which is equal to the number of dependent variables, that is, $m = 1$, a is a constant from 1 to 10.

2.5.2. Model verification

The evaluation metrics for model performance primarily include the accuracy, efficacy and stability of the model [47]. The coefficient of determination R^2 provides insight into the level of fit and stability of the model. However, there are limitations when used to assess the fitting model, which ignores the statistical significance that R^2 grows higher with the increase of variables [48]. In order to scientifically evaluate the model performance for heavy metals, therefore, the adjusted coefficient of determination $Adjust\_R^2$ was calculated to evaluate the stability and precision of the inversion models. The closer the $Adjust\_R^2$ value is to 1, indicates a stronger degree of fit between the measured and predicted values, signifying higher accuracy in the inversion and greater stability of the model. The assessment indicator known as Root Mean Square Error (*RMSE*) was used to measure the deviation between the predicted value and the actual value and it is sensitive to outliers. *RMSE* exhibits the predictive ability of the model, the smaller the value, the better of the predictive effect of the model. The model evaluation index formula are as follows (Eq. (2), Eq. (3), Eq. (4)).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - S_i)^2} \tag{2}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(S_i - \bar{S_i}\right)^2 * \left(P_i - \bar{P_i}\right)^2}{\sqrt{\sum_{i=1}^{n}\left(S_i - \bar{S_i}\right)^2 * \sum_{i=1}^{n}\left(P_i - \bar{P_i}\right)^2}} \tag{3}$$

$$Adjust\_R^2 = 1 - \frac{\left(1 - R^2\right)(n-1)}{n-p-1} \tag{4}$$

where S_i and P_i are the measured and predicted values of soil heavy metal content, $\bar{S_i}$ and $\bar{P_i}$ are the average of the observed values and the mean of the predicted values, respectively. n is the number of samples, p is the number of selected independent variables of the model.

Table 2
The statistical characteristics of soil samples in Beiminghe tailings area.

Element	Minimum (mg/kg)	Maximum (mg/kg)	Mean (mg/kg)	Standard Deviation (mg/kg)	Skewness	Kurtosis	Coefficient of variation	Background value (mg/kg)	Hyper background rate
Cu	28.296	4218.23	1034.47	1112.293	1.202	0.881	1.075	21.8	100.00%
Cd	0	10	2.697	1.909	1.682	3.403	0.708	0.094	93.94%
Cr	21.638	160	64.91	32.133	1.151	0.794	0.495	68.3	34.85%
Ni	9.987	511.667	126.586	131.728	1.323	0.965	1.041	30.8	62.12%
Pb	3.327	823.333	131.381	162.94	1.897	4.59	1.24	21.5	57.58%
Zn	24.983	266.134	73.652	46.479	1.936	4.613	0.631	78.4	37.88%
As	24.983	101.396	43.988	18.983	1.43	0.98	0.432	13.6	100.00%
Hg	0.003	0.073	0.019	0.016	1.483	2.272	0.828	0.036	13.64%

3. Results and discussion

3.1. Statistical analysis of soil heavy metal content

Table 2 shows the statistical characteristics of soil heavy metals content. From the table, it is evident that the average concentrations of Cu, Cd, Pb, Ni and As far exceed the background values observed in Hebei Province in1990, making it representative of a heavily multi-source polluted region for heavy metals [49]. The hyper background rate of Cu and As samples reached up to 100%, and Hg with the lowest ratio of 13.64%. Theoretically, the skewness and kurtosis of the standard normal distribution are both 0, but the actual data is not an absolute normal distribution. It is considered that the data can be basically regarded as normally distributed if the absolute value of the data kurtosis is less than 10 and the absolute value of the skewness is less than 3 [50]. The accumulation of soil components in environmental geochemistry typically coincides with an increase in variability. As a parameter representing the fluctuation features of the soil environment, the coefficient of variation can be employed to some extent to describe the accumulation status of soil heavy metals [51]. In this study, all elements exhibited approximately normal distributions and their spatial distribution demonstrated significant variability and heterogeneity, with coefficients of variation exceeding 30%. It is evident that soil metal elements have been enriched to varied degrees under the influence of long-term mining activities and other anthropogenic factors. Although this matter has been acknowledged for a while, it has not yet been fully resolved.

3.2. Characteristics of soil spectral

Firstly, it is essential to apply spectral data smoothing to minimize noise interference and enhance the signal-to-noise ratio. The approach of Savitzky-Golay convolution was performed for spectral smoothing. The comparison of the reflectance spectral curves of soil samples before and after splice correction and smoothing treatment are shown in Fig. 2(a) and (b). It is visible that the spectral curves are smoother at wavelengths of 1000 nm and 1800 nm, and the original spectral characteristics were preserved. The maximum spectral reflectance of dark soil samples is 0.6, whereas the maximum reflectance of light soil samples is 0.4, indicating a consistent spectral trend between the two. Generally, excluding the 900 nm band, the absorption characteristics of the soil spectrum in the visible and near-infrared bands were mainly caused by the electronic transition of metal ions such as Fe^{2+}, Fe^{3+}, Cu^{2+}, and Mn^{3+}, as well the frequency doubling and harmonic frequency generated by the bending vibration of molecules, such as –OH, CO_3^{2-}, OH-and NH^{4+} [52–54]. For example, there are obvious curve absorption characteristics around 1400, 1900 and 2200 nm, which are mainly related to the OH− contained in iron oxides and kaolin-like clay minerals [55]. While absorption peak near 2450 nm is generated by the vibration of CO_3^{2-} groups in soil carbonates, and the vicinity of 1450 nm was an evident absorption valley caused by the stretching vibration of water molecular –OH in soil silicate minerals [47]. A deep “V" shape can be detected in the spectrum curve between 1875 and 2130 nm as a result of the metal hydroxyl group stretching [56]. There are distinct characteristic absorption bands and reflection peaks around 1350 nm and 2450 nm, which significantly affected by atmospheric water vapor absorption [57].

3.3. Correlation analysis between heavy metal content and soil spectrum

Spectral preprocessing is crucial in spectral analysis as it can draw attention to spectral feature bands and reduce spectral noise [13]. Thus, eight spectral transformation indices were calculated for spectral feature screening and model inversion. Consequently, the correlation analysis between soil heavy metals and spectral transformations were performed to further determined the soil spectral characteristics, and the correlation coefficients *r* can be seen from Fig. 3(a–h). Nearly all of the correlation coefficients between the

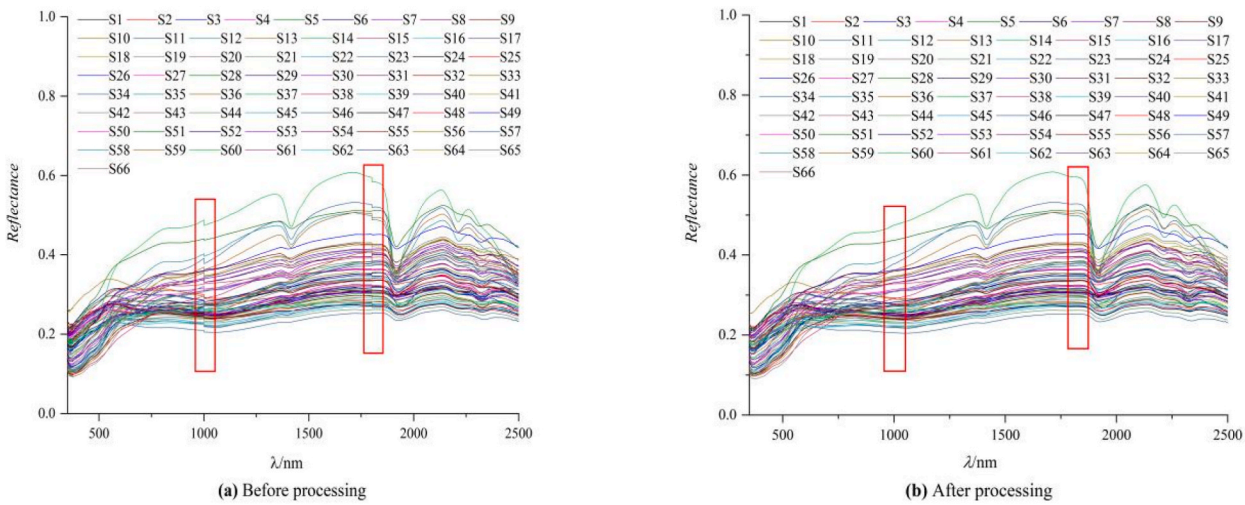

Fig. 2. Spectral curves of soil samples before and after treatment.

concentration of soil heavy metal elements and their corresponding spectral index curves were below 0.65. The peak positions of the correlation coefficients exhibit a consistent pattern, however the extreme and significant correlation bands were distinct. Compared to the original spectral *R*, the transformed spectra exhibited a significant improvement in correlation with heavy metals, effectively highlighting the spectral response characteristics. As compared to the first derivative spectrum, which showed a more moderate correlation variation, the second derivative spectrum showed an intense and compact correlation change across the band range. The correlation between spectral index of *R*, *LR*, *Sqrt(R)* and soil heavy metals present a horizontal wine glass shape, with obvious trend consistency. Meanwhile the correlation between the first derivative spectrum, and *CR* with heavy metals have clear characteristics in the range of 350~1500 nm with a gently change trend, while the correlation changes violently at 1500–2500 nm. As showed a relatively low correlation with the soil spectra, while Hg exhibited a favorable response and demonstrated the highest correlation with the soil spectra, followed by Cu, Cd, Cr, Ni, Pb and Zn. In the wavelength range of 350~460 nm and 700–1000 nm, the concentrations of Zn and Hg showed an increasing trend in correlation with the spectral *R*, *Sqrt(R)* and *CR*. Specifically, there was a negative correlation observed in 350~460 nm and a positive correlation in 700~1000 nm. However, the correlation between the concentrations of Zn and Hg and the spectral *LR* displayed an opposite trend compared to the correlation between Zn and Hg concentrations and spectral *R* and *CR* Furthermore, in the wavelength range of 700~2500 nm, the concentrations of soil heavy metals, including Cu, Cd, Cr, Ni, Pb, and As, were found to be negatively correlated with the spectral transformation *R* and *CR*, and positively correlated with the spectral *LR*, and both correlations exhibited an increasing trend. The correlation coefficients are significantly correlated near 460 nm, which is mainly due to the weak absorption peak of soil manganese oxides.

