Journal of Hazardous Materials 401 (2021) 123288

Contents lists available at ScienceDirect

Journal of Hazardous Materials

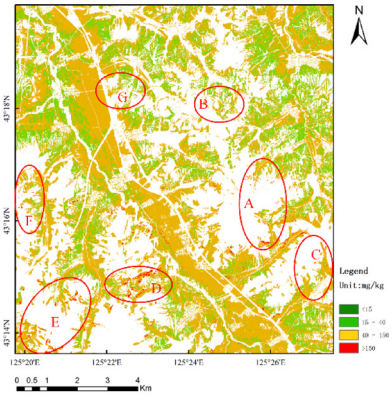
journal homepage: www.elsevier.com/locate/jhazmat

Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning

Kun Tan[a,b,c,*,1], Weibo Ma[c,d,*,1], Lihan Chen[c], Huimin Wang[c], Qian Du[e], Peijun Du[f,*], Bokun Yan[g], Rongyuan Liu[g], Haidong Li[d]

[a] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
[b] School of Geographic Sciences, East China Normal University, Shanghai 200241, China
[c] Key Laboratory for Land Environment and Disaster Monitoring of NASG, China University of Mining and Technology, Xuzhou 221116, China
[d] Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment, Nanjing 210042, China
[e] Department of Electrical and Computer Engineering, Mississippi State University, MS 39762, USA
[f] Key Laboratory for Satellite Mapping Technology and Applications of NASG, Nanjing University, Nanjing 210023, China
[g] China Aero Geophysical Survey&Remote Sensing Center for Natural Resources, Beijing 100083, China

GRAPHICAL ABSTRACT



a. As

ARTICLE INFO

Editor: Deyi Hou

Keywords:
Airborne hyperspectral remote sensing
Soil heavy metal estimation
Heavy metal spectral characteristics
Overfitting
Ensemble learning

ABSTRACT

The problem of heavy metal pollution of soils in China is severe. The traditional spectral methods for soil heavy metal monitoring and assessment cannot meet the needs for large-scale areas. Therefore, in this study, we used HyMap-C airborne hyperspectral imagery to explore the estimation of soil heavy metal concentration. Ninety five soil samples were collected synchronously with airborne image acquisition in the mining area of Yitong County, China. The pre-processed spectrum of airborne images at the sampling point was then selected by the competitive adaptive reweighted sampling (CARS) method. The selected spectral features and the heavy metal data of soil samples were inverted to establish the inversion model. An ensemble learning method based on a stacking strategy is proposed for the inversion modeling of soil samples and image data. The experimental results show that this CARS-Stacking method can better predict the four heavy metals in the study area than other methods. For arsenic (*As*), chromium (*Cr*), lead (*Pb*), and zinc (*Zn*), the determination coefficients of the test data

* Corresponding authors at: Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
E-mail addresses: tankuncu@gmail.com (K. Tan), weibo_ma@126.com (W. Ma), dupjrs@126.com (P. Du).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jhazmat.2020.123288
Received 12 August 2019; Received in revised form 10 March 2020; Accepted 20 June 2020
Available online 26 June 2020

set (R_P^2) are 0.73, 0.63, 0.60, and 0.71, respectively. It was found that the estimated results and the distribution trend of heavy metals are almost the same as in actual ground measurements.

1. Introduction

Soil is an important part of the terrestrial ecosystem, and the quality of soil is directly related to the health of the organisms in it. Human activities, such as mining, industrial waste, and the irrational use of pesticides, increase the concentration of heavy metals in soils. This can seriously threaten the life and health of human beings through the heavy metal enrichment of crops (Jia et al., 2018; Wei and Yang, 2010; Chen et al., 2015a). In order to accurately estimate the concentrations and distributions of heavy metals in soil, hyperspectral remote sensing has been employed in recent years, which is both time- and labor-saving (Zhao et al., 2017; Brevik et al., 2016; Bendor et al., 2009; Chen et al., 2015b; Wang et al., 2018a). Due to the rich spectral information, hyperspectral remote sensing data can capture the weak discriminative information of heavy metals (Gholizadeh et al., 2018). An inversion model can then be stablished to estimate the concentrations of the heavy metals in the larger scope of the study area (Wu et al., 2007; Kinoshita et al., 2012; Brown et al., 2006). Most of the research on detecting heavy metals in soil by remote sensing have concentrated on the visible–near-infrared portion of the spectrum (i.e., 350–2500 nm), and the mid-infrared and far-infrared wavelengths have been used far less frequently (Wang et al., 2018a; Shi et al., 2014a; Soriano-Disla et al., 2014). This is because the visible–near-infrared spectrum has distinctive features for soil, and rich data resources are obtained by hundreds of sensors.

In late 20th century, some scholars have tried to detect the concentration of heavy metals in soil by imaging spectroscopy. Farrand and Harsanyi, using Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) data, mapped exposed concentrations of sediments on the floodplain of the Coeur d'Alene River in northern Idaho in 1997 based on Constrained Energy Minimization technique (Farrand and Harsanyi, 1997). Ferrier discovered that tailings dump material consists of a variety of ferruginous materials that often contain trace elements which have distinctive spectral features and make them amenable to detection and mapping by airborne imaging spectrometer data (Ferrier, 1999). Later scholars were able to measure the concentration of heavy metals in soil indirectly via intercorrelation with the soil attributes that are spectrally active in this region and through their complex-action, by Fe, Fe_2O_3, organic matter (OM), Clay etc. (Wu et al., 2007; Sares et al., 2004). In general, better prediction results were found for the total elemental concentrations of heavy metals in soil than for extractable and exchangeable fractions (Soriano-Disla et al., 2014).

Previous studies show that it is feasible to estimate soil heavy metal concentration by imaging hyperspectral. However, according to the review literature (Wang et al., 2018a), there are few successful cases of estimating soil heavy metal concentration by hyperspectral image data. The main reason is that when the imaged area is large, spatial heterogeneity is significant, and imbalanced samples are caused by the areas where heavy metals exceed the standard severely. As a consequence, traditional estimation methods cannot effectively overcome this imbalance problem, so it is necessary to develop new methods to improve the stability and accuracy of model evaluation and prediction.

At present, the most successful application of heavy metals estimation in soil uses ground-based hyperspectral data, which generally includes spectral pretreatment, spectral enhancement, feature selection and modeling (Wang et al., 2018a). Remote sensing spectroscopy was used to estimate the content of heavy metals in freshwater sediment by their association with organic matter (Malley and Williams, 1997). Subsequently, many other scenarios which may cause high heavy metal concentrations in soil have also been explored, including mining areas (Ma et al., 2016; Choe et al., 2008; Kemper and Sommer, 2002; Gannouni et al., 2012), reclaimed mining areas (Tan et al., 2014), river and lake sediments (Moros et al., 2009; Liu et al., 2011a; Ji et al., 2010), and agricultural soils (Wang et al., 2014). Most of these attempts are successful (Liu et al., 2016; Sun and Zhang, 2017), which demonstrated that estimation of heavy metal concentrations using spectral analysis of hyperspectral remote sensing imagery is a feasible approach.

Many methods of spectral feature enhancement and selection have been developed. Meanwhile, the models involved with spectral inversion have also been continually advanced. The heavy metal content in soils is usually low, so it is necessary to preprocess the spectra to enhance weak spectral information. A number of preprocessing methods (Rinnan et al., 2009; Asadzadeh and de Souza Filho, 2016; Asmaryan et al., 2014)—such as Savitzky-Golay (SG) smoothing, first derivative (FD) preprocessing (Dehaan and Taylor, 2002), second derivative (SD) preprocessing, standard normal variate (SNV) preprocessing (Fearn et al., 2009), multiplicative scatter correction (MSC), and wavelet and continuum removal (CR)—are now widely used, and they can smooth the spectra, eliminate the signal error caused by instrument itself, and suppress the noise in data acquisition, thereby enhancing the weak spectral information related the heavy metals.

After spectral pretreatment, it is necessary to select or extract spectral features for modeling, which can improve the explanatory power of the model and reduce the amount of calculation (Balabin and Smirnov, 2011; Xiaobo et al., 2010; Jain and Zongker, 1997). The Pearson method is accepted by most researchers, as the features selected by this method have a certain statistical basis and explanatory power (Wilford et al., 2016). Other feature selection methods (Jain and Zongker, 1997; Stańczyk, 2015; Yu and Liu, 2004; Guyon, 2003), such as terrain features or vegetation indices, may be limited in generalization. Among these methods, competitive adaptive reweighted sampling (CARS) is well-received (Duan et al., 2017; Li et al., 2009; Vohland et al., 2014; Tan et al., 2018). Moreover, these methods has limited capacity in coping with variations from multiple sites. An improved estimation model, CARS-PLS-SVM, to cope with the nonlinear problem in multiple sites with support vector machine (SVM) has proposed (Tan et al. 2018). The nonlinear CARS-PLS-SVM produces the highest accuracy in soil heavy metal (loid) estimation.

In the exploration of modeling methods, the related research progress is significant. Based on traditional statistical regression (such as multiple linear regression, stepwise regression), the partial least squares (PLS) method is found to be very effective in modeling spectral information (Soriano-Disla et al., 2014; Sun and Zhang, 2017; Haaland and Thomas, 1988; Shi et al., 2014b; Pandit et al., 2010). The PLS method can deal with very high dimensional data. Through iteration, the importance of each characteristic variable can be estimated, thus giving a specific expression of the inversion model. PLS can also be applied to feature selection, and on this basis, synergy interval partial least squares (siPLS) (Jiang et al., 2012) and other variants were developed.

Recently, machine learning and pattern recognition techniques are applied to the inversion of heavy metals in soils (Bishop, 2006; Mitchell, 2003), such as support vector machine (SVM) (Thissen et al., 2004; Devos et al., 2009; Jie, 2012), least squares support vector machine (LS-SVM) (Balabin and Lomakina, 2011), artificial neural networks (ANNs) (Tan et al., 2014; Rodriguez-Galiano et al., 2015), fuzzy neural networks (FNNs) (Liu et al., 2011b; Chen and Wu, 2017a), decision tree (DT)-based methods (Rodriguez-Galiano et al., 2015), random forest (RF)-based methods (Ma et al., 2016; Wang et al., 2015), the gradient boosting decision tree (GBDT) (Wang et al., 2015), extreme learning machine (ELM) (Chen and Wu, 2017a), etc. In particular, Goodarzi et al. (2015) estimated the lead (*Pb*) concentration in a mining

