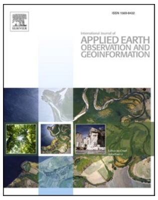
Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



Retrieving heavy metal concentrations in urban soil using satellite hyperspectral imagery

Nannan Yang [a], Liangzhi Li [a,b,c,*], Ling Han [a,b,c,*], Kyle Gao [d], Songjie Qu [a], Jonathan Li [d,e]

[a] School of Land Engineering, Chang'an University, Xi'an, SX 710064, China
[b] Xi'an Key Laboratory of Territorial Spatial Information, School of Land Engineering, Chang'an University, Xi'an, SX 710064, China
[c] Shaanxi Key Laboratory of Land Reclamation Engineering, Chang'an University, Xi'an, SX 710064, China
[d] Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[e] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Soil heavy metals
Hyperspectral imagery
2D-SSI
Spectral calibration
Inversion model

ABSTRACT

Efficient prediction and precise depiction of heavy metal concentrations in urban soil are essential for mitigating non-point source pollution and safeguarding public health. Therefore, this research investigated the estimation of soil heavy metal concentrations derived from Gaofen-5 (GF-5) hyperspectral images calibrated by the direct standardization (DS) algorithm. The inversion strategy for soil heavy metal concentrations in response to the two-dimensional soil spectral index (*2D-SSI*) was proposed by coupling Pearson correlation coefficient (*r*) and competitive adaptive reweighting algorithm (CARS) for feature selection. The results indicated that the optimal models based on *2D-SSI* outperform the models based on calibrated, filtered original spectral bands. For Pb, Cu, Cd, and Hg, the optimal model determination coefficients for the validation data set (R_V^2) were 0.871 (SVM), 0.883 (BPNN), 0.834 (PLSR), and 0.907 (PLSR), respectively. The spectral features were highlighted in the two-dimensional feature space, and the predicted distribution of heavy metal concentrations was aligned with the observed ground measurements. This study revealed that the prediction strategy based on DS-corrected GF-5 AHSI images with constructed *2D-SSI* features can serve as a reliable technical approach for soil heavy metal prediction and pollution prevention.

1. Introduction

Soil contamination continues to worsen in light of anthropogenic instances like speedy industrialization and urbanization (Tao et al., 2019; Qin et al., 2021; Nyarko et al., 2022). Thus, detecting and visualizing of heavy metal contamination in urban soil can provide strategic guidance for the rational development and use of urban land. However, conventional monitoring techniques that rely on laboratory chemical analysis and discrete sampling approaches have difficulty in accurately obtaining continuous distributions with a limited number of samples, which are expensive, labor-intensive, and prone to secondary pollution (Chen et al., 2015). Consequently, credible and green approaches are desperately needed to identify potential contaminated areas and develop remediation measures.

Hyperspectral remote sensing, leveraging its spectral continuity, broad spectral range, imaging characteristics, and non-invasive advantages, finds extensive application in numerous domains such as environment monitoring, geology, and soil science (Bonifazi et al., 2018; Cheng et al., 2019). However, achieving high-accuracy metal retrieval remains challenging due to inevitable influences from natural and anthropogenic factors, such as spatial scale, soil physicochemical properties, environmental details, and timeliness (Wang et al., 2018). It is of paramount significance to comprehensively explore data from hyperspectral imagery to obtain trace responses and then develop new methodologies, especially for estimating with limited samples. Extensive research suggests that high-precision estimation can be achieved using laboratory spectra under well-controlled experimental conditions (Tan et al., 2018). Therefore, employing laboratory-measured spectra to adjust image spectra and thereby mitigating the impact of interference components is an essential approach to enhancing spectral quality. Environmental factor removal algorithms exhibit promise in hyperspectral inversion with special respect to soil attributes. The DS algorithm has been applied to calibrate field-acquired spectra from specific sampling sites with laboratory- measured spectra in early study (Ji et al.,

* Corresponding authors.
E-mail addresses: liliangzhi@chd.edu.cn (L. Li), hanling@chd.edu.cn (L. Han).

https://doi.org/10.1016/j.jag.2024.104079
Received 31 March 2024; Received in revised form 15 July 2024; Accepted 2 August 2024
Available online 9 August 2024