The maximum correlation and response wavelengths for the concentration of soil heavy metals and different spectral are showed in Table 3. Results showed that spectral transformation effectively improves the spectral response of soil heavy metals, especially the second derivative spectral, which might help extract relevant information for the rare components [58]. After the continuum removed treatment, the soil spectral absorption characteristics were highlighted, and the absolute values of the maximum correlation

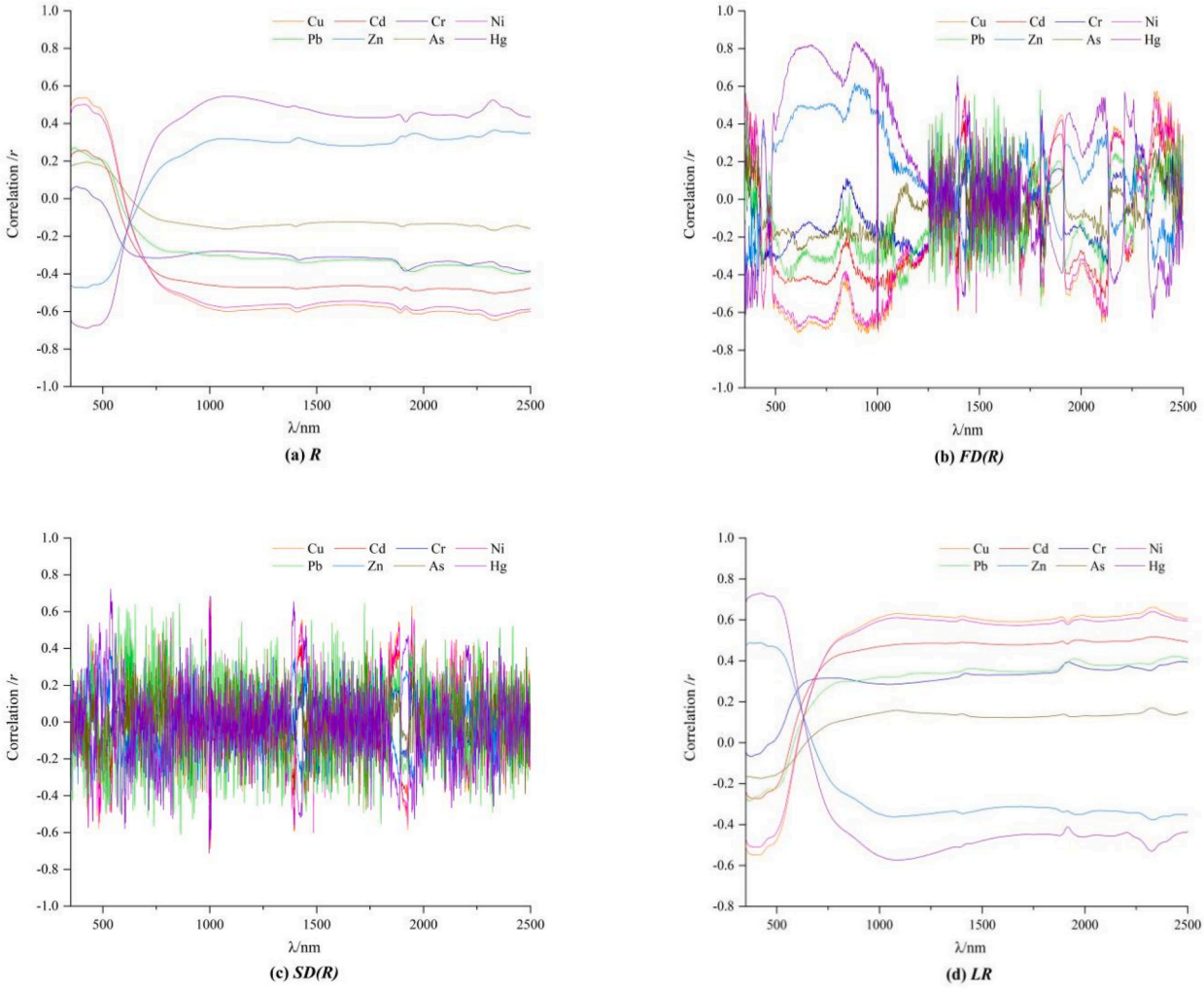

Fig. 3. Correlation coefficient between soil heavy metals contents and spectral transformation,

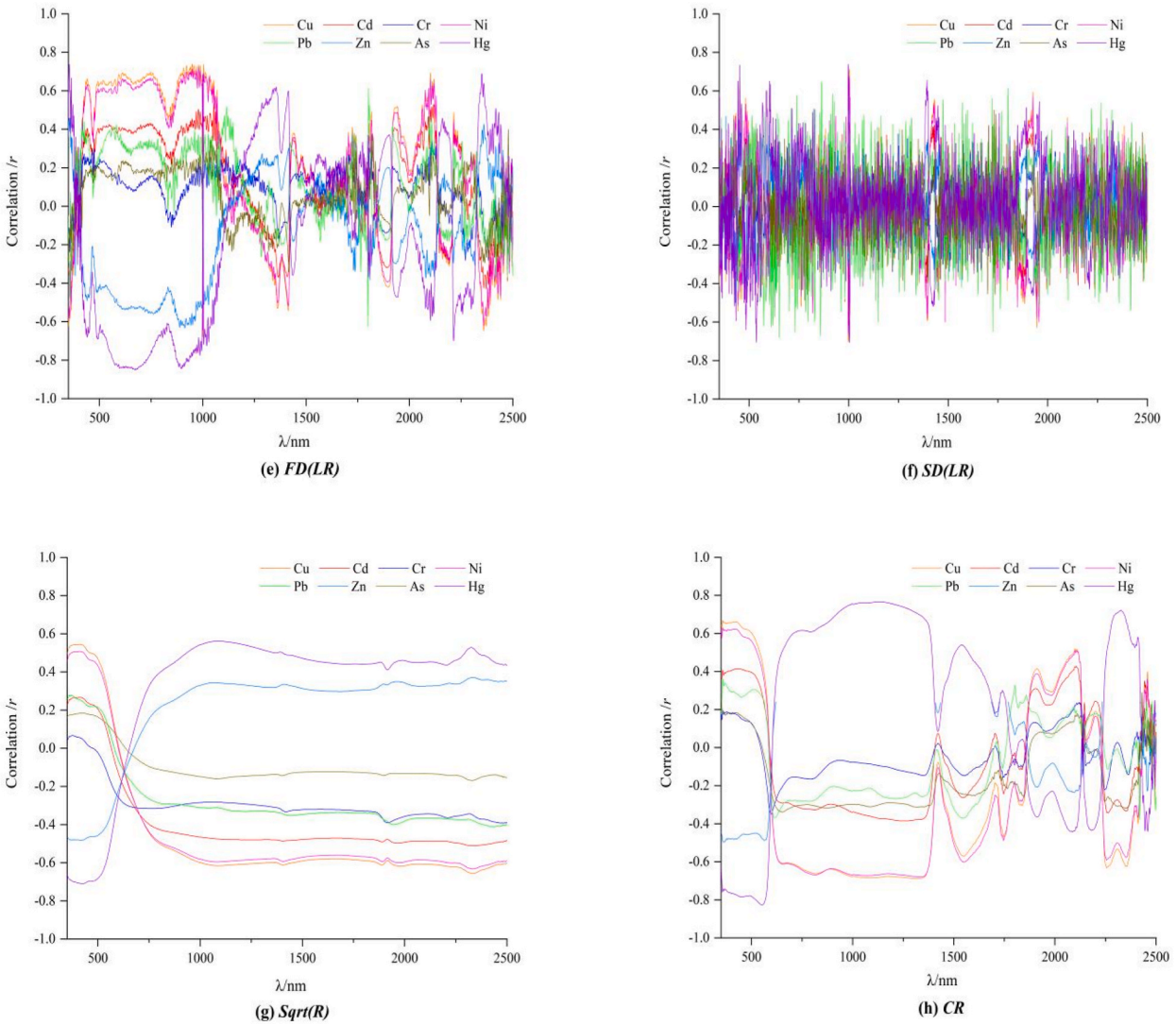

Fig. 3. (*continued*).

coefficients of Hg and Cu with the spectra in visible and near-infrared band ranges reached to 0.827 and 0.688, respectively. The maximum correlation response bands of Cu, Cd, and Ni with the spectrum *R*, *LR*, and *Sqrt(R)* were primarily distributed in the wavelength range b2326~b2346, with an overlap, indicating that there were similar spectral response characteristics. The maximum correlations coefficient of Cu, Ni and Hg with the spectrums in the were all greater than 0.6, while As and Cr were all lower than 0.5. The spectrum *FD(R)*, *SD(R)*, and *SD(LR)* have single maximum correlation band with soil heavy metals, while the other transformed spectrum have maximum correlation group bands with soil heavy metals.

The correlation between soil heavy metals content and simulated multispectral *R*_landsat8-OLI, as well its characteristic vegetation index can be seen from Table 4. There were considerable and strong connections between the simulated spectral variables and the concentration of soil heavy metals, with the exception of CMR and the simulated B3 band. Cu, Ni, Zn, and Hg all demonstrated correlation coefficients that were greater than 0.5 with MNDWI, DVI, EVI, NDVI, and greenness, from which Hg has the strongest correlation with these variables and the correlation coefficient reached to 0.81. EVI, NDVI, and B7 were the greatest correlation index between soil heavy metal concentrations and simulated spectral *R*_landsat8-OLI. With the exception of As and Zn, all other soil heavy metals exhibited significant correlations with B5, B6, B7, and Wetness. Specifically, Hg showed a positive and significant correlation with B5, B6, and B7, while displaying a negative and significant correlation with Wetness. Conversely, Cu, Cd, Cr, Ni, and Pb showed significant negative correlations with B5, B6, and B7, while demonstrating a significant negative correlation with Wetness.

