



OPEN Hyperspectral inversion of heavy metal content in farmland soil under conservation tillage of black soils

Yanan Chen^{1,3✉}, Wanying Shi¹, Guzailinuer Aihemaitijiang³, Feng Zhang³, Jiquan Zhang^{4✉}, Yichen Zhang¹, Dianqi Pan² & Jinying Li¹

Globally, heavy metal (HM) soil pollution is becoming an increasingly serious concern. Heavy metals in soils pose significant environmental and health risks due to their persistence, toxicity, and potential for bioaccumulation. These metals often originate from anthropogenic activities such as industrial emissions, agricultural practices, and improper waste disposal. Once introduced into the soil, they can bind to soil particles, making them difficult to remove, while potentially entering the food chain through plant uptake or water contamination. Rapid access to reliable data on HM viscosity in soils is necessary to efficiently monitor remediated soils. Visible and near-infrared reflectance spectroscopy (350–2500 nm) is an economical and zero-pollution method that can evaluate multiple HM concentrations in soil simultaneously. Black soil is a valuable agricultural resource that helps guarantee food security worldwide and can serve as a soil carbon reservoir, but its protection faces several challenges. Due to long-term high-intensity development and utilization and the severe over-exploitation of groundwater, the arable land in China's black soil area has been degraded. Using hyperspectral inversion of heavy metal content in soil can reduce the destructive sample collection and chemical pollution of soil, better protect black land resources, and steadily restore and improve the basic fertility of black land. Focusing on the black area region of Jilin Province, this study explored the correlation between three HMs, namely copper, zinc, and cadmium, and organic substances, clay minerals, and ferromanganese oxides through an in-depth analysis of soil samples using soil reflectance spectrometry. The spectra were transformed using first- and second-order derivatives, multiple scattering corrections, autoscales, and Savitzky–Golay smoothing. The successive projection algorithm was used to screen characteristic bands (Table S1) to establish the link between HM content in soil and soil spectra. By employing the support vector machine (SVM), random forest (RF), and partial least squares (PLS) models, feature band-based soil HM inversion modeling was established. Moreover, the optimal combinations of spectral transforms and inversion models were also examined. The findings indicate that the RF model ($R^2 > 0.8$, RPIQ > 0) outperformed the SVM and PLS models in anticipating the three soil HMs, thus demonstrating superior accuracy. Understanding the behavior of heavy metals in soils and developing effective management strategies are essential for ensuring sustainable land use and protecting public health. This study contributes to the development of large-scale monitoring systems for the HM content of soil and assessments of HM contamination.

Keywords Spectral pretreatment, Soil heavy metal, Random forest model, Black soil area, Spectral transformation, Conservation tillage

Abbreviations

HM Heavy metal
SPA Successive projection algorithm

¹College of Jilin Emergency Management, Changchun Institute of Technology, Changchun 130012, China.

²College of Graduate Studies, Changchun Institute of Technology, Changchun 130012, China. ³College of Resources and Environment, Jilin Agricultural University, Changchun 130118, China. ⁴Institute of Natural Disaster Research, School of Environment, Northeast Normal University, Changchun 130024, China. ✉email: chenyn061@nenu.edu.cn; zhangjq022@nenu.edu.cn

PLS	Partial least squares
SVM	Support vector machine
RF	Random forest
DT	Derivative transformation
MSC	Multiplicative scatter correction
FD	First-order derivative
SD	Second-order derivative
SG	Savitzky-Golaysa2

Soil is a material necessary for life on Earth, representing the basis for biological survival and plant growth and serving as a valuable resource for human development¹. Soil chemistry, soil quality, and environmental status are influenced by a range of biological, physical, and structural factors^{2,3}. Northeast black soil is China's valuable land resources; the Northeast black soil area, known as the "golden corn belt" and "home of soybeans", is China's most important grain producing area and the largest high-quality commodity grain production base. Among different soil types, black soil has the highest fertility, is the most suitable for farming, and has the most productive potential; thus, it is known as the "giant panda of arable land"^{4,5}. As one of the only three black soil areas in the Northern Hemisphere, the northeastern region of China is an important area for grain production and the largest commercial grain production base in China^{6,7}.

Increases in mineral development, pesticide use, and industrial waste discharge in the past decade have caused large amounts of heavy metal (HM) elements to be discharged into the soil. HMs have harmful compounds that negatively affect human and wildlife health, with potentially long-lasting effects^{8–10}. For humans, exposure to such HMs can result in the development of neurological and immune system imbalance, allergies, osteoporosis, asthma, and an increased lifetime risk of cancer^{11–13}. Common methods of human exposure to HM include contaminated food and drinking water (especially from foods such as vegetables, meat, and milk), skin contact, and inhalation through the respiratory tract¹⁴. HMs are slow to break down and migrate in agricultural fields, where they negatively impact soil health conditions. HM ions in agricultural soils are highly susceptible to crop uptake and enrichment and thus pose a serious threat to human health. In addition, methane in agricultural land may accumulate and move higher up the food chain, further aggravating the global crises in human health and food safety¹⁵. The toxicity, abundance, and persistence of HMs make them a global concern¹⁶. Therefore, developing a method that can rapidly and accurately determine soil HM contents and reveal the changes in the soil environment is important for safeguarding its ecosystems and ensuring sustainable agriculture development¹⁷.

The spatial distribution of HMs is generally determined through chemical analysis, which is environmentally polluting, costly and time-consuming¹⁸. In contrast, hyperspectral remote sensing technology utilizes the electromagnetic wave reflections of different wavelengths for analysis, producing images that provide spectral and spatial information in the visible and infrared realms. The spectral continuity of the sampling interval ensures that the reflectance reflects the features of different objects. High-resolution image data can be promptly monitored using hyperspectral remote sensing technology, which is already widely used in various fields. However, using hyperspectral data over sample measurements in the inversion of HM concentrations in the visible and infrared bands is highly challenging, and the integration of data from different hyperspectral sources becomes critically important. Nevertheless, given the richness of spectral information, hyperspectral remote sensing techniques have been progressively used to monitor the chemical and physical causalities of soils^{19,20}. Recently, several studies have successfully applied deep learning techniques such as Convolutional Neural Networks (CNN)²¹ Long Short-Term Memory Networks (LSTM)²², migration learning²³ and other deep learning²⁴ in hyperspectral soil properties prediction.

Currently, many studies have estimated HMs in soil. Due to the lack of a clear mechanism for HM estimation, many studies have resorted to statistical methods or organic matter correlation analyses, which increases the difficulty of revealing the characteristics and mechanisms of soil HM uptake. Meanwhile, due to factors such as the scarcity of HM content in soils, the complexity of environmental imaging, and the redundancy of spectroscopic data, HM estimation is ineffective and cannot effectively outline the trend of their distribution. The close relationship between heavy metals and spectrally active components²⁵ and several other studies explored the combination of concentration information with spectral features related to spectrally active components to improve the estimation accuracy of soil heavy metals. Kästner et al.²⁶ proposed that the addition of a laboratory soil preparation step and a further spectral pre-processing step can improve the performance of PTE estimation for non-homogeneous soil samples. In soil spectral measurements, various factors such as light scattering and different distributions of particle size and density within the soil may lead to the appearance of noise in the spectral data, thereby requiring adequate preprocessing of the spectral data prior to spectral estimation²⁷. The commonly used preprocessing methods include derivative transform (DT), smoothing, standard normal variate correction, multiplicative scatter correction (MSC), baseline correction, normalization, wavelet transform, and continuum removal^{28,29}. Several recent studies²⁴ have shown that combining spectral preprocessing techniques gives better training results than using a single preprocessing technique. Therefore, we apply a combination of spectral preprocessing methods to find the best combination of features.

An important method for determining the HM content of soils, based on the characterization of HM contamination, is hyperspectral soil analysis³⁰. The influence of HMs on spectral reflectance is considered a nonlinear process; therefore, nonlinear models exhibit better inverse performance³¹. As nonlinear models, deep learning algorithms have been widely used in hyperspectral technology. Various studies have been conducted on deep learning-based hyperspectral inversion^{32,33}, and the results show that deep learning can provide an effective method for realizing HM concentration inversion using hyperspectral reflectance. Two types of inversion models have been used for hyperspectral inversion: linear inversion models and nonlinear inversion models, such as random forests and gradient-enhanced decision trees^{34–36}.

In this study, the northeastern black earth region in Jilin Province was chosen as the research area and the inversion and adsorption mechanisms of HMs in the soil were explored using laboratory hyperspectral data. An inversion model was established to accurately determine the HM content in the research area. The objectives of this study are as follows:

- (i) Develop a rapid method for estimating HM content based on the hyperspectral correlation of soil fractions;
- (ii) Examine the Zn, Cu, and Cd contents in the black earth region; and (iii) Create a model that can accurately determine the Zn, Cu, and Cd contents using hyperspectral data.

Traditional soil heavy metal detection methods usually require complex chemical analysis or sampling, while hyperspectral technology can provide a faster and non-invasive method, so the innovativeness of this paper is reflected in the application of cutting-edge remote sensing technology for soil heavy metal monitoring. By combining the hyperspectral inversion technique, the distribution characteristics of soil heavy metals under conservation tillage can be quantitatively analyzed, revealing the complex relationship between soil environmental management measures and heavy metal pollution.

Materials and methods

Research area

The research area is situated in the black soil zone of Jilin Province, with a geographic location of 123°23'–127°38'E and 41°58'–45°22' N, covering four cities and six counties (Fig. 1). This area is part of the continental temperate monsoonal climate zone, with an annual average temperature of 5.2 °C and annual average rainfall quantity of 612.2 mm. The terrain is mainly characterized by mountains, hills, terraces and plains, and the soil types are black soil, black calcium soil, meadow soil, and brown soil. The Jilin black soil area is the most important cereal production base in China due to its favorable geographic location, suitable climatic conditions, and organic matter-rich soil. However, simultaneously with the fast economic development and increased anthropogenicity, the HM content of the soil is on the rise, seriously threatening the local ecological environment and human health. Therefore, the accurate assessment and effective supervision of changes in HM content in the black soil areas is vital.