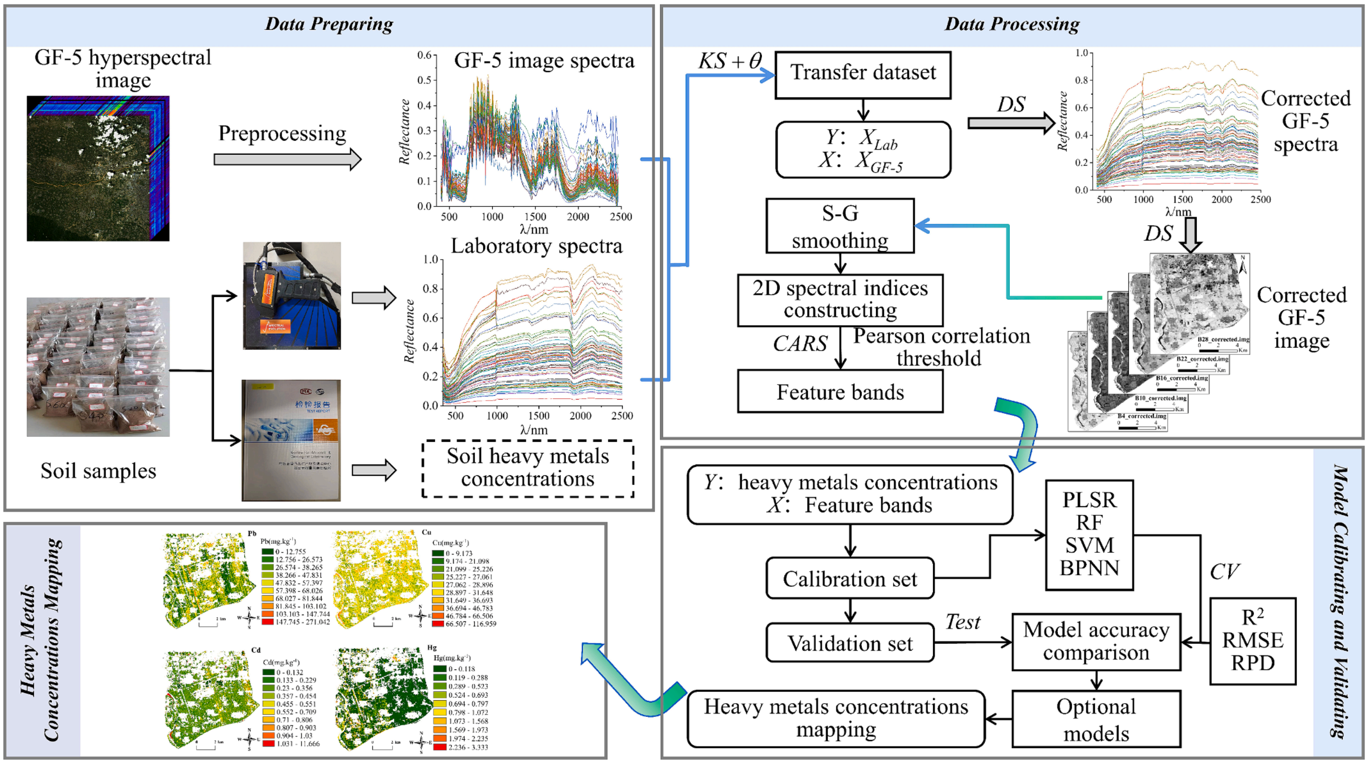


Fig. 1. Flowchart of our study.

2015a). Researchers have only recently begun to explore the application of the DS algorithm for the calibration of continuous hyperspectral images (Zhang et al, 2022). However, image calibration for different land use types in soil heavy metal inversion has rarely been touched, and the potential and capabilities of the DS algorithm in correcting continuous hyperspectral images acquired from satellites have yet to be evaluated (Zou et al., 2020). Meanwhile, model prediction accuracy can be improved by spectral mathematical transformation and by simultaneously incorporating significant bands from different transformed spectra into inversion, as corroborated by numerous scholars (Dai et al., 2022; Meng et al., 2020; Yang et al., 2023). Nonetheless, characteristics supplied by only a few sensitive bands might be inadequate since the trace nature of concentrations, which limits the model's accuracy to a certain extent (Husnizar et al., 2018; Wang et al., 2018). Hence, unlike prior research, where *2D-SSI* was commonly applied to multi-spectral images with fewer bands, such as those from Landsat and Sentinel. In this study, the *2D-SSI* was constructed with the calibrated hyperspectral image spectral bands, which is an approach that has been relatively less explored to achieve feature expansion and fully utilize spectral information. What's more, *PI* constructed based on the soil line analysis is rarely studied, and the previous *2D-SSI* can not indicate the spectral quality, while *PI* has the ability to discuss the reliability of the spectral.

Correlation analysis is widely preferred by scholars for feature selection due to its statistical foundation and interpretability (Wilford et al., 2016; Bolón-Canedo et al., 2015). Additionally, with the innovation and advancement of machine learning algorithms, the CARS algorithm (Li et al., 2019) has been well received for feature selection (Jiang et al., 2018; Wang and Wang, 2022). Selecting the optimal combination of hyperspectral characteristics introduces challenges since the complex interplay of spectral and spatial data dimensions. The combination of the Pearson correlation and the CARS algorithm ensures reliable input selection and significantly reduces computational complexity compared to conventional single-feature selection algorithms. For model inversion, conventional statistical regression techniques often fail to capture the nonlinear linkage between soil spectral variables and concentration dependent variable, resulting in low-precision predictions (Guo et al., 2021). Partial least-squares regression (PLSR) is efficient in handling high-dimensional data, as well as providing a precise expression for the inversion model and revealing the significance of each feature (Sun and Zhang, 2017). With the introduction of machine learning algorithms, multivariable models, such as back propagation neural networks (BPNN), support vector machines (SVM), random forests (RF), have exhibited the capability to overcome the linear limitations for soil attribute estimation (Odebiri et al., 2021; Wang et al., 2020; Xavier and Yoshua, 2010). Moreover, the diversity and high dimensionality of remote sensing data pose additional challenges to traditional machine learning algorithms.

