Journal of Environmental Management 347 (2023) 119196

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Estimation of the distribution patterns of heavy metal in soil from airborne hyperspectral imagery based on spectral absorption characteristics

Kun Tan [a,b,c,*], Lihan Chen [d], Huimin Wang [d,e], Zhaoxian Liu [f], Jianwei Ding [f], Xue Wang [a,b,c]

[a] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, 200241, China
[b] 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
[c] School of Geographic Sciences, East China Normal University, Shanghai, 200241, China
[d] Key Laboratory of Land Environment and Disaster Monitoring of MNR, China University of Mining and Technology, Xuzhou, 221116, China
[e] Xi'an Meihang Remote Sensing Information Co., Ltd, Xi'an, 710199, China
[f] The Second Surveying and Mapping Institute of Hebei, Shijiazhuang, 050037, China

ARTICLE INFO

Handling Editor: Jason Michael Evans

Keywords:
Airborne hyperspectral image
Retrieval mechanism
Soil heavy metals estimation
Spectral absorption characteristics

ABSTRACT

Though soil is widely known as one of the most valuable resources for the world, its quality is going to be lower because of unsustainable economic development and social progress. Therefore, it is important for us to monitor and evaluate the quality of soil, especially its heavy metal contents which is too scarce to identify in soil spectra easily but poisonous enough to affect human health in a long run. Most of the existing estimation methods have based the characteristic bands on statistical analysis to a large extent, which is hard to accurately explain the retrieval mechanism. In this paper, the absorption characteristics of heavy metal are studied based on the soil spectra, and the distribution pattern is mapped in a large-scale continuous space, for environmental monitoring and further decision support. Taking Yitong County, China as the study area. After spectra continuum removal, the heavy metal contents were estimated by 11 features including the absorption depth, absorption area, and band ratio around 2200 nm, which showed the best performance. For arsenic (*As*), the best model yields R_p^2 value of 0.8474, and the $RMSE_P$ value is 36.1542 (mg/kg). It is concluded that *As* is adsorbed by organic matter, clay minerals, and iron/manganese oxides in soil, and the adsorption of *As* by first two components is greater than that of the last. For airborne spectra after continuum removal, combining the spectral absorption characteristic parameters and the highly correlated bands is more accurate than using the spectral absorption characteristic parameters or bands alone. AdaBoost is presented for the heavy metal estimation, and the fitting ability of the method is found to be stronger than that of the traditional classical methods, with the R_p^2 values of 0.6242 and the $RMSE_P$ value of 43.6481 (mg/kg). In summary, these results will provide a prospective basis for the rapid estimation of soil heavy metals, the risk assessment of soil heavy metals and soil environmental monitoring in a large scale.

1. Introduction

The excess of heavy metals caused by various reasons and it influences the essential characteristics of the soil ecosystem. Many heavy metals mine, smelt, process and commercial manufacture activities have resulted in serious environmental pollution (Merdy et al., 2006; Wang et al., 2015; Yao et al., 2015). Heavy metal ions, easily absorbed and enriched by crops in farmland soils, significantly threat human health. Therefore, it is essential to acquire the soil heavy metals content. Hyperspectral technology has widely applied to soil monitoring and decision making by its repetitive coverages in large scale areas (Ma et al., 2016a,b; Liu et al., 2017; Tan et al., 2018).

During the spectral measurement, light scattering, various particle sizes, and the different density distributions in soil all cause spectra noise Therefore, the preprocess on the spectra before estimation is necessary. There are several commonly-used preprocessing methods, including standard normal variate preprocessing, derivative transformation, normalized processing, Savitzky-Golay smoothing,

* Corresponding author. Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai, 200241, China.
E-mail addresses: tankuncu@gmail.com (K. Tan), tb19160016b2@cumt.edu.cn (L. Chen), wanghm9423@126.com (H. Wang), 975511756@qq.com (Z. Liu), 840085650@qq.com (J. Ding), wx_ecnu@yeah.net (X. Wang).

https://doi.org/10.1016/j.jenvman.2023.119196
Received 29 May 2023; Received in revised form 16 September 2023; Accepted 30 September 2023
Available online 4 October 2023

continuum removal, wavelet preprocessing, and multiplicative scatter correction (Shi et al., 2013; Zhang et al., 2017). Continuum removal is commonly utilized in the soil heavy metals estimation. Peng et al. (2014) found that the application of the partial least squares regression (PLSR) model after continuum removal can achieve estimate the content of *As* in soil quickly and accurately. It is confirmed that continuum removal can help highlight the absorption characteristics of the spectra. Therefore, in this study, continuum removal was chosen as the spectral pretreatment method.

The spectra absorption characteristics can reflect the structure and motion state of the soil components. Though the validity of these characteristics are proved in many research areas, such as mineral mapping (Meer, 2004; Wei et al., 2015), the chlorophyll-a of algae inversion (Zhang et al., 2011), and soil water content prediction (Jin et al., 2016), they are rarely used in soil heavy metal estimation. It (Choe et al., 2008) was reported that statistical features such as Depth500, Area2200 Asym2200 and $R_{610,500}$ were of high sensitive with the heavy metals content, which was successfully applied to the soil heavy metals estimation in airborne hyperspectral image. Jin et al. (Jin and Zhou, 2017) found the model based on bands depth performed better than spectral reflectance or reciprocal logarithm for Inner Mongolia chestnut soil *Cd* estimation. The absorption relationships between organic matter, iron oxides, and clay minerals, and soil heavy metals can be explored by analyzing the spectral absorption characteristics, providing theoretical support to the estimation mechanism of soil heavy metals.

Some scholars have tried their best to explore the influencing mechanism between soil heavy metals and soil spectra. Furthermore, the absorption by organic matter has been extensively investigated in many heavy metal researches. Cao et al. (2007) found, through the analysis of black soil in polluted farmland in Northeast China, that *Pb* is easily adsorbed by organic matter, while the adsorption of *Cd*, *Cu*, and *Zn* is relatively difficult. Chen et al. (2016) found that the concentration of *As*, *Pb*, *Cr*, and *Cd* in soil along the Bortala River in China were highly correlated with the organic matter content. The assessment of heavy metals is reflected in the absorption of heavy metal ions by the other components. Xu et al. found that the most significant relative components are organic matter, iron/manganese oxides, and clay minerals (Wang et al., 2007; Xu et al., 2011a,b). Therefore, we studied the spectra adsorption characteristics of these three main components in the black soil of Northeast China, as well as a further exploration of soil heavy metal estimation mechanism.

Researches on modeling methods have made significant progress, including univariate regression (Qi et al., 2007), multiple linear regression (Kokaly and Clark, 1999; Song et al., 2015), partial least squares regression (Pandit et al., 2010; Xia et al., 2015), ridge regression, and principal component regression (Lu et al., 2007) models. In addition, machine learning techniques have also been widely utilized to the soil heavy metals estimation, including artificial neural networks (ANNs) (Anagu et al., 2009), support vector machine (SVM) (Balabin and Lomakina, 2011; Ma et al., 2016a,b), and decision tree (DT) (Rodriguez-Galiano et al., 2015), which are widely used in soil heavy metal estimation. However, the conventional statistical approaches are of poor applicability because of the low content of soil heavy metals and the limited samples.