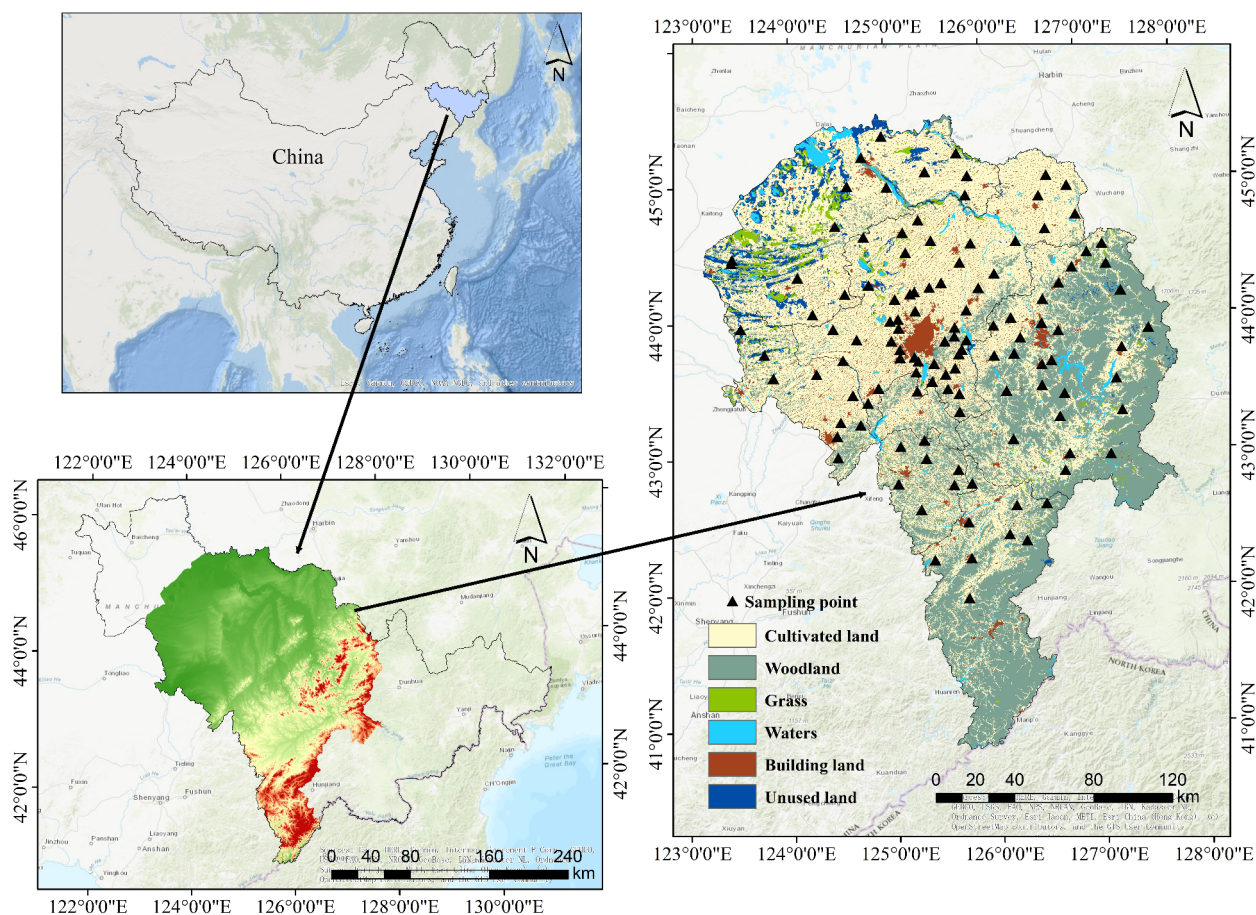


Fig. 1. Overview of the black soil area in Jilin Province (from ArcGIS Vision 10.8; <https://www.esri.com/en-us/arcgis/products/arcgis-desktop/overview>).

Soil sample collection and preparation

A total of 119 soil samples were collected for this study. All sampling points were located at distances greater than 100 m from other features. The topsoil was sampled at depths between 10 and 20 cm using tools such as shovels. For each soil sample, the temperature, humidity, pH, and other relevant soil properties were recorded. In this study, the five-point sampling method of O, A, B, C, and D was selected, and five soil samples of 500 g each with a surface thickness of 5 cm were collected. These five samples were mixed homogeneously and put in a sealed plastic bag. Subsequently, the sampling sites were numbered. All the samples were then taken to a laboratory for natural drying to remove excess moisture. Large particles were removed by filtering through a 2 mm sieve, followed by appropriate polishing to achieve the desired fineness. During this process, strict control of each step was required to ensure that the samples were not contaminated and to guarantee their purity.

Soil spectrum laboratory test

Using an ASD fieldspec4 spectrometer (Fieldspec 4 Hi-Res, Malvern Panalytical), visible to near-infrared (-NIR) spectra in the 350–2500 nm wavelength realm were measured to ascertain the spectral reflectance. Near-sunlight from the incident light source probe was used as the light source in the darkroom, the instrument was warmed up for 30 min, the spectrometer probe was standardized at 10 cm directly above the reference white board, and the spectral reflectance of the soil samples was measured vertically downward using the spectrometer probe at an angle of no more than $\pm 10^\circ$ from the normal vector of the horizontal plane. To avoid measurement errors, 5 spectral measurements were averaged for each soil sample. The final spectrum for each sample was calculated by averaging the remaining spectra after removing the anomalous spectra³⁷. To mitigate the errors that originated from equipment, we calibrated the instrument after measuring 10 specimens using the white panel³⁸. Out of the 119 soil samples, 84 were chosen as the training set (computing), and 35 were the validation set³⁹.

Soil spectral pretreatment

Noise distorts the raw spectral curves of indoor data, which somewhat reduces the accuracy of the spectral estimation model. Thus, the original spectral curves need to be preprocessed, including data conversion and denoising, before building the spectral estimation model. Using the splicing correction tool of the ViewSpec software, the five spectral curves of each soil sample were processed by arithmetic averaging to acquire the real reflectance spectra before calibrating the soil reflectance spectra. To minimize the changes in the laboratory light environment and sample grinding, the spectral data were also preprocessed using multiple scattering corrections, standard normal variables, normalization, autoscales, and multiplicative scattering correlation.

Spectral differential transform is a method widely used in spectral preprocessing. Its main role is to eliminate background interference and resolve overlapping spectral information, greatly minimizing the impact of external differences imposed by the drift of the original spectral baseline, thereby effectively improving the precision of spectral analysis. According to Shi et al.⁴⁰, first-order derivatives (FDs) and second-order derivatives (SDs) are frequently used in MATLAB preprocessing methods. First-order differentiation is primarily used to remove the effects of particular spectral problems and external factors while better highlighting the high frequency spectrum messages and reflecting the fluctuating specificities of raw spectra, whereas second-order differentiation primarily removes context signals and base line offset, improving the response performance of the spectra, which is frequently one of the reasons why spectral waveforms are more complex⁴¹. To remove multiplicative interference in particle size and scattering, multiplicative scattering correction was used to linearly transform the average spectral match of the entire set of spectra. This requires close linear relationships between the spectra and component concentrations. The standardization algorithm converts the data to a standard normal distribution with mean 0 and variance 1 by subtracting the mean and dividing by the standard deviation of each spectral point, thus eliminating the differences in magnitude between different spectra and facilitating subsequent data processing and analysis. SG smoothing is a digital filtering method for noise reduction that smoothes the data by polynomial fitting and preserves the main features of the spectrum. The method not only effectively removes high-frequency noise (e.g., instrument noise, sensor errors, etc.), but also minimizes the loss of spectral information, which contributes to the improvement of the signal-to-noise ratio, which is especially important when analyzing heavy metals at low concentrations in soils. The process of normalization exposes the spectrum variety, removes superfluous information, and streamlines the calibration model.

The goal of all these data preprocessing techniques is to improve the quality of hyperspectral data and the robustness of inversion models. By removing noise, reducing background interference, correcting for scattering effects and standardizing the data, spectral features associated with soil heavy metal content can be captured more accurately, thereby improving inversion accuracy and model prediction.

Inversion method

The equilibrium of HM distribution and adsorption is greatly influenced by soil pH, texture and organic matter content. Certain ground elements, such as organic substances and clay minerals, have a great ability to adsorb particular metals. The carboxyl and phenolic functional groups in organic matter can form strong complexes with Cu, and the adsorption capacity of clay minerals for copper is also stronger, especially under acidic conditions, the surface of the clay minerals has a higher negative charge. As there are more oxygen-containing functional groups (such as carboxyl groups and phenolic hydroxyl groups) in organic matter, their adsorption of cadmium ions is more significant. Zinc ions have strong water solubility, but are also easily adsorbed with organic matter and clay minerals. Compared with copper and cadmium, the adsorption capacity of zinc varies less in organic matter and clay minerals. One method of evaluating the HM concentrations in soil using spectral reflectance is to analyze the sorption process of HMs by soil composition. Because soil organic matter has the ability to adsorb and retain Zn, Cu, and Cd, these three metals were utilized as examples in this investigation.