The study endeavors to delve into concentration prediction by accurately calibrating hyperspectral imagery and *2D-SSI* construction. Specifically, our objectives are as follows: (1) Utilizing the DS algorithm to calibrate the GF-5 Advanced Hyperspectral Imagery (AHSI) with laboratory spectra, from which the representative spectra were chosen by the Kennard Stone (KS) algorithm. (2) Constructing *2D-SSI* to thoroughly explore spectral information, enhancing and highlighting the spectral signature. (3) Extracting features via coupling the Pearson correlation coefficient with the CARS algorithm. (4) Mapping the spread of soil heavy metal concentrations with ultimate model, which was identified through contrasting of the inversion accuracy of PLSR, RF, SVM, and BPNN.

2. Datasets and methods

Fig. 1 illustrates the methodology based on *2D-SSI* generated from the DS-corrected GF-5 imagery for urban soil heavy metal content estimation. The overall research methodology comprises five main steps: (1) Spectral data collection and pre-processing, including image spectra and laboratory spectra. (2) DS calibration of the GF-5 hyperspectral imagery with laboratory spectra. (3) Construction of *2D-SSI* informed by the DS-corrected GF-5 images and feature extraction by coupling the Pearson correlation coefficient with the CARS algorithm. (4) Establishment and

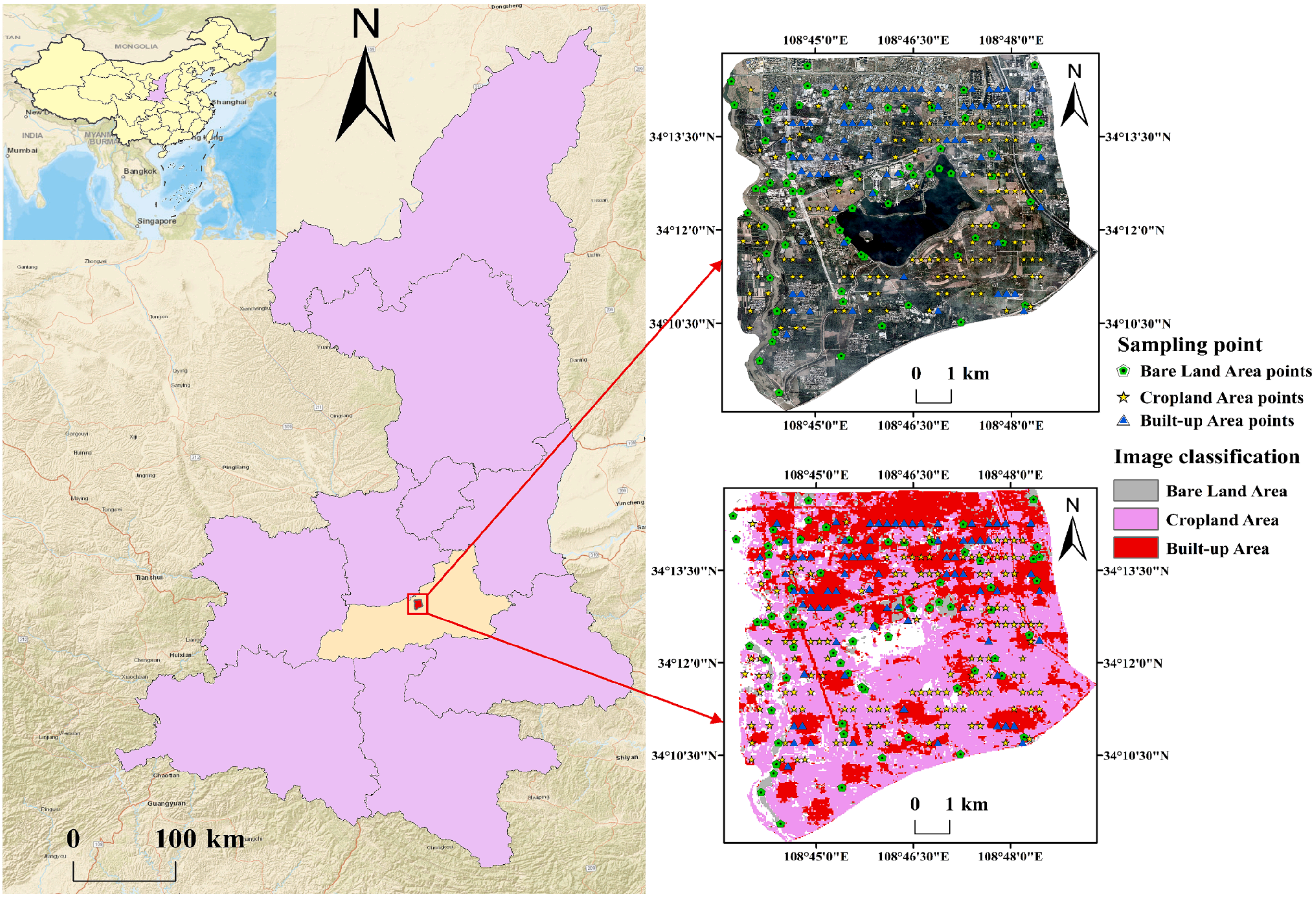


Fig. 2. Study area overview and distribution of sampling sites.

evaluation of PLSR, RF, SVM, and BPNN inversion models. (5) Mapping the concentrations spatially in accordance with the optimal model picked from (4).

2.1. Study area

The investigation area depicted in Fig. 2 is situated in Fengdong New District, which is subordinate to the Xixian development zone in Xi`an, Shaanxi Province, China (34°12′21″N, 108°46′10″E). The region experiences an annual average precipitation of 600 to 700 mm, with a yearly average temperature of 13.4 °C. The region features a flat topography with a mean elevation of 388 m. The predominant soil types are collapsible loess and Lou soil, which are conducive to the growth of various crops. The land use pattern is complex, characterized by cropland, built-up areas, and bare/sparsely vegetated land as the predominant land types. As a transitional area connecting urban and rural areas, it has a high population density and is marked by well-developed industrial and transportation infrastructure. However, the growth of industries has contaminated and degraded the soil, and the accumulation of heavy metals has substantially exceeded regulatory thresholds due to insufficient environmental protection awareness. The region exhibits typical characteristics commonly observed in other developing urban fringe areas.