Table 3
Maximum correlation coefficients and response wavelengths for heavy metals contents and various spectral indices.

Maximum Correlation	*R*	*FD(R)*	*SD(R)*	*LR*	*FD(LR)*	*SD(LR)*	*Sqrt(R)*	*CR*	*R*_Landsat8-OLI
Cu	−0.647** b2328~b2341	−0.713** b949	−0.688** b1002	0.663** b2329~b2332	0.739** b948	0.713** b1002	−0.656** b2330~b2333	−0.688** b1301~b1327	−0.703** EVI
Cd	−0.504** b2327~b2339	−0.525** b2116	0.556** b1945	0.518** b2326~b2346	0.533** b1053, b1054	−0.569** b1945	−0.511** b2325~b2346	0.427** b2105,b2106	−0.485** B7
Cr	−0.388** b2480	−0.354** b474	0.427** b468	0.398** b2480	−0.343** b1800	−0.401** b468	−0.393** b2480	−0.348** b596~b599	−0.343** B7
Ni	−0.623** b2328~b2341	−0.681** b948	−0.641** b1002	0.641** b2328~b2333	0.723** b979	0.673** b1003	−0.633** b2330~b2333	−0.679** b1332~b1343	−0.679** EVI
Pb	−0.4** b2416,b2417,b2435~b2437	0.58** b1799	0.644** b858	0.423** b2436	−0.625** b1799	−0.679** b651	−0.412** b2436	−0.376** b2417	−0.358** B7
Zn	−0.475** b414	0.617** b888	0.511** b430	0.488** b367~b370, b412~b416, b422,b423	−0.633** b913	−0.536** b1003	−0.482** b412~b416, b421~b423	0.541** b998,b999	0.551** NDVI
As	0.194 b417~b433	−0.454** b1697	0.485** b779	−0.174 b2329	0.398** b2478	0.501** b1731	0.185 b420~b429, b431	−0.339** b654	−0.249* EVI
Hg	−0.689** b425	0.838** b896	0.723** b537	0.73** b422~b427	−0.85** b671, b675	0.739** b997	−0.71** b417~b430	−0.827** b550~b554	0.81** NDVI

Attention: **. At the level of 0.01 (two tails), the correlation is significant; *. At the level of 0.05 (two tails), the correlation is significant.

Table 4
The correlation coefficients between soil heavy metals contents and simulated landsat8-OLI spectral variables.

Correlation	Cu	Cd	Cr	Ni	Pb	Zn	As	Hg
B1	0.513**	0.238	0.025	0.477**	0.220	−0.464**	0.190	−.682**
B2	0.477**	0.210	−0.011	0.441**	0.208	−0.452**	0.177	−0.668**
B3	0.219	0.011	−0.174	0.193	0.063	−0.345**	0.106	−0.469**
B4	−0.209	−0.269*	−0.305*	−0.216	−0.164	−0.062	−0.032	0.012
B5	−0.520**	−0.436**	−0.303*	−0.507**	−0.284*	0.229	−0.132	0.430**
B6	−0.566**	−0.465**	−0.310*	−0.546**	−0.326**	0.282*	−0.125	0.440**
B7	−0.610**	−0.485**	−0.343**	−0.588**	−0.358**	0.327**	−0.142	0.450**
MNDWI	0.661**	0.422**	0.171	0.628**	0.348**	−0.503**	0.183	−0.723**
DVI	−0.672**	−0.430**	−0.132	−0.638**	−0.296*	0.514**	−0.199	0.784**
CMR	0.052	0.001	0.106	0.046	0.107	−0.092	0.063	0.102
EVI	−0.703**	−0.426**	−0.141	−0.679**	−0.291*	0.545**	−0.249*	0.773**
NDVI	−0.669**	−0.406**	−0.104	−0.637**	−0.296*	0.551**	−0.192	0.810**
Greenness	−0.576**	−0.305*	−0.019	−0.540**	−0.240	0.505**	−0.198	0.784**
Brightness	−0.395**	−0.388**	−0.330**	−0.390**	−0.249*	0.103	−0.073	0.206
Wetness	0.639**	0.458**	0.252*	0.609**	0.353**	−.417**	0.157	−.573**

Attention: **. At the level of 0.01 (two tails), the correlation is significant; *. At the level of 0.05 (two tails), the correlation is significant.

3.4. The establishment and analysis of the spectral inversion model

3.4.1. Feature selection and model construction

According to the statistical characteristics analysis that the long-term mining has significantly enriched the heavy metal concentration in the soil of the study area. Therefore, the direct inversion method was adopted in this paper, which was in line with objective facts. Direct inversion models for soil heavy metals can be established based on the correlations between soil heavy metals and spectra [59,60]. The top 15 significant band groups with the highest correlation at levels 0.05 and 0.01 were chosen based on the correlation between the concentration of soil heavy metals and different spectral indicators. Average values of each band group were calculated as characteristic variables of each spectral indexes, and the significant simulated spectral variables were identified as the characteristic variables of the spectral simulation model. According to the order of heavy metal content from small to large, one sample was obtained as a verification sample for every three samples, and the data was finally split into a traing set with 44 samples and a testing set with 22 samples. The spectral features and heavy metals content were taken as independent variables and dependent variables respectively, thus the SMLR, PLSR and BPNN inversion models of soil heavy metals were established, which include the SMLR and PLSR models based on spectral *R* and spectral transformation, BPNN models based on the spectral *R*, and the SMLR and PLSR models base on simulate multispectral *R*_landsat8-OLI. For the purpose of facilitating model comparison, we identify the models using symbols that correspond to their spectral.

3.4.2. Modeling evaluation

Fig. 4(a–f) shows the *Adjust*_R^2 comparison of the SMLR training models based on each spectral transformation. It indicates that the spectral transformations have enhanced the correlation between the spectra and soil heavy metals, and consequently the prediction ability of the models based on spectral transformation have been effectively improved. With the exception of Zn, the second derivative models of soil heavy metals show better performance to some extent compared to the first derivative models. When the two kinds of second-order derivative SMLR training models were compared, it demonstrated that, except for Ni, the training *Adjust*_R^2 of *SD(LR)* models of soil heavy metals were higer than those of the *SD(R)* training models. With the exception of Cr, the stability of the first-order derivative training models is better than the continuum removal models. Additionally, for most soil heavy metals, the *Adjust*_R^2 of *CR* training models were higher than those of the *LR* training models. By comparing the *Adjust*_R^2 of the training models based on *R* and *R*_Landsat8-OLI, it was observed that the former exhibited higher *Adjust*_R^2 values for Cu, Zn, and Hg. In conclusion, the recommended order of indices for establishing spectral models of heavy metals with SMLR model is as follows: *SD(LR)* > *SD(R)* > *FD(LR)* > *FD(R)* > *CR* > *LR* > *Sqrt (R)* > *R*.

The larger of the *Adjust*_R^2 and the smaller the of *RMSE*, the better of the model, according which the statistical information of the optimal SMLR models of soil heavy metals in Table 5 were obtained. From the table, it can be observed that, except for Zn, the optimal SMLR models with smaller *RMSE* for soil heavy metals are all based on *SD(LR)*, indicating that this spectral SMLR model exhibits good predictive capability. The optimal SMLR models show *Adjust*_R^2 values within the range of 0.395–0.795 for training and 0.247–0.544 for testing, respectively. Among them, both the training and testing *Adjust*_R^2 for Hg exceed 0.5. The optimal training models of soil heavy metals ranked from high to low according to the training *Adjust*_R^2 are Hg > Pb > Cu > Ni > As > Zn > Cd > Cr. The input variables and their band composition of the optimal SMLR model for soil heavy metals are shown in Table 6. The arrangement order of the input variables represents the degree of correlation between their corresponding composition characteristic bands and heavy metal content, and the higher the ranking, the greater the correlation. It can be seen that the spectral response characteristics of soil heavy metals are mainly concentrated in the wavelength of b440~b540, b610~b660, b760~b890, b950~b1010, b1140~b1290, b1690~b1740 and b1940~b1960, which are the key spectral features of heavy metal inversion models. Pb exhibits the strongest spectral response to soil spectra with the largest number bands that have been chosen, whereas Zn exhibits the weakest spectral response to soil spectra with only one band has been selected.