A growing number of researchers have introduced ensemble learning methods into the inversion of soil components, which are famous for having a more stable, accurate, and faster performance. Ensemble learning methods combine individual learners to obtain a better model. The framework algorithms commonly utilized in ensemble learning methods are bagging (Dietterich, 2000), boosting (Svetnik et al., 2005), and stacking algorithms (Chen et al., 2014). Boosting algorithms include AdaBoost (Liao and Zhou, 2012) and gradient boosted decision trees (GBDT). The typical representative bagging method is random forest (RF) (Breiman, 2001). Stacking is a combination of various basic models. Extremely randomized trees (ET) is an ensemble learning model that is a further improvement of RF, which is both more random and robust than RF. Recently, XGBoost, GBDT, RF and ELM are utilized to analyze and estimate soil heavy metals. Wang et al. (2023) proposed a estimation model combining SMA and RF to invert heavy metal concentrations. Ye et al. (2023) integrated spatial correlation with XGBoost algorithm to improve the soil arsenic concentration estimation accuracy. Bian et al. (2023) reported that ELM-based spectral estimation models were able to predict metal concentrations with high accuracy and efficiency. Saha et al. (2022) found ground spectral data can improve the precision and stability of the inversion of soil As with XGBoost. Ma et al. (2016a,b) found that RF was superior to PLS, SVM, and ET heavy metals inversion and analysis. However, AdaBoost and ET are less utilized in heavy metal estimation.

Due to the complexity of the imaging process, the difficulty of spectral feature extraction, and the fact that most of the current research focuses on ground spectra, there are fewer studies to quantitatively estimate heavy metals based on imaging spectra. Wu et al. (2011) used simulated HyMap, Landsat Thematic Mapper (TM), and QuickBird spectra to estimate soil heavy metals, and found that the heavy metals with a high correlation with iron (*Fe*) had a better effect; Yang et al. (2016) used the spectral reflectance of Hyperion imagery to estimate the concentration of *Zn* and *Cd* in soil; and Tan et al. (2019) utilized airborne hyperspectral data to create a spectral analysis model for the heavy metal concentrations retrieval. Compared to the ground spectra, imaging spectroscopy can achieve a larger-scale estimation of soil heavy metal. Therefore, the use of hyperspectral imagery for heavy metal mapping was the focus of our research.

The motivation for this work is as follows:

At present, there are many studies on soil heavy metals estimation. Since lacking of clear mechanism for heavy metals estimation, most of works utilized the statistical methods or correlation analysis between the heavy metals and organic matter, which is hard to reveal the heavy metals absorption characteristics and explain the absorption mechanism of soil heavy metals estimation. Meanwhile, the low content of heavy metals in soil, complex imaging environment, and spectral data redundancy result in the poor performance of heavy metals estimation, which conducts heavy metals distribution trend cannot be effectively mapped.

The main contributions of this work are as follows.

(1) We analyze and derive the spectral absorption features to explore the heavy metal estimation mechanism.
(2) We utilize ensemble learning methods to improve model reliability and generalization.

The model can lay a foundation for exploring a universal soil heavy metal estimation model and provide an important prerequisite for environmental quality monitoring in soil on a large scale quickly and efficiently.

2. Data acquisition

2.1. Study area

The study region of 139 km^2 is located in Siping city, Jilin province, China (see Fig. 1). The climate of Yitong County belongs to the cold temperate monsoon climate. The annual average temperature, the annual average precipitation and the frost-free period are 5.5 °C, 651.7 mm, and 138 days, respectively. There is a river crossing the sampling area from northwest to southeast. In Yitong County, there are more than 30 kinds of metallic, non-metallic, and energy minerals, including gold, silver, copper, silica, fluorite, kaolin, coal, oil, natural gas, and more than 200 mineral deposits have been discovered. This study area was also researched by Ou (Ou et al., 2021) and Tan (Tan et al., 2020). Fig. S1 shows the slope map of the study area. Advanced Spaceborne Thermal Emission and Reflection Global Digital Elevation Model (ASTER GDEM) data, with a spatial resolution of 30 m, is utilized to generate the slope map.

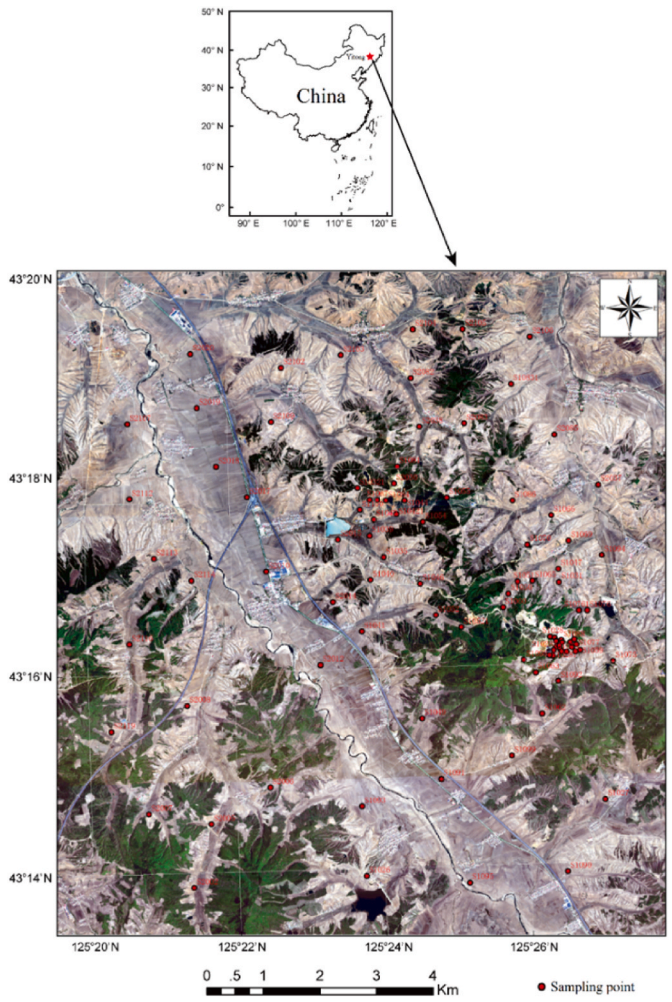

Fig. 1. The locations of the sampling sites in study region.

2.2. Datasets

2.2.1. Soil samples collection and analysis

93 ploughed soil points were sampled based on grid sampling method in late April 2017 in this study. The sample depth is 0~20 cm. Sampling sites were densely laid near potential pollution sources such as mining areas, factories and residential areas in the study area to better analyze the sources of hazardous substances in the soil. Soil samples were chosen as pure as possible to reduce the impact of spectral mixing.

Real-time kinematic (RTK) positioning acquired high-precision coordinate of each sampling point. After the coordinates of the sampling point were selected, 400 g of soil were sampled at the sampling point and four points around it with a diameter of 1 m. The surface soil samples were collected and mixed into sealed bags. Records and photos were taken during the soil samples collection for subsequent viewing of the geographical conditions. The samples were removed from debris, air dried, ground, and 100-mesh screened. The chemical analysis of the soil components and the laboratory spectroscopy were conducted for each sample.