2.2. Data acquisition

2.2.1. Sample data gathering

The 800 × 800 m equidistant grid method was employed for systematic sampling, and a cumulative count of 500 soil samples was gathered between April and June 2020 using the five-point sampling approach. Soil specimens were amassed at an approximate depth of 10–20 cm and filtered through a nylon sieve measuring 0.15 mm. Prepared soil samples were delivered to a specialized testing institution for the concentration detection of Pb, Cu, Cd, and Hg. GPS coordinates were obtained at same time as sampling, and the surrounding environment was documented. The spectra were recorded by the spectrometer SR-2500. A 100 W halogen lamp was fixed and maintained at a 30° angle to the vertical direction, serving as the light source for the soil spectrum measurements. Subsequently, soil spectral readings were performed with the probe positioned 15 cm above the leveled soil surface vertically. Reference whiteboard calibration was performed initially and repeated after every three soil samples were measured to mitigate systematic errors arising from environmental changes. Eight spectral profiles were acquired for every sample, and the effective spectrum was derived by averaging the data with erroneous readings removed.

2.2.2. GF-5 AHSI imagery acquisition and processing

The GF-5 AHSI imagery adopted by research was acquired on April 9, 2020, with center coordinates of 108.590° E and 34.269° N, and was acquired via the platform of Land Satellite Observation Data Service (China Centre for Resources Satellite Data and Application, 2024). The image spatial resolution, swath width, and spectral range of the GF-5 AHSI imagery are 30 m, 60 km, and 390–2513 nm, respectively, with a total of 330 spectral bands. The solar altitude angle for the image acquisition is 23.103°. Comprehensive specifications of the AHSI sensor were documented in the literature of Zhang et al. (2022). Pre-processing of the GF-5 images involved the removal of bands significantly affected by water vapor, bands with poor imaging quality, and overlapping bands in the shortwave infrared, near-infrared, and visible ranges. Additionally, some bands with stripe noise and bad lines were repaired. Finally, 277 bands remained for further analysis. Radiometric correction and atmospheric correction were performed to mitigate the interference from atmosphere and other environmental variables. Elevation data was incorporated into the orthorectification process to enable precise ground reflectance retrieval (Tan et al., 2021; Ye et al., 2020). Furthermore, the SVM algorithm was employed for image classification, from which areas of built-up, cropland, and bare land were extracted.