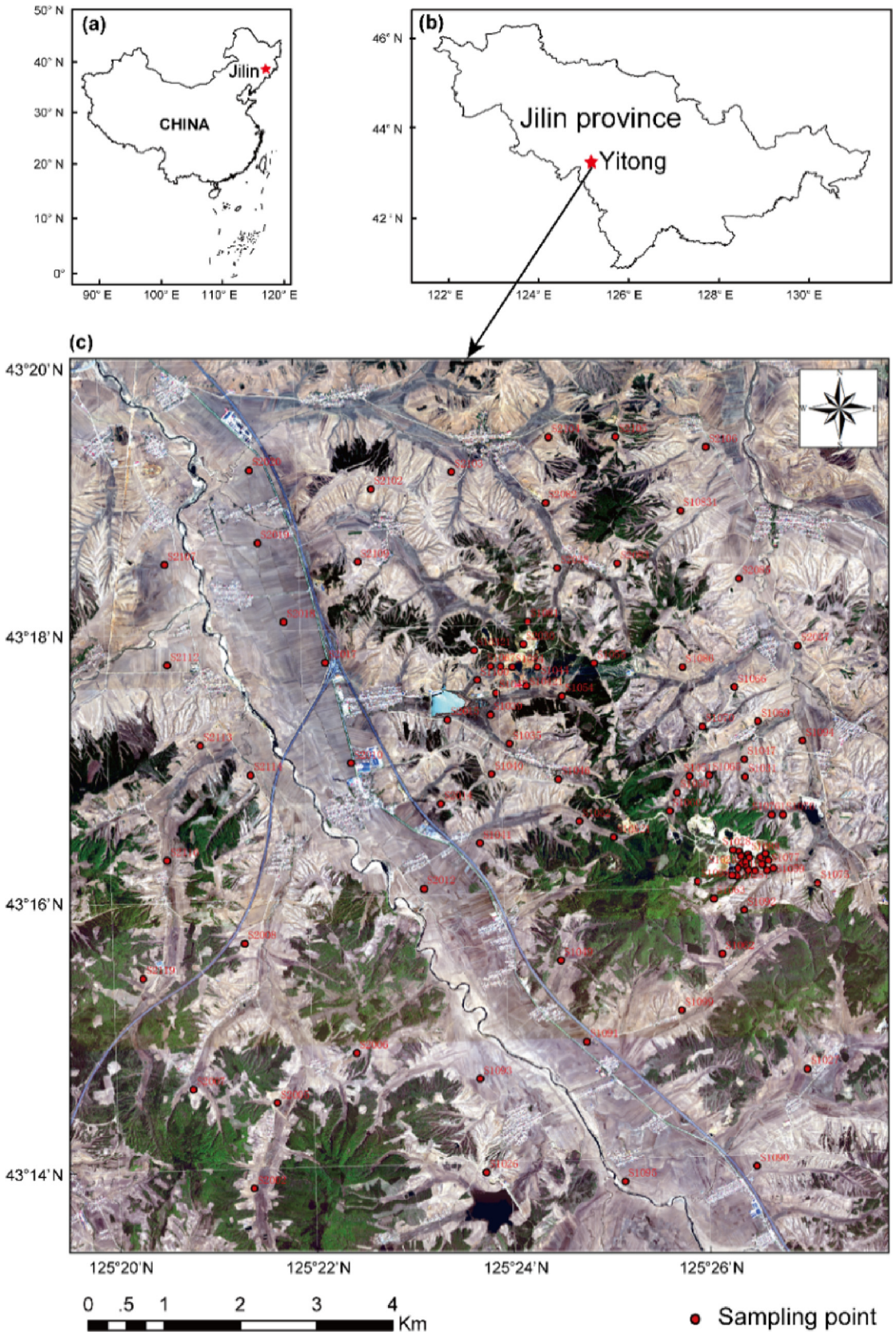


Fig. 1. The study area is located in Yitong County, Jilin Province, Northeast China. The image true color band from HyMAP-C data. The red channel is band 17(central wavelength: 0.68 μm), the green channel is band 8(central wavelength: 0.56 μm), and the blue channel is band 1 (central wavelength: 0.47 μm). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

area by an FNN method, and the decision coefficient reached 0.98, which was higher than PLS and ANN. Wang et al. (2014) predicted the concentrations of *Pb*, zinc (*Zn*), and copper (*Cu*) more accurately than Principle Component Regression (PCR)through the use of partial least squares regression with genetic algorithm (GA-PLSR). In 2015, Rodriguez-Galiano et al. (2015) used Hyperion data to predict and evaluate mineral resources in the Rodalquilar mining area of Spain. During this study, ANN, RF, DT, stochastic forest, and SVM algorithms were tested and evaluated. After analyzing the size sensitivity of the training data, the sensitivity of the hyperparameters, and the explanatory ability of the model parameters, it was concluded that the stochastic forest algorithm had the highest stability and robustness.

However, most of the previous studies are based on ground spectra. Quantitative estimation of heavy metals using imaging spectroscopy has been explored by few researchers, and the performance is limited due to the coarse spatial resolution. They could not provide the spatial distribution of heavy metals in the study area. In this research, we intend to estimate the distribution of heavy metals using airborne hyperspectral image and propose a CARS-Stacking model for efficient estimation. This proposed strategy can map the soil heavy metal concentration in a large area.

In this paper, an airborne hyperspectral image cube of the study area in Yitong County, Jilin province, China, was collected via the HyMap-C airborne hyperspectral imaging system. Most of the study area is farmland, and there are several mining areas and Yitong river flow in the surrounding. After preprocessing analysis, we propose a CARS-Stacking ensemble learning method based on a stacking strategy to select feature, estimate, analyze, and map the soil heavy metal concentration in the study area.

2. Materials and methods

2.1. Study area

The study area is located in Yitong County, Jilin province, China. This area belongs to the humid monsoon climate zone in the middle temperate zone of China. The region is hilly, with an average annual temperature of 5.5 ℃. The annual average precipitation is 651.7 mm, and the sunshine is sufficient. To the north of the research area are the suburbs of Changchun, and the Yitong River runs through the whole research area from southeast to northwest. Yitong County is rich in mineral resources, including gold, silver, copper, and iron. The whole research area (125.33 °E–125.47 °E, 43.22 °N–43.33 °N) covers an area of 139 km^2. The average altitude of the study area is 305 m, the lowest altitude is 262 m, and the highest altitude is 446 m. The location of the research area is shown in Fig. 1.

2.2. Datasets

2.2.1. Soil samples collection and testing

At the end of April and early May in 2017, after the fields had been ploughed, the surface soil was easily accessible. It was therefore a good time to collect surface soil samples. The principle of the soil sampling undertaken in this study was to densely lay sampling points in the farmland areas around the mines, while the sampling points in the other farmland areas were arranged as evenly as possible.

The collection of surface soil samples was basically synchronized with the airborne hyperspectral data acquisition. The method of soil sampling is described as follows. The location of each sampling point is expressed by the letter O. A soil sample was composed of five individual soil samples, each with a thickness of 5 cm (Fig. 2, ABCDO), which were then mixed together and sealed in a bag. In the process of sampling, the physical quantity of the sample should be no less than 2 kg, which was strictly followed to avoid the influence of random error on the soil samples.

In order to obtain high-precision sampling point position data, real-time kinematic (RTK) mobile station positioning technology was used to obtain high-precision coordinates of the sampling points in real-time. Finally, a total of 95 soil samples were collected in the study area. The sample locations are shown in Fig. 1. We fully mix the tested soil samples, then store them in polytetramethylene sealed bags, take them back to the pretreatment room, air-dry the soil samples, grind 100 mesh nylon sieves by quartering method according to the national standards, and mix them into sealed bags for testing. After the sample is digested by the method of total decomposition of hydrochloric acid-nitric acid-hydrodecanoic acid-perchloric acid, the heavy metal concentrations of the soil samples were determined by inductively coupled plasma-mass spectrometry (ICP-MS).

2.2.2. Analysis of spatial heterogeneity

In general, the areas of previous heavy metal inversion estimations have been small, but this study area is more than 100 km^2. The larger the study area, the more complex the spatial heterogeneity of the heavy metals is likely to be. To analyze the spatial heterogeneity, the global autocorrelation of each heavy metal was measured by the global Moran's I index is adopted which is defined as

$$I = \frac{\sum_{i=1}^{n}\sum_{j\neq 1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{S^2 \sum_{i=1}^{n}\sum_{j\neq 1}^{n} w_{ij}} \quad (1)$$

where x_i and x_j are concentration values of a certain heavy metal in the i-th and j-th sampling points, respectively, $\overline{x}$ is the concentration average of all the sampling points, S^2 is the concentration variance of a certain type o, n is the number of samples, and w_{ij} is the spatial weight for the i-th and j-th sampling points.

2.2.3. Hyperspectral image data

The HyMap-C imaging spectrometer (Kruse et al., 2000) is a data acquisition and analysis system developed by HyVista corporation, Australia. In addition to the host computer (optical scanner, electronic components, control components, data transmission and storage components), it is equipped with a position and orientation system (POS) (inertial measurement unit (IMU)/differential global positioning system (DGPS)), a three-axis stabilized gyroscope platform, a calibration system, an advanced data preprocessing system, and data processing software. The technical specifications of the HyMap-C imaging spectrometer are listed in Table 1.

2.2.4. Image data preprocessing

After obtaining the data of the study area, geometric correction,

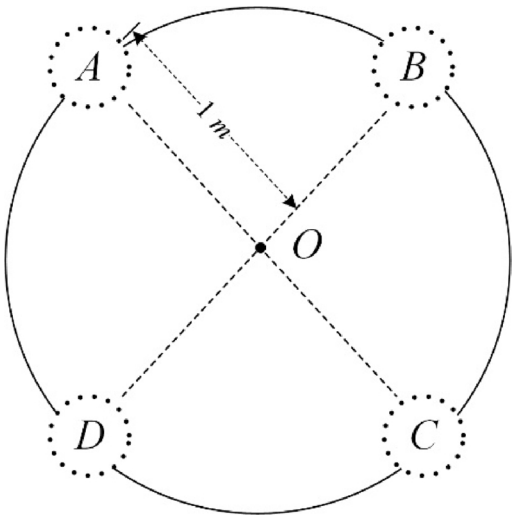


Fig. 2. Soil samples at each soil sampling site consist of 5 cm-thick soil at 5 sites of OABCD. The distance between O point and other points is 1 m.

Table 1
Main technical details of the HyMap-C imaging spectrometer.

Main technical details	Data
Spectral range	450–2500 nm
Channels	135
Field of view	60°
Instantaneous field of view	2.5 mrad
Scan rate (lines/s)	5–25 Continuous adjustable
Pixel registration	Less than 0.1 pixels
Row pixel number	668
Radiance accuracy (in flight)	95 %
Dynamic range	16 bit
Signal-to-noise ratio (SNR)	Visible and near-infrared mean SNR: 1000 Shortwave infrared mean SNR: 600
Working temperature (℃)	−10 ~ +40
Integrated POS	Integrated IMU/DGPS system (Novatel/SPAN SE)

radiation correction and atmospheric correction were carried out.

(1) Geometric correction

The process of image geometric correction is divided into systematic geometric correction and geometric precise correction. Firstly, the geometric correction of the system is to automatically calculate the geographic coordinates of each pixel according to the geographic position, attitude parameters, terrain information (DEM) of the flight platform acquired by the Position and Orientation System (POS) system, and the relative position between the optical axis of the hyperspectral sensor and the flight platform. Geometric precise correction is accomplished by selecting 160 uniformly distributed control points on the ground, taking some control points as known points, using least squares fitting, the fitting equation between image coordinates and geodetic coordinates is established, and finally, the image is registered in geodetic coordinate system. In present study, the Geometric correction work was completed specifically through HyMap system supporting software HyMapTMGeo. The spatial resolution of the hyperspectral images is 4.5 m.

(2) Radiometric correction

Before the flight experiment, the imaging system completed the indoor spectral calibration. The instrument central wavelength and spectral response parameters were obtained accurately, and the radiation correction parameters at each central wavelength under each pixel were obtained. On the basis of geometric correction, the radiation calibration is completed, so that the gray value of the image can be restored to the radiant brightness value at the pupil entrance with physical meaning. Specifically, this paper completed the radiation correction work by HyDn2Rad, which is the supporting software of HyMap system.

(3) Atmospheric correction

HyCorr program is used to atmospherically correct georectified HyMap-C radiance flight strip data to surface reflectance by ATREM model (CSES, 1999). HyCorr processes HyMap-C radiance to surface reflectance in two stages. The first stage is rigorous atmospheric correction of a geometrically corrected radiance dataset. The second stage is polishing or 'smoothing' of the atmospheric correction to tie the reflectance values, of the four independent detectors, to generate a contiguous spectral reflectance profile. The main parameters of atmospheric correction are shown in Table 2.

The study area was made up of nine strips after geometric correction and radiation correction. After atmospheric correction, the spectral curves show obvious distortion or even negative value near the wavelength of 1.4 μm and 1.9 μm. The main reason is that the absorption of water vapor is very strong. Therefore, it is necessary to remove the distorted wavelength in the spectral curve. The 11 related bands are 1.37 μm, 1.38 μm, 1.39 μm, 1.41 μm, 1.42 μm, 1.84 μm, 1.85 μm, 1.87 μm, 1.88 μm, 1.90 μm, 1.91 μm. The results of image spectrum extraction at 95 sampling points are shown in Fig. 3. Because most of the sampling sites are homogeneous soil surface, the spectral characteristics at sampling points are closer to pure soil spectra.

2.2.5. Soil information extraction

The objective of this study is to estimate the concentration of heavy metals in bare farmland soil, so it is necessary to extract bare soil from the image scene. For this purpose, we chose the fully constrained least squares unmixing method, which guarantees abundances are positive and their sum is 1. Before unmixing, several representative pure spectra of six kinds of objects were selected, which are bare soil, vegetation, water, buildings, highlight, and roads. Then, we utilize vertex component analysis (VCA) to automatically select endmembers. For some complex areas, we determine the endmember candidates manually. The fully constrained least squares (FCLS) is used to estimate fractional abundances (Heinz and Chang, 2001).