Vis–NIR spectra were measured by an Analytical Spectral Devices (ASD) field spectrometer with a wavelength range of 350–2500 nm. This device has a spectral resolution of 1.4 nm over the 350–1000 nm region and 2 nm over the 1700–2500 nm region, and records the reflectance at 1 nm intervals. Soil samples were loaded in petri dishes and distributed evenly to reduce the effect of unequal particle size. The measurement probe was positioned perpendicular to each sample surface 10 times. For each sample, after anomalous spectra removing, the remaining spectra were averaged as the final spectrum. The laboratory spectra are shown in Fig. S1.

At 350–1300 nm, the spectral curve shows an upward trend and a large slope. There exist a small absorption peak at 480 nm. The increase trend of the spectra is slower over 1500–1900 nm than that over 350–1300 nm. The spectra begin to decrease with an obvious absorption peak near 2200 nm. 940 nm, 1400 nm and 1900 nm are water vapor absorption bands.

We included the inductively coupled plasma-mass spectrometry (ICP-MS) to obtain the soil heavy metal concentration chemically. The concentrations of heavy metal were analyzed statistically. Table 1 shows the statistical results.

Data dispersion can be reflected in CV. The CV of *As* is larger than 1, indicating that its variation is very high. The maximum and minimum values show that the *As* content data set has a large degree of dispersion and sample unbalance Meanwhile, there may also be a high-concentration accumulation zone of *As*.

Total of 93 soil samples are divided into training set and testing set, which is composed according to the 2:1 segmentation rule at each soil samples group. Finally, 62 samples are selected into the training set and 31 samples are selected into the test set.

2.2.2. Imagery acquisition and preprocessing

The HyMap-C imaging spectrometer is an airborne hyperspectral imaging system with a wavelength range of 400–2500 nm, which is operated by the HyVista corporation of Australia. It has a spectral range of 400–2500 nm and four detectors covering the VNIR (400–905 nm and 880–1440 nm) and SWIR (1400–1960 nm and 1950–2500 nm) regions with a spectral resolution of 13–17 nm and 36 bands per detector. The spectrometer provides 144 spectral bands of the target. The flight altitude is 1800m, with a spatial resolution of 3.5m.

During the image acquisition process, due to the instability of the airborne platform and the disturbance of airflow, terrain, and other factors, the images inevitably contain geometric distortion. The geometric correction process included rough geometric correction and fine correction. The rough geometric correction aimed to correct the hyperspectral image distortion according to the RTK data, attitude data, eccentricity vector, and digital elevation model; that is, using the principle of collinear equations in photogrammetry to inversely construct the local geodetic coordinates of the pixels. Fine geometric correction was completed by image-to-image matching in ENVI using orthophotos with a high spatial resolution.

The radiation correction of the images involved radiometric calibration and atmospheric correction. Radiometric calibration was taken to convert the brightness value of the image into an absolute radiance value. The process involved wavelength calibration using monochromator and energy accuracy calibration using integrating sphere. The atmospheric correction is performed to obtain the true reflectance values. The atmospheric radiation transmission model used for the atmospheric correction was MODTRAN (Berk et al., 1999).

Airborne hyperspectral spectra of the 93 samples were extracted from the imagery shown in Fig. S3. In the vicinity of 1400 nm and 1900 nm, the curves are distorted or even show negative values as a result of water vapor influence. In the modeling process of the airborne hyperspectral data, the bands ranging over 1369–1418 nm and 1835–1912 nm with serious water vapor influence were removed.

For this study, bare soil needs to be extracted from the imagery. The spectral unmixing method involves extracting a certain feature component and a proportion of each component from a spectrum in which a

Table 1
Basic statistics of *As* content.

Element	Max (mg/kg)	Min (mg/kg)	Mean (mg/kg)	Std. (mg/kg)	CV
As	419.9602	6.3509	42.7515	67.9779	1.5901

plurality of components is mixed. There are two main steps in spectral unmixing. First of all, select several representative pure spectra of the four objects: pure bare soil, buildings, water, and vegetation. Then estimate fractional abundances of the four kinds of objects by the fully constrained least squares (FCLS) method (Heinz and Chang, 2001). Each pixel whose soil abundance was larger than 0.65 was defined as bare soil. Bare soil image in the study area is shown in Fig. S4.

2.3. Feature analysis

2.3.1. Continuum removal

A spectral curve that highlights absorption and reflection characteristics was obtained by removing from the continuum (Van der Meer, 2006). The continuum is connected by the extreme value points on the spectral curve, which is utilized to divide the original curve to obtain the continuum removed curve.

Fig. 2 takes a laboratory spectral curve as an example to show the spectral curve and the continuum-removed spectral curve, where the solid blue line is the laboratory spectral curve and the red solid line is its continuum spectral curve. The green dashed line is the curve after continuum removal. After continuum removal, the absorption characteristics caused by the soil components can be obviously reflected on the spectral curve, which is a foundation for the following analysis on the components' occurrence state in soil based on absorption characteristic. The absorption peaks on the spectral curve, such as 500 nm, 1400 nm, 1900 nm, and 2200 nm, are more obvious after continuum removal.

2.3.2. Extraction of spectral absorption parameters

The parameters reflecting spectral absorption characteristic in this study included absorption depth (D) and absorption area (A) shown in Fig. S5.

2.4. Modeling methods

2.4.1. Traditional methods

In linear regression, least squares (LS) is the most basic method, which is useful to bridge the dependent variables and independent variables (Markovsky and Huffel, 2007). Partial least squares (PLS) (Leone et al., 2012) extracts the latent variables, i.e., principal components, from the independent variables and the dependent variables, respectively. A linear model then is built by the latent variables of the independent and dependent variables. Ridge regression (RR) is an improved least square estimation method and can handle the non-full rank coefficient matrix. Support vector machine (SVM) is a kernel-based method(Vapnik, 1997). It can come to a better settlement on small samples, over-learning, non-linear, high-dimensional numbers and other practical situations.

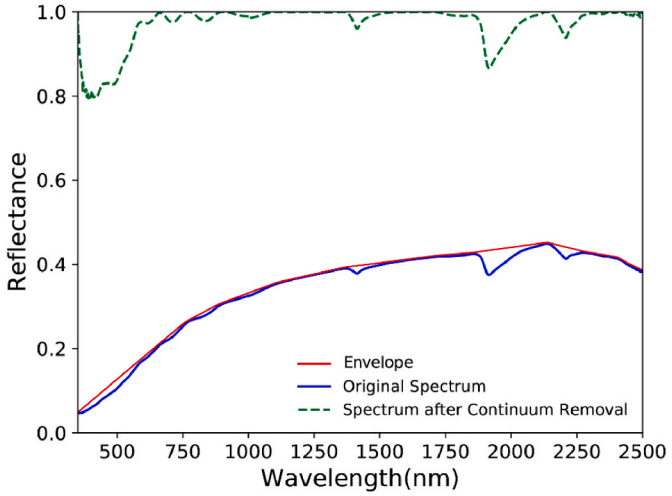

Fig. 2. Schematic diagram of the continuum removal process.

2.4.2. Decision tree (DT)

DT, a base learner, can be regarded as a tree model. Each node in the DT represents the attribute of the object. Starting from the following node, it passes through several intermediate nodes to the leaf node, and the path represents the predicted rule.

2.4.3. Random forest (RF)

RF, an improvement of the DT algorithm, is proposed by Breiman (2001). This parallel ensemble algorithm consisting of DT is an extended variant of Bagging. It combines multiple weak classifiers and introduces random attribute selection, which makes the overall model result with high accuracy, excellent generalization performance, and good stability.