RF

RF is a category of integrated learning arithmetic that utilizes multiple weak classifiers to increase the accuracy and generalization of the overall model. Random forest is constructed by selecting n data from the training data as training data input, after selecting the input training data, the decision tree is constructed, and the splitting attribute is determined according to the strategy of Gini exponential reduction until it cannot be split or reaches the set threshold, at which time 1 decision tree is established, and each decision tree grows as much as it can, without pruning process^{42,43}. The formula for the Gini index is as follows:

$$Gini(T) = 1 - \sum_k^K p_k^2 \quad (1)$$

where T is the number of samples; K is the number of categories in the sample; p_k is the frequency of the category, k is appearing in sample T . Repeat the above steps until a predetermined number of trees is reached. All the generated decision trees are formed into a random forest, which is then used to predict the new input data.

The flowchart of the random forest model used in this study is shown in Fig. 2. Figure 2 was produced by Visio. It operates by combining the outputs of these weak classifiers through either voting or averaging. This approach is a significant enhancement of the decision tree algorithm, as suggested by Breiman⁴⁴. The RF arithmetic extends the parallel integration algorithm, which combines weak classifiers and randomly selects attributes. Therefore, it shows better advantages in accuracy, generalization performance, and model stability. RF can achieve good results, mainly due to the “random” and “forest” components, with one making it resistant to overfitting ability and the other making it more accurate. RF is an algorithm that integrates multiple models to improve accuracy. This integration results in a higher level of accuracy than most individual algorithms. In addition, RF improves the performance of the test set because the two randomization methods reduce the likelihood of overfitting (random samples and random features). RF is capable of handling high dimensional data without the need for a feature selector. It also fits different datasets, handles continuous and discrete data, and does not require dataset normalization. Owing to the implementation simplicity, high accuracy, and high resistance to overfitting, RF is suitable for use as a baseline model when confronted with nonlinear data. In this study, the hyperparameters of the random forest were set, the number of decision trees was 500, the minimum number of leaf nodes per tree was set to 1, and OBB prediction and feature importance calculation were enabled.

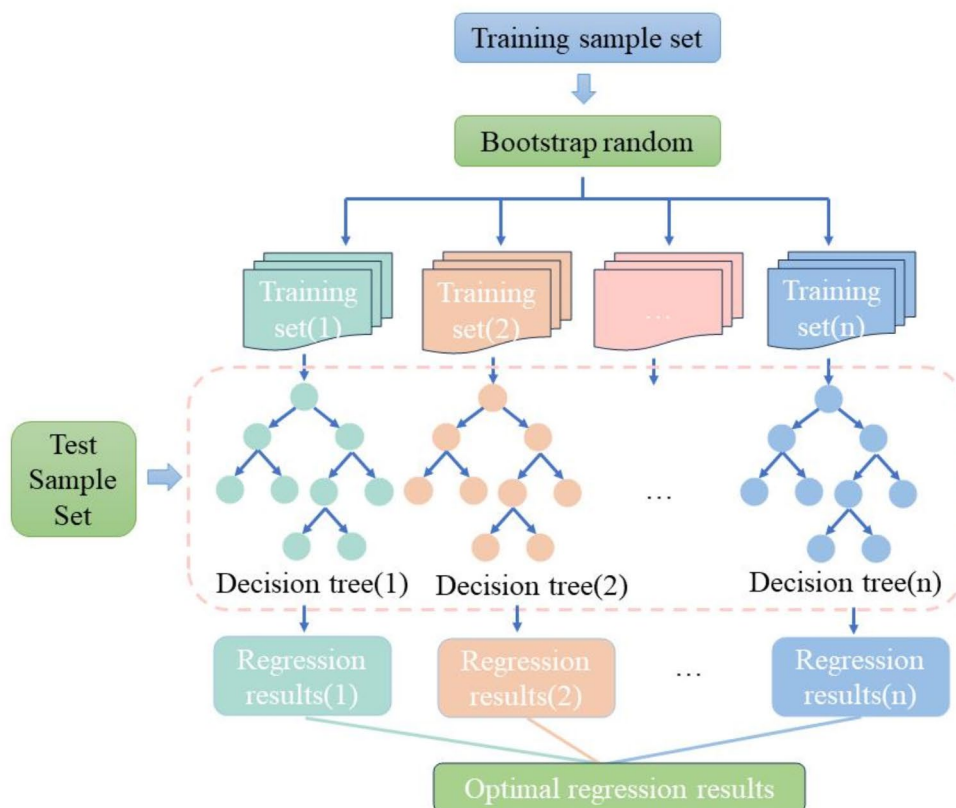


Fig. 2. The RF regression model.

Partial least squares (PLS) regression

The PLS methodology is commonly utilized for the inversion of soil spectroscopic data. To enhance the accuracy of inversion in the soil spectroscopy field, researchers have attempted to integrate nonlinear methods of mathematical analysis, including neural network techniques, genetic algorithms, and support vector machines (SVMs)^{45–47}. In previous studies, PLS has been employed as a benchmark model, and findings suggest that the machine language inversion model can notably enhance its inversion precision^{48,49}. Moreover, linear regression employs least squares as its primary technique for connecting dependent and independent variables. PLS, as described by Leone et al.⁵⁰, constructs a linear model by isolating the latent variables or main components in both the dependent and independent variables. Conversely, ridge regression is a modified version of least squares estimation that is primarily utilized to address non-full rank coefficient matrices. PLS regression is a multi-variate demographic methodology that addresses the problem of covariates and enables the synthesis and analysis of multiple dependent Y variables while also accounting for the impact of small samples in relational studies. PLS regression integrates main component analysis, multiple linear regression, and typical correlation analysis is performed by compressing multiple X and Y variables into principal components through the principle of principal component analysis. In this method, X corresponds to the main component U, and Y corresponds to the main component V. The connection between X and V is analyzed by combining the principles of multiple linear regression, which facilitates the study of the association between X and Y. The hyperparameter of PLS mainly selects the number of principal components, and the complexity and prediction effect of the model can be controlled by adjusting this parameter. In this study, the optimal number of principal components was selected for regression through cross-validation and standardized and normalized before the data were input into the PLS model to ensure that the scales of individual features were consistent, which improved the stability and prediction accuracy of the model.

SVM

SVM is a demographics learning approach based on the theory of structural risk minimization (SRM)⁵¹, which is able to effectively deal with small samples, nonlinear and high-dimensional data compared to the local minimization problem of neural networks. In remote sensing applications, SVM can be successfully applied to small-sample training sets and reduces the limit of the model generalization errors while reducing the sampling error, which in turn improves the generalization model capacity and accuracy⁵². In various disciplines, SVMs, as a widely adopted binary classification model, are basically constructed on interval-maximizing linear classifiers in the feature space with unique interval-maximizing properties, which are significantly different from other classifiers such as perceptual machines. An important class of SVMs, the kernel SVMs, are essentially nonlinear classifiers. SVM formalizes problems as convex quadratic programming solution problems, i.e., the problem of minimizing a regularized hinge loss function, by employing an interval maximization learning strategy. Optimization algorithms for solving convex quadratic programming problems are mainly involved in SVM learning algorithms. In this study, the RBF-kernel was chosen for the SVM regression process because it provides a reasonable trade-off between the number of kernel parameters to be optimized and the adaptability and flexibility of the nonlinear data⁵³. The support vector machine RBF-kernel has two hyperparameters related to the regularization parameter C and the kernel width γ . In this paper, we use grid search and cross-validation to tune C and γ .

Model evaluation and optimization

To assess the forecasting capabilities and robustness of the models, the three key indicators, the root mean square error (RMSE), the coefficient of determination (R^2) and the ratio of performance to interquartile (RPIQ), were used in this study for reliability analysis. The formulae are as follows:

$$R^2 = 1 - \sum_{i=1}^n (y_i - Y_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^n (y_i - Y_i)^2 / n} \quad (3)$$

$$RPIQ = \frac{RMSE_{baseline} - RMSE_{improved}}{RMSE_{baseline}} \quad (4)$$

Uncertainty analysis

The random forest model constructed on the basis of the bagging method can effectively reduce the variance of a single model and improve the accuracy and stability of prediction by integrating multiple decision trees for prediction. However, although the random forest shows strong robustness in many applications, its prediction results still have some uncertainty⁵⁴. Random forests consist of multiple decision trees, and the training data for each tree is generated by self-sampling, so the training set for each tree is different. Although the trees are generated randomly, they are not completely independent because their training sets are drawn from the same dataset. In addition, the tree structure, depth, and splitting criteria may also affect the final model performance. Cross-validation is a commonly used method to assess model stability and generalization ability. By dividing the dataset into multiple subsets and performing multiple training and testing, cross-validation can effectively assess the performance and stability of the model on different datasets, thus quantifying the uncertainty of the model.

Metal	Mean	Std	Max	Min	CV	CF	Jilin Province	National
Zn	157.11	184.85	579	17	1.18	2.34	67.14	300
Cu	31.87	14.02	63.5	7.75	0.44	1.69	18.88	100
Cd	0.25	0.15	0.91	0.05	0.59	2.27	0.11	1.0

Table 1. Statistical results of the soil heavy metal contents (unit: mg/kg).

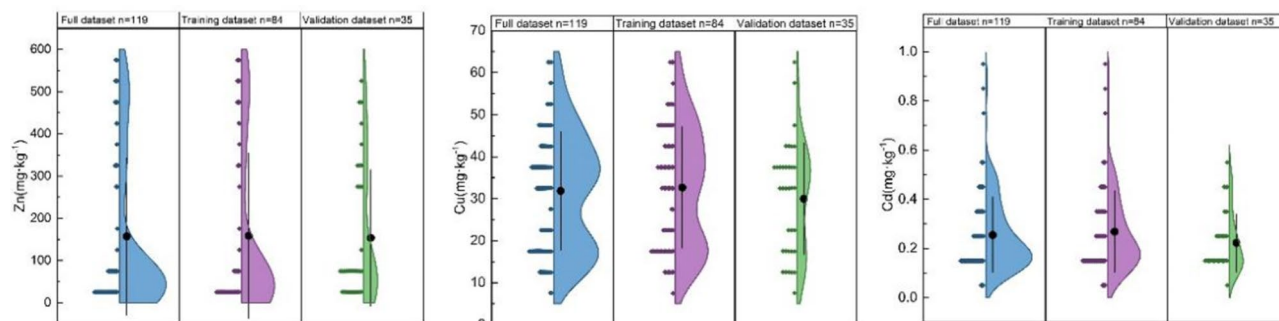


Fig. 3. Descriptive statistics of the soil heavy metal concentrations in the research area. Note: Black dots indicate mean values.