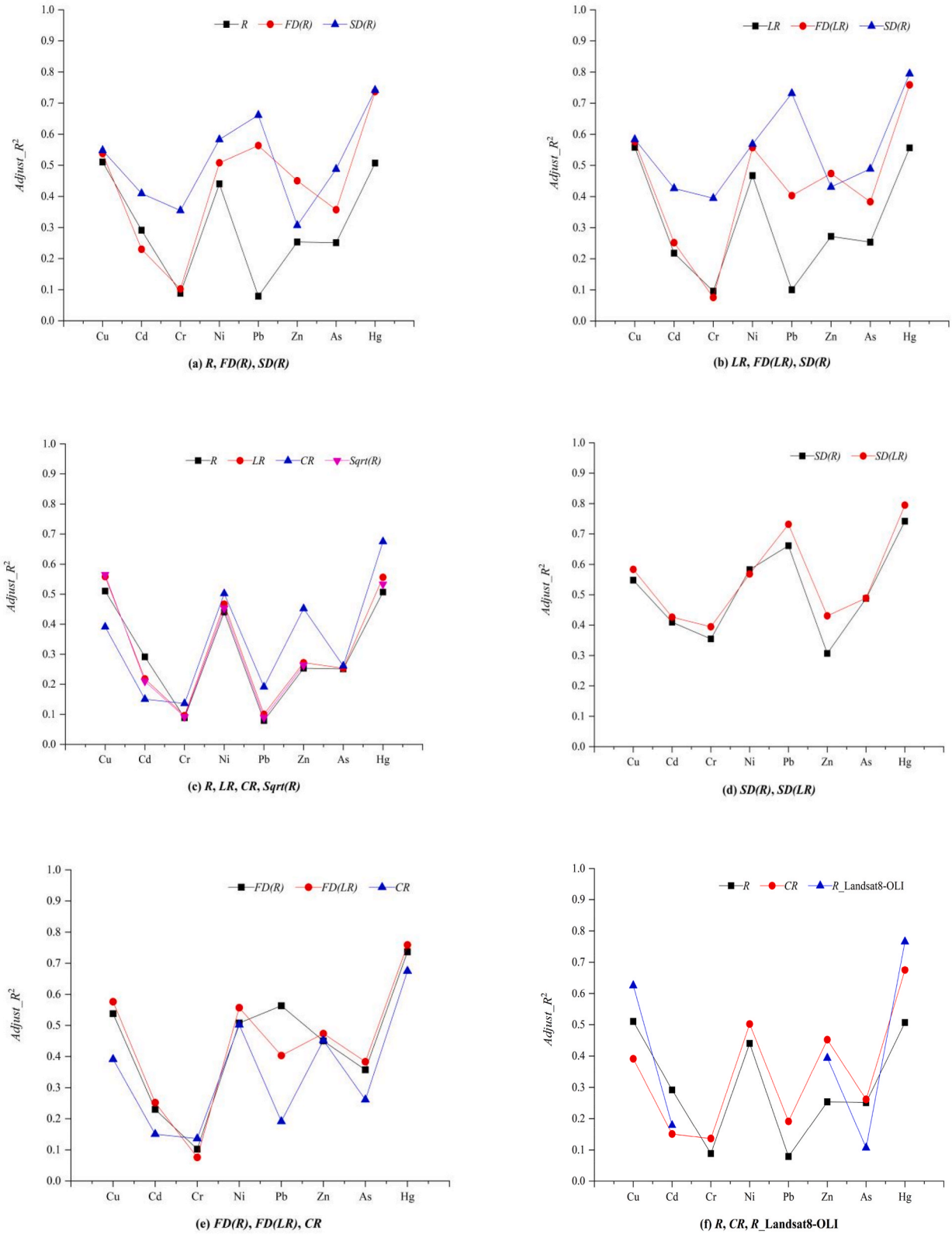

Fig. 4. Comparison of the *Adjust_R*^{2} of SMLR training models based on diffrent spectral index.

Table 5
Information statistics of the optimal SMLR model of soil heavy metals.

Heavy metal element	Model index	the number of the entered variables	Training		Testing	
			Adjust_R^2	RMSEC (mg/kg)	*Adjust_R*^{2}	*RMSEP* (mg/kg)
Cu	*SD(LR)*	2	0.576	6.916	0.484	7.739
Cd	*SD(LR)*	3	0.426	3.303	0.285	1.641
Cr	*SD(LR)*	3	0.395	3.318	0.325	2.564
Ni	*SD(LR)*	2	0.568	8.055	0.383	10.459
Pb	*SD(LR)*	4	0.732	7.093	0.449	12.798
Zn	*FD(LR)*	1	0.474	3.0024	0.247	4.519
As	*SD(LR)*	3	0.488	2.506	0.286	5.754
Hg	*SD(LR)*	3	0.795	0.007	0.544	0.011

Table 6
Feature band statistics of the optimal SMLR model of soil heavy metals.

Heavy metal element	the entered variables	Band composition
Cu	*B*1, *B*15	b_{1002}, b_{1955}
Cd	*B*5, *B*7, *B*8	b_{980}, b_{443}, b_{820}
Cr	*B*14, *B*9, *B*5	b_{1738}, b_{980}, b_{1248}
Ni	*B*6, *B*3	$b_{1485,1956}$, b_{997}
Pb	*B*2, *B*1, *B*13, *B*4	b_{858}, b_{651}, b_{796}, b_{1724}
Zn	*B*2	b_{888}
As	*B*6, *B*8, *B*14	$b_{1697,761}$, b_{1735}, b_{1269}
Hg	*B*5, *B*2, *B*6	b_{537}, b_{453}, b_{1002}

Fig. 5(a–d) shows the comparison of the *Adjust_R*^{2} about the PLSR training models based on each spectral transformation. Similar to the conclusion of the heavy metal SMLR inversion models, the models based on derivative spectral indices demonstrated significantly better performance than the original spectral *R*, with the second derivative indices are particularly enhancing the stability and accuracy of the training models. Except for Hg and Cu, the *Adjust_R*^{2} of the second derivative training models are larger than those of the first derivative training models. The findings also demonstrate that, with the exception of Cd and Cr, the *Adjust_R*^{2} of the PLSR training models based on second derivative spectra are all greater than 0.5. Moreover, the PLSR training models based on *FD(LR)* show higher *Adjust_R*^{2} compared to those based on *CR*. It is observed that the model performance is similar by comparing the *Adjust_R*^{2} of the two kinds of second derivative PLSR training and testing models based on *SD(R)* and *SD(LR)*.

Further, Table 7 provides the statistical information of the optimal PLSR models of soil heavy metals. It is discovered that the best PLSR models for soil heavy metals were all second derivative models, with an equal distribution of the two types of second derivative models. The *Adjust_R*^{2} of the optimal PLSR training models are all greater than 0.5, which means a better inversion. When comparing the training *Adjust_R*^{2}, the stability of heavy metal modeling is as follows: Hg > As > Pb > Cu > Ni > Zn > Cd > Cr. The *Adjust_R*^{2} of the testing models are ranged from 0.329 to 0.606, and the testing *Adjust_R*^{2} of Cu, Cr, As and Hg all exceeded 0.5, showing good model prediction ability. The optimal PLSR model expressions of each heavy metal element are listed in Table 8.

The result of BPNN models based on denoised spectral *R* demonstrated that the range of training *Adjust_R*^{2} values is from 0.421 to 0.822, while the range of testing *Adjust_R*^{2} values is from 0.163 to 0.827. The *Adjust_R*^{2} and *RMSE* of the BPNN model for soil heavy metals are shown in Fig. 6(a) and (b). It is clear that the training and testing *Adjust_R*^{2} for Cu, Cd, Ni, and Hg are all greater than 0.5, proving the effectiveness and applicability of the BPNN inversion models. With a maximum difference of 10 mg/kg, the modeling *RMSE* of the BPNN model is relatively similar to the predicted *RMSE*. According to the model evaluation criteria, the performance of heavy metal inversion models was compared, and it is obvious that Hg > Ni > Pb, Cu > Pb, Cd > Cr, and Cd > Zn.

Fig. 7(a) and (b) compares the *Adjust_R*^{2} of training and testing models of soil heavy metals based on the landsat8-OLI simulated multispectral characteristics. Cr, Ni, and Pb were unable to be modeled using SMLR since no characteristic variables could be selected into the SMLR process, while PLSR models for all of the investigated elements were all successfully established. The *Adjust_R*^{2} of PLSR training models of Cu, Ni, Hg are all greater than 0.5 with small *RMSE*, while only the testing *Adjust_R*^{2} of Hg exceeded 0.5. The *Adjust_R*^{2} of SMLR and PLSR training models of Hg are 0.766 and 0.833, respectively, and the testing *Adjust_R*^{2} are 0.513 and 0.553, indicating that the PLSR model based on simulated Landsat8-OLI spectral provides the best prediction for Hg. Formula 5 (Eq. (5)) illustrates how this model is calculated, and the *P*-value test of the variable coefficients in formula 5 showed that the coefficients of the variables have all reached the 0.05 significant level.

$$\begin{aligned} &-0.0067015b_{coastal} + 0.087487b_{blue} - 0.208961b_{green} + 0.339939b_{NIR} - 0.467545b_{SWIR1} \\ &+0.410681b_{SWIR2} - 0.0847886b_{MNDWI} + 0.382773b_{DVI} - 0.57085b_{EVI} + 0.427975b_{NDVI} \\ &+0.0680955b_{greenness} + 0.208389b_{wetness} - 0.0268428 \end{aligned} \quad (5)$$

3.4.3. Comparison of three inversion models

The training and testing *Adjust_R*^{2} of the optimal SMLR, PLSR, and BPNN inversion models were compared on the basis that the *RMSE* of the whole are not significantly different, which was shown in Fig. 8(a) and (b). According to the statistics of the optimal SMLR