2.2.6. Feature selection

The specific process of CARS (Duan et al., 2017; Li et al., 2009; Tan et al., 2018) is to use the adaptive weighted sampling technique to retain the spectral wavelengths with large absolute coefficients of the PLS model and to delete those with small coefficients. In this way many subsets of wavelength variables can be obtained. Then, each subset of wavelength variables is modeled by the PLS method with Monte Carlo cross-validation, and the optimal subset is selected by the root-mean-square error of the model in cross-validation.

Let the data set matrix be denoted as X_{m*p}, where m is the number of samples, p is the number of variables, and Y_{m*1} is the dependent variable. Let T be an X-segment matrix, which is a linear combination of X and W, and W is a combination coefficient. Then

$$T = XW \tag{2}$$

$$Y = Tc + e = XWc + e = Xb + e \tag{3}$$

where c is the regression coefficient vector, e is the prediction residual, $b = Wc = [b_1, b_2, b_3, \ldots, b_p]$, and p represents a dimension coefficient vector. The absolute value of the i th element in b $\mid b_i \mid (1 \le i \le p)$ represents the contribution of the i th band or the independent variable to Y. The larger the value of $\mid b_i \mid$, the more important the corresponding independent variable is. We then define weights for evaluating the importance of independent variables as

$$\omega_i = \frac{|b_i|}{\sum_{i=1}^{p} |b_i|} \tag{4}$$

for $i = 1, 2, \ldots, p$. The variables are removed by the CARS algorithm, and their weights ω_i are set to 0. The main process is shown in Fig. 4.

The variable retention rate is calculated by the reference formula $r_i = ae^{-ki}$, where a and e are constants. At the 1 st and Nth Monte Carlo cross-validation samplings, all p variables and only two variables in the sample set participate in the modeling. That is, $r_1 = 1$ and $r_N = 2/p$.

Table 2
Main parameters of atmospheric correction.

Parameters	Parameters setting
Aerosol Model	Continental
Total ozone (atm-cm)	0.34
Atmospheric Model	Mid Latitude Summer
Visibility (Km)	100.0
Gases	H_2O, CO_2, O_3, N_2O, CO, CH_4, O_2
H_2O Vap. Modeling	Rock, Soil & Minerals

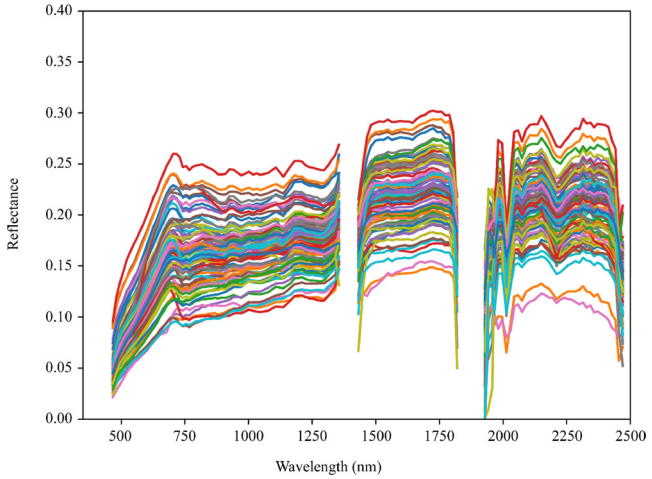


Fig. 3. Imaging spectra of the 95 sampling points.

$$a = \left(\frac{p}{2}\right)^{1/(N-1)} \tag{5}$$

$$k = \frac{In(p/2)}{N-1} \tag{6}$$

2.3. Modeling methods

2.3.1. Partial least squares (PLS)

PLS is a classical statistical method, and its modeling ability is stronger than other multiple linear regression methods in general (Haaland and Thomas, 1988; Wold et al., 2001; Abdi and Williams, 2013). The PLS model attempts to map an independent variable X projection into a new learning space Y, and explain the direction of the maximum multidimensional variance in the Y space. It performs well when multiple collinearity exists between independent variables.

Soil element concentration is correlated with corresponding spectral reflectance values using PLSR models (Malmir et al., 2019). The potential of vis-NIR spectroscopy and PLSR for prediction of chemical and physical properties is evaluated and the accuracy of the calibrations and validations for the different soil properties are assessed (Antonio et al., 2012; Rossel, 2007). Partial least square regression (PLSR) models are commonly utilized to correlate data extracted from hyperspectral images to their corresponding chemical concentrations (Axelsson et al., 2013). PLSR is underpinned by the assumption that the dependent variable can be estimated via a linear combination of explanatory variables (Wang et al., 2018a).

PLSR can be considered as a sum of regression analysis, principal component analysis (PCA), and correlation analysis. The prediction matrix $X_{m\times p}$ (m samples and p variables) is decomposed as

$$X = SP^T + E \tag{7}$$

where a score matrix S and a loading matrix P^T are derived from PCA, and E is the error matrix. Similarly, response matrix $Y_{m\times 1}$ is also decomposed into a score matrix U and a loading matrix Q^T plus the error matrix F as

$$Y = UQ^T + F \tag{8}$$

Here, S and P have a dimension of $m \times k$ and $n \times k$, respectively, the dimension of U and Q is $m \times l$ and $1 \times l$, respectively, and k and l are the number of principal components for reconstructing X and Y, respectively. E is derived from the summary variance of k principal

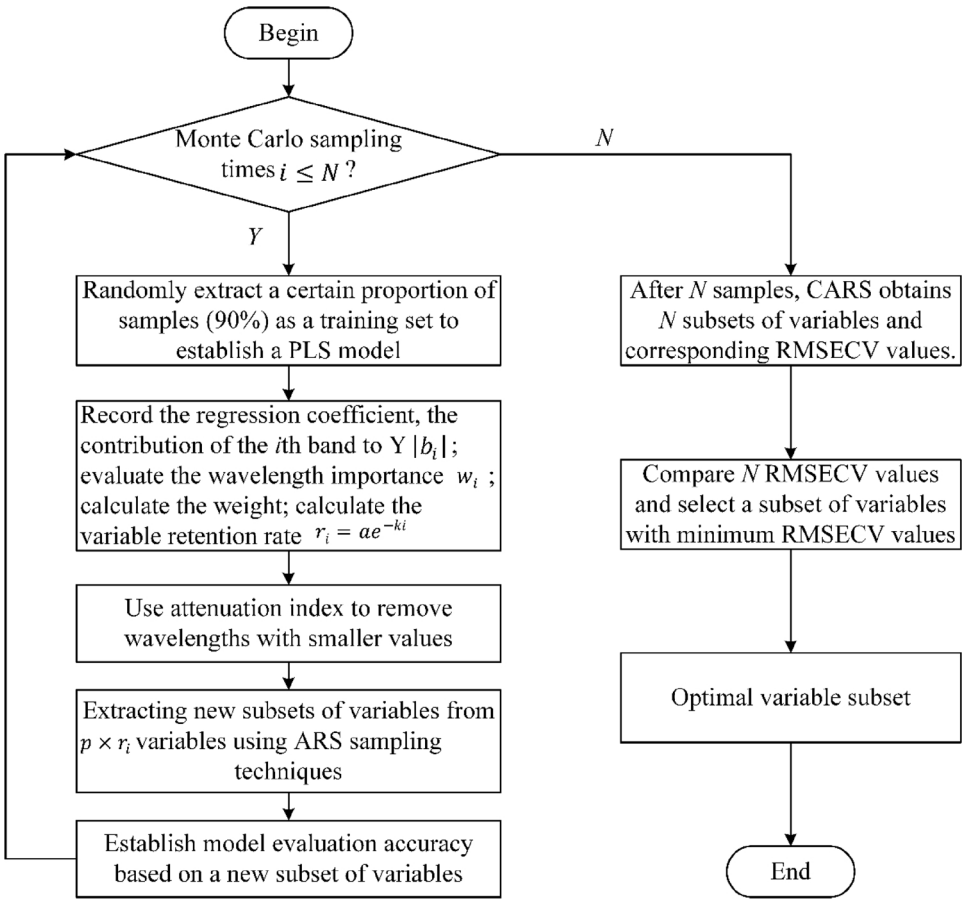


Fig. 4. The CARS method flowchart.

components, and F is calculated from subtraction of l principal components from the Y matrix. Then replacing matrix X and Y with residual matrix E and F, respectively. The score matrix and loading matrix are derived iteratively.

2.3.2. k-nearest neighbor (k-NN)

The *k*-NN algorithm (Mitchell, 2003) is a nonparametric algorithm. The distance relationship between the test sample data and the previously stored training sample data is then measured, and the dependent variable value of the new sample data is estimated based on this distance relationship. The similarity between the test data and training data is used to estimate the dependent variable values of the test data. The core of the algorithm is training data storage, *k* value selection, and distance measurement between training data and test data.

2.3.3. Support vector machine (SVM)

The SVM model is a kernel-based method proposed by Vapnik (Devos et al., 2009; Smola and Schölkopf, 2004). It is a nonlinear modeling method based on statistical learning theory. SVM can use support vectors in training samples to design a optimal decision boundary. It can handle both linear and nonlinear problems, and solve regression modeling problems.

2.3.4. Random forest (RF)

The RF algorithm (Belgiu and Drăguţ, 2016; Breiman, 2001) is a predictive modeling algorithm based on classification and regression trees (CART) and the bagging learning strategy. In bagging, a decision tree is generated from all of the properties each time, while in RF, it is randomly generated from a fixed-size subset of all the attributes, resulting in a reduced computational cost (Bauer and Kohavi, 1999). Specifically, by the bootstrap resampling technique, random sampling is repeated K times to generate a fixed number of subset training samples (in general, the subset sample size is two-thirds of the training samples) from all the samples (where K is the number of trees in the forest). Meanwhile, for each sample, only a fixed number of sub-attributes are selected. Each randomly selected subsample with its corresponding sub-attributes can then be used to generate a classification tree or regression tree, and all the trees make up the forest. Finally, the results are obtained according to the scores of the class voting from all the trees (certain algorithms can be implemented to determine the average of each tree, mostly for regression trees). The trained forest $\widehat{F}_{RF}^{K}(x)$ with K trees can be expressed as:

$$\widehat{F}_{RF}^{K}(x) = \frac{1}{K}\sum_{K=1}^{K} T(x_s) \tag{9}$$

where $T(x)$ is a single tree, x is all the training samples, and x_s is each tree's training sample data obtained with the bootstrap sampling method. Another parameter that is not noted in Eq. (8) is the number of sub-attributes selected from all the attributes with bootstrap sampling.

2.3.5. ExtraTrees

The ExtraTrees (Geurts et al., 2006) method is a further improvement to the RF method, but there are two main differences. Firstly, the training process of the RF method uses the bootstrap method with only some of the samples, and ExtraTrees trains each individual decision tree with all the training samples. Secondly, in the RF method, the learning of each tree is divided by the random subset of the characteristic, while ExtraTrees completely randomly obtains the specific bifurcation value to realize split learning of the decision tree.

In ExtraTrees, the bifurcation attribute of a tree is determined randomly, so the fitting ability of a single tree is very weak, but the predictive ability of the model increases rapidly after the aggregation of multiple decision trees. The measurement of the fitting ability of the aggregation model can also be tested in all data sets. Because the best bifurcation property is randomly chosen, the predictive ability of the same data set may result in different predictions. From the perspective of data learning, ExtraTrees further enhances the randomness of the sample space.

2.3.6. XGBoost

XGBoost (Chen and Guestrin, 2016; Schapire, 2003) is a tree-based boosting algorithm. The major difference from other boosting tree algorithms is that its objective function introduces a regularization term as

$$\begin{aligned}\psi(y, F(X)) &= \sum_{i=1}^{N} \psi(y_i, F(x_i)) + \sum_{m=0}^{T} \Omega(f_m) \\ &= \sum_{i=1}^{N} \psi(y_i, F(x_i)) + \sum_{m=0}^{T} (\gamma L_m + \frac{1}{2}\lambda \|\omega_m\|^2)\end{aligned} \tag{10}$$

where L_m represents the number of leaf nodes of the tree model f_m generated in the mth iteration, and $\omega_{m1} = (\omega_m, \omega_{m2}, \cdots, \omega_{mL_m})$ represents the output value of each leaf node of f_m. Here, γ and λ are regularization coefficients, which provide strong control over the complexity and output of the model. When both γ and λ are zero, the size and output value of the generated tree are not limited.