Results and analysis

Description of the soil samples

The successive projection algorithm (SPA) was applied to select the feature band selection. 70% of the dataset was classified as training samples ($n=84$) and the remaining data were used for testing. ($n=35$). Statistical information on the Zn, Cu, and Cd content is given in Table 1. Notably, the average Zn content was above the average background values in Jilin Province, and the maximum value was higher than the national standard, suggesting potential contamination. In contrast, the Cu and Cd levels did not exceed the secondary national environmental quality standard (GB 15618–1995). However, the HM contents in the soil were both higher than the background values in the black soil area of Jilin Province. The pollution coefficient (CF) of Cu ranged from 1 to 2, indicating slight pollution. The CFs of Zn and Cd were > 2 , indicating moderate pollution. The investigation and dynamic monitoring of the soil quality in the research zone should be strengthened to promptly detect and control soil HM pollution. Figure 3 shows the descriptive statistics of the Zn, Cu, and Cd concentrations by Origin 2022.

Soil spectral characterization

As shown in Fig. 4, the reflectance spectra of the soil samples from the black soil zone in Jilin Province show a similar trend, indicating relatively stable reflectance in the visible band and roughly the same location of the features in the absorption bands. (In this study, the pre-processing of the spectra and the modeling were done in MATLAB R2023b.) The spectroscopy reflectance increases monotonically with wavelength in the ranges of 400–787, 1000–1238, 1323–1474, 1801–1881, and 1912–1990 nm and decreases with wavelength in the ranges of 787–1000, 1238–1323, 1474–1801, 1881–1912, and 1990–2500 nm. In general, the absorption features of soil spectra in the visible and NIR bands are principally influenced by the electron-leaping processes of the metal ions Mn^{3+} , Fe^{2+} , and Fe^{3+} , whereas in the NIR band they are mainly influenced by the octave and harmonic flexural vibrations of molecules such as CO_3^{2-} , OH^- , Fe-OH , and $\text{NH}_4^{+55,56}$.

Iron oxides, organic substances, clay minerals and manganese oxides in soil play powerful adsorption roles in the detection of HM content and are some of the indispensable factors in the enrichment process of HMs. By evaluating the coefficients of correlation among the organic substance content, iron and clay minerals, and manganese oxides in soil and HMs, their central supportive roles in soil element accumulation and synergistic effects during the determination of HM content can be revealed, thereby providing a rationale for indirectly determining the content of soil HMs. The required bands were obtained by various means to acquire the spectral properties of the soil elements, the required bands were obtained by various means such as FD, SD, normalization, mean centering, and MSC. The band information is shown in Fig. 5.

Different soil samples exhibited similar trends under the same spectral transformation form. According to previous studies, spectral absorption near 1238 nm and 1881 nm has a clear valley known to be due to the absorption caused by the residual moisture in the soil as well as metal OH binding to soil clay minerals. In the vicinity of 1990 nm, a decreasing trend was observed along the longwave direction. The spectra changed from aggregation to divergence from the short-wave to long-wave direction, indicating that different soil organic matter contents may affect the spectra in the NIR band.

Figure 6 shows a two-by-two correlation plot of the spectroscopic curves and reflectance spectral bands of the soil samples under different spectral transformation forms. This provides accurate results and is thus suitable for modeling. In the reflectance spectral correlation plots under different transformed forms, the correlation

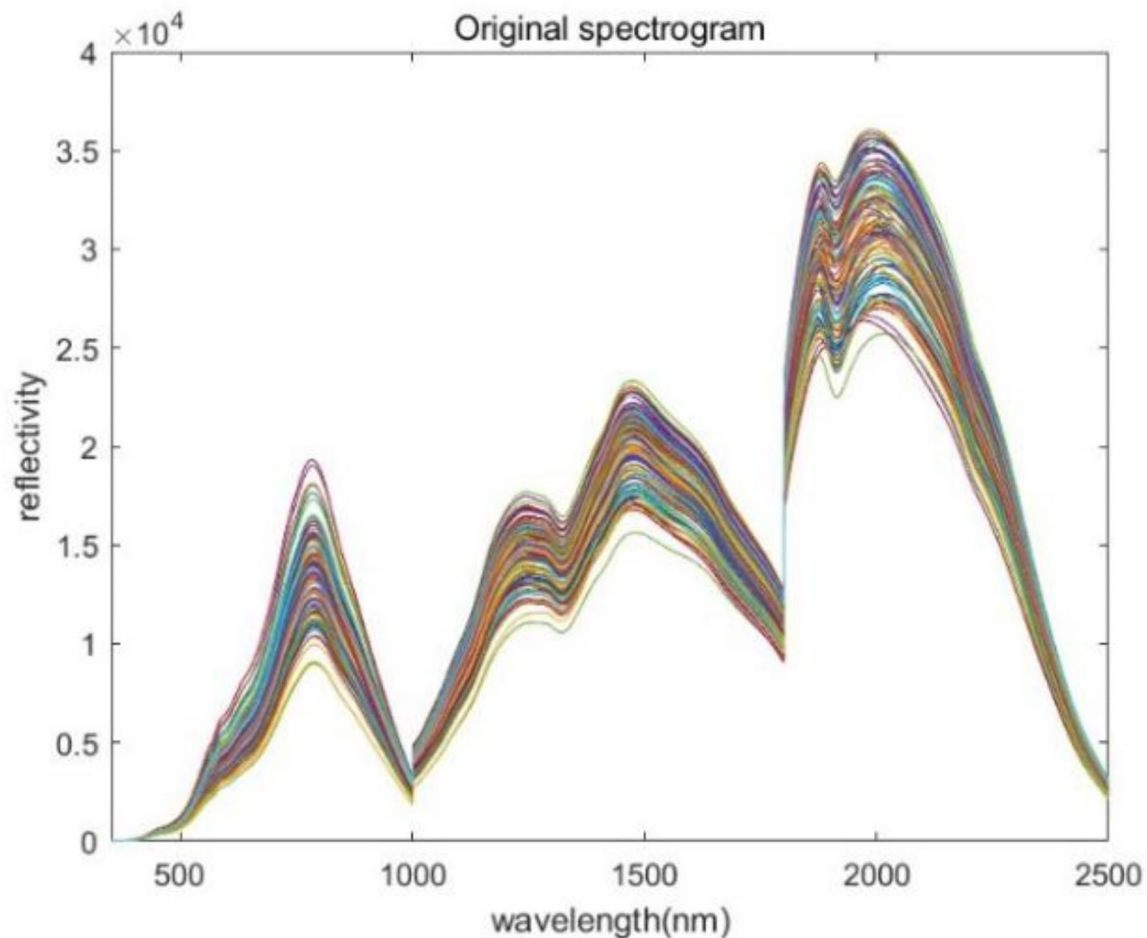


Fig. 4. Reflectance spectra of soil samples from the black soil area of Jilin Province, China.

coefficients of neighboring bands were larger and the correlation coefficients of the short-wave bands, such as at 350–400 nm, were lower because of the effect of the spectral noise generated by the spectrometer. After the first-order and second-order differentiation processes, the covariance between the spectra was partially eliminated. Moreover, bands with small relative coefficients can represent more information about the dependent variable when constructing the regression model and can be used as a basis for further modeling. Analysis of variance (ANOVA) was used to access the variance between means of analyzed pollutants⁵⁷. We used ANOVA in Table S2 to understand the variation of parameters associated with soil heavy metal concentrations.

Model building and evaluation

Spectral feature selection

The SPA can effectively reduce and minimize the covariance of high-dimensional spectral data due to its forward variable selection property, and its undeniable advantage is that it can extract only a few representative characteristic wavelengths of the whole band from the original spectral matrix, thus eliminating the numerous bands attached to the matrix. In this paper, the SPA was used in the MATLAB environment to successfully extract the characteristic wavelength bands of the three HMs Zn, Cu, and Cd from the spectroscopic data (Fig. 7), and the inversion of the soil samples was performed using the RF model.

Modeling evaluation

Specific spectral bands and HM content were selected as model input data, the model parameters were constructed and tuned, and the validation set was used to assess the accuracy and generalization ability of the model. The three machine learning models, SVM, PLS and RF, were used in the study to build the HM prediction model. We have performed hyperparameter optimization to improve the accuracy, efficiency and applicability of the model, so that it performs more stably and efficiently in the face of complex and variable data. The estimation results of the RF model are shown in Fig. 8.

Model accuracy analysis

The FD spectral variable is suitable for modeling HMs with hyperspectral remote sensing, mainly because the processed spectra exclude the context noise, which helps to highlight the spectroscopic information of soil HMs.