2.3. Method

The methods employed in this study primarily include DS spectral correction, construction of *2D-SSI*, feature selection, model inversion and evaluation. The details are as follows:

2.3.1. Spectroscopic calibration

Sampling points in bare land, cropland and built-up areas were extracted and classified first based on the sampling coordinates and the classified image. Subsequently, soil sample spectra of the GF-5 AHSI image and laboratory in various areas were prepared. A subset with dimensions of $m \times p$ of laboratory-measured spectra ($\boldsymbol{X}_{lab}$) and GF-5 AHSI image spectra ($\boldsymbol{X}_{GF-5}$) needs to be chosen to serve as the DS algorithm transfer matrix, where the size is equal to or greater than two-thirds of aggregate sample size, which signifies the disparity between the two sets of spectra (Eq. (1)).

$$\boldsymbol{X}_{lab} = \boldsymbol{X}_{GF-5}\boldsymbol{B} + \boldsymbol{E} \tag{1}$$

where $\boldsymbol{B}$ is a spectral transformation matrix, and $\boldsymbol{E}$ means the residual matrix. DS correction obtains the parameter B through baseline difference adjustment, spectral centering, and least squares transformation (Eq. (2)), and parameter E is then calculated based on B.

$$\boldsymbol{B} = \boldsymbol{H}^{+}_{GF-5}\boldsymbol{H}_{lab} \tag{2}$$

where $\boldsymbol{H}_{lab}$ is the centralized matrices of laboratory spectra and $\boldsymbol{H}^{+}_{GF-5}$ represents the generalized inverse matrix of the centralized matrices of image spectra. The details concerning the process can be seen from the published paper (Ji et al., 2015). Once the parameters $\boldsymbol{B}$ and $\boldsymbol{E}$ are determined, the corrected field spectra $\boldsymbol{X}^{C}_{GF-5}$ can be obtained from the high-resolution imagery set according to the scale invariance principle (Eq. (3)).

$$\boldsymbol{X}^{C}_{GF-5} = \boldsymbol{X}_{GF-5}\boldsymbol{B} + \boldsymbol{E} \tag{3}$$

However, given that the representativeness and quantity of transferred samples also affects the DS algorithm's performance, the KS algorithm was initially employed to determine the sample set. The procedure involved selecting the two samples which have the maximum Euclidean distance serving as the initial dataset, subsequently, the remaining were sequentially added to the transformation set based on their maximum distance from the existing sample set until the predetermined quantity was reached (Zou et al., 2019). Subsequently, the DS capability was assessed using the cosine similarity, which is expressed as follows, according to which the optimal sample set was determined.

$$\theta = \cos^{-1}\frac{\sum_{i=1}^{n}\boldsymbol{X}^{C}_{GF-5}\boldsymbol{X}_{lab}}{\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{X}_{lab}\right)^{2}}\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{X}^{C}_{GF-5}\right)^{2}}} \tag{4}$$

where n indicates the total amount of wavelengths.

2.3.2. Construction of two-dimensional spectral indices

The traditional vegetation index (VI) relies on regions with abundant vegetation cover, which often results in poor estimation of soil properties in areas with bare land and sparse vegetation coverage. Therefore, *2D-SSI* has been developed to identify optimal feature combinations from image spectral data (Table 1). According to the *PI* calculation formula, the slope and intercept of the soil line, as described as below, can indicate the quality of remote sensing images or spectral data (Bellinaso et al., 2021). Therefore, analyses of the soil line were conducted to reflect the quality of the image spectral data in various areas.

Table 1
2D-SSI constructed for feature expansion.

2D-SSI	Abbreviation	Equation	Reference
Difference	DI	$R_i - R_j$	(Ge et al., 2019)
Sum	SI	$R_i + R_j$	(Liu et al., 2022)
Ratio	RI	R_i/R_j	(Ge et al., 2019)
Normalized difference	NDI	$(R_i - R_j)/(R_i + R_j)$	(Ge et al., 2019)
Re-normalized difference	RNDI	$(R_i - R_j)/\sqrt{(R_i + R_j)}$	(Roujean and Breon, 1995)
Derived ratio	DRI	$\log\left(\frac{R_i}{R_j}\right)$	(Bao et al., 2021)
Derived simplified ratio	DSRI	$\log R_i/\log R_j$	(Liu et al., 2022)
Reflectivity	ARI	$\left\|R_i^2 - R_j^2\right\|/\sqrt{(R_i + R_j)}$	(Gitelson et al., 2001)
Brightness	BI	$\sqrt{R_i^2 + R_j^2}/2$	(Escadafal, 1989)
Perpendicular	PI	$(R_i - \alpha R_j - \beta)/\sqrt{(1 + \alpha^2)}$	(Ge et al., 2019)
Attention: α and β for *PI* in various regions correspond to different values.	PI_Bare Land Area	$\alpha = 1.030, \beta = 0.082$	$R^2 = 0.675$
	PI_Cropland Area	$\alpha = 0.783, \beta = 0.087$	$R^2 = 0.581$
	PI_Built-up Area	$\alpha = 0.685, \beta = 0.187$	$R^2 = 0.632$

$$NIR = \alpha \times Red + \beta \tag{5}$$

where *NIR* and *Red* symbolize the reflectance of the near-infrared and red bands, respectively, α and β are the slope and intercept of the soil line.