After introducing the regularization term, the algorithm chooses a simple and well-performing model. The regularization term $\sum_{m=0}^{T} \Omega(f_m)$ in Eq. (9) is only used to suppress the overfitting of the weak learner $f_m(X)$ in each iteration, and does not participate in the integration of the final model.

2.3.7. Extreme learning machine (ELM)

ELM (Chen and Wu, 2017b) is a novel training algorithm for single-hidden-layer feedforward networks (SLFNs), in which only needs to set the number of hidden layer nodes of the network, and it does not need to adjust the input weight of the network and the offset of hidden elements in the process of execution. No parameters need to be manually tuned except predefined network architecture. Therefore, it maintains faster training speeds and has higher generalization performance.

2.3.8. AdaBoost

AdaBoost (Rätsch et al., 2001) is an excellent boosting algorithm. The principle of the algorithm is that, the best weak learner is selected from the trained weak learners, then the best weak learners combine into a final strong learner by adjusting the weight of samples and the weight of the weak learners. The advantages of AdaBoost are that it fully considers the weight of each learner and it has few parameters, so there is no need to adjust too many parameters in practical application.

2.3.9. Back Propagation Neural Netwok (BPNN)

A Back Propagation Neural Netwok (BPNN) (Zhang et al., 2018) is an effective learning method for multilayer neural networks. Its learning rules are constantly adjusting the weights and thresholds of the network through back propagation to minimize the sum of the squared errors of the network. By constantly adjusting the network weight value, the final output of the network is as close as possible to the expected output, so it can achieve the purpose of training.

2.3.10. CARS-Stacking

From the perspective of data mining, the characteristics of inversion analysis can be summarized as follows: 1) the high-dimensional feature space; and 2) the small sample size that may lead to model overfitting. The first problem can be solved by effective feature selection method (CARS). For the second problem, it is necessary to improve the generalization ability of the prediction model. The most representative way to improve the model predictive and generalization ability is to use an ensemble learning method (Dietterich, 2000; Zhou, 2012). Therefore, in this paper, we construct an integrated model based on a stacking ensemble strategy (Breiman, 1996; Wolpert, 2011). The CARS-Stacking method of this paper will be described below, and its flowchart is

shown in Fig. 5.

The feature data are extracted from CARS and put into the stacking method. In the selection of learners in Level 0 and Level 1 of the stacking ensemble strategy, due to the realistic problems of airborne hyperspectral data with high-dimension, a small and unbalanced samples of soil heavy metal concentrations, and weak hyperspectral

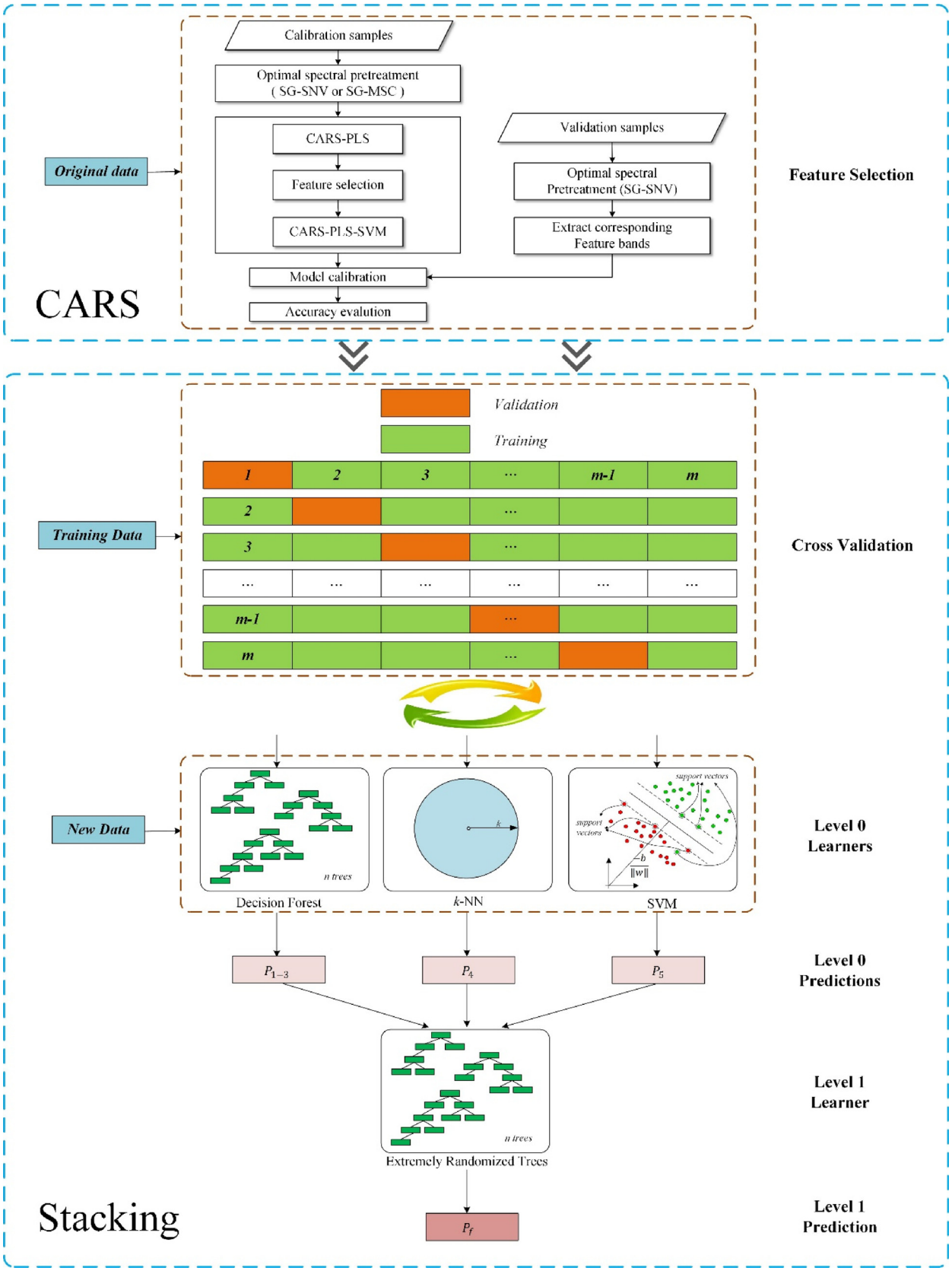


Fig. 5. Flowchart based on the CARS-Stacking method.

response capabilities of soil heavy metal concentrations (Wang et al., 2018b), we strive to find a way to overcome the above problems and use Stacking method for efficient integration (Table 3).

We construct the stacking ensemble strategy to overcome the disadvantages of each individual sole method such as RF, ExtraTrees, XGBoost, SVM and k-NN. These machine learning algorithms offer overall good performance in soil heavy metals estimation from airborne hyperspectral imagery, especially when there is limited or no knowledge about data distrib ution or the form of relationship (Koushik et al., 2020). RF, ExtraTrees, XGBoost, SVM and k-NN are selected as the Level 0 of the Stacking ensemble strategy models, and ExtraTrees is selected as the Level 1 of the Stacking ensemble strategy model. ExtraTrees has many advantages such as relatively stable performance, strong prediction ability, simple parameter adjustment, and fast implementation speed. The Level 1 learner attaches more importance to the characteristics of the learner with stable and fast performance.

In order to simplify the process, three methods (RF, ExtraTrees, and XGBoost) are unified with decision forest, but the five methods in the process of implementation are independent and do not affect each other. As shown in Fig. 5, firstly, the training data set is divided into m parts by cross-validation (Kohavi, 1995). Taking the SVM as an example, it is trained by $m-1$ data sets, and the remaining one is used as the test set to estimate the accuracy of the model. After completing the m cycles, the accuracy of the test set is obtained by averaging the prediction results. During the whole process, the accuracy of the test set is observed to adjust the SVM parameters, such as gamma and C. Similarly, the other decision forest method and k-NN method are carried out independently. After determining the parameters of the Level 0 layer of each base learner, the estimated results of each learner in Level 0 is connected in parallel to form a joint feature dataset and input to Level 1, and further training and study are done by the method of ExtraTrees. Similarly, we use the remaining part of the data set to verify and adjust the model parameters, and finally determine the parameters of the Level 1 learner in the whole model. P_{1-5} is the result of Level 0, and P_f is the final prediction result.

The training process of the whole stacking method is completed in *Training Data*, and the overall parameters of the model are determined after the training is completed. The fitting degree of the model is then evaluated by New Data. New Data in Fig. 5 refers to validation data sets or other unlabeled spectral data that need mapping prediction. These data are consistent with the structure and organization of model training data.

2.4. Model evaluation method

In order to evaluate the fitting and generalization ability of the model, three determinant indicators are selected: coefficient of determination (R^2), root-mean-square error (*RMSE*), and mean absolute error (*MAE*), which are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(observed_i - predicted_i)^2}{\sum_{i=1}^{N}(observed_i - \bar{observed})^2} \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(observed_i - predicted_i)^2}{N}} \tag{12}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|observed_i - predicted_i| \tag{13}$$

where *observed* is the true value, and *predicted* is the predicted value, $\bar{observed}$ is the average of the true value, and N is the number of samples. The three evaluation indices are distinguished by the letters C and P at the end of the right-lower corner of the model training data set and the prediction data set, respectively (from the initials of Calibration and Prediction, respectively). That is, the training data set evaluation is represented as R_C^2, $RMSE_C$, and MAE_C, and the prediction data set evaluation is respectively expressed as R_P^2, $RMSE_P$, and MAE_P.

3. Results

3.1. Data acquisition or preprocessing results

(1) Soil heavy metal data

The basic statistical analysis of heavy metal concentrations, including mean, standard deviation, minimum value, maximum value, and coefficient of variation (C.V.), is provided in Table 4. It can be seen that the maximum value of arsenic (*As*) is far beyond the minimum, and the gap is two orders of magnitude. There are few samples with very high concentrations, and the concentration of most of the samples is very low. Overall, the sample appears imbalanced. The coefficient of variation of *As* and chromium (*Cr*) is very high, indicating that the variation in the sample set is also very high. The coefficient of variation of *Pb* and *Zn* is relatively low, indicating that the spatial distribution of *Pb* and *Zn* is more random and even, and is less affected by human activities. *As* and *Cr* clearly show sample imbalance. Among the four heavy metals, *As* and *Cr* exceeded the national standard level 2 and 3 most, and As exceeded the national standard most obviously.

(2) Pearson correlation between the four heavy metals

The correlation analysis between the heavy metals can provide some reference for the explanation of their physical distribution. The correlation between the four types of heavy metals is analyzed in terms of Pearson correlation coefficients in Table 5. *Pb* and *Zn* have a certain correlation with *As*, at 0.29 and 0.24, respectively, while the other heavy metals have a very weak correlation. This indicates that there is a weak linear relationship between *Pb*, *Zn*, and *As* in the soil.

(3) Spatial Heterogeneity Analysis

Firstly, the spatial weights were generated according to the spatial position of the sampling points, and then the Moran's I index of each heavy metal was calculated according to Eq. (1). Meanwhile, the evaluation index z-score and the p-value of the Moran's I index were

Table 3
Stacking method selection strategy.

Problem	Methods	Advantages	Combination strategy
High dimensional characteristics of airborne hyperspectral data	SVM,	SVM: strong learning ability of small samples	the Level 0 of the Stacking ensemble strategy models are RF, ExtraTrees, XGBoost, SVM and k-NN, and the Level 1 model is ExtraTrees
	k-NN	k-NN: nonparametric learning strategy	
Small and unbalanced samples of soil heavy metal concentrations	RF,	RF: strong adaptability to data sets, good anti- noise performance.	
	SVM	SVM: high generalization performance	
Weak spectral characteristic response capabilities of soil heavy metal concentrations	XGBoost,	XGBoost: data mining	
	ET	ET: more stable and higher prediction accuracy than RF	

Table 4
Descriptive statistics of heavy metal concentration (Unit: mg/kg).