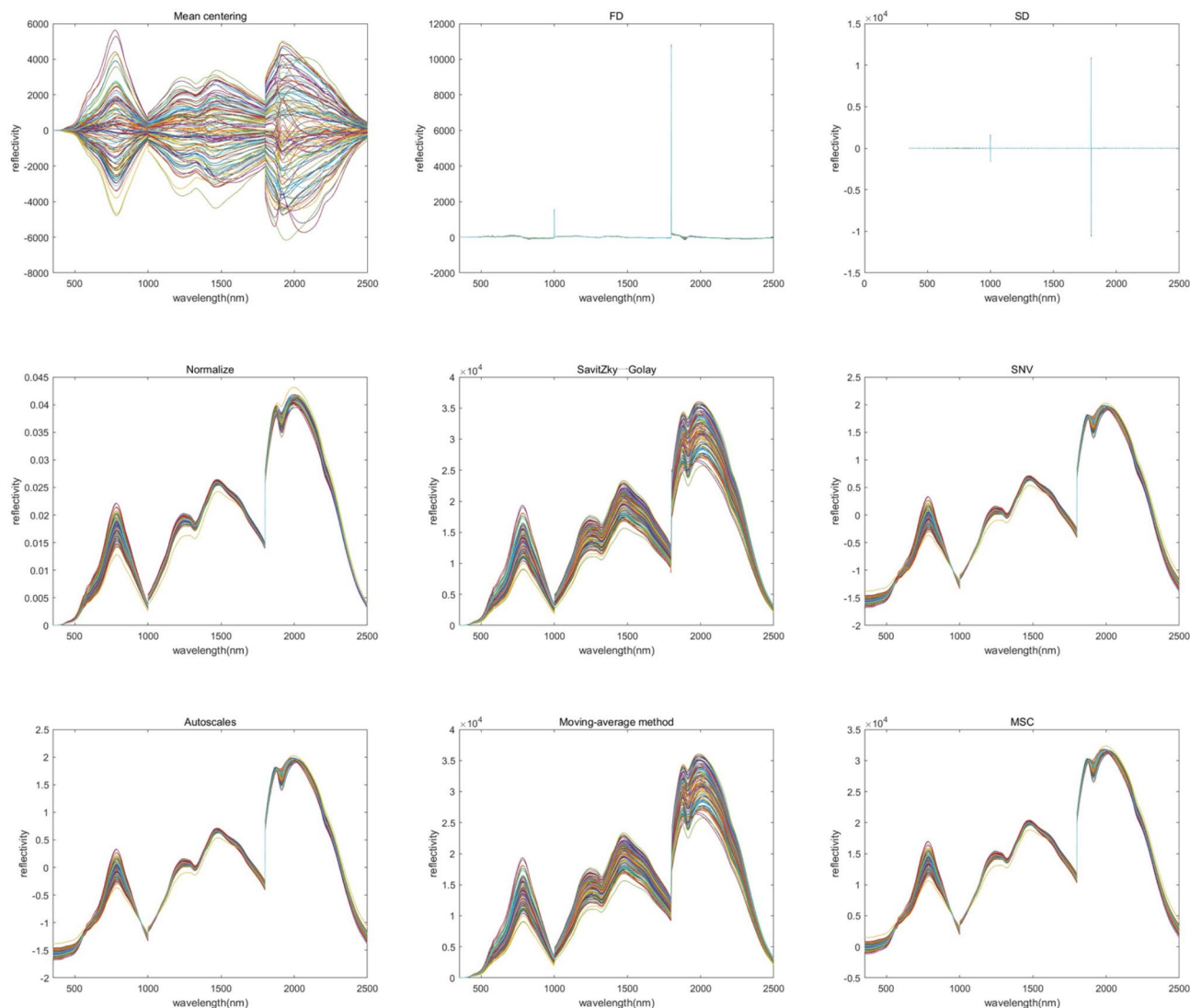


Fig. 5. Spectral transformation curves of soil.

In addition, spectral and RF modeling after FD processing improved the inversion accuracy of Zn, Cu, and Cd. Accordingly, a scatter diagram of the measured and forecasted HM contents based on the peak combination of RF modeling and spectral transformation was plotted in Fig. 9.

Comparison and discussion

Model performance comparison

Comparative results (Table 2) indicate that the RF model has better inversion accuracy than SVM and PLS and the best inversion results for all HMs ($R_{RF}^2 > 0.8$). This indicates that the RF model is highly stable and can effectively invert the HM quantity contained in the black soil area of Jilin Province. The differences in the simulation accuracy of Zn ($R_{RF}^2 = 0.93$, $R_{SVM}^2 = 0.04$, $R_{PLS}^2 = 0.06$) and Cu ($R_{RF}^2 = 0.96$, $R_{SVM}^2 = 0.07$, $R_{PLS}^2 = 0.25$) were the most significant. In addition, the inversion results for Cd show that the RMSE of the RF model (0.07) was significantly smaller than that of SVM (0.09) and PLS (29.71).

Discussion

In the black soil area in Jilin Province, HM pollutants such as Cu, Zn, and Cd have been successfully identified. However, given the low content of HMs in the soil, spectroscopic information is relatively scarce, making it challenging to directly utilize the characteristic spectral analysis for the determination of HM content in the soil. Nevertheless, the mechanism of HM enrichment in soils can be determined and has been found to be closely related to the adsorption of different minerals, such as clays, Fe-Mn oxides, and organics⁵⁸. These not only influence the morphology and reflectance of soil spectra but also exhibit particular absorption features, thus providing a theoretical basis for extracting hyperspectral soil contamination information⁵⁹.

It is urgent to improve the blackland management model and formulate corresponding protection measures to strengthen the quality of blackland. There is an urgent need to follow the laws of nature, adhere to ecological

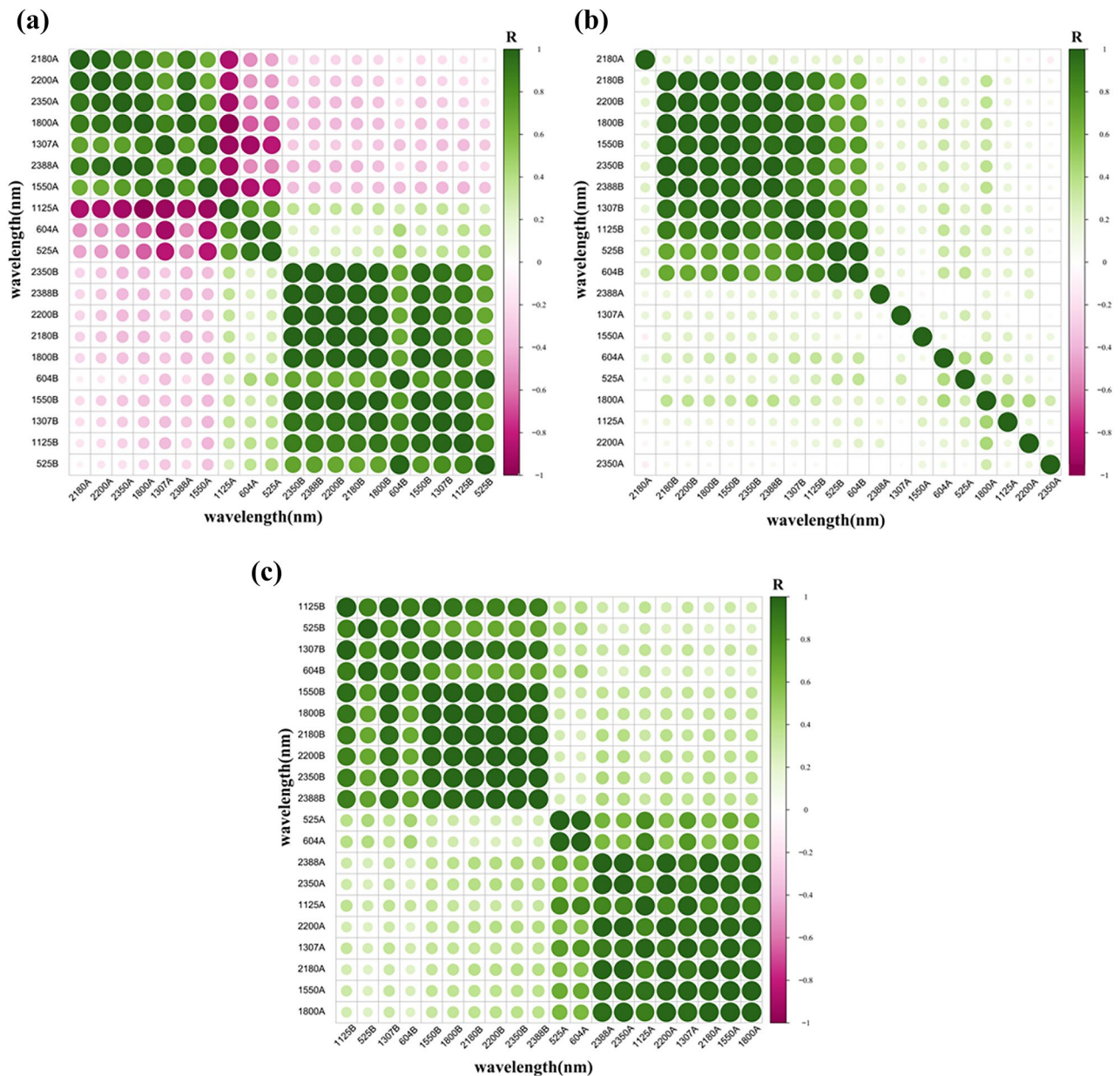


Fig. 6. (a) Correlation analysis of first-order differentials with raw spectra; (b) Correlation analysis of second-order differentials with raw spectra; (c) Correlation analysis of SG smoothing with raw spectra.