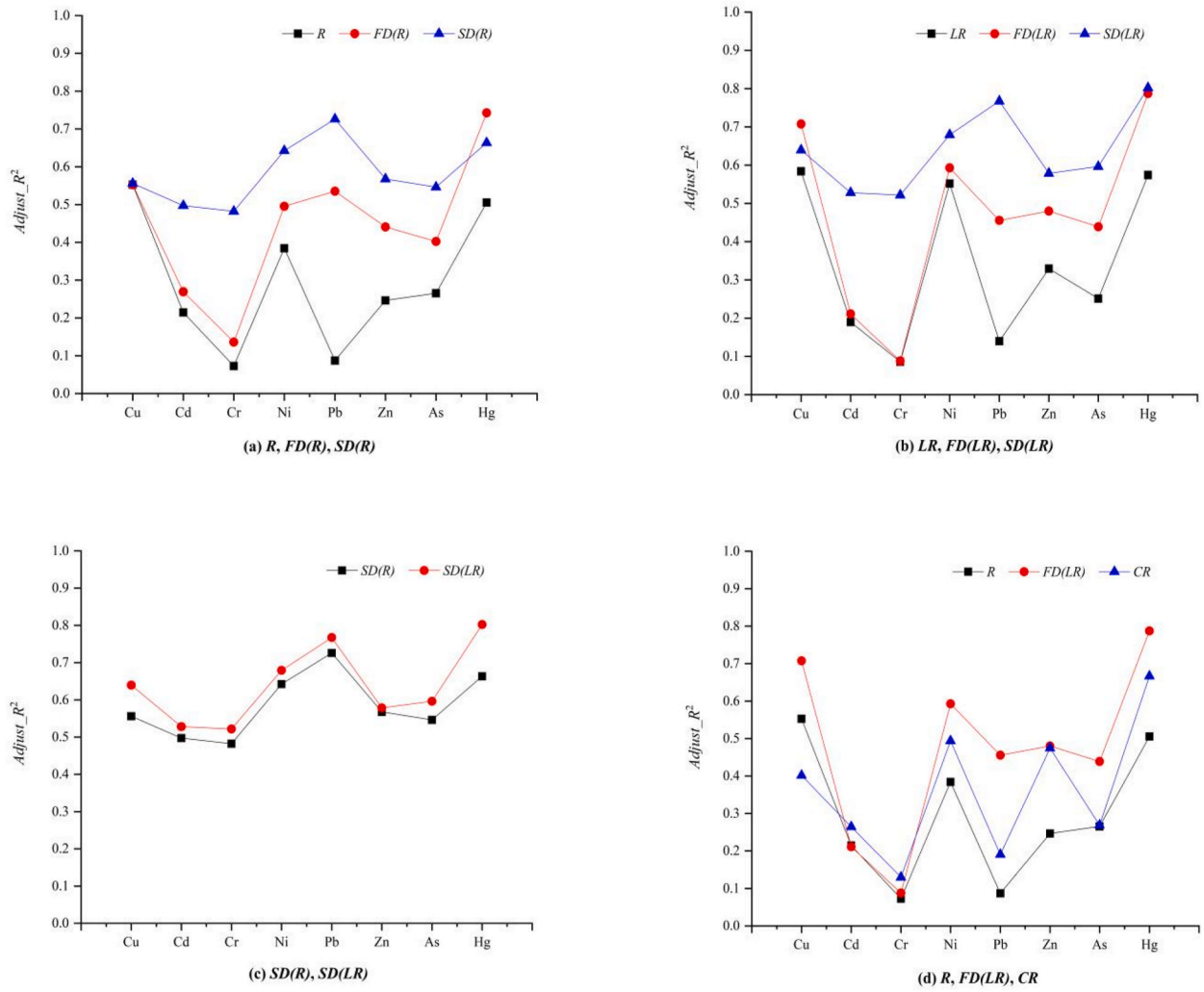

Fig. 5. Comparison of the *Adjust_R*^{2} of PLSR training models based on diffrent spectral index.

Table 7
Information statistics of the optimal PLSR model of soil heavy metals.

Heavy metal element	Model index	the number of principal components	Training		Testing	
			Adjust_R^2	*RMSEC* (mg/kg)	*Adjust_R*^{2}	*RMSEP* (mg/kg)
Cu	*SD(LR)*	3	0.639	12.459	0.508	12.098
Cd	*SD(R)*	4	0.528	3.072	0.329	2.809
Cr	*SD(R)*	4	0.517	7.119	0.606	6.429
Ni	*SD(LR)*	3	0.590	15.174	0.415	15.560
Pb	*SD(LR)*	6	0.767	15.185	0.482	20.478
Zn	*SD(R)*	4	0.564	8.101	0.455	8.534
As	*SD(R)*	3	0.787	4.718	0.583	4.483
Hg	*SD(LR)*	4	0.802	0.021	0.522	0.022

and PLSR models of soil heavy metals, it seems that the PLSR model appears to have better performence than the SMLR model overal. It turns out that the testing *Adjust_R*^{2} were always lower than the training *Adjust_R*^{2} when comparing the two ideal models for soil heavy metals. Models based on the second derivative spectral have better prediction ability than other spectral models. The findings suggested that the BPNN models based on *R* for Cu, Cd, Ni, and Hg are superior to the ideal PLSR model based on spectral transformation. Therefore, the recommended ranking of inversion models for Cu, Cd, Ni, and Hg is: BPNN > PLSR > SMLR. The comparison results also indicate that the BPNN inversion models based on *R* for Cr, Pb, As, and Zn are inferior to the best SMLR models based on spectral transformation. Hence, BPNN < SMLR < PLSR is the suggest order of the inversion models for Cr, Pb, As, and Zn. According to the

Table 8
Expressions of the optimal PLSR models about soil heavy metals.

Heavy metal element	Optimal PLSR model expression
Cd	$Y = 5.62b_{1945} + 52.39b_{819} - 34.50b_{1944} + 23.70b_{1485} - 9.38b_{432} + 28.62b_{443} + 111.05b_{820} + 37.62b_{997} + 78.99B_{1451,980} - 0.98b_{2152} + 84.75b_{800} - 7.61b_{998} - 34.88b_{1448} + 56.464b_{760} + 32.18b_{1003,533} - 0.17$
Hg	$Y = 29.75b_{997} + 38.65b_{453} - 2.15b_{998} + 7.17b_{1003} - 170.93b_{537} - 56.35b_{1002} - 11.51b_{536} + 13.67b_{1004} - 13.29b_{999} - 29.90b_{1393} + 8.23b_{444} + 19.12b_{603} + 6.71b_{485} - 17.07b_{606} - 46.99b_{1392} - 0.01$
As	$Y = -0.13b_{779} - 0.18b_{1731} - 0.19b_{1739} + 0.30b_{1697} + 0.33b_{2009} + 0.92b_{1735,761} + 0.32b_{792} + 0.39b_{2010} - 0.010.92b_{1736} - 0.09b_{1698} + 0.28b_{1740} + 0.01b_{2480} - 0.76b_{1269} + 0.10b_{1399,791} - 0.38b_{1001} - 0.01$
Cu	$Y = 4.33b_{1002} + 2.15b_{1003} - 1.81b_{997} + 0.04b_{998} - 1.11b_{1004}\overset{1004}{b} - 3.31b_{1945} - 0.39b_{1944} - 9.11b_{1485} + 0.85b_{1926} - 6.42b_{1956} - 1.86b_{1396} + 5.73b_{1393} + 3.51b_{999} - 1.71b_{1397} - 5.47b_{1955} - 0.20$
Pb	$Y = -0.83b_{651} - 1.29b_{858} - 0.73b_{729} - 1.17b_{1724} - 0.14b_{863} + 1.29b_{864} - 0.32b_{612} - 0.03b_{573} + 1.38b_{621} + 1.51b_{1801} - 1.22b_{1149} - 0.37b_{829} - 2.21b_{796} + 0.30b_{859} + 0.03b_{613} - 0.99$
Cr	$Y = 0.32b_{468} + 0.38b_{820} + 0.04b_{800} + 0.16b_{819} - 0.91b_{1248} - 2.38b_{1738} + 0.77b_{760} + 0.25b_{1944} - 1.90b_{1295} + 0.46b_{1806} + 0.41b_{1358,467} + 0.51b_{1550} + 0.49b_{1837} - 0.97b_{1256} + 0.85b_{1124} + 0.004$
Zn	$Y = 0.26b_{430} + 0.05b_{1003} + 0.69b_{1002} - 0.05b_{999} - 0.72b_{1000} - 0.57b_{1004} + 0.19b_{998} + 0.26b_{431} - 0.20b_{1453,997} + 0.02b_{602} - 0.06b_{1810} + 0.80b_{536} - 0.02b_{1432} - 1.45b_{1809} + 3.68b_{1570} + 0.02$
Ni	$Y = 0.10b_{1003} + 0.49b_{1002} - 0.36b_{997} + 0.09b_{998} - 0.08b_{1004} - 1.51b_{1956,1485} + 0.30b_{999} - 0.34b_{1396} - 0.25b_{1945} - 0.21b_{1926} - 0.04b_{1944} + 0.60b_{1393} - 1.16b_{1955,819} - 0.47b_{1397} + 0.02b_{1925} - 0.06$

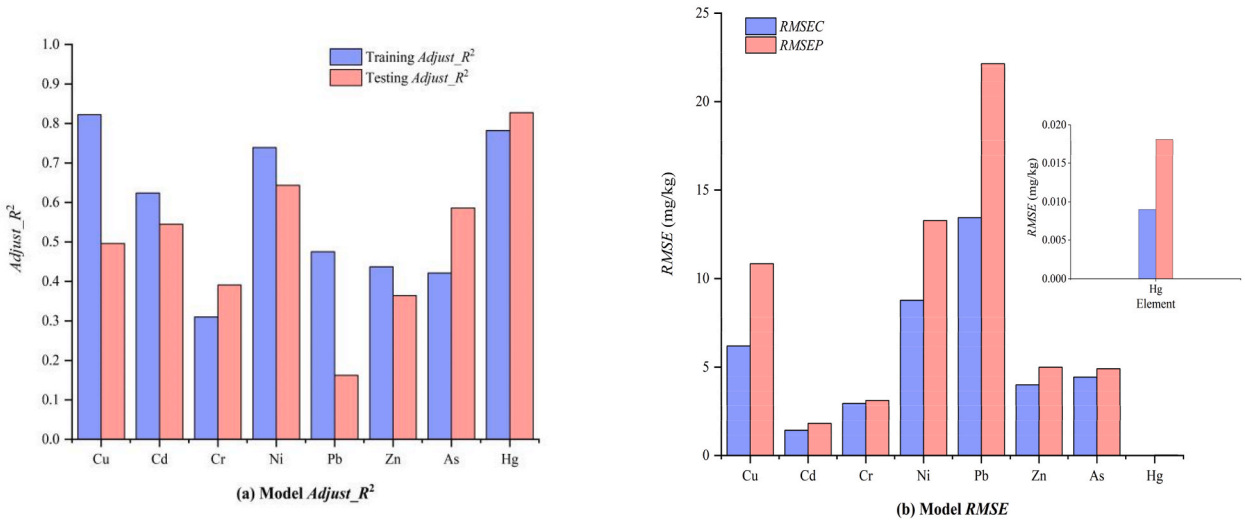

Fig. 6. BPNN model coefficient comparison based on *R*.

*Adjust*_R^2 of the training models based on *R*_landsat8-OLI and *R*, it is evident that, except for Cd, the former models outperform the latter in terms of model performance. Therefore, it is feasible to perform heavy metal model inversion based on the simulation multispectral of Landsat8-OLI sensor. The significant improvement in the correlation between spectral data and soil heavy metal content, especially through spectral differentiation, highlights the potential for selecting appropriate spectral indices and model inversion methods to enhance soil heavy metal prediction models. As a result, it is essential to use various model prediction techniques for various heavy metal study items.