Metal	Mean	Std	Min	Max	C.V	Siping City[a]	Jilin Province[b]	National[c]
As	42.00	67.45	6.35	419.96	1.61	9.46	11.6	15
Cr	399.43	864.10	36.04	4617.56	2.16	49.64	42.4	90
Pb	15.30	4.00	9.04	36.79	0.26	17.87	14.96	35
Zn	51.83	10.75	38.48	117.18	0.21	58.44	49.95	100

[a] Bao Xinhua, Jilin-Changchun-Siping Urban Economic Zone Surface Soil Environmental Quality Assessment asnd Ecogeochemical Zoning. Jilin University, 2011.
[b] China Environmental Monitoring Station. Background Value of Soil Elements in China. China Environmental Science Press, 1990.
[c] National Environmental Protection Agency. Environmental quality standard for soils Beijing; National Environmental Protection Agency. 1995: 1–5. (GB15618-1995).

Table 5
Pearson correlation between the four heavy metals.

	As	*Cr*	*Pb*	*Zn*
As	1	−0.07	0.29	0.24
Cr	−0.07	1	−0.02	−0.13
Pb	0.29	−0.02	1	0.14
Zn	0.24	−0.13	0.14	1

Table 6
Statistics of the global Moran's I index of the heavy metals in the soil.

Metal	Moran's *I*	z	*p*
As	0.20	5.31	0.00
Cr	−0.07	−1.32	0.19
Pb	0.01	0.48	0.63
Zn	−0.01	0.07	0.95

obtained. The *p*-value represents the probability that the observed spatial pattern is created by a stochastic process. The z-score is a multiple of the standard deviation. The higher the absolute value of the Moran's *I* index, the stronger the spatial aggregation pattern. The absolute value of the z-score reflects the degree of dispersion. When the absolute value of *p* is small, the reliability of the Moran's *I* index is high.

As shown in Table 6, the Moran's *I* index of *As* is about 0.2, indicating that the concentration distribution of *As* has positive spatial correlation and spatial aggregation characteristics. Moreover, the z-score is extremely high, and the *p*-value is close to 0, showing that the confidence of this model is very high. The Moran's I index values of the other three heavy metals are low or less than zero, the *p*-values are all greater than 0.1, and the z-scores are low. This shows that the spatial aggregation patterns of *Cr*, *Pb*, and *Zn* are extremely weak, and the spatially random patterns are obvious.

(4) Unmixing result

After unmixing, the abundances and residual of each pixel were obtained, as shown in Fig. 6.

The soil pixels have abundance values of greater than 0.65, accounting for 48.73 % of the total number of pixels in the entire study area. The soil distributed area were obtained by mask processing.

3.2. Spectral feature

Based on the principle of "survival of the fittest", CARS can eventually choose characteristic variables with strong adaptability. The input data of this method are the original reflectance spectra. The optimum number of iterations is determined at the minimum cross-validation error in the PLS model, and the selection of the CARS results corresponding to heavy metals is determined under this iteration number. Finally, a statistical rendering table for the four CARS features of heavy metals is given in Table 7. In the later stage, the validation and analysis of the heavy metal spectral validity and inversion modeling are based on the spectral characteristics listed in Table 7. The simple statistics show that the number of variable sets (14–16) after CARS feature

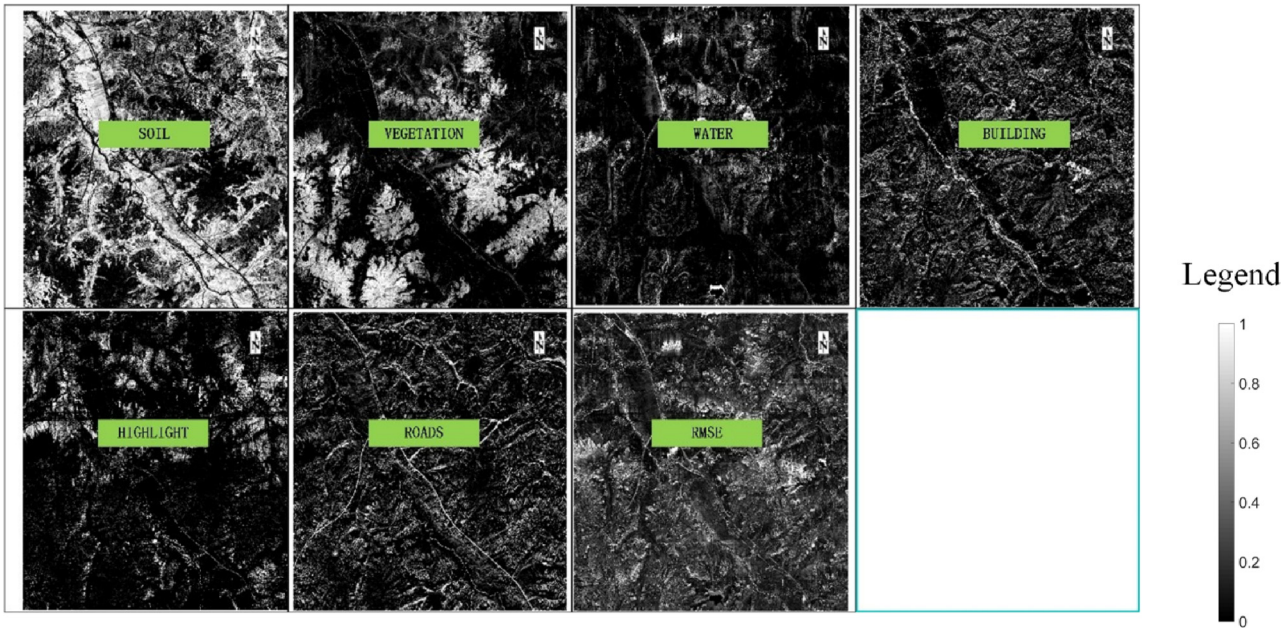


Fig. 6. Results of the pixel unmixing.

Table 7
Feature band statistics based on the CARS method.

Metal	Wavelength (Unit: μm)	Sum
As	1.14, 1.45, 1.66, 1.69, 1.98, 2.06, 2.14, 2.19, 2.21, 2.23, 2.40, 2.41, 2.43, 2.44	14
Cr	0.48, 0.57, 0.63, 0.67, 1.08, 1.27, 1.51, 1.56, 1.58, 1.98, 2.09, 2.12, 2.15, 2.21, 2.36, 2.40	16
Pb	0.51, 0.55, 0.57, 0.67, 0.89, 0.97, 1.11, 1.53, 1.79, 2.06, 2.11, 2.23, 2.28, 2.33, 2.36	15
Zn	0.47, 0.51, 0.59, 0.74, 0.78, 0.88, 1.14, 1.50, 1.53, 1.77, 2.06, 2.09, 2.15, 2.21	14

Table 8
Regression results of PLS, SVM, ExtraTrees, RF, XGBoost, *k*-NN, and stacking.

Metal	Method	R_C^2	$RMSE_C$	MAE_C	R_P^2	$RMSE_P$	MAE_P
As	PLS	0.73	37.20	58.24	0.63	43.01	57.86
	SVM	0.90	23.30	10.68	0.64	42.57	27.78
	ExtraTrees	0.90	30.52	20.17	0.21	63.10	39.01
	RF	0.97	25.67	15.16	0.10	68.35	40.38
	XGBoost	1.00	4.62	1.22	0.08	73.76	46.03
	k-NN	1.00	0.00	0.00	0.24	62.82	34.93
	AdaBoost	1.00	0.87	0.21	0.58	12.57	6.47
	ELM	0.70	24.33	17.62	0.42	27.44	19.86
	BPNN	0.58	12.54	9.23	0.47	20.78	13.41
	CARS-Stacking	**0.91**	**22.35**	**12.73**	**0.73**	**37.09**	**23.70**
Cr	PLS	0.41	24.11	29.17	0.21	26.27	28.68
	SVM	0.83	13.41	5.03	0.61	16.68	11.74
	ExtraTrees	1.00	0.63	0.41	0.60	17.29	12.80
	RF	0.95	11.24	8.45	0.46	20.08	16.56
	XGBoost	1.00	0.03	0.01	0.40	21.00	16.85
	k-NN	1.00	0.00	0.00	0.38	21.11	15.78
	AdaBoost	1.00	1.22	0.17	0.35	23.68	12.38
	ELM	0.54	37.20	29.33	0.40	39.90	32.07
	BPNN	0.48	20.96	15.85	0.46	22.94	14.06
	CARS-Stacking	**0.68**	**18.97**	**13.85**	**0.63**	**16.47**	**12.76**
Pb	PLS	0.54	2.24	3.23	0.17	3.20	3.23
	SVM	0.99	0.30	0.15	0.56	2.24	1.64
	ExtraTrees	1.00	0.12	0.07	0.55	2.34	1.67
	RF	0.97	0.98	0.74	0.52	2.45	1.81
	XGBoost	1.00	0.00	0.00	0.54	2.27	1.72
	k-NN	1.00	0.00	0.00	0.41	2.61	1.73
	AdaBoost	1.00	0.05	0.00	0.57	2.21	1.35
	ELM	0.59	4.12	3.23	0.32	4.47	3.55
	BPNN	0.44	3.49	1.88	0.33	2.76	2.15
	CARS-Stacking	**0.65**	**2.03**	**1.52**	**0.60**	**2.17**	**1.53**
Zn	PLS	0.41	5.98	7.34	0.14	6.97	6.95
	SVM	1.00	0.10	0.10	0.70	4.38	3.34
	ExtraTrees	1.00	0.12	0.08	0.66	4.34	3.15
	RF	0.96	2.49	1.95	0.47	5.35	4.06
	XGBoost	1.00	0.00	0.00	0.48	5.10	3.83
	k-NN	1.00	0.00	0.00	0.26	4.24	3.20
	AdaBoost	1.00	0.10	0.01	0.54	4.89	2.98
	ELM	0.52	9.51	7.62	0.40	9.06	7.30
	BPNN	0.50	0.54	4.29	0.30	6.22	4.94
	CARS-Stacking	**0.71**	**4.34**	**3.25**	**0.71**	**4.03**	**3.03**

selection accounts for 11–13 % of the total variables (124), which greatly reduces the computational complexity of the subsequent modeling. However, the effectiveness of such selection needs to be evaluated.

3.3. Modeling evaluation

The heavy metal concentrations and the spectral characteristics selected by the CARS method are modeled in this section. Firstly, all the samples are divided into training sets and test sets according to the 2:1 ratio and the method of decreasing the concentration gradient of heavy metals. The training set is used to adjust the parameters and establish the model, and the test set is used to test and evaluate the generalization ability of the model. The accuracies of the seven models are listed in Table 8.

The R_C^2 of the SVM, XtraTrees, XGBoost, *k*-NN, AdaBoost, ELM, BPNN reach one, which appear in the accuracy evaluation of the training data set but not in the testing data set. It also illustrates that the traditional models show serious overfitting. In contrast, the ensemble learning method based on a stacking strategy has better performance on the training set, which reflects the strong generalization performance of the method. For *Pb* and *Zn*, the R_C^2 of the SVM model reaches one. The R_C^2 of the ExtraTrees model is one for *Cr*, *Pb* and *Zn*.The R_C^2 of the XGBoost and *k*-NN models reach one for all four metals. As for the *k*-NN, it is to preprocess the training data set efficiently before inputting the test set. Each testing sample searches for k nearest training samples according to distance measurement; in the regression, its estimated value is obtained in a weighted manner of the k samples. The stacking-based method is superior to other methods in inversion accuracy of all the heavy metals. For *As*, the R_C^2 of different methods are more than 0.9 except PLS, which is a linear model and prone to overfitting in heavy metal retrieval. However, the R_P^2 of the ExtraTrees model reaches 0.2, and the R_P^2 of the RF and XGBoost models are less than 0.1. The accuracy of the training data set is abnormally high. This indicates that, based on the characteristic subset of *As* selected by CARS, the decision forest inversion model shows serious overfitting. These three models cannot estimate heavy metals although the R_C^2 values are better. The SVM inversion model performs better than the decision forest model. Based on the proposed stacking model, the R_P^2 of *As* is 0.73, the $RMSE_P$ is 37.09, and the MAE_P is 23.70. The R_P^2 is the highest while the $RMSE_P$ and the MAE_P are the least among all the methods. Therefore, the performance and predictive ability of the model based on stacking are the best. For other heavy metals such as *Cr, Pb, Zn*, there are the same phenomena. The R_C^2 of different methods are almost larger than 0.9 except PLS. Based on the proposed CARS-Stacking model, the R_P^2 reaches the highest value. Finally, the CARS-Stacking model is stable and robust using the small training set. It can been utilized to estimate heavy metals in practical applications. The inversion scatter diagram of CARS-Stacking model is shown in Fig. 7.