2.4.4. Extremely randomized trees (ET)

In addition to RF method, ET [49] method improves on it. Compare with RF, the randomness of the split point calculation method is further enhanced. RF divides each tree's learning by a random subset of characteristic, while ET obtained the threshold randomly for each candidate feature, and the best of these is selected as the segmentation rule.

2.4.5. AdaBoost

AdaBoost is an excellent boosting algorithm, and a high prediction accuracy can be achieved by improving several weak learners with low prediction accuracy through it. AdaBoost assigns a weight to each sample during the training process, to control the learning degree of the sample by the weak learner. It then assigns a weight to each weak learner, calculates the inner product between the weak learner and weight, and finally obtains a strong learner. The sample of the AdaBoost algorithm is (x_i, y_i), and the set of weak learners is $H = (h_1, h_2, h_3, \ldots, h_n,)$. When n weak learners are integrated, the expression of the strong learner is as follows:

$$f(x) = sign\left(\sum_{t=1}^{n} \omega_t h_t(x)\right) \quad \omega_t > 0, t = 1, 2, \ldots, n \tag{2}$$

where $f(x)$ is the strong learner, $h_t(x)$ is the weak learner, and ω_t is the weight of the weak learner. ω_t is normalized as:

$$\sum_{t=1}^{n} \omega_t = 1 \tag{3}$$

The inner product of the weak learner and the weight of the AdaBoost algorithm are calculated by *sign*() to obtain the final prediction result.

2.5. Model evaluation method

Five accuracy indicators are employed, the mean absolute error (MAE), the root-mean-square error (RMSE), the coefficient of determination (R^2), the ratio of prediction performance to interquartile range (RPIQ) and the residual prediction deviation (RPD). Each indicator is calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|obs_i - pred_i| \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(obs_i - pred_i)^2}{N}} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(obs_i - pred_i)^2}{\sum_{i=1}^{N}(obs_i - \overline{obs})^2} \quad (6)$$

$$RPIQ = \frac{IQ}{RMSE_P} \quad (7)$$

$$RPD = \frac{SD}{RMSE_P} \quad (8)$$

$$IQ = Q_3 - Q_1 \quad (9)$$

where *obs* denotes the observed value, *pred* denotes the predicted value, and IQ is the interquartile distance of the prediction set. R_C^2, $RMSE_C$, and MAE_C represent the calibration data set evaluation, and R_P^2, $RMSE_P$, and MAE_P express the prediction data set evaluation.

3. Results

3.1. Study of the retrieval mechanism based on the laboratory spectra

3.1.1. Characteristics analysis

By continuum removal, absorption characteristics can be effectively highlighted. These characteristics belong to the bands which are near 420 nm, 480 nm, 610 nm, 700 nm, 840 nm, 1400 nm, 1780 nm, 1900 nm and 2200 nm. Among them, 1400 nm and 1900 nm are the water vapor absorption bands. The red circles indicate the locations of the selected absorption features in Fig. 3.

The ions (Fe^{2+}, Fe^{3+} and Mn^{3+}) have electron transition which caused the spectral absorption over the 380–780 nm. Iron/manganese oxides caused the spectral absorption at 420 nm and 480 nm(Xu et al., 2011a,b). The harmonic frequency and double frequency of the molecular groups of organic matter and clay minerals resulted in the absorption over 780–2500 nm (Viscarra Rossel and Behrens, 2010; Knadel et al., 2013).

The absorption characteristics at 1780 nm and 2200 nm are concerned in organic matter (BenDor et al., 1997; Viscarra Rossel and Behrens, 2010; Knadel et al., 2013). One thousand and nine hundred nm and 2200–2400 nm features are relevant to the double frequency of the molecular hydroxyl (-OH) stretching vibration of the clay minerals (Clark et al., 1990; Knadel et al., 2013; Chen et al., 2022). Fig. S6. illustrates the correlation coefficients. The red circles indicate the regions located in the absorption features that have a high correlation with *As*.

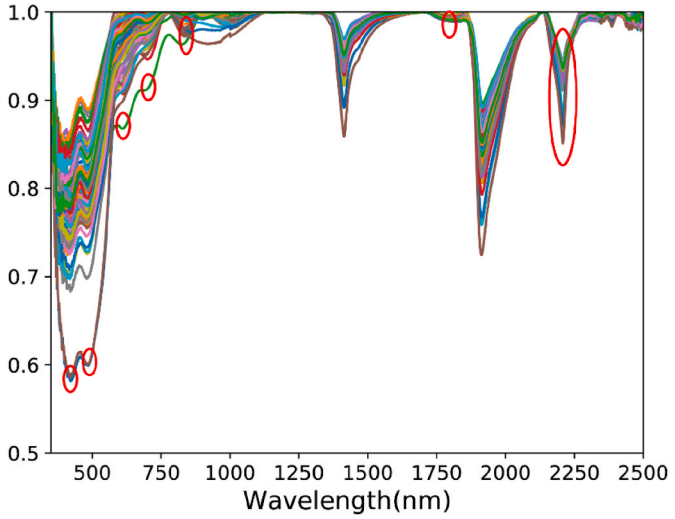

Fig. 3. Laboratory spectra after continuum removal.

In Fig. S6, with regard to *As*, there are highly correlated bands around 420 nm, 480 nm, 1780 nm, and 2200 nm. For the estimation of *As*, 11 spectral absorption parameters were extracted, including D_{420}, D_{480}, D_{610}, D_{700}, D_{840}, D_{1780}, D_{2200}, A_{2200}, and lgA_{2200}. $R_{(2153\text{-}2102),(2153+2102)}$ and $R_{2153,2102}$ were also selected due to there being a sudden drop from 2102 nm to 2153 nm. Shi et al. (2016) reported the wavebands at 480 nm, 600 nm and 810 nm have high correlation with *As* concentration. Xu et al. (2017) found that the wavebands at 1778 nm have significant relationships with *As*. These characteristics bands are similar to our findings.

3.1.2. Modeling using a single variable

The heavy metals are sorted by concentration gradient, two out of every three samples are used as the training dataset, and the remainder is set as the training dataset.

As concentration was estimated using the 11 spectral absorption characteristic parameters, and the regression coefficients were obtained by the LS method (Table S1).

In the visible range, the estimation performance for *As* by D_{480} is better than by the other parameters, indicating that the iron/manganese oxide in soil has an adsorption effect on *As*. In the near-infrared range, the effect of estimating *As* with D_{2200} is better than that for D_{1780}, and the R_p^2 is 0.4394. Moreover, the R_p^2 of $R_{(2153\text{-}2102),(2153+2102)}$ can reach over 0.6. The 2200 nm band is the response region for clay minerals and organic matter. It is obvious that iron/manganese oxide, organic matter and clay minerals all have adsorption effect on *As*. However, there is obvious underfitting of the model estimated by a single variable, and the R_C^2 is clearly smaller than the R_p^2. Therefore, we introduce multivariate variables to improve the model performance.

3.1.3. Modeling using multivariate variables

According to the results shown in Table S1, several single spectral characteristic parameters were utilized to estimate the concentrations of *As*, and to fully explore the effective characteristics. Since it is inconvenient to display the number of dependent variables in the table, only the variable number is written, which is the same as the numbering in Table S1.

Table 2 reports the estimation results for *As* using multiple variables.