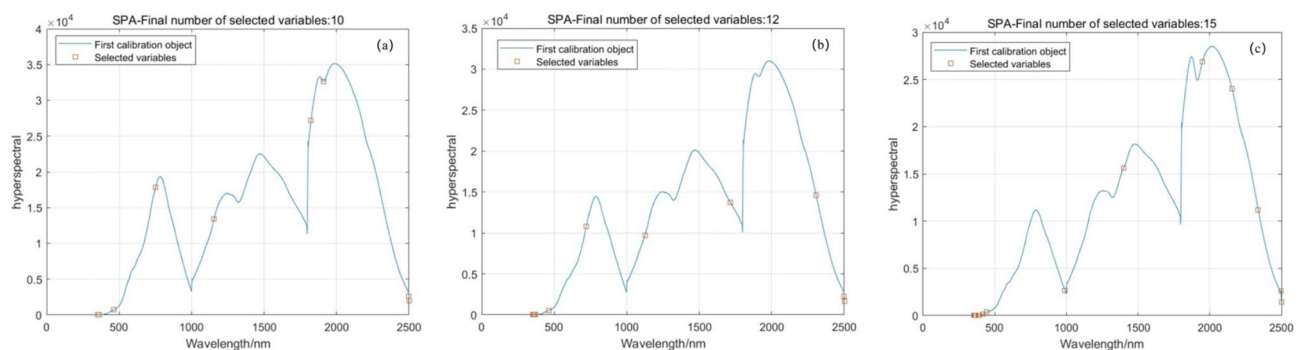


Fig. 7. (a) Characteristic bands of Zn (b) characteristic bands of Cu (c) characteristic bands of Cd.

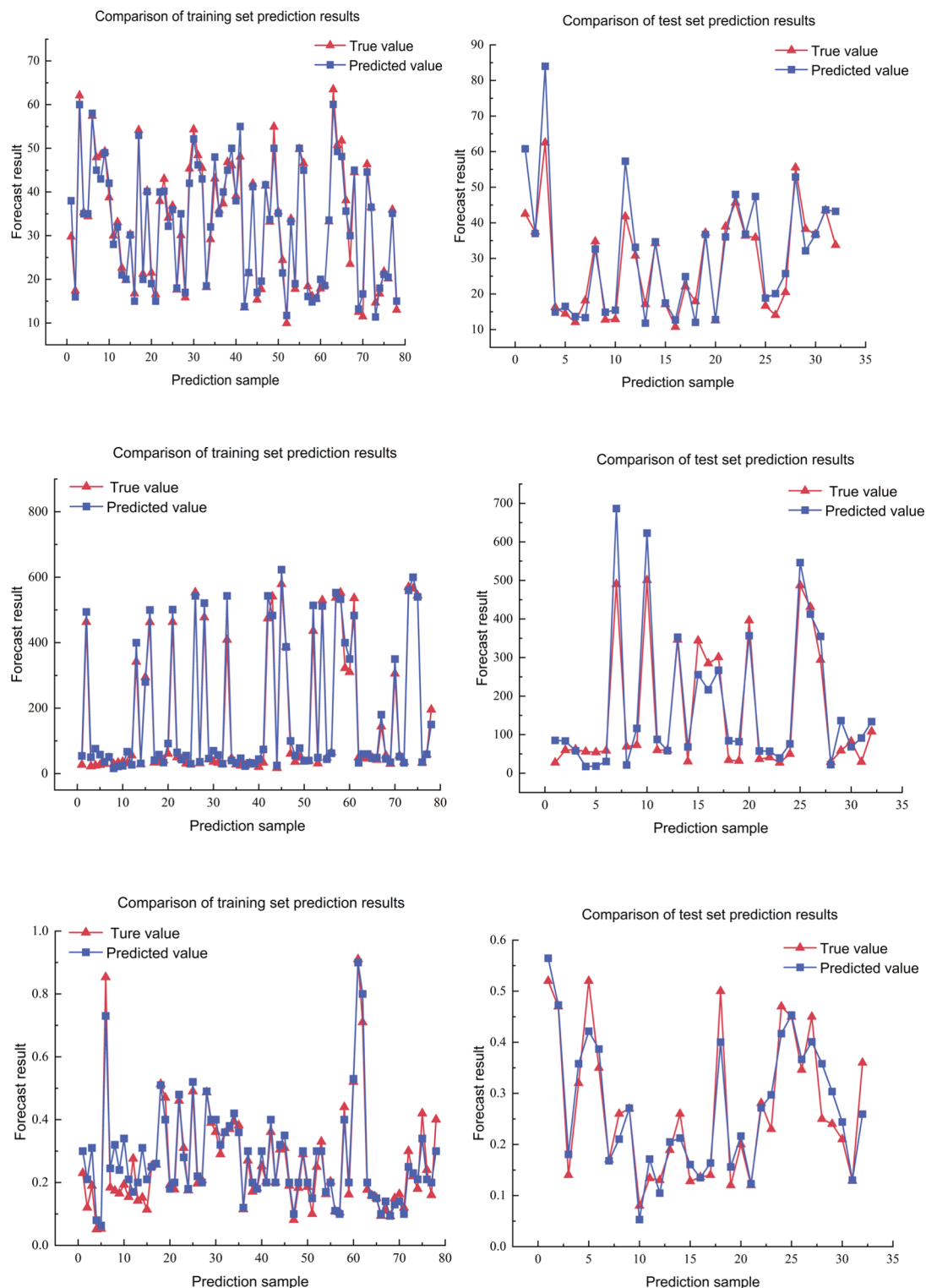


Fig. 8. Plot of the inversion results for Cu, Zn, and Cd.

priorities, combine the use of nutrients, strengthen the management mechanism and science and technology, focus on the development of low-carbon sustainable agriculture, promote carbon sequestration and sinks in the northeastern black soil, and realize the win-win goal of carbon neutrality and food security. Therefore, the use of hyperspectral technology to estimate the content of heavy metals in soil is crucial for the protection of black soil farming, without the need for large-scale destructive collection of soil samples, which greatly reduces the impact of chemical substances on the environment. With low-carbon development as the goal-oriented, coordinated

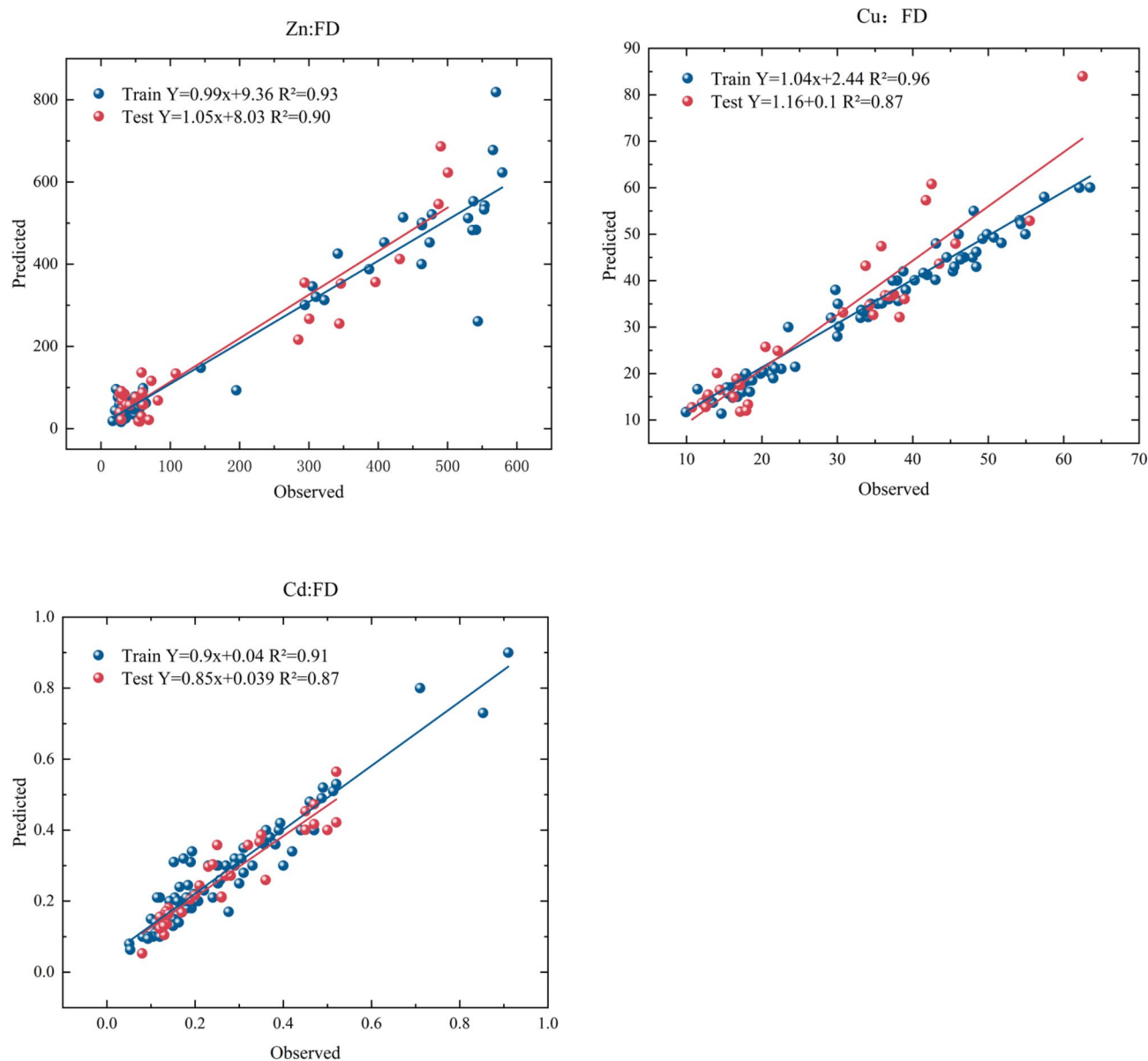


Fig. 9. Scatterplot showing the measured and predicted heavy metal content.