4. Discussion

4.1. Spectral feature summary

The common characteristics of heavy metal spectra are discussed in this section. All the features are drawn on the average spectral line of the sample sets, to analyze their spectral characteristics more clearly.

The intersections between the spectral characteristics of the four heavy metals are marked by the black vertical dashed lines. It can be observed that the characteristic spectrum of *As* is densely distributed in the range of 2–2.4 μm in Fig. 8. In this investigation (Ren et al., 2009) spectral regions around 460, 1400, 1900, and 2200 nm were jointly used to build the prediction models of *As*. Meanwhile, some other spectral feature regions like those around 550, 760, and 2300–2500 nm were also used in the construction of the prediction models of *As* concentrations. Compared with Ren's results, the bands between 0.4–0.75 μm has not been selected as the characteristic bands in this study. *As* is not strictly a heavy metal, but is a metallic element. Therefore, the spectral characteristics of *As* are different from those of the other metals.

For other heavy metals, their characteristic spectral bands are scattered. In addition to the presence of characteristic bands between 2–2.4 μm, there are also some characteristic bands in the visible band of the 0.5–0.75 μm spectrum. *Cr*, *Pb*, and *Zn* are strictly heavy metals. According to the spectrochemical analysis, the spectral characteristics

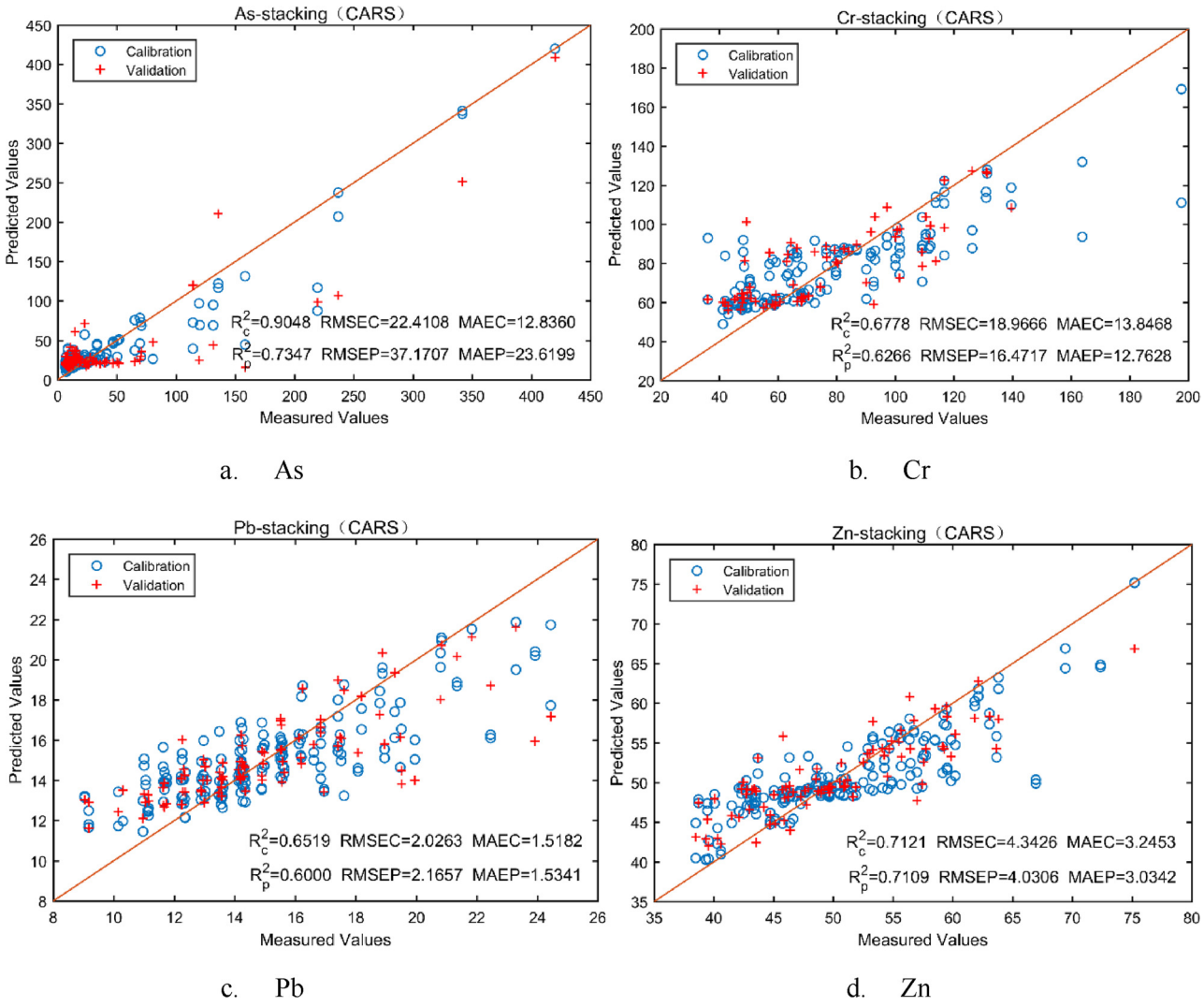


Fig. 7. Inversion Scatter Diagram of CARS-stacking Model (Unit: mg/kg).

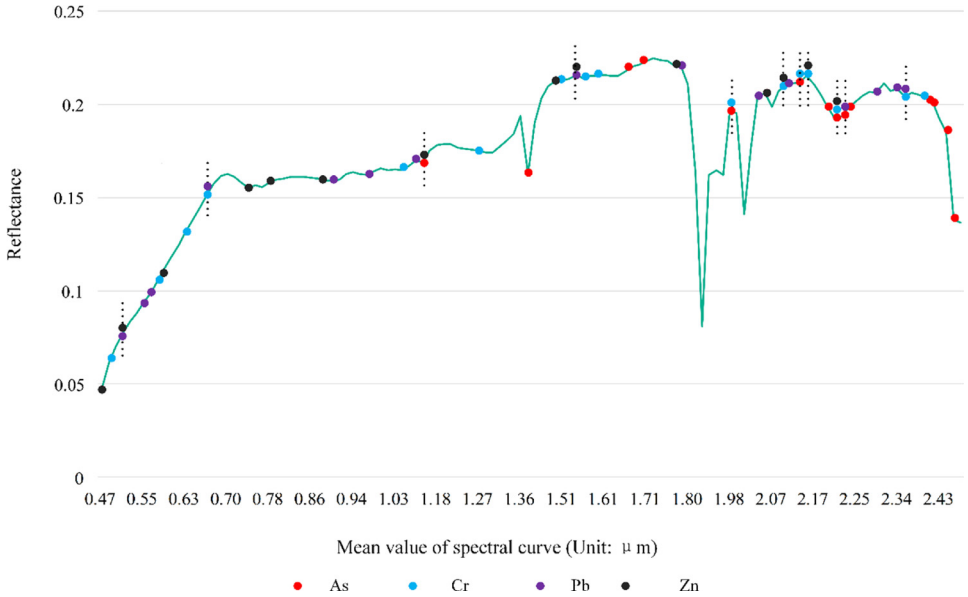


Fig. 8. Reflectance spectrum characteristics of heavy metals based on the CARS selection results.

Table 9
CARS feature wavelength intersection.

	As	*Cr*	*Pb*	*Zn*
As				
Cr	**1.98, 2.21**, 2.40			
Pb	**2.06, 2.23**	0.57, 0.67, **2.36**		
Zn	1.14, **2.06, 2.21**	**2.09, 2.15, 2.21**	0.51, 1.53,**2.06**	

Table 10
Soil environmental quality Risk control standard for soil contamination of agricultural land (Agency, 1995; Regulation, 2018) (Unit: mg/kg).

Metal	Background	Risk screening values	Risk intervention values
As	15	40	150
Cr	90	150	850
Pb	35	90	500
Zn	100	200	NA

of the heavy metals in the soil are closer in the ultraviolet wavelengths, where the spectral characteristics are more significant. This explains why the three heavy metals (*Cr, Pb, Zn*) have characteristic bands between 0.4–0.75 μm, but *As* has no distinct spectral characteristics. In order to accurately analyze the common spectral characteristics of the four heavy metals, their intersection is shown in Table 9.

It is found that the spectral wavelengths of 2.0–2.3 μm are

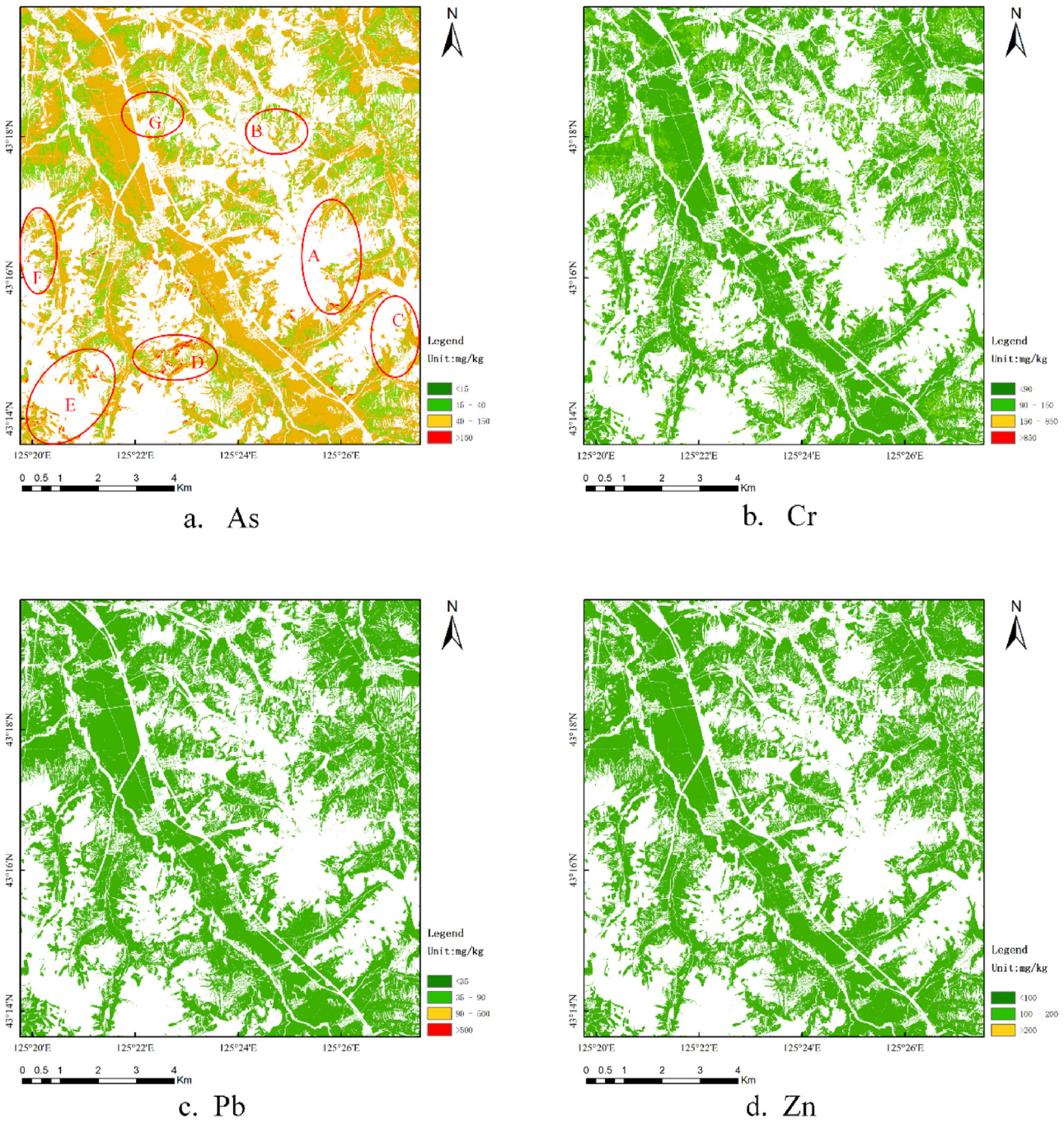


Fig. 9. The estimation results for *As, Cr*, *Pb*, and *Zn* in the study area. The locations marked A–G are affected by human activities. Colors are used to show different areas, divided by the soil environmental quality risk control standard for soil contamination of agricultural land. (For interpretation of the references to colour in this figure text, the reader is referred to the web version of this article.)