The first set of variables (2, 6–11) is the seven variables whose R_p^2 is higher than 0.2, according to the results in Table S1. The performance after combination of the variables is better than that of a single variable. In the first set, the R_p^2 improves from 0.6356 to 0.7826, while the R_C^2 is also greatly improved from 0.1763 to 0.4311. These results indicate that the multiple variables can promote the model performance. The second set of variables (1–9) is a combination of absorption depth and absorption area, which performs slightly worse than the model established by the first set. In the second set, the R_C^2 and R_p^2 are 0.3817 and 0.6502, respectively. The fourth set of variables (1–11) includes the 11 characteristics of absorption depth, absorption area, and band ratio. The model for *As* estimation performs the best using this set of variables, in which the R_C^2 is 0.4892, the R_p^2 is 0.8474, and *RPD* is 2.2665. In summary, the combination of the 11 spectral absorption parameters, and especially the band ratio, can promote the estimation performance.

The fourth set of variables is based on the following models: $Y = 391.84 * D_{420} - 812.11 * D_{480} + 324.99 * D_{610} + 208.32 * D_{700} + 1699.88 * D_{840} - 10{,}970.67 * D_{1780} + 4789.35 * D_{2200} - 744.53 * A_{2200} + 6066.31 * lgA_{2200} - 5{,}269{,}473.06 * R_{(2153\text{-}2102),(2253+2102)} + 2{,}611{,}461.33 * R_{2153,2102} - 2{,}612{,}220.65$.

The estimation scatter diagram for the fourth set is shown in Fig. S7.

In Fig. S7, the points of the predicted and observed values are concentrated in the lower-left corner, and several larger values are loosely distributed, indicating the imbalance of the samples. The values of the predicted set are concentrated near the 1:1 line, which is consistent with the fact that the R_p^2 is larger than the value of the R_C^2.

Table 2
Estimation results using multiple variables.

X Number	R_C^2	$RMSE_C$	MAE_C	R_p^2	$RMSE_P$	MAE_P	RPD	RPIQ
2,6–11	0.4311	44.4691	29.8063	0.7826	41.7166	30.0040	1.9643	0.7709
1–9	0.3817	46.3626	31.5444	0.6502	51.6368	34.0138	1.5869	0.6228
10,11	0.3196	48.6337	32.6184	0.5334	56.9930	39.0300	1.4378	0.5643
1–11	0.4892	42.1372	29.6831	**0.8474**	36.1542	27.6927	2.2665	0.8895

3.2. Estimation based on airborne hyperspectral imagery

3.2.1. Characteristics analysis

The spectra absorption characteristics could be effectively highlighted by continuum removal, and were 600 nm, 1300 nm, 1600 nm, 2000 nm, and 2200 nm (Fig. S8). The spectral absorption characteristics that were the same as the laboratory spectra were 600 nm, 1600 nm, and 2200 nm. Due to the differences in the spectral resolution between the laboratory spectra and imagery spectra, the absorption characteristics differ from the absorption characteristics of the laboratory spectra.

The spectral absorption parameters of the airborne hyperspectral imagery were D_{600}, D_{1300}, D_{1630}, D_{2000}, D_{2200}, A_{2200}, and lgA_{2200}. The selection of band and band ratio depends on the actual conditions of heavy metal. The selection method for the band ratio was the same as for the laboratory spectra analysis method. According to the relationship between spectra and heavy metal content after continuum removal, two bands with a sharp increase and decrease were selected as a band ratio. This method combines the bands with high positive correlation and high negative correlation, which can achieve a better performance of the heavy metal estimation.

The correlation coefficients between the *As* concentrations and the image spectra after continuum removal are shown in Fig. S9. The red circles indicate the regions located in the absorption features that have high correlation with *As*.

The bands which have a high correlation with *As* are around 500 nm, 1300 nm, 1600 nm, and 2200 nm, which is consistent with the characteristic bands of the laboratory spectral analysis at 500 nm, 1780 nm, and 2200 nm. The correlation shows a sharp increase from 1449 nm to 1628 nm, as well as 2210 nm and 2240 nm, so the two pairs of bands were selected as band ratios. Therefore, there were 11 spectral absorption characteristic parameters selected to estimate *As*: D_{600}, D_{1300}, D_{1630}, D_{2000}, D_{2200}, A_{2200}, lgA_{2200}, $R_{(1628-1449),(1628+1449)}$, $R_{1628,1449}$, $R_{(2240-2210),(2240+2210)}$, and $R_{2240,2210}$. It has been reported that the wavebands at 591 nm, 1299 nm, 2015 nm, and 2260 nm are associated with *As* (Wu et al., 2021). In Ou et al. (2021) research, 1290 nm, 1640 nm and 2210 nm are selected to conduct *As* distribution map. These characteristics bands are similar to our findings.

3.2.2. Model evaluation

In order to establish a better estimation model, we introduced the ensemble methods of ET and AdaBoost. LS, PLS, RR, SVM, and DT were compared to explore the superiority of the two methods.

In the analysis of *As* concentration estimation by laboratory spectra, the spectral absorption characteristic parameters combination can perform better than just the LS method. *As* was estimated using the 11 spectral absorption characteristic parameters extracted from the airborne imagery spectra. Table S2 lists the results.

The performance of the eight models established by only the 11 spectral absorption characteristic parameters is poor. The R_p^2 values of the RF and ensemble methods are all superior to those of the other methods, but the three ensemble methods all show overfitting. Table S3 shows the results of *As* estimation using the 30 highly correlated characteristic bands after continuum removal.

The performance when using the 30 bands with high correlation to establish an estimation model is better than when using the 11 spectral absorption characteristic parameters. The three ensemble methods all perform well. Based on the AdaBoost model, the R_P^2 of *As* is 0.5878, the $RMSE_P$ is 44.5263, and the MAE_P is 21.1311.

A total of 41 characteristics were composed of the 11 spectral absorption parameters and 30 bands. These characteristics were utilized to estimate *As*. The estimation results of the eight models are reported in Table 3.

The eight models established with 41 characteristics are all superior to those using 11 spectral absorption characteristic parameters or 30 bands. For *As*, the R_C^2 and R_p^2 of the linear methods (LS, PLS, Ridge) are similar. Compared with the other methods, the accuracies of the three ensemble models are poor, which shows that the relationship between airborne hyperspectral spectra and *As* concentration is complex, so the relationship cannot be expressed by a linear model. SVM performs slightly better than the linear methods. For *As*, the R_C^2 of DT is more than 0.9, while the R_p^2 is less than 0.1, so it can be seen that the training dataset has an abnormally high accuracy. This shows that the DT estimation model shows serious overfitting, and it cannot estimate *As*, even though the R_c^2 values are better. Based on the AdaBoost model, the R_P^2 for *As* is 0.6242, the $RMSE_P$ is 43.6481, and the MAE_P is 20.5760. The R_P^2 is the highest, while the $RMSE_P$ and the MAE_P are the lowest among all the methods. Therefore, the introduced methods AdaBoost are stable and able to predict the concentration of *As* well.

For *As*, the performance of the AdaBoost is the best, and Fig. 4 draws the estimation scatter diagram of it.

In Fig. 4, the measured-predicted points of the calibration set are distributed well around the 1:1 line, but the accuracy of the predicted set is lower. The smaller value distribution in the lower-left corner is denser, and the larger value distribution in the upper-right corner is sparser, indicating that there are several high leverage values in the *As* data set.