2.3.3. Feature selection methods

The Pearson significant correlation bands corresponding to the confidence thresholds of $P = 0.01$ and $P = 0.05$ were chosen as the initial screening feature set. Subsequently, the CARS algorithm was employed for fine selection based on significant correlation bands, which addressed the issue of combinatorial explosion in variable selection to some extent (Li et al., 2009). The CARS algorithm functions in the following specific manner:

Firstly, the Monte Carlo (MC) sampling method was employed, where, in each iteration, samples were randomly divided by the ratio of 8:2 to serve as the datasets used for modeling and calibration of the PLS model. The weights (w_i) which assigned to the absolute values of regression coefficients ($|l_i|$) were computed in each sampling process.

$$w_i = |l_i|/\sum_{i=1}^{p}|l_i| \tag{6}$$

where p represents the quantity of variables utilized in each sampling procedure.

Secondly, variables with relatively small weights were forcibly removed by the exponentially decreasing function (EDF). The variable retention rate was determined according to EDF and can be expressed as

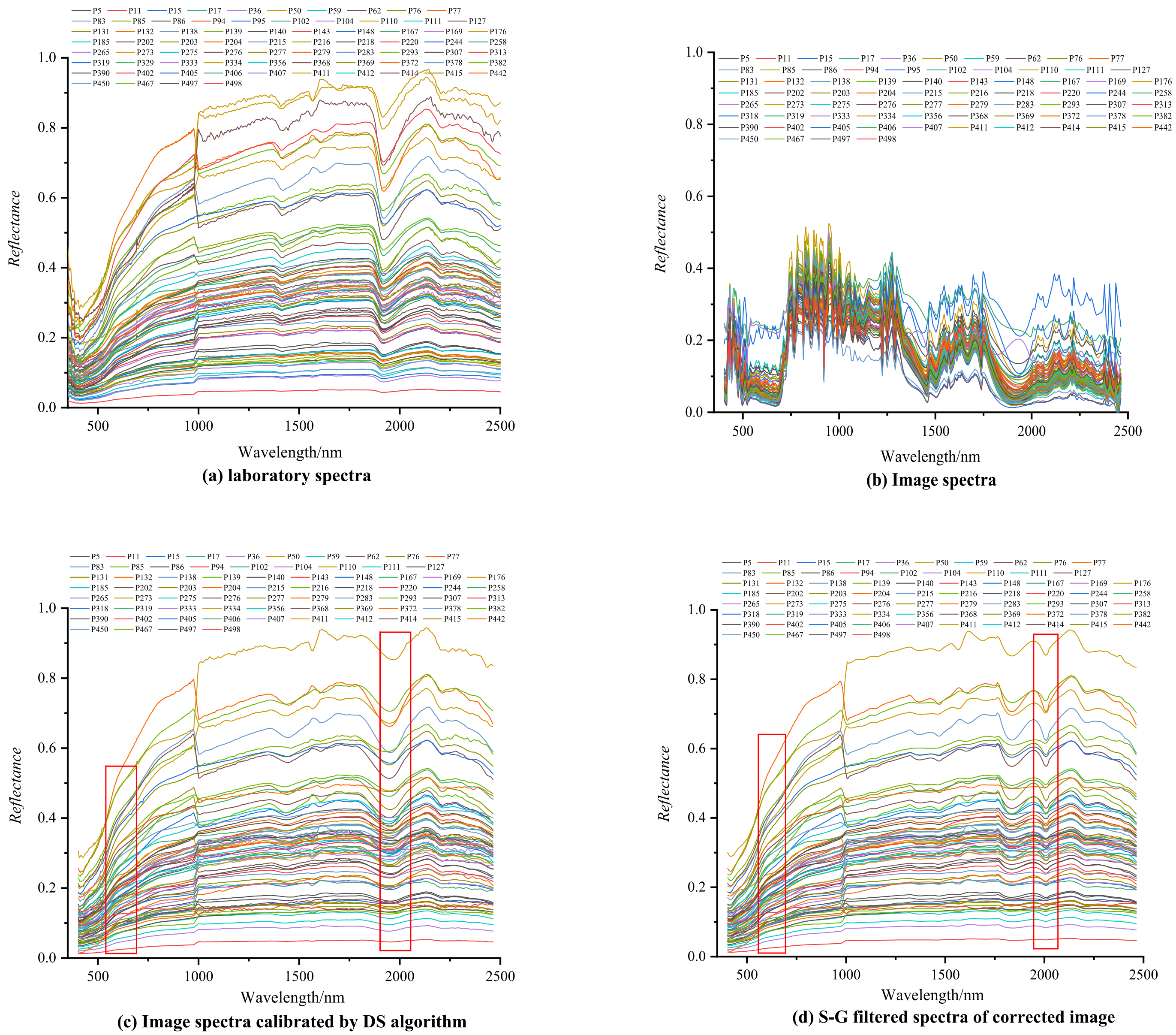
$$R_i = \mu e^{-ki} \tag{7}$$

where μ and k are constants. In the first and last iteration (N) of MC cross-validation sampling, the variable retention rate was 1 and $2/p$. The

Table 2
Descriptive analysis of soil heavy metal concentrations (unit: mg kg^{-1}).