Table 11
The area proportion of different soil environmental quality.

Metal	Background	Risk screening values	Risk intervention values	Pollution
As	0	27.477 %	71.545 %	0.978 %
Cr	82.318 %	17.677 %	0.005 %	0
Pb	100 %	0	0	0
Zn	100 %	0	0	0

characteristic bands for a variety of heavy metals, and are generally representative. In addition, there are some common features near the wavelength of 0.5 μm. Also, Song et al. has found that the common characteristic bands of *Cu* and *Cr* in the soil of Chongqing Wan-sheng mining area are around 480 nm, 500 nm, 610 nm, 750 nm, 1430 nm, 1920 nm and 2260 nm (Song et al., 2015). There are the characteristic bands of *Cr* same as our finding.

4.2. CARS-Stacking modeling summary

When comparing the accuracy evaluation indicators of all the models, the CARS-Stacking model performs well. It shows an improvement in the ability to overcome the various problems caused by a small sample set and imbalanced data. In the estimation result of CARS-Stacking, the R_P^2 of the prediction data set is 0.73, 0.63, 0.60, and 0.71 for the four metals (As, Cr, Pb, Zn). HyMAP data has utilized to map heavy metal distribution in stream sediments of the Rodalquilar mining area (Choe et al., 2008). This study is to derive parameters from spectral variations associated with heavy metals in soil and use these parameters to map the distribution of areas affected by heavy metals. The reliable nature of results obtained by multiple linear regressions (generally, $R^2 > 0.5$) between the ground-derived spectral parameters and heavy metal concentrations. Moreover, the proposed CARS-Stacking model in this paper has better performance which indicates the feasibility of retrieving heavy metal concentrations based on spectral features. The statistical evaluation of the comprehensive training data set shows that the CARS-Stacking model does not result in overfitting, and the overall performance is stable.

After field analysis and verification, it is confirmed that the distribution of heavy metal concentrations (especially the heavy metals with high spatial heterogeneity) based on model inversion is consistent with the actual distribution trend, and the model is reliable.

4.3. Heavy metal estimation from the imagery

Based on the characteristics of the four heavy metals selected by the CARS method, the image features were constructed and input into the stacking model. The inversion of the four heavy metals in the whole study area was carried out to further analyze the spatial distribution of heavy metals.

Fig. 9 shows the estimation results of the CARS-Stacking model for *As* in the study area. The criterion is refer to the national standard (GB15618-1995, GB15618-2018) in China (Agency 1995; Regulation 2018). These standards are used to define the background and anomalies in Chinese soil in Table 10. In the study area, *Pb* and *Zn* do not have anomalous distribution in the chemical dataset.

For the purpose of analysis and interpretation, the areas of great interest and the areas with higher estimated concentration values in the estimation map are marked with red ellipses, and the numbers are indicated by the letters A–G. According to the previous analysis, there is spatial clustering of *As*. In this paper, the estimation results are related to the possible occurrences of foreign heavy metal sources (see Fig. 9) for in-depth verification analysis.

In Table 11, the area proportion of different soil environmental quality has been calculated. For *As,* the risk intervention area are more than 70 %. Moreover, it is about 1 % soil which is serious pollution. Special care should be taken for this area in red plot of Fig. 9. For *Cr*, it is about 17 % under risk screening. *Pb* and *Zn* are within the normal

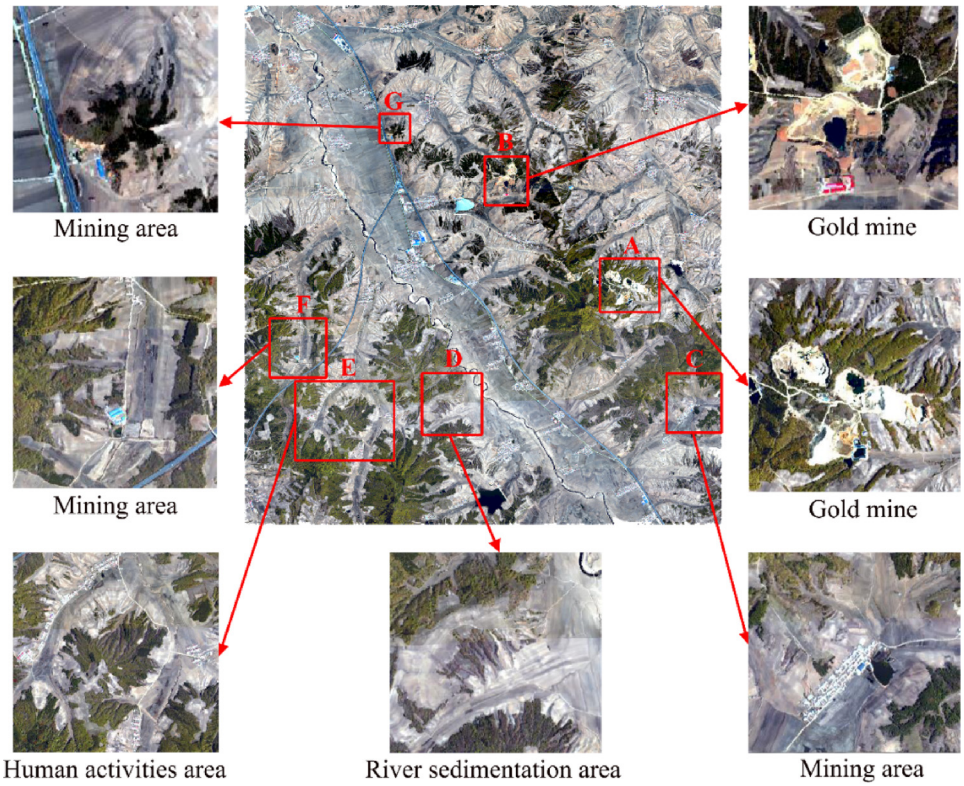


Fig. 10. Analysis of *As* with regard to human activity. (For interpretation of the references to colour in this figure text, the reader is referred to the web version of this article.)

scope.

The A and B areas in Fig. 9 are gold mining areas. It can be observed that the concentration of *As* in the local areas of A/B is high. Area A is a gold mining area with no strong environmental protection measures for tailings and mining production. Therefore, the gold mining area in Area A has a significant impact on the surrounding environment. Area B was once a gold mine, and some of the surrounding area has been partially reclaimed and restored. Therefore, the *As* pollution in Area B does not spread as much as in Area A. Area C is at the boundary of the entire study area, and a mining area is found in the adjacent area on the east side of the study area. This area is likely to be the main factor causing the high concentration of *As* in Area C. There is also a mining area in the west side of Area F. The high concentrations of *As* in Area E are found close to the residential area, as various wastes from the residential area can also cause high concentrations of heavy metals. Area D is a small hill, and some areas are planted with crops. No significant human activities take place in this area. The reason for this may be that the concentration of *As* is historically high. Area G is near the road, surrounded by farmland, and the concentration of *As* in some parts of the area is high due to the presence of a concentrating mill. It can also be seen from Fig. 8 that *As* exists to different degrees on both sides of the river sedimentation zone, from upstream to downstream.

From the correlation analysis results, there is no linear relationship between *As* and *Cr*. The interpretation of the results of *As* estimation is thus difficult to overlap with the interpretation of *Cr*. Due to a large number of abnormal samples removed for *Cr*, the prediction of the model may be low for high-concentration regions. Although the model's ability to fit *Cr* is stronger than the traditional method, there is room for improvement in the fitting ability. From Fig. 10, it can be seen that *Cr* shows an abnormal concentration near the mining areas. There are also abnormalities in the sedimentation area on both sides of the river, and the other areas are basically normal. The concentration of *Pb* is high in some areas. As indicated by the red ellipse in Fig. 10, the west side of the Yitong River is the road. The local geological structure is also very complex, and a large number of mineral types are densely distributed in a small area. Therefore, the local geological environment has a certain explanatory power for the estimation results of *Pb*. The estimation results for *Pb* are extremely low compared with the previous global Moran's I index, the z-score is higher, the p-value is higher, and the spatially random distribution is more intense. Only for a very few pixels is the estimated concentration of *Zn* high. The distribution has a random pattern.

5. Conclusions

In this paper, in order to overcome the problems of overfitting and model instability, we have proposed a CARS-Stacking method for estimating soil heavy metals. Comparing the accuracy indices of all the models using the combined features selected by the CARS method, the accuracy and stability of the CARS-Stacking method is the best. The CARS method is also simple and efficient in selecting spectral characteristics. The spectrum in 2–2.3 μm is the common characteristic band for the four heavy metals. The CARS-Stacking method can overcome the overfitting problems caused by imbalanced data and the small training sample set. Moreover, even for *As* with high spatial heterogeneity, the heavy metal concentration distribution is consistent with the actual verification analysis. The reliability of the CARS-Stacking estimation model is high. However, the CARS-Stacking method has a high complexity. It is therefore necessary to explore low-complexity estimation methods, such as semi-supervised active learning and strategies based on GIS-based spatial analysis. This proposed strategy can actually map the soil heavy metal concentrations in a large spatial area. Such mapping provides the guidance for in-field sampling and precise measurement as needed if some place is under risk screening or risk intervention. Also, these results provide early warning reference for people in polluted areas. In this paper, the analysis of the factors affecting the heavy metals is relatively deficient. In practice, the geological environment, topography, crops, and even seasonal changes have certain effects on the transport and transfer of heavy metals in soil. Therefore, we will give full consideration to these factors and make a thorough analysis of their impacts in the future study.

CRediT authorship contribution statement

Kun Tan: Conceptualization, Methodology, Writing - review & editing. **Weibo Ma:** Writing - review & editing, Software. **Lihan Chen:** Data curation, Writing - original draft. **Huimin Wang:** Data curation, Writing - original draft. **Qian Du:** Visualization, Investigation, Validation. **Peijun Du:** Writing - review & editing, Software. **Bokun Yan:** Visualization, Investigation, Validation. **Rongyuan Liu:** Visualization, Investigation, Validation. **Haidong Li:** Visualization, Investigation, Validation.