In Fig. 5, the measured-predicted points of the calibration set are distributed well around the 1:1 line, but the accuracy of the predicted set is lower. The smaller value distribution in the lower-left corner is denser, and the larger value distribution in the upper-right corner is sparser, indicating that there are several high leverage values in the *As* data set.

3.3. Mapping heavy metal using airborne hyperspectral imagery

Based on the estimation characteristics and models, *As* was estimated to analyze their distribution patterns and influencing factors in the whole study area.

Fig. 5 draws the estimation results for *As* concentration of the study area. The regions with high concentration values indicated by A-D are marked. The criterion of *As* referred to is the National Standard (GB15618–2018, GB15618-1995) in China, which is shown in Table 4 (Agency, 1995; Regulation, 2018).

In the early experiment period, the situation of the study region was investigated. There are four mining areas and one concentrating mill in the study area and nearby. The river running from the northwest to the southeast is the Yitong River, for which the flow direction is from south to north. There is also a small river channel in the northeast corner. The Yingsong Expressway is running nearly parallel to the right side of the Yitong River. The main bare soil in the study area exists in farmland. Fig. 6 shows the detailed landscape.

We have conducted a field survey of the study area. In Fig. 6, area A is artificial activity areas. The farming and human activities are likely to have made the concentrations of heavy metal differ from the background values. The inconsistent management of cultivated land and

Table 3
Estimation results using all 41 features extracted from the image spectra.

Element	Method	R_C^2	$RMSE_C$	MAE_C	R_p^2	$RMSE_P$	MAE_P	RPD	RPIQ
As	LS	0.5373	46.4199	30.9838	0.3562	56.4412	40.5609	1.2090	0.5445
	PLS	0.4651	82.5993	53.4361	0.3529	84.1353	56.7201	0.8111	0.3653
	Ridge	0.5365	46.4599	30.8491	0.3537	56.3917	40.2052	1.2101	0.5450
	SVM	0.5438	51.3722	21.2991	0.4491	53.3276	29.1093	1.2796	0.5763
	DT	0.9835	8.7607	4.7337	0.0630	96.2933	52.3934	0.7087	0.3191
	RF	0.9379	21.5117	13.2719	0.4471	51.1683	32.9125	1.3336	0.6006
	ET	0.9992	2.0033	1.2821	0.5540	46.2785	29.0105	1.4745	0.6640
	AdaBoost	0.9978	3.1872	1.0896	**0.6242**	43.6481	20.5760	1.5634	0.7041

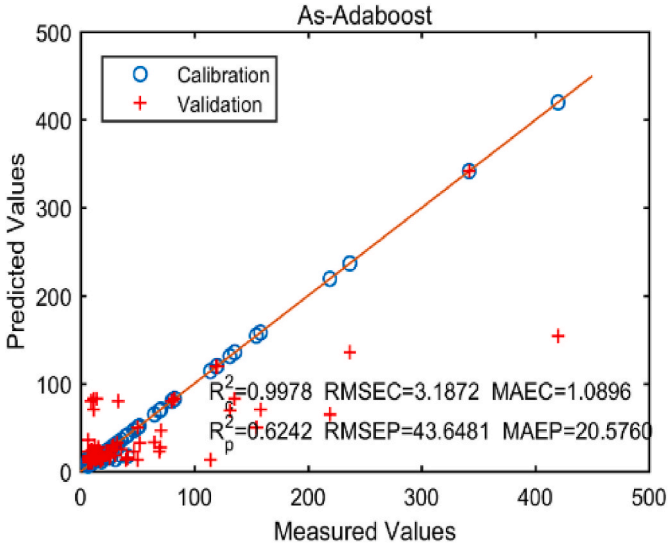

Fig. 4. The scatter plot of *As* estimation using AdaBoost. (unit: mg/kg).

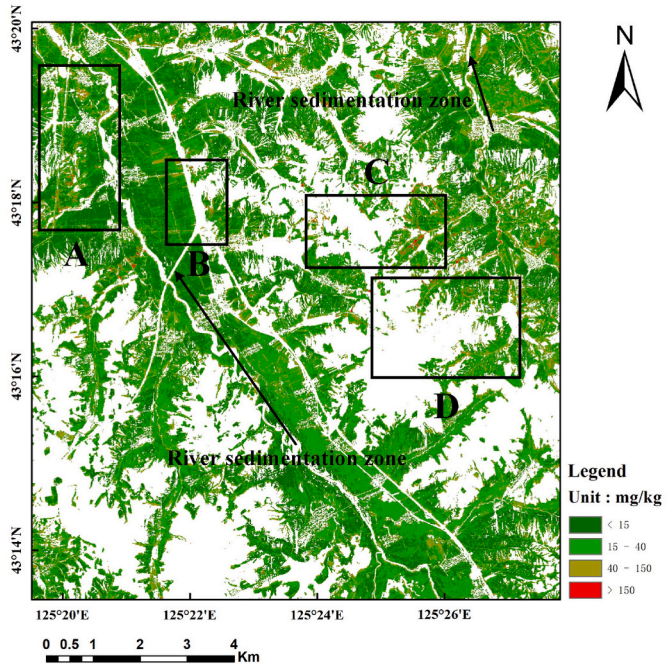

Fig. 5. Estimation map for *As*.

livestock excrement pollution further exacerbate the As pollution of cultivated land in the study area. Area B is a concentrating mill. The pollutants such as waste water and waste slag discharged from the concentrating mill will have likely resulted in an increasement of the soil heavy metal concentration. This area is located next to the national highway, the field survey reveals that the road is in subpar condition. The dust and the truck exhausts contribute to the As pollution in the surrounding farmland. Both areas C and D are gold deposits. There will be some residual heavy metal ions in the tailings of the gold deposits, which are likely to have entered the farmland or groundwater via rainwater. Moreover, mineral powder will also exist in the tailings, which can be blown into the farmland and villages by wind, posing a threat to the health of residents. Area D is a gold mining area that is being mined without strong environmental protection measures in the production process. Area C used to be a gold mine, but is no longer mined, and the surrounding environment has been restored.

Table 4
Soil environmental quality risk control standard for the soil contamination of agricultural land (unit: mg/kg).

Metal	Background	Risk screening values	Risk intervention values
As	15	40	150

Fig. 5 shows that the content of *As* is higher in the mining areas (areas C and D), the concentrating mill (area B), the artificial activity areas (area A), and near the river sedimentation area, but are generally less than 95 mg/kg. Areas C and D, which are near the gold deposits, have the highest content of *As*, with values above 300 mg/kg. According to the National Standard, there may be a risk of soil pollution if the content of *As* in soil is more than 40 mg/kg. When it is more than 150 mg/kg, edible agricultural products are not in accordance with the safety standards. Therefore, serious *As* pollution exists in the study area.

As the most serious polluted areas of heavy metal (*As*), in the vicinity of areas C and D, the gold deposits have a bad impact on the soil. *As* is a companion of many non-ferrous metals, and is transferred from the depth of formation to the surface in the mining process of the gold mine. Therefore, the *As* content is high in the tailings. *As* does not migrate, but it does become a substance that can migrate after oxidation, hydrolyzing, and weathering. The surface water flowing through the tailings dissolves the *As*, which then flows into the farmland and rivers nearby, resulting in pollution of the soil and water area. Moreover, the particles containing *As* are disseminated into the air by wind, which can cause *As* poisoning via inhalation. Therefore, the supervision of the production in mining areas should be strengthened to protect the environment. In the human activity areas of A, the long-term use of chemical fertilizers in the cultivated land has led to the heavy metal accumulation. The content of *As* in soil has reached 50–60 mg/kg, because of the long-term application of fertilizers. In the process of cultivation, land resources should be rationally utilized, and fertilized reasonably.