3.4.4. Mapping of Hg concentrations using Landsat8_OLI image data

According to the *Adjust*_R^2 of the SMLR training models based on *R*_Landsat8-OLI and *R*, it is observed that the *Adjust*_R^2 values for Cu, Zn, and Hg are higher in the model based on *R*_Landsat8-OLI compared to the model based on *R*. Therefore, it is feasible to perform heavy metal model inversion based on the simulation multispectral of Landsat8-OLI sensor. Furthermore, through comparing the ranking of the optimal models, it is consistently observed that Hg consistently exhibits the best model fitting performance among the three models. In light of this, this investigation used Hg as the exploration object and the multispectral mapping of Hg contamination distribution, which performed by the best model according Eq. (5) was shown in Fig. 9. Numerous active mines and slag dumping sites have been discovered in the research area through visual interpretation of photographs and field investigation. The pollution sources are mainly located in the southwest of the research region and have directly contaminated the soil there. The contamination map of Hg demonstrates that the worst Hg pollution were the nearby settlements, tailings sites, and sewage outlets, which are the key factors for the high concentrations.

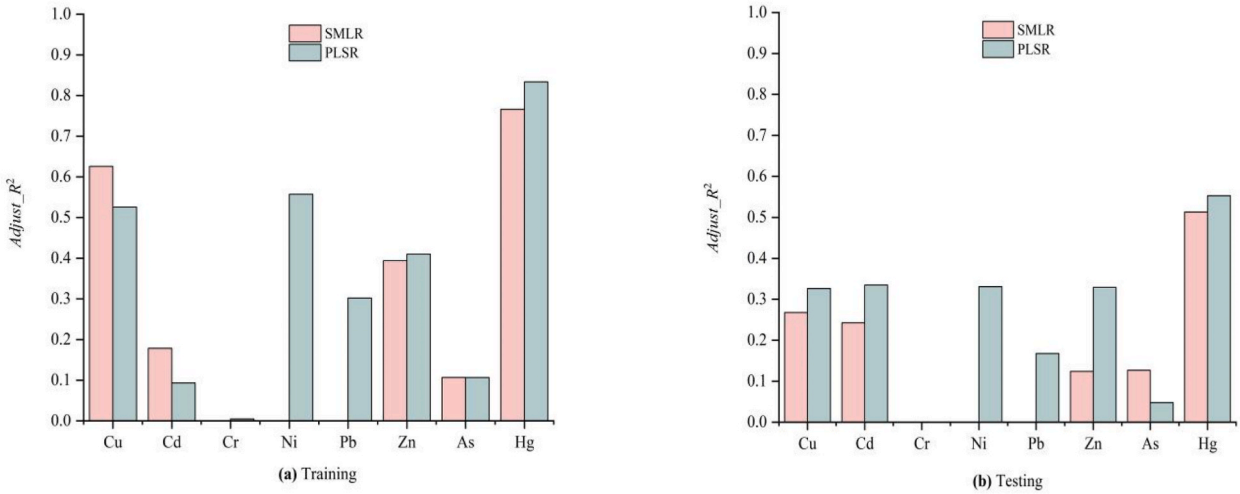

Fig. 7. Comparison of *Adjust_R*^{2} of the SMLR and PLSR models based on *R*_Landsat8-OLI.

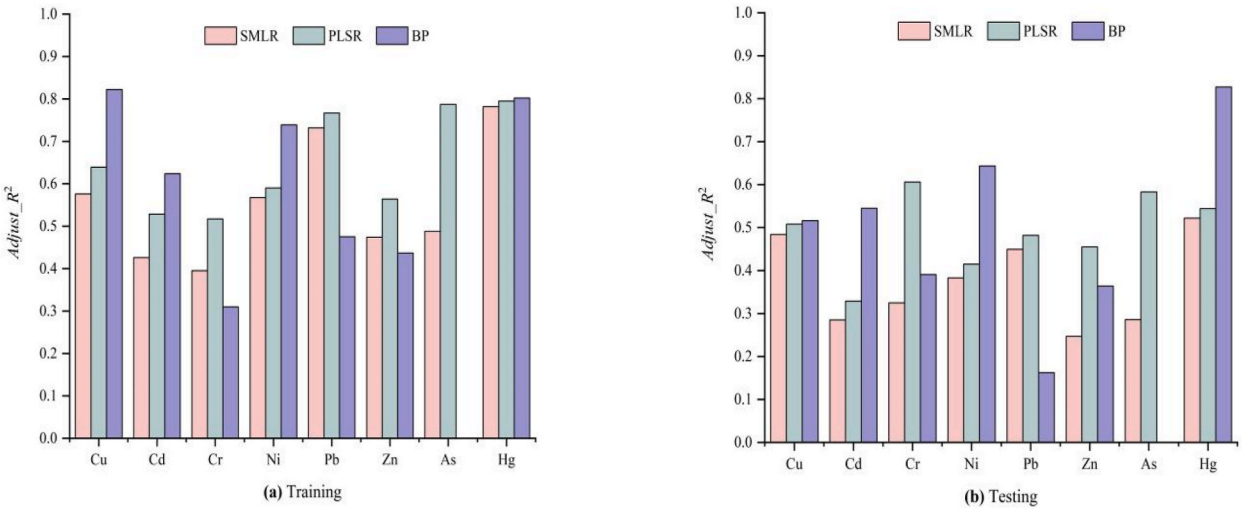

Fig. 8. Comparison of the *Adjust_R*^{2} between the optimal SMLR, PLSR models and BPNN models of soil heavy metals.

4. Discussions and conclusions

4.1. Discussions

Although soil heavy metals are difficult to invert directly through spectral characteristic analysis due to their low contents, there are studies have been conducted to achieve direct inversion of heavy metals by analyzing the correlation between heavy metal content and soil spectrum [17,61]. The study analyzed the relationship between the spectral transformation and the soil heavy metal content, and a set of technical processes for predicting soil heavy metal content using hyperspectral data was developed, and finally the optimal inversion models for soil heavy metals were selected out. Statistical analysis demonstrated that there were severe heavy metal pollution in the study area. Soil heavy metals have specific spectral absorption characteristics in the soil spectrum, which provides a theoretical basis for the extraction of soil pollution information [9]. Savitzky-Golay convolution smoothing was proved to be a method that can effectively reduce spectral noise and preserve the original spectral features, which beneficial to improve the spectral information of soil heavy metals. Furthermore, the goal of spectral transformation is to eliminate or weaken the change in soil spectral signal intensity caused by random factors, as well as to reduce noise influence and enhance soil heavy metal spectral information. Correlation analysis results demostrated that spectral transformation, especially the second derivative spectrum, effectively improved the correlation between heavy metals and spectra, which highlights the spectral response characteristics of heavy metals, and this is consistent with previous studies [9,11]. Hyperspectral data have many bands and high redundancy, however, traditional feature selection methods based on correlation screening usually ignore the impact of significant bands on inversion accuracy. The PLSR

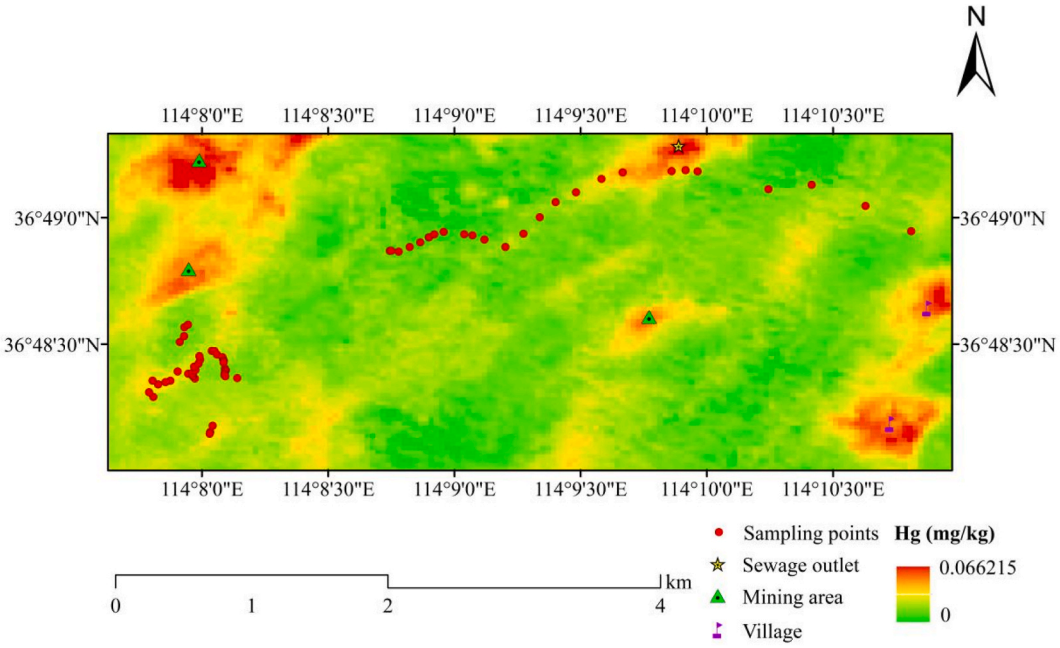

Fig. 9. Simulated multispectral mapping of Hg contamination distribution in metal tailings area.

method can be applied to a continuous spectrum with many bands and serious autocorrelation, which can effectively improve the quantitative inversion accuracy of metal elements.

Although the accuracy of the hyperspectral inversion model in this study was somewhat less accurate than that of the conventional detection techniques, the soil samples and the obtained data were handled, and the main factors that could affect the results were excluded. Therefore, the established model precision and credibility were comparatively higher. The inversion models summary results indicated that the training models based on spectral derivative indices show significantly higher $Adjust\_R^2$ compared to the models based on the original spectral R. Moreover, the second-order derivative models particularly stand out in improving the stability and accuracy of the models. With the exception of Hg, the optimal PLSR model outperforms the optimal SMLR model in terms of training and testing accuracy for other heavy metals. We speculate that the superior performance of the PLSR model compared to the SMLR model is attributed to its ability to address multicollinearity, prevent overfitting, and account for nonlinear relationships between the independent and dependent variables. The BPNN inversion results indicating a better model effect and good practical extrapolation for Cd, Ni, and Hg. The simulated Landsat8-OLI multispectral inversion model demonstrates the feasibility of heavy metal content retrieval for Landsat8-OLI multispectral data, particularly for Hg.