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.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Nos. 41871337, 41471356), the Geological Survey Project of China (Grant no. DD20160068) the research and development fund for the central level scientific research institutes, Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment (GYZX190101 and NIES 2011); Key Laboratory for National Geographic Census and Monitoring, National Administration of Surveying, Mapping and Geoinformation (2018NGCM08) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Abdi, H., Williams, L.J., 2013. Partial least squares methods: partial least squares correlation and partial least square regression. In: In: Reisfeld, B., Mayeno, A.N. (Eds.), Computational Toxicology II. Humana Press, Totowa, NJ, pp. 549–579.
Agency, N.E.P., 1995. Environmental Quality Standard for Soils. National Environmental Protection Agency, Beijing, pp. 1–5.
Antonio, P.L., Raphael, A.V.-R., Pietro, A., Andrea, B., 2012. Prediction of soil properties with PLSR and vis-NIR spectroscopy: application to Mediterranean soils from southern Italy. Curr. Anal. Chem. 8, 283–299.
Asadzadeh, S., de Souza Filho, Carlos Roberto, 2016. A review on spectral processing methods for geological remote sensing. Int. J. Appl. Earth Obs. Geoinf. 47, 69–90.
Asmaryan, S., Muradyan, V., Sahakyan, L., Saghatelyan, A., Warner, T., 2014. Development of Remote Sensing Methods for Assessing and Mapping Soil Pollution with Heavy Metals. Taylor & Francis Group, London.
Axelsson, C., Skidmore, A.K., Schlerf, M., Fauzi, A., Verhoef, W., 2013. Hyperspectral analysis of mangrove foliar chemistry using PLSR and support vector regression. Int. J. Remote Sens. 34, 1724–1743.
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, 1703–1712.
Balabin, R.M., Smirnov, S.V., 2011. Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. Anal. Chim. Acta 692, 63–72.
Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36, 105–139.
Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31.
Bendor, E., Chabrillat, S., Demattê, J.A.M., Taylor, G.R., Hill, J., Whiting, M.L., Sommer, S., Ustin, S.L., Schaepman, M.E., 2009. Using Imaging Spectroscopy to study soil properties. Remote Sens. Environ. 113, S38–S55.
Bishop, Christopher M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc, New York.
Breiman, L., 1996. Stacked regressions. Mach. Learn. 24, 49–64.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: history and future

directions. Geoderma 264, 256–274.
Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290.
Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794.
Chen, Y., Wu, W., 2017a. Mapping mineral prospectivity using an extreme learning machine regression. Ore Geol. Rev. 80, 200–213.
Chen, Y.L., Wu, W., 2017b. Mapping mineral prospectivity using an extreme learning machine regression. Ore Geol. Rev. 80, 200–213.
Chen, H., Teng, Y., Lu, S., Wang, Y., Wang, J., 2015a. Contamination features and health risk of soil heavy metals in China. Sci. Total Environ. 512–513, 143–153.
Chen, T., Chang, Q., Clevers, J.G.P.W., Kooistra, L., 2015b. Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy. Environ. Pollut. 206, 217–226.
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. Remote Sens. Environ. 112, 3222–3233.
CSES, 1999. Atmosphere REMoval Program (ATREM) User's Guide, Version 3.1. Boulder, Colorado. pp. 1–31.
Dehaan, R.L., Taylor, G.R., 2002. Field-derived spectra of salinized soils and vegetation as indicators of irrigation-induced soil salinization. Remote Sens. Environ. 80, 406–417.
Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., Huvenne, J.P., 2009. Support vector machines (SVM) in near infrared (NIR) spectroscopy: focus on parameters optimization and model interpretation. Chemom. Intell. Lab. Syst. 96, 27–33.
Dietterich, T.G., 2000. Ensemble Methods in Machine Learning. Springer, Berlin Heidelberg.
Duan, H.W., Zhu, R.G., Xu, W.D., Qiu, Y.Y., Yao, X.D., Xu, C.J., 2017. Hyperspectral imaging detection of total viable count from vacuum packing cooling mutton based on GA and CARS algorithms. Spectrosc. Spectr. Anal. 37, 847–852.
Farrand, W.H., Harsanyi, J.C., 1997. Mapping the distribution of mine tailings in the Coeur d'Alene River Valley, Idaho, through the use of a constrained energy minimization technique. Remote Sens. Environ. 59, 64–76.
Fearn, T., Riccioli, C., Garrido-Varo, A., Guerrero-Ginel, J.E., 2009. On the geometry of SNV and MSC. Chemom. Intell. Lab. Syst. 96, 22–26.
Ferrier, G., 1999. Application of imaging spectrometer data in identifying environmental pollution caused by mining at Rodaquilar, Spain. Remote Sens. Environ. 68, 125–137.
Gannouni, S., Rebai, N., Abdeljaoued, S., 2012. A spectroscopic approach to assess heavy metals contents of the mine waste of Jalta and Bougrine in the north of Tunisia. J. Geogr. Inf. Syst. 4, 242–253.
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.
Gholizadeh, A., Saberioon, M., Ben-Dor, E., Boruvka, L., 2018. Monitoring of selected soil contaminants using proximal and remote sensing techniques: background, state-of-the-art and future perspectives. Crit. Rev. Environ. Sci. Technol. 48, 243–278.
Goodarzi, R., Mokhtarzade, M., Zoej, M.J.V., 2015. A robust fuzzy neural network model for soil lead estimation from spectral features. Remote Sens. 7, 8416–8435.
Guyon, I., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Haaland, D.M., Thomas, E.V., 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60, 1193–1202.
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. Remote. Sens. 39, 529–545.
Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19, 153–158.
Ji, J.F., Song, Y.X., Yuan, X.Y., Yang, Z.F., Gilkes, R.J., Prakongkep, N., 2010. Diffuse reflectance spectroscopy study of heavy metals in agricultural soils of the Changjiang River Delta, China. In: Proceedings of the 19th World Congress of Soil Science: Soil Solutions for a Changing World. Brisbane, Australia. pp. 47–50.
Jia, Z., Li, S., Wang, L., 2018. Assessment of soil heavy metals for eco-environment and human health in a rapidly urbanization area of the upper Yangtze Basin. Sci. Rep. 8, 3256.
Jiang, H., Liu, G., Mei, C., Yu, S., Xiao, X., Ding, Y., 2012. Measurement of process variables in solid-state fermentation of wheat straw using FT-NIR spectroscopy and synergy interval PLS algorithm. Spectrochim. Acta A: Mol. Biomol. Spectrosc. 97, 277–283.
Jie, L., 2012. Hyperspectral remote sensing estimation model for Cd concentration in rice using support vector machines. J. Appl. Sci. 30, 105–110.
Kemper, T., Sommer, S., 2002. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environ. Sci. Technol. 36, 2742–2747.
Kinoshita, R., Moebiusclune, B.N., Es, H.M.V., Hively, W.D., Bilgilis, A.V., 2012. Strategies for soil quality assessment using visible and near-infrared reflectance spectroscopy in a Western Kenya Chronosequence. Soil Sci. Soc. Am. J. 76, 1776–1788.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence 1137–1143.
Koushik, A.N., Manoj, M., Nezamuddin, N., 2020. Machine learning applications in activity-travel behaviour research: a review. Transp. Rev. 40, 288–311.
Kruse, F.A., Boardman, J.W., Lefkoff, A.B., Young, J.M., Kierein-Young, K.S., Cocks, T.D., Jenssen, R., Cocks, P.A., 2000. HyMap: an Australian hyperspectral sensor solving global problems-results from USA HyMap data acquisitions. Proc. of the 10th Australasian Remote Sensing and Photogrammetry Conference 18–23.
Li, H., Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta 648, 77–84.
Liu, Y., Li, W., Wu, G., Xu, X., 2011a. Feasibility of estimating heavy metal contaminations in floodplain soils using laboratory-based hyperspectral data—a case study along Le'an River, China. Geo-Spatial Inf. Sci. 12, 10–16.
Liu, M., Liu, X., Wu, M., Li, L., Xiu, L., 2011b. Integrating spectral indices with environmental parameters for estimating heavy metal concentrations in rice using a dynamic fuzzy neural-network model. Comput. Geosci. 37, 1642–1652.
Liu, K., Zhao, D., Fang, J.Y., Zhang, X., Zhang, Q.Y., Li, X.K., 2016. Estimation of heavy-metal contamination in soil using remote sensing spectroscopy and a statistical approach. J. Indian Soc. Remote. Sens. 45, 1–9.
Ma, W., Tan, K., Du, P., 2016. Predicting soil heavy metal based on Random Forest model. Geoscience and Remote Sensing Symposium 4331–4334.
Malley, D.F., Williams, P.C., 1997. Use of near-infrared reflectance spectroscopy in prediction of heavy metals in freshwater sediment by their association with organic matter. Environ. Sci. Technol. 31, 3461–3467.
Malmir, M., Tahmasbian, I., Xu, Z., Farrar, M.B., Bai, S.H., 2019. Prediction of soil macro- and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique. Geoderma 340, 70–80.
Mitchell, 2003. Machine Learning. China Machine Press; McGraw-Hill Education (Asia).
Moros, J., Fdezortiz, D.V.S., Gredilla, A., De, D.A., Madariaga, J.M., Garrigues, S., De, L.G.M., 2009. Use of reflectance infrared spectroscopy for monitoring the metal content of the estuarine sediments of the Nerbioi-Ibaizabal River (Metropolitan Bilbao, Bay of Biscay, Basque Country). Environ. Sci. Technol. 43, 9314–9320.
Pandit, C.M., Fieippelli, G.M., Li, L., 2010. Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression. Int. J. Remote Sens. 31, 4111–4123.
Rätsch, G., Onoda, T., Müller, K.-R., 2001. Soft margins for AdaBoost. Mach. Learn. 42, 287–320.
Regulation, S.Af.M., 2018. Soil Environmental Quality Risk Control Standard for Soil Contamination of Agricultural Land. Ministry of Ecological Environment of the people's Republic of China, Beijing, pp. 1–7.
Ren, H.-Y., Zhuang, D.-F., Singh, A.N., Pan, J.-J., Qiu, D.-S., Shi, R.-H., 2009. Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: a case study. Pedosphere 19, 719–726.
Rinnan, Å., Berg, F.V.D., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. Trac Trends Anal. Chem. 28, 1201–1222.
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.
Rossel, R.A.V., 2007. Robust modelling of soil diffuse reflectance spectra by "Bagging-Partial least squares regression". J. Near Infrared Spectrosc. 15, 39–47.
Sares, A., Hauff, P.L., P.D. C, 2004. Characterizing sources of acid rock drainage and resulting water quality impacts using hyperspectral remote sensing—examples from the upper Arkansas River Basin. In: Geospatial Conference. Colorado. pp. 7–9.
Schapire, R.E., 2003. The Boosting Approach to Machine Learning: An Overview. Springer, New York.
Shi, T., Chen, Y., Liu, Y., Wu, G., 2014a. Visible and near-infrared reflectance spectroscopy-an alternative for monitoring soil contamination by heavy metals. J. Hazard. Mater. 265, 166–176.
Shi, T., Chen, Y., Liu, Y., Wu, G., 2014b. Visible and near-infrared reflectance spectroscopy—an alternative for monitoring soil contamination by heavy metals. J. Hazard. Mater. 265, 166–176.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.
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.
Soriano-Disla, J.M., Janik, L.J., Rossel, R.A.V., Macdonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49, 139–186.
Stańczyk, U., 2015. Feature selection for data and pattern recognition. Stud. Comput. Intell. 584, 1–7.
Sun, W., Zhang, X., 2017. Estimating soil zinc concentrations using reflectance spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 58, 126–133.
Tan, K., Ye, Y., Cao, Q., Du, P., Dong, J., 2014. Estimation of arsenic contamination in reclaimed agricultural soils using reflectance spectroscopy and ANFIS model. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 2540–2546.
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, 2008–2022.
Thissen, U., Pepers, M., Üstün, B., Melssen, W.J., Buydens, L.M.C., 2004. Comparing support vector machines to PLS for spectral regression applications. Chemom. Intell. Lab. Syst. 73, 169–179.
Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection. Geoderma 223–225, 88–96.
Wang, J., Cui, L., Gao, W., Shi, T., Chen, Y., Gao, Y., 2014. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216, 1–9.
Wang, Q., Xie, Z., Li, F., 2015. Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. Environ. Pollut. 206, 227–235.

Wang, F., Gao, J., Zha, Y., 2018a. Hyperspectral sensing of heavy metals in soil and vegetation: feasibility and challenges. ISPRS J. Photogramm. Remote Sens. 136, 73–84.
Wang, F., Gao, J., Zha, Y., 2018b. Hyperspectral sensing of heavy metals in soil and vegetation: feasibility and challenges. ISPRS J. Photogramm. Remote Sens. 136, 73–84.
Wei, B., Yang, L., 2010. A review of heavy metal contaminations in urban soils, urban road dusts and agricultural soils from China. Microchem. J. 94, 99–107.
Wilford, J., De Caritat, P., Bui, E., 2016. Predictive geochemical mapping using environmental correlation. Appl. Geochem. 66, 275–288.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130.
Wolpert, D.H., 2011. Stacked Generalization. Springer U.S.
Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., Ma, H., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71, 918–926.
Xiaobo, Z., Jiewen, Z., Povey, M.J., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667, 14–32.
Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224.
Zhang, X., Zhang, F., Kung, H., Shi, P., Yushanjiang, A., Zhu, S., 2018. Estimation of the Fe and Cu contents of the surface water in the Ebinur Lake Basin Based on LIBS and a machine learning algorithm. Int. J. Environ. Res. Public Health 15.
Zhao, S., Wang, Q., Li, Y., Liu, S., Wang, Z., Zhu, L., Wang, Z., 2017. An overview of satellite remote sensing technology used in China's environmental protection. Earth Sci. Inform. 10, 137–148.
Zhou, Z.H., 2012. Ensemble Methods: Foundations and Algorithms. Taylor & Francis.