	RF			PLS			SVM		
		Model	Test		Model	Test		Model	Test
Cu	R ²	0.96	0.87	R ²	0.25	0.36	R ²	0.07	0.01
	RMSE	6.08	6.88	RMSE	9.71	9.13	RMSE	13.78	15.63
Zn	R ²	0.93	0.90	R ²	0.06	-1.22	R ²	0.04	0.09
	RMSE	53.44	57.92	RMSE	270.84	424.91	RMSE	194.78	261.09
Cd	R ²	0.91	0.87	R ²	-0.29	0.017	R ²	0.71	0.003
	RMSE	0.07	0.05	RMSE	29.71	35.40	RMSE	0.09	0.14

Table 2. Model validation and accuracy statistics for all samples.

delineation of permanent basic farmland, strictly abide by the red line of arable land protection to reduce blind reclamation and overuse of arable land, so as to reduce the increase in the source of carbon.

Currently, studies on the inversion of HMs rely on the intrinsic correlations between HM adsorbents and different HMs. The indirect inversion approach for HM content in soils is based on the spectral properties of HM adsorbents, which are mainly related to various absorption features in soils, including iron oxides, clay minerals,

water, and organic substances^{60–62}. These absorption features are mainly reflected in the visible/ NIR spectral region of 350–2500 nm, and their formation mechanism originates from the transfer of vibrational energy of major molecular bonds. Specifically, in soils, most of the iron oxide molecules in soils exhibit absorption characteristics in the visible (350–780 nm) and short-wave NIR (780–1100 nm) regions of the spectrum⁶³. In addition, the absorption properties of clay minerals in the long wave NIR (1100–2500 nm) region are primarily because of the overtones and combined vibrations of OH, H₂O and CO₃. Water shows strong absorption properties in the visible/NIR region, with peaks occurring near 1400 and 1900 nm and relatively weak absorption properties in other frequency bands⁶⁴.

The application of hyperspectral technology can effectively help obtain spectral information of chemical indicators and conduct real-time monitoring; however, factors such as external interference and instrument noise may affect the spectral quality^{65,66}. These interfering factors may present a prominent “burr jump” phenomenon in the raw soil spectral data, and the preprocessing of the primary spectrum are indispensable to improve the validity, sensitivity, and resolution of the spectral information and accurately estimate the follow-up analytical models^{67–69}. In this study, the “step jump” and whiskers phenomenon can be eliminated completely by applying Savitzky-Golay (SG) smoothing (Fig. 5) to the spectral data, which is a method that processes high-order polynomials by weighted filtering and least squares fitting techniques. Filtering and smoothing are intended to improve the processing effect by effectively retaining the variation information of the signal, eliminating the effect of noise, and maintaining the feature of the raw data. In addition, the nature of spectroscopic information can be significantly improved by using the differential processing method in the processing relative to the multiple scattering correction processing method.

Applicable soil spectral processing can improve the effectiveness of hyperspectral simulation accuracy. Using suitable pre-processing methods, the feature spectral wave bands of the soil can be highlighted more effectively, and the reflected wave peaks and absorbed troughs in the spectroscopic curves can be enhanced. In this study, various soil HM hyperspectral remote sensing quantitative models were established for nondestructive soil HM concentration monitoring and HM contamination assessment. Different mathematical correction methods were used in the study to analyze the connection between the spectroscopic data and the HM content measured⁷⁰, and effective bands were identified to extract HM using SPA. The results of the research revealed that the developed model using differential preprocessing could accurately predict Zn and Cu concentrations (Fig. 9). Multiple linear regression of the spectral reflectivity of SG-FD processed soils was able to accurately forecast the Cd content in the research area (Fig. 9). This contrasts significantly with the use of different pre-processing techniques to improve the spectral characteristics that increased the prediction accuracy rate of the retrieval model.

The results of this research advance the applications of hyperspectral remote sensing technology in the area of integrated surveillance and evaluation of ecosystems and furnish foundational support for the management and early warning of soil HM contamination. This research further provides technological support for the nondestructive monitoring and pollution assessment of soil HMs and promotes the application of hyperspectral remote sensing in the field of ecological environmental monitoring and assessment. This study is conducive to the effective implementation of conservation tillage technology to protect the resource utilization of black soil from the perspective of reducing damage and chemical pollution. For the conservation and utilization of black soil, the integration and research and development of advanced technology should be strengthened to improve the core of conservation tillage technology. However, the accuracy of the established quantitative hyperspectral remote sensing model for soil HMs needs to be further validated because it is limited by the sampling location setup, sample collection and measurements, and the number of samples, which may have an impact on accuracy.

Conclusions

The average value of Zn exceeded the mean background values for Jilin Province and national standards, suggesting contamination. The CF of Cu ranged from 1 to 2, indicating slight pollution, and the CFs for Zn and Cd were >2, indicating moderate pollution. Spectral conversion was achieved using FD, SD, mean centering, normalization and MSC and the relationship between soil HM content and soil spectra was further validated. Based on the correlation coefficients, the characteristic spectral bands of the three HMs differed in location and number. These bands effectively reflect the spectral properties of organics, iron, clay minerals and manganese oxides in the soil. The associations between the three HMs and the spectral variables were modeled and analyzed using PLS regression, SVM, and RF models. The results indicate that among the three models, RF had the highest accuracy in HM inversion, with R_{RF}^2 values >0.8, demonstrating high stability and the ability to effectively invert the HM content.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Received: 1 July 2024; Accepted: 16 December 2024

Published online: 02 January 2025

References

- Guerra Sierra, B. E., Muñoz Guerrero, J. & Sokolski, S. Phytoremediation of heavy metals in tropical soils an overview. *Sustainability* **13**(5), 2574 (2021).
- Jia, Z., Li, S. & Wang, L. 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 (2018).
- Liu, H., Zhang, Y. & Zhang, B. Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China. *Environ. Monit. Assess.* **154**, 147–154 (2009).

4. Ma, S., Wang, L. J., Wang, H. Y., Zhao, Y. G., & Jiang, J. Impacts of land use/land cover and soil property changes on soil erosion in the black soil region, China. *Journal of Environmental Management*, 328, 117024 (2023).
5. Miao, X., Li, Z., Wang, M., Mei, J., & Chen, J. Measurement of cultivated land ecosystem resilience in black soil region of Northeast China under the background of cultivated land protection policy in China: Case study of Qiqihar City. *Journal of Cleaner Production*, 434, 140141 (2024).
6. Xiao, D., Yang, X., Bai, H., Tang, J., & Tao, F. Trends and climate response in the yield of staple crops across Northeast China. *Frontiers in Sustainable Food Systems*, 7, 1246347 (2024).
7. Jiang, F., Peng, X., Yao, S., Islam, M. U., Zhang, Z., Chen, B., & Meng, F. A six-site field study on assessing the suitability of conservation and conventional tillage in the black soil region, Northeast China. *Soil and Tillage Research*, 248, 106379 (2025).
8. Chen, X. et al. Characterization and risk assessment of heavy metals in road dust from a developing city with good air quality and from Shanghai, China. *Environ. Sci. Pollut. Res.* **26**, 11387–11398 (2019).
9. Hong, N. et al. Quantitative source tracking of heavy metals contained in urban road deposited sediments. *J. Hazard. Mater.* **393**, 122362 (2020).
10. Mirzaei Aminiyan, M. et al. The ecological risk, source identification, and pollution assessment of heavy metals in road dust: a case study in Rafsanjan, SE Iran. *Environ. Sci. Pollut. Res.* **25**, 13382–13395 (2018).
11. Goudarzi, G. et al. Health risk assessment on human exposed to heavy metals in the ambient air PM 10 in Ahvaz, Southwest Iran. *Int. J. Biometeorol.* **62**, 1075–1083 (2018).
12. Lisiak-Zielińska, M. et al. How polluted are cities in central Europe?—Heavy metal contamination in *Taraxacum officinale* and soils collected from different land use areas of three representative cities. *Chemosphere* **266**, 129113 (2021).
13. Zhang, H. et al. Multiple exposure pathways and health risk assessment of heavy metal (loid)s for children living in fourth-tier cities in Hubei Province. *Environ. Int.* **129**, 517–524 (2019).
14. Jabbo, J. N., Isa, N. M., Aris, A. Z., Ramli, M. F. & Abubakar, M. B. Geochemometric approach to groundwater quality and health risk assessment of heavy metals of Yankari Game Reserve and its environs, Northeast Nigeria. *J. Clean. Prod.* **330**, 129916 (2022).
15. Xu, D. M., Fu, R. B., Liu, H. Q. & Guo, X. P. Current knowledge from heavy metal pollution in Chinese smelter contaminated soils, health risk implications and associated remediation progress in recent decades: a critical review. *J. Clean. Prod.* **286**, 124989 (2021).
16. Fei, X., Lou, Z., Xiao, R., Ren, Z. & Lv, X. Source analysis and source-oriented risk assessment of heavy metal pollution in agricultural soils of different cultivated land qualities. *J. Clean. Prod.* **341**, 130942 (2022).
17. Jiang, H. H. et al. An integrated approach to quantifying ecological and human health risks from different sources of soil heavy metals. *Sci. Total Environ.* **701**, 134466 (2020).
18. Gholizadeh, A. & Kopačková, V. Detecting vegetation stress as a soil contamination proxy: a review of optical proximal and remote sensing techniques. *Int. J. Environ. Sci. Technol.* **16**, 2511–2524 (2019).
19. Ding, S., Zhang, X., Sun, W., Shang, K. & Wang, Y. Estimation of soil lead content based on GF-5 hyperspectral images, considering the influence of soil environmental factors. *J. Soils Sedim.* **22**, 1431–1445 (2022).
20. Liu, Z. et al. Estimation of soil heavy metal content using hyperspectral data. *Remote Sens.* **11**, 1464 (2019).
21. Padian, J., Minasny, B. & McBratney, A. B. Using deep learning to predict soil properties from regional spectral data. *Geoderma Reg.* **16**, e00198 (2019b).
22. Singh, S. & Kasana, S. S. Estimation of soil properties from the EU spectral library using long short-term memory networks. *Geoderma Reg.* **18**, e00233 (2019).
23. Padian, J., Minasny, B. & McBratney, A. B. Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma* **340**, 279–288 (2019a).
24. Tsakiridis, N. L., Keramaris, K. D., Theocharis, J. B. & Zalidis, G. C. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* **367**, 114208 (2020).
25. Wang, J. et al. Assessing toxic metal chromium in the soil in coal mining areas via proximal sensing: prerequisites for land rehabilitation and sustainable development. *Geoderma* **405**, 115399 (2022).
26. Kästner, F. et al. Estimating heavy metal concentrations in Technosols with reflectance spectroscopy. *Geoderma* **406**, 115512 (2022).
27. Shen, L. et al. Hyperspectral estimation of soil organic matter content using different spectral preprocessing techniques and PLSR method. *Remote Sens.* **12**(7), 1206 (2020).
28. Shi, T. et al. Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. *Plant. Soil.* **366**, 363–375 (2013).
29. Zhang, Q., Zhang, H., Liu, W. & Zhao, S. Inversion of heavy metals content with hyperspectral reflectance in soil of well-facilitated capital farmland construction areas. *Trans. Chin. Soc. Agric. Eng.* **33**, 230–239 (2017).
30. Arif, M., Qi, Y., Dong, Z. & Wei, H. Rapid retrieval of cadmium and lead content from urban greenbelt zones using hyperspectral characteristic bands. *J. Clean. Prod.* **374**, 133922 (2022).
31. Tan, K. et al. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. *J. Hazard. Mater.* **382**, 120987 (2020).
32. Khosravi, V., Ardejani, F. D., Yousefi, S. & Aryafar, A. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. *Geoderma* **318**, 29–41 (2018).
33. Pyo, J., Hong, S. M., Kwon, Y. S., Kim, M. S. & Cho, K. H. Estimation of heavy metals using deep neural network with visible and infrared spectroscopy of soil. *Sci. Total Environ.* **741**, 140162 (2020).
34. Li, Y., Yang, K., Gao, W., Han, Q. & Zhang, J. A spectral characteristic analysis method for distinguishing heavy metal pollution in crops: VMD-PCA-SVM. *Spectrochim. Acta - A: Mol. Biomol. Spectrosc.* **255**, 119649 (2021).
35. Yaseen, Z. M. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: review, challenges and solutions. *Chemosphere* **277**, 130126 (2021).
36. Zhou, W. et al. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. *Catena* **202**, 105222 (2021).
37. Zhang, Z. H., Guo, F., Xu, Z., Yang, X. Y. & Wu, K. Z. On retrieving the chromium and zinc concentrations in the arable soil by the hyperspectral reflectance based on the deep forest. *Ecol. Ind.* **144**, 109440 (2022).
38. Chen, T. et al. A flexible and efficient machine learning library for heterogeneous distributed systems. *Arxiv*. **1512**, 01274 (2015).
39. Sun, Y. et al. Coupled retrieval of heavy metal nickel concentration in agricultural soil from spaceborne hyperspectral imagery. *J. Hazard. Mater.* **446**, 130722 (2023).
40. Shi, T., Chen, Y., Liu, Y. & Wu, G. Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. *J. Hazard. Mater.* **265**, 166–176 (2014).
41. Tanaka, S. et al. Spectral index for quantifying leaf area index of winter wheat by field hyperspectral measurements: a case study in Gifu Prefecture, Central Japan. *Remote Sens.* **7**, 5329–5346 (2015).
42. Hassan-Esfahani, L., Torres-Rua, A., Jensen, A., & McKee, M. Assessment of surface soil moisture using high-resolution multi-spectral imagery and artificial neural networks. *Remote Sensing*, **7**(3), 2627–2646 (2015).
43. Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, **27**, 111–125.
44. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
45. Gao, Y. et al. Feasibility of estimating heavy metal concentrations in wetland soil using hyperspectral technology. In: *2017 IEEE International Geoscience and Remote Sensing Symposium* 4886–4889 (2017).