Fig.S10 shows the distribution trend (a) and interpolated image (b) of As. In Fig.S10 (b), the red area is gold deposit which cause serious pollution.

4. Conclusion and prospects

In this paper, we have analyzed soil hyperspectral data, combined

Fig. 6. Geomorphological analysis of the study region.

this with the soil physicochemical properties, and extracted the spectral absorption characteristic parameters, to model the estimation heavy metal (*As*). In the laboratory spectral analysis, it was found that the LS method for *As* estimated by 11 features (including absorption depth, absorption area, and the band ratio around 2200 nm) had the highest accuracies in the multivariate variables analysis. It was concluded that *As* is adsorbed by organic matter, iron/manganese oxides, and clay minerals in soil, and the adsorption of *As* by first two components is greater than that of the last. When estimating *As* content based on airborne spectra, it is impossible to accurately make an estimation only using the absorption features, but the estimation could be achieved by combining the parameters and the high-correlation bands. AdaBoost was found to have the best prediction accuracy. The *As* content in the study area exceeds the National Standard level. After analyzing the influencing factors, it was concluded that the mining areas and concentrating mill in the study area have had the most serious impacts on the soil environment. The content of *As* around farmland soil is also significantly higher than the natural background value. Furthermore, *As* has also been deposited in the sedimentation areas along the rivers. Human agricultural behavior has impacted on the soil heavy metal content of the study area.

This study conducts on soil heavy metals estimation and traceability analysis with hyperspectral image, which provides a solid foundation for heavy metals pollution risk assessment, heavy metal pollution prevention and control, and crop health risk assessment. Meanwhile, it is also an important premise for government departments to carry out soil environmental governance and food security.

The retrieval mechanism was analyzed from the spectral absorption characteristics, and it was not analyzed from the physical model, so the model of soil radiation transfer could be considered in future work. In future work, we will attempt to investigate the correlation between soil heavy metal concentration and anisotropy reflectance characteristics of soil to estimate heavy metal content.

Credit author statement

Kun TAN: Conceptualization, Methodology, Writing-Reviewing. Lihan CHEN: Data curation, Writing – original draft. Huimin WANG: Methodology, Visualization, Writing-Reviewing. Zhaoxian LIU: Supervision. Jianwei DING: Investigation. Xue WANG: Writing-Reviewing, Editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

This work is jointly supported by the National Natural Science Foundation of China (grant nos. 42171335, 42001350), the Shanghai Municipal Science and Technology Major Project (grant no. 22511102800), the Postdoctoral Science Foundation of China (grant nos. 2023T160218), the National Civil Aerospace Project of China (grant no. D040102) and Supported by the International Research Center of Big Data for Sustainable Development Goals (CBAS2022GSP07).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2023.119196.