Air-dried ground soil samples were used in the research, thus the accuracy of the inversion models was less affected by objective conditions such as soil moisture and particle size. However, the accuracy of the heavy metal regression model was affected by many factors, such as the number of measured soil samples, remote sensing image resolution, the significance of the correlation of selected modeling factors, and the rationality of the modeling method. Therefore, the influence of these variables such as the type of soil, textural variations according to depth, organic matter content and even the simple volumetric moisture content of the samples, on the spectral response mechanism of heavy metal soils should not ignored if the experimental conditions allowed. And it is essential to choose suitable characteristic bands to establish models in light of the complex mechanism of spectral response for soil heavy metals.

4.2. Conclusions

In this study, the statistical analysis of soil heavy metal content was carried out through the measured sample data, and the correlation between the soil heavy metals content and the hyperspectral were analyzed to establish the direct estimation models based on different spectral transformations. In addition, by combining the simulated Landsat8-OLI multispectral and PLSR model, the high-precision multispectral remote sensing mapping of Hg has been accomplished. The current evidence suggests that the conventional mathematical transformations applied to the spectral data in the paper have effectively improved the spectral response correlation of heavy metals, thereby demonstrating an advantage in the direct inversion model of soil heavy metals. The combination of simulated Landsat8-OLI multispectral and PLSR model seems to have better performence in some cases, proving that the simulated multispectral is feasible for soil heavy metals inversion. However, there are certain limitations, such as the unexplored potential of spectral wavelet transformations and other possible enhancements. Furthermore, the study only employed SMLR, PLSR, and BPNN models to validate the performance of heavy metal inversion based on correlation-based band combinations. Therefore, it is essential to further investigate their performance in mainstream models, such as the random forest model and other similar models. As an exploratory attempt

in the inversion of heavy metal models, this study aims to provide a reference for the model inversion of soil heavy metals in mining locations.

Author contribution statement

Nannan Yang and Ling Han - Conceived and designed the experiments; Nannan Yang - Performed the experiments; Nannan Yang - Analyzed and interpreted the data. Ming Liu - Contributed reagents, materials, analysis tools or data; Nannan Yang and Ming Liu - Wrote the paper.

Data availability statement

The authors do not have permission to share data.

Ethics declarations

The authors declare that review and/or approval by an ethics committee was not needed for this study because the research does not involve medical treatment, drug trials or biomedical research. In addition, informed consent was not required for this study because all participants decided to participate in the study independently, voluntarily and without any external pressure on the basis of fully understanding the research purpose, participation conditions, implementation process, possible risks and expected benefits.

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

This work was supported by the Fundamental Research Funds for the Central Universities, Chang'an University (300102352901), the Key research and development project of Shaanxi Province(2022ZDLSF-07-05).

Useful suggestions given by professors of the Hebei University of Engineering are acknowledged.

References

[1] R. Samiei Fard, H.R. Matinfar, Capability of vis-NIR spectroscopy and Landsat 8 spectral data to predict soil heavy metals in polluted agricultural land (Iran), Arabian J. Geosci. 9 (2016) 745, https://doi.org/10.1007/s12517-016-2780-4.

[2] P. Zeng, L. Liang, Z. Duan, Ecological and environmental impacts of mineral exploitation in urban agglomerations, Ecol. Indicat. 148 (2023), 110035, https://doi.org/10.1016/j.ecolind.2023.110035.

[3] A. Falahi-Ardakani, Contamination of environment with heavy metals emitted from automotivees, Ecotoxicol. Environ. Saf. 8 (1984) 152–161, https://doi.org/10.1016/0147-6513(84)90057-5.

[4] T. Shi, Y. Chen, Y. Liu, G. Wu, Visible and near-infrared reflflectance spectroscopy—an alternative for monitoring soil contamination by heavy metals, J. Hazard Mater. 265 (2014) 166–176, https://doi.org/10.1016/j.jhazmat.2013.11.059.

[5] W. Zhou, H. Yang, L. Xie, H. Li, L. Huang, Y. Zhao, T. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model, J. Catena. 202 (2021), 105222, https://doi.org/10.1016/j.catena.2021.105222.

[6] K. Yang, G. Wang, P. Fu, W. Zhang, X. Wang, A model on extracting the pollution information of heavy metal copper ion based on the soil spectra analyzed by HHT in timefrequency, Spectrosc. Spectr. Anal. 38 (2018) 564–569, https://doi.org/10.3964/j.issn.1000-0593(2018)02-0564-06.

[7] H. Yang, H. Xu, X. Zhong, Prediction of soil heavy metal concentrations in copper tailings area using hyperspectral reflectance, Environ. Earth Sci. 81 (2022) 183, https://doi.org/10.1007/s12665-022-10307-x.

[8] Q. Shen, K. Xia, S. Zhang, C. Kong, Q. Hu, S. Yang, Hyperspectral indirect inversion of heavy-metal copper in reclaimed soil of iron ore area, Spectrochim. Acta Mol. Biomol. Spectrosc. 222 (2019), 117191, https://doi.org/10.1016/j.saa.2019.117191.

[9] S. Zhang, Q. Shen, C. Nie, Y. Huang, J. Wang, Q. Hu, X. Ding, Y. Zhou, Y. Chen, Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods, Spectrochim. Acta Mol. Biomol. Spectrosc. 211 (2019) 393–400, https://doi.org/10.1016/j.saa.2018.12.032.

[10] T.M. Phuong, Z. Lin, R.B. Altman, Choosing SNPs using feature selection, J. Bioinf. Comput. Biol. 4 (2006) 241–257, https://doi.org/10.1142/S0219720006001941.

[11] F. Wang, J. Gao, Y. Zha, Hyperspectral sensing of heavy metals in soil and vegetation: feasibility and challenges, ISPRS J. Photogrammetry Remote Sens. 136 (2018) 73–84, https://doi.org/10.1016/j.isprsjprs.2017.12.003.

[12] H. Yu, B. Kong, Q. Wang, X. Liu, X. Liu, Hyperspectral Remote Sensing Applications in Soil: a Review, Hyperspectral Remote Sensing, Elsevier, 2020, pp. 269–291, https://doi.org/10.1016/B978-0-08-102894-0.00011-5.

[13] L. Galvez-Sola, R. Moral, M.D. Perez-Murcia, A. Perez-Espinosa, M.A. Bustamante, E. Martinez-Sabater, C. Paredes, The potential of near infrared reflectance spectroscopy (NIRS) for the estimation of agroindustrial compost quality, Sci. Total Environ. 408 (2010) 1414–1421, https://doi.org/10.1016/j.scitotenv.2009.11.043.

[14] X. Gu, Y. Wang, Q. Sun, G. Yang, C. Zhang, Hyperspectral inversion of soil organic matter content in cultivated land based on wavelet transform, Comput. Electron. Agric. 167 (2019), 105053, https://doi.org/10.1016/j.compag.2019.105053.

[15] C.M. Pandit, G.M. Filippelli, L. Li, Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression, Int. J. Rem. Sens. 31 (2010) 4111–4123, https://doi.org/10.1080/01431160903229200.

[16] K. Tan, H. Wang, L. Chen, Q. Du, P. Du, C. Pan, Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest, J. Hazard Mater. 382 (2020), 120987, https://doi.org/10.1016/j.jhazmat.2019.120987.

[17] R. Wang, S. Wu, K. Wu, S. Huang, R. Wu, B. Liu, M. Lin, L. Li, D. Zhou, X. Diao, Estimation and spatial analysis of heavy metals in metal tailing pond based on improved PLS with multiple factors, IEEE Access 9 (2021) 64880–64894, https://doi.org/10.1109/ACCESS.2021.3073933.

[18] T. Kemper, S. Sommer, Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy, Environ Sci Ttechnol 36 (2002) 2742–2747, https://doi.org/10.1021/es015747j.

[19] G. Bonifazi, G. Capobianco, S. Serranti, Asbestos containing materials detection and classifification by the use of hyperspectral imaging, J. Hazard Mater. 344 (2018) 981–993, https://doi.org/10.1016/j.jhazmat.2017.11.056.

[20] H. Wang, Y. Wang, Inversion of Landsat 8 for soil heavy metals after terrain correction, Agric. Res. Arid Areas 37 (2019) 11–17, https://doi.org/10.7606/j.issn.1000-7601.2019.01.02.

[21] D. Dessalegn, S. Beyene, N. Ram, F. Walley, T.S. Gala, Effects of topography and land use on soil characteristics along the toposequence of Ele watershed in southern Ethiopia, Catena 115 (2014) 47–54, https://doi.org/10.1016/j.catena.2013.11.007.

[22] Y. Xian, M. Wang, W. Chen, Quantitative assessment on soil enzyme activities of heavy metal contaminated soils with various soil properties, Chemosphere 139 (2015) 604–608, https://doi.org/10.1016/j.chemosphere.2014.12.060.

[23] L. Shao, Geological disaster prevention and control and resource protection in mineral resource exploitation region, Int. J. Low Carbon Technol. 14 (2019) 142–146, https://doi.org/10.1093/ijlct/ctz003.

[24] R. Fernández-Martínez, I. Rucandio, A simplified method for determination of organic mercury in soils, ANAL METHODS-UK 5 (2013) 4131, https://doi.org/10.1039/c3ay40566d.

[25] Analytical Spectral Devices, Inc., FieldSpec 4 User Guide[User Manual], 2016. https://www.malvernpanalytical.com/en/learn/knowledge-center/user-manuals/fieldspec-4-user-guide,, 2023-07-26.

[26] Y. He, R. Deng, Q. Chen, L. Chen, Y. Qin, Diffuse attenuation coefficient of suspended sediment based on ASD spectrometer, Acta Sci. Nat. Univ. Sunyatseni 50 (2011) 134–140.