Sample type	Element	Concentration range	Mean	Standard Deviation	Skewness	Kurtosis	Coefficient ofVariation (*CV*)
Bare Land Area	Pb	17.800 ~ 79.900	36.569	14.106	1.205	1.258	0.386
	Cu	17.400 ~ 43.100	28.868	5.414	0.582	0.574	0.188
	Cd	0.072 ~ 3.308	0.254	0.402	6.683	48.219	1.584
	Hg	0.014 ~ 0.242	0.069	0.048	1.337	1.651	0.693
Cropland Area	Pb	17.500 ~ 61.000	36.535	10.448	0.604	0.421	0.286
	Cu	18.100 ~ 38.800	28.903	5.769	0.278	0.627	0.146
	Cd	0.105 ~ 0.898	0.285	0.246	1.638	3.503	0.527
	Hg	0.019 ~ 0.587	0.091	0.403	4.794	31.713	0.790
Built-up Area	Pb	20.200 ~ 167.000	48.516	24.830	2.167	6.792	0.512
	Cu	18.000 ~ 475.700	34.442	51.677	8.523	73.665	1.500
	Cd	0.121 ~ 0.802	0.259	0.127	1.923	4.772	0.492
	Hg	0.018 ~ 0.978	0.097	0.120	5.788	40.152	1.231

Note: Natural background levels for soil elements in the A layer of Shaanxi Province (mg.kg^{-1}): Pb (21.4), Cu (21.4), Cd (0.094), and Hg (0.030); GB15618-2018 (PH>7.5) (mg.kg^{-1}): Pb (1 7 0), Cu (1 0 0), Cd (0.6), and Hg (3.4).



(a) laboratory spectra

(b) Image spectra

(c) Image spectra calibrated by DS algorithm

(d) S-G filtered spectra of corrected image

Fig. 3. Comparison of various kinds spectra in cropland area.

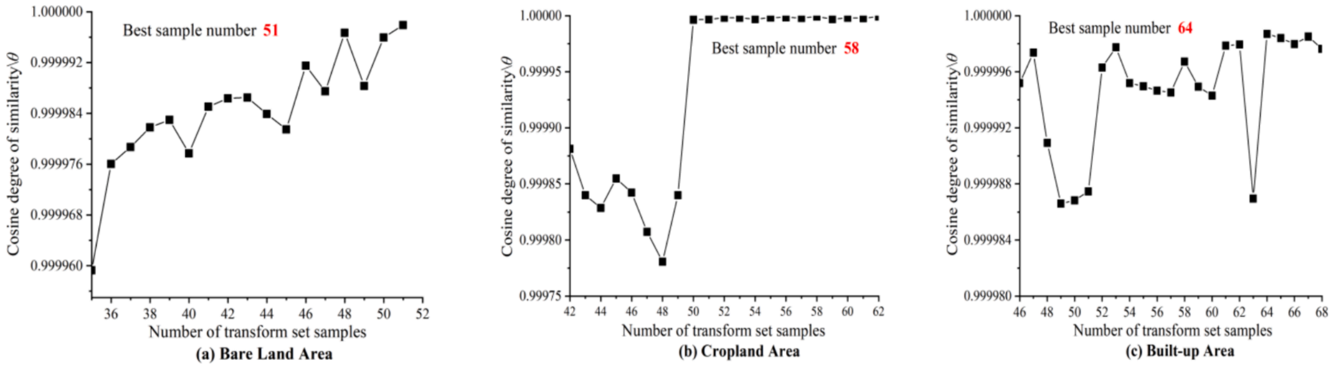


Fig. 4. Change curves of cosine similarity θ in various areas.