References

Agency, N.E.P., 1995. Environmental Quality Standard for Soils. National Environmental Protection Agency,Beijing. 1–5.
Anagu, I., Ingwersen, J., Utermann, J., Streck, T., 2009. Estimation of heavy metal sorption in German soils using artificial neural networks. Geoderma. 152 (1–2), 0-112.
Balabin, R.M., Lomakina, E.I., 2011. Support vector machine regression (SVR/LS-SVM)–an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst. 136 (8), 1703–1712.
BenDor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400-2500 nm) during a controlled decomposition process. Rem. Sens. Environ. 61 (1), 1–15.
Berk, A., Anderson, G.P., Bernstein, L.S., Acharya, P.K., Hoke, M.L., 1999. MODTRAN4 radiative transfer modeling for atmospheric correction. Proc. SPIE-Int. Soc. Opt. Eng. 3756, 348–353.
Bian, Z., Sun, L., Tian, K., Liu, B., Huang, B., Wu, L., 2023. Estimation of multi-media metal (loid) s around abandoned mineral processing plants using hyperspectral technology and extreme learning machine. Environ. Sci. Pollut. Res. 30 (7), 19495–19512.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Cao, H.C., Wang, J.D., Zhang, X.L., 2007. Study on the association between heavy metals and organic matter in polluted black soil in Northeastnortheast China. Research of Environmental Sciences 1, 36–41.
Chen, H., Jilili, A., Liu, W., Chen, J., 2016. Correlation between heavy metals organic Matter,pH value in the soils along the Bortala River. Res. Soil Water Conserv. 23 (5), 210–213.
Chen, L., Lai, J., Tan, K., Wang, X., Chen, Y., Ding, J., 2022. Development of a soil heavy metal estimation method based on a spectral index: combining fractional-order derivative pretreatment and the absorption mechanism. Sci. Total Environ. 813, 1–12.
Chen, Y.J., Wong, M.L., Li, H., 2014. Applying Ant Colony Optimization to configuring stacking ensembles for data mining. Expert Syst. Appl. 41 (6), 2688–2702.
Choe, E., Meer, F.V.D., Ruitenbeek, F.V., Werff, H.V.D., Smeth, B.D., Kim, K.W., 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. Rem. Sens. Environ. 112 (7), 3222–3233.
Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G.A., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. 95 (B8), 12653–12680.
Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. 40 (2), 139–157.
Heinz, D.C., Chang, C.-I., 2001. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Rem. Sens. 39, 529–545.
Jin, H.N., Zhang, X.L., Liu, H.J., Kang, R., Fu, Q., Ning, D.H., 2016. Soil moisture predicting model based on spectral absorption characteristics of the soil. Acta Pedol. Sin. 83 (3), 627–635.
Jin, J., Zhou, X.P., 2017. Quantitative inversion of heavy metal Cd content in middle-west area of inner Mongolia chestnut Soil. J. Inner Mongolia University of Sci. Technol. 36 (3), 280–284.
Knadel, M., Viscarra Rossel, R.A., Deng, F., Thomsen, A., Greve, M.H., 2013. Visible–near infrared spectra as a Proxy for topsoil texture and glacial boundaries. Soil Sci. Soc. Am. J. 77 (2), 568–579.
Kokaly, R.F., Clark, R.N., 1999. Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. Rem. Sens. Environ. 67, 267–287.
Leone, A.P., Rossel, R.A.V., Amenta, P., Buondonno, A., 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.
Liao, H.W., Zhou, D.L., 2012. Review of AdaBoost and its improvement. Computer Systems & Applications 21 (5), 240–244.
Liu, K., Zhao, D., Fang, J.Y., Zhang, X., Zhang, Q.Y., Li, X.K., 2017. Estimation of heavy-metal contamination in soil using remote sensing spectroscopy and a statistical approach. J. Indian Society of Remote Sensing 45 (5), 805–813.
Lu, Y., Bai, Y., Yang, L., Wang, H., 2007. Hyperspectral extraction of soil organic matter content based on principal component regression. N. Z. J. Agric. Res. 50 (5), 1169–1175.
Ma, W.B., Tan, K., Du, P.J., 2016a. Predicting soil heavy metal based on Random Forest model. IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 4331–4334.
Ma, W.B., Tan, K., Li, H.D., Yan, Q.W., 2016b. Hyperspectral inversion of heavy metals in soil of a mining area using extreme learning machine. J. Ecol. Rural Environ. 32 (2), 213–218.
Markovsky, I., Huffel, S.V., 2007. Overview of total least-squares methods. Signal Process. 87 (10), 2283–2302.
Meer, F.V.D., 2004. Analysis of spectral absorption features in hyperspectral imagery. Int. J. Appl. Earth Obs. Geoinf. 5 (1), 55–68.
Merdy, P., Huclier, S., Koopal, L., 2006. Modeling METAL-PARTICLE INTERACTIONS with an Emphasis on natural organic matter. Environ. Sci. Technol. 40 (24), 7459–7466.
Ou, D., Tan, K., Lai, J., Jia, X., Li, J., 2021. Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction. Geoderma 385, 114875.
Pandit, C.M., Filippelli, G.M., Lin, L.I., 2010. Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression. Int. J. Rem. Sens. 31 (15), 4111–4123.
Peng, X.T., Gao, W.X., Wang, J.J., 2014. Inversion of soil parameters from hyperspectra based on continuum removal and partial least squares regression. Geomatics Inf. Sci. Wuhan Univ. 39 (7), 862–866.
Qi, X.X., Qian, M.Y., Jun Feng, J., Hong, R., Jun, C., Lin, L.Q., 2007. Reflectance spectroscopy study of Cd contamination in the sediments of the Changjiang River, China. Environ. Sci. Technol. 41 (10), 3449–3454.
Regulation, S.A.f.M., 2018. Soil Environmental Quality Risk Control Standard for Soil Contamination of Agricultural Land, GB1vols. 5618–2018. Ministry of Ecological Environment of the people's Republic of China, Beijing, pp. 1–7.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818.
Saha, A., Sen Gupta, B., Patidar, S., Martínez-Villegas, N., 2022. Identification of soil arsenic contamination in rice paddy field based on hyperspectral reflectance approach. Soil Systems 6 (1), 1–26.
Shi, T., Cui, L., Wang, J., Fei, T., Chen, Y., Wu, G., 2013. Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. Plant Soil 366 (1–2), 363–375.
Shi, T.Z., Wang, J.J., Chen, Y.Y., Wu, G.F., 2016. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. Int. J. Appl. Earth Obs. Geoinf. 52, 95–103.
Song, L., Jian, J., Tan, D.J., Xie, H.B., Luo, Z.F., Gao, B., 2015. Estimate of heavy metals in soil and streams using combined geochemistry and field spectroscopy in Wan-sheng mining area, Chongqing, China. Int. J. Appl. Earth Obs. Geoinf. 34, 1–9.
Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R.P., Song, Q., 2005. Boosting: an ensemble learning tool for compound classification and QSAR modeling. J. Chem. Inf. Model. 45 (3), 786–799.
Tan, K., Ma, W., Chen, L., Wang, H., Du, Q., Du, P., Yan, B., Liu, R., Li, H., 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.
Tan, K., Wang, H., Chen, L., Du, Q., Du, P., Pan, C., 2019. 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., Wang, H., Zhang, Q., Jia, X., 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.
Van der Meer, F., 2006. Indicator kriging applied to absorption band analysis in hyperspectral imagery: a case study from the Rodalquilar epithermal gold mining area, SE Spain. Int. J. Appl. Earth Obs. Geoinf. 8 (1), 61–72.
Vapnik, V.N., 1997. The nature of statistical learning theory. IEEE Trans. Neural Network. 8 (6), 1564-, 1564.
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, J., Bai, J., Gao, Z., Lu, Q., Zhao, Q., 2015. Soil as levels and bioaccumulation in Suaeda salsa and Phragmites australis Wetlands of the Yellow river Estuary, China. BioMed Res. Int. 2015, 301898.
Wang, L., Lin, Q.Z., Dong, J., Shi, H.S., Huang, X.H., 2007. Study on the prediction of soil heavy metal Elements content based on reflectance spectra. Journal of Remote Sensing 6, 906–913.
Wang, Y., Niu, R., Hao, M., Lin, G., Xiao, Y., Zhang, H., Fu, B., 2023. A method for heavy metal estimation in mining regions based on SMA-PCC-RF and reflectance spectroscopy. Ecol. Indicat. 154, 110476.
Wei, J., Ming, Y., Liu, F., 2015. Hyperspectral mineral mapping method based on spectral characteristic parameter combination. Earth Sci. 40 (8), 1432–1440.
Wu, F.Y., Wang, X., Liu, Z.X., Ding, J.W., Tan, K., Chen, Y., 2021. Assessment of heavy metal pollution in agricultural soil around a gold mining area in Yitong County, China, based on satellite hyperspectral imagery. J. Appl. Remote Sens. 15 (4), 042613.
Wu, Y., Zhang, X., Liao, Q., Ji, J., 2011. Can contaminant Elements in soils Be assessed by remote sensing technology: a case study with simulated data. Soil Sci. 176 (4), 196–205.
Xia, F., Peng, J., Wang, Q.L., Zhou, L.Q., Shi, Z., 2015. Prediction of heavy metal content in soil of cultivated land:Hyperspectral technology at provincial scale. J. Infrared Millim. Waves 34 (5), 593–598 and 605.
Xu, L., Li, Q., Zhu, X., Liu, S., 2017. Hyperspectral inversion of heavy metal content in coal gangue Filling reclamation land. Spectrosc. Spectr. Anal. 37, 3839–3844.
Xu, M., Wu, S., Zhou, S.L., Liao, F.Q., ChunMei, M.A., Zhu, C., 2011a. Hyperspectral reflectance models for retrieving heavy metal content:application in the archaeological soil. J. Infrared Millim. Waves 30 (30), 109–114.
Xu, M., Zhou, S., Ding, W., Wu, S., Wu, W., 2011b. Hyperspectral reflectance models for predicting soil organic matter content in coastal tidal land area, northern Jiangsu. Trans. Chin. Soc. Agric. Eng. 27 (2), 219–223.
Yang, L.Y., Gao, X.H., Wei, Z., Shi, F.F., Wei, J., 2016. Estimating heavy metal concentrations in topsoil from vegetation reflectance spectra of Hyperion images: a case study of Yushu County,Qinghai,China. J. Appl. Ecol. 27 (6), 1775–1784.
Yao, X., Xiao, R., Ma, Z., Xie, Y., Zhang, M., 2015. Distribution and contamination assessment of heavy metals in soils from tidal flat, oil exploitation zone and restored wetland in the Yellow River Estuary. Wetlands 36 (S1), 153–165.
Ye, M., Zhu, L., Li, X., Ke, Y., Huang, Y., Chen, B., Yu, H., Li, H., Feng, H., 2023. Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. Sci. Total Environ. 858, 159798.
Zhang, J., Shen, Y.T., Wang, X.J., 2011. Hyperspectral quantitative models for chlorophyll-a of algae based on spectral absorption feature parameters and spectral absorption index. Journal of Agro-Environment Science 30 (8), 1622–1629.
Zhang, Q., Zhang, H., Liu, W., Zhao, S., 2017. Inversion of heavy metals content with hyperspectral reflectance in soil of well-facilitied capital farmland construction areas. Trans. Chin. Soc. Agric. Eng. 33 (12), 230–239.