[27] M. Vidal, J.M. Amigo, Pre-processing of hyperspectral images. Essential steps before image analysis, Chemometr Intell Lab 117 (2012) 138–148, https://doi.org/10.1016/j.chemolab.2012.05.009.

[28] Exelis Visual Information Solutions, ENVI 5.3. 2015, Computer software. L3Harris Geospatial. Web, 2023. Retrieved from, https://www.l3harrisgeospatial.com/software/envi.

[29] H. Xu, Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery, Int. J. Rem. Sens. 27 (14) (2006) 3025–3033, https://doi.org/10.1080/01431160600589179.

[30] C.F. Jordan, Derivation of leaf-area index from quality of light on the forest floor, Ecology 50 (4) (1969) 663–666, https://doi.org/10.2307/1936256.

[31] A. Alasta, Using remote sensing data to identify iron deposits in central western Libya, in: International Conference on Emerging Trends in Computer and Image Processing (ICETCIP' 2011), Bangkok, Thailand, 2011, pp. 56–61. http://psrcentre.org/images/extraimages/122.%201211924.pdf, 2023-07-26.

[32] Z. Jiang, A.R. Huete, K. Didan, T. Miura, Development of a two-band enhanced vegetation index without a blue band, Remote Sens. Environ. 112 (10) (2008) 3833–3845, https://doi.org/10.1016/j.rse.2008.06.006.

[33] J.A. Schnell, Monitoring the Vernal Advancement and Retrogradation (Greenwave Effect) of Natural Vegetation, Nasa/gsfct Type Final Report, 1974.

[34] M.H.A. Baig, L. Zhang, T. Shuai, Q. Tong, Derivation of a tasselled cap transformation based on landsat 8 at-satellite reflectance, Remote Sensing Letters 5 (5) (2014) 423–431, https://doi.org/10.1080/2150704X.2014.915434.

[35] X. Dai, Z. Wang, S. Liu, Y. Yao, R. Zhao, T. Xiang, T. Fu, H. Feng, L. Xiao, X. Yang, S. Wang, Hyperspectral imagery reveals large spatial variations of heavy metal content in agricultural soil - a case study of remote-sensing inversion based on Orbita Hyperspectral Satellites (OHS) imagery, J. Clean. Prod. 380 (2022), 134878, https://doi.org/10.1016/j.jclepro.2022.134878.

[36] Y. Dong, S. Yang, C. Xu, Y. Li, W. Bai, Z. Fan, Y. Wang, Q. Li, Determination of soil parameters in apple-growing regions by near-and mid-infrared spectroscopy, Pedosphere 21 (2011) 591–602, https://doi.org/10.1016/S1002-0160(11)60161-6.

[37] W. Ma, K. Tan, H. Li, Q. Yan, Hyperspectral inversion of heavy metals in soil of a mining area using Extreme Learning Machine, J. Ecol. Rural Environ. 32 (2016) 213–218, https://doi.org/10.11934/j.issn.1673-4831.2016.02.007.

[38] L. Xu, D. Xie, Estimation of soil organic matter content based on regional feature bands, Advances in Intelligent Systems and Computing,Springer International Publishing 1075 (2020) 1063–1070, https://doi.org/10.1007/978-3-030-32591-6_116. Cham.

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

[40] Z. Liu, X. Ma, Y. Wen, Y. Wang, W. Cai, X. Shao, A practical approach for near infrared spectral quantitative analysis of complex samples using partial least squares modeling, Sci. China, Ser. B: Chem. 52 (2009) 1021–1027, https://doi.org/10.1007/s11426-009-0110-3.

[41] K. Tan, Y. Ye, P. Du, Estimation of heavy metal concentrations in reclaimed mining soils using reflectance spectroscopy, Spectroscopy & Spectral Analysis 34 (2014) 3317–3322, https://doi.org/10.3964/j.issn.1000-0593(2014)12-3317-06.

[42] S. Wang, Y. Zhao, R. Hu, Y. Zhang, X. Han, Analysis of Near-Infrared spectra of coal using deep synergy adaptive moving window partial least square method based on genetic algorithm, Chinese J Anal Chem 47 (2019) 19034–19044, https://doi.org/10.1016/S1872-2040(19)61150-3.

[43] B. Zhang, B. Guo, B. Zou, W. Wei, Y. Lei, T. Li, Retrieving soil heavy metals concentrations based on GaoFen-5 hyperspectral satellite image at an opencast coal mine, Inner Mongolia, China, Environ Pollut 300 (2022), 118981, https://doi.org/10.1016/j.envpol.2022.118981.

[44] Y. Yang, Q. Cui, P. Jia, J. Liu, H. Bai, Estimating the heavy metal concentrations in topsoil in the Daxigou mining area, China, using multispectral satellite imagery, Sci. Rep. 11 (2021), 11718, https://doi.org/10.1038/s41598-021-91103-8.

[45] Y. Yan, L. Dong, Y. Han, W. Li, A general inverse control model of a magneto-rheological damper based on neural network, J. Vib. Control 28 (2021) 952–963, https://doi.org/10.1177/1077546320986380.

[46] Y. Yao, J. Li, C. He, X. Hu, L. Yin, Y. Zhang, J. Zhang, H. Huang, S. Yang, H. He, F. Zhu, S. Li, Distribution characteristics and relevance of heavy metals in soils and colloids around a mining area in Nanjing, China, Bull. Environ. Contam. Toxicol. 107 (2021) 996–1003, https://doi.org/10.1007/s00128-021-03350-0.

[47] N. Lin, R. Jiang, G. Li, Q. Yang, D. Li, X. Yang, Estimating the heavy metal contents in farmland soil from hyperspectral images based on Stacked AdaBoost ensemble learning, Ecol. Indicat. 143 (2022), 109330, https://doi.org/10.1016/j.ecolind.2022.109330.

[48] D. Figueiredo, S. Júnior, E. Rocha, What is R2 all about? Leviathan-Cadernos de Pesquisa Polútica 3 (2011) 60–68, https://doi.org/10.11606/issn.2237-4485.lev.2011.132282.

[49] J. Chen, F. Wei, C. Zheng, Y. Wu, D.C. Adriano, Background concentrations of elements in soils of China, Water, Air, and Soil Pollution 57 (1991) 699–712, https://doi.org/10.1007/BF00282934.

[50] S.P. Smith, A.K. Jain, A test to determine the multivariate normality of a data set, IEEE T Pattern Anal 10 (1988) 757–761, https://doi.org/10.1109/34.6789.

[51] D.S. Manta, M. Angelone, A. Bellanca, R. Neri, M. Sprovieri, Heavy metals in urban soils: a case study from the city of Palermo (Sicily), Italy, Sci Total Environ. 300 (2002) 229–243, https://doi.org/10.1016/S0048-9697(02)00273-5.

[52] W.R. Roper, W.P. Robarge, D.L. Osmond, J.L. Heitman, Comparing four methods of measuring soil organic matter in North Carolina soils, Soil Sci. Soc. Am. J. 83 (2019) 466–474, https://doi.org/10.2136/sssaj2018.03.0105.

[53] J. Liu, Y. Zhang, H. Wang, Y. Du, Study on the prediction of soil heavy metal elements content based on visible near-infrared spectroscopy, Spectrochim. Acta Mol. Biomol. Spectrosc. 199 (2018) 43–49, https://doi.org/10.1016/j.saa.2018.03.040.

[54] F. Gan, R. Wang, A. Ma, Spectral identification tree (SIT) for mineral extraction using AVIRIS data, Proc. SPIE-Int. Soc. Opt. Eng. (2002), 4897, https://doi.org/10.1117/12.466877.

[55] A.C. Dotto, R.S.D. Dalmolin, A. Ten Caten, S. Grunwald, A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra, Geoderma 314 (2018) 262–274, https://doi.org/10.1016/j.geoderma.2017.11.006.

[56] X. Jia, D. O Connor, Z. Shi, D. Hou, VIRS based detection in combination with machine learning for mapping soil pollution, Environ Pollut 268 (2021), 115845, https://doi.org/10.1016/j.envpol.2020.115845;
(a) A. Gholizadeh, M. Saberioon, E. Ben-Dor, R.A. Viscarra Rossel, L. Borůvka, Modelling potentially toxic elements in forest soils with vis–NIR spectra and learning algorithms, Environ Pollut 267 (2020), 115574, https://doi.org/10.1016/j.envpol.2020.115574.

[57] C. Yang, M. Feng, L. Song, C. Wang, W. Yang, Y. Xie, B. Jing, L. Xiao, M. Zhang, X. Song, M. Saleem, Study on hyperspectral estimation model of soil organic carbon content in the wheat field under different water treatments, SCI REP-UK 11 (2021), 18582, https://doi.org/10.1038/s41598-021-98143-0.

[58] M. Zhou, B. Zhou, Y. Tu, J. Xia, Hyperspectral modeling of Pb content in mining area based on spectral feature band extracted from near standard soil samples, Spectrosc. Spectr. Anal. 40 (2020) 2182–2187, https://doi.org/10.3964/j.issn.1000-0593(2020)07-2182-06.
[59] B. Hu, X. Jia, J. Hu, D. Xu, F. Xia, Y. Li, Assessment of heavy metal pollution and health risks in the soil-plant-human system in the Yangtze River Delta, China, Int. J. Environ. Res. Publ. Health 14 (2017) 1042, https://doi.org/10.3390/ijerph14091042.
[60] X. Ma, K. Zhou, J. Wang, S. Cui, S. Zhou, S. Wang, G. Zhang, Optimal bandwidth selection for retrieving Cu content in rock based on hyperspectral remote sensing, J Arid Land 14 (2022) 102–114, https://doi.org/10.1007/s40333-022-0050-8.
[61] N. Mezned, F. Alayet, B. Dkhala, S. Abdeljaouad, Field hyperspectral data and OLI8 multispectral imagery for heavy metal content prediction and mapping around an abandoned Pb–Zn mining site in northern Tunisia, Heliyon 8 (2022), e9712, https://doi.org/10.1016/j.heliyon.2022.e09712.