46. Tan, K., Ma, W., Wu, F. & Du, Q. Random forest-based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data. *Environ. Monit. Assess.* **191**, 1–14 (2019).
47. Wang, J., Liu, X., Huang, F., Tang, J. & Zhao, L. Salinity forecasting of saline soil based on ANN and hyperspectral remote sensing. *Trans. Chin. Soc. Agric. Eng.* **25**, 161–166 (2009).
48. Guo, Y. K., Liu, N., Liu, L., Li, D. & Zhu, S. Hyper-spectral inversion of soil cu content based on BP neural network model. *Sci. Surv. Mapp.* **43**, 135–139 (2018).
49. Tan, K., Zhang, Q. Q., Cao, Q. & Du, P. J. Hyperspectral retrieval model of soil organic matter content based on particle swarm optimization-support vector machines. *Earth Sci.* **40**, 1339–1345 (2015).
50. Leone, A. P., Viscarra-Rossel, A., Amenta, R., Buondonno, A. & P. and Prediction of soil properties with PLSR and vis-NIR spectroscopy: application to mediterranean soils from Southern Italy. *Curr. Anal. Chem.* **8**, 283–299 (2012).
51. Du, Y. G. et al. Spatial characteristics of soil organic matter and as content in source regions of Yangtze River and Yellow River. *Chin. J. Grassl.* **34**, 24 (2012).
52. Melgani, F. & Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **42**, 1778–1790 (2004).
53. Zhang, H. et al. Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. *J. Clean. Prod.* **273**, 122858 (2020).
54. Gasmi, A., Gomez, C., Chehbouni, A., Dhiba, D. & Gharous, E. Using PRISMA hyperspectral satellite imagery and GIS approaches for soil fertility mapping (FertiMap) in northern Morocco. *Remote Sens.* **14**(16), 4080 (2022).
55. Gan, F. P., Wang, R. S. & Ma, A. N. Spectral identification tree (sit) for mineral extraction based on spectral characteristics of minerals. *Earth Sci. Front.* **10**, 445–454 (2003).
56. Wang, J. N., Zheng, L. F. & Tong, Q. X. The spectral absorption identification model and mineral mapping by imaging spectrometer data. *Remote Sens. Environ.* **11**, 20–31 (1996).
57. Trifi, M. et al. Machine learning-based prediction of toxic metals concentration in an acid mine drainage environment, northern Tunisia. *Environ. Sci. Pollut. Res.* **29**(58), 87490–87508 (2022).
58. Garcia-Sánchez, A., Alastuey, A. & Querol, X. Heavy metal adsorption by different minerals: application to the remediation of polluted soils. *Sci. Total Environ.* **242**, 179–188 (1999).
59. Galvão, L. S., Pizarro, M. A. & Epiphany, J. C. N. Variations in reflectance of tropical soils: spectral-chemical composition relationships from AVIRIS data. *Remote Sens. Environ.* **75**, 245–255 (2001).
60. Junliang, H., Shuyuan, Z., Yong, Z. & Jianjun, J. Review of retrieving soil heavy metal content by hyperspectral remote sensing. *Remote Sens. Technol. Appl.* **30**, 407–412 (2015).
61. Kemper, T. & Sommer, S. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. *Environ. Sci. Technol.* **36**, 2742–2747 (2002).
62. Xiao, J. Y. et al. Review on methods of monitoring soil heavy metal based on hyperspectral remote sensing data. *Hubei Agric. Sci.* **52**, 1248–1253 (2013).
63. Torrent, J. & Barrón, V. Diffuse reflectance spectroscopy of iron oxides. *Encycl. Surf. Colloid Sci.* **1**, 1438–1446 (2002).
64. He, T., Wang, J., Lin, Z. & Cheng, Y. Spectral features of soil organic matter. *Geo-Spat. Inf. Sci.* **12**, 33–40 (2009).
65. Ou, D. et al. Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction. *Geoderma* **385**, 114875 (2021).
66. Yang, H., Xu, H. & Zhong, X. Prediction of soil heavy metal concentrations in copper tailings area using hyperspectral reflectance. *Environ. Earth Sci.* **81**, 183 (2022).
67. Ding, S., Keal, C. A., Zhao, L. & Yu, D. Dimensionality reduction and classification for hyperspectral image based on robust supervised ISOMAP. *J. Ind. Prod. Eng.* **39**, 19–29 (2022).
68. Guo, F. et al. Estimating chromium concentration in arable soil based on the optimal principal components by hyperspectral data. *Ecol. Indic.* **133**, 108400 (2021).
69. Liu, W., Yu, Q., Niu, T., Yang, L. & Liu, H. Inversion of soil heavy metal content based on spectral characteristics of peach trees. *Forests* **12**, 1208 (2021).
70. Yang, X., Lei, S., Zhao, Y. & Cheng, W. Use of hyperspectral imagery to detect affected vegetation and heavy metal polluted areas: a coal mining area, China. *Geocarto Int.* **37**, 2893–2912 (2022).

Acknowledgements

This study was supported by the Jilin Science and Technology Development Planning Project [20210203006SF, 20210203150SF], the Sixth Batch of Young Science and Technology Talents Promotion Program of Jilin Province [QT202216]. We would like to thank Editage (www.editage.cn) for English language editing.

Author contributions

Yanan Chen: Methodology, writing-editing, and funding acquisition. Wanying Shi: writing-original draft, formal analysis, and visualization. Guzailinuer Aihemaitijiang: Investigation, Data curation. Feng Zhang: Conceptualization and funding acquisition. Jiquan Zhang: Conceptualization, and methodology. Yichen Zhang: Conceptualization and supervision. Dianqi Pan: Supervision and resources. Jinying Li: Software and validation.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-83479-0>.

Correspondence and requests for materials should be addressed to Y.C. or J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024