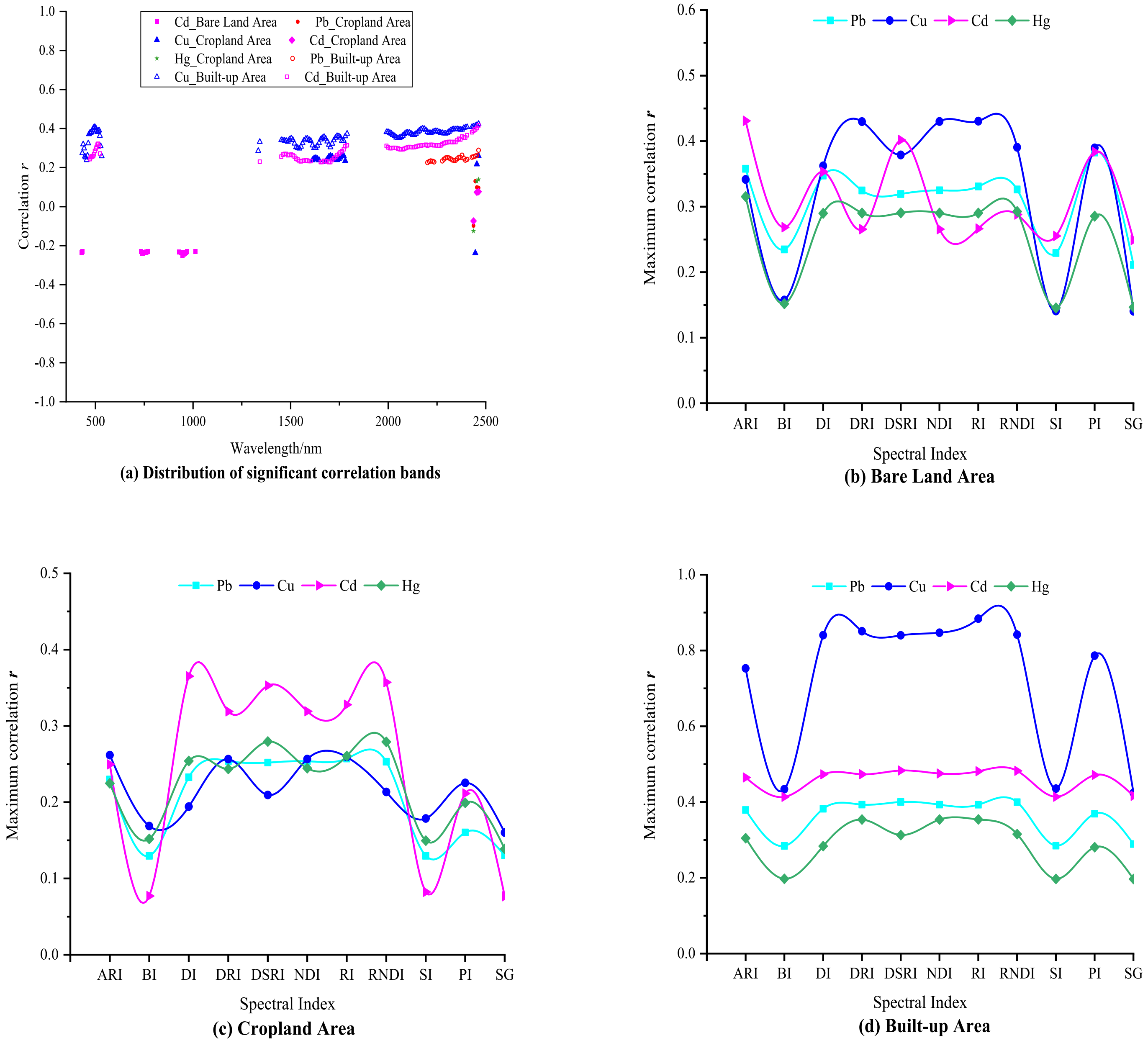


Fig. 5. Significant correlation analysis of various spectra.

derivation process for μ and k is given by

$$\begin{cases} R_1 = \mu e^{-k} = 1 \\ R_N = \mu e^{-kN} = \dfrac{2}{P} \end{cases} \Rightarrow \begin{cases} \mu = (P/2)^{\frac{1}{N-1}} \\ k = \dfrac{\ln(P/2)}{N-1} \end{cases} \tag{8}$$

Thirdly, a certain quantity of variables, denoted as $R_i \times p$, was chosen out of the variable set in the preceding sampling period by adaptive weighted sampling (ARS). Subsequently, PLS modeling was conducted, and the cross-validation error (RMSECV) was acquired. The wavelength subset that matched the minimal RMSECV was then recognized as the final feature set.

2.3.4. Inversion and model evaluation of soil heavy metal concentration

PLSR, RF, SVM, and BPNN models were implemented to predict the soil heavy metal content. Sample data was separated into training and validation partitions with a split proportion of 3:1, employing a concentration gradient reduction technique. The model determination coefficient (R^2), root mean square error (*RMSE*), and residual predictive deviation (*RPD*) were metrics used to evaluate model accuracy. The denominator of the traditional goodness of fit R^2 is the sum of squares of the dependent variable values. Adding another explanatory variable to the model does not change the denominator but does affect the numerator, this may appear to improve the model fit, but it can be misleading (Yang et al., 2023). To address this issue, we use the adjusted determination coefficient as given by Eqn. (9). We can see that the adjusted determination coefficient normalizes the numerator and denominator by their respective degrees of freedom. This effectively compensates for the number of variables in the model. Since parameter tuning is of great significance for model performance, the optimal parameter combinations were obtained through iterative model training using grid search and cross validation.

$$R^2 = 1 - \sum_{i=1}^{n} \left(y_i - y_p\right)^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{9}$$

where n represents the samples amount, and y_i and y_p are the actual and estimated concentrations, respectively; $\bar{y}$ indicates the average of the actual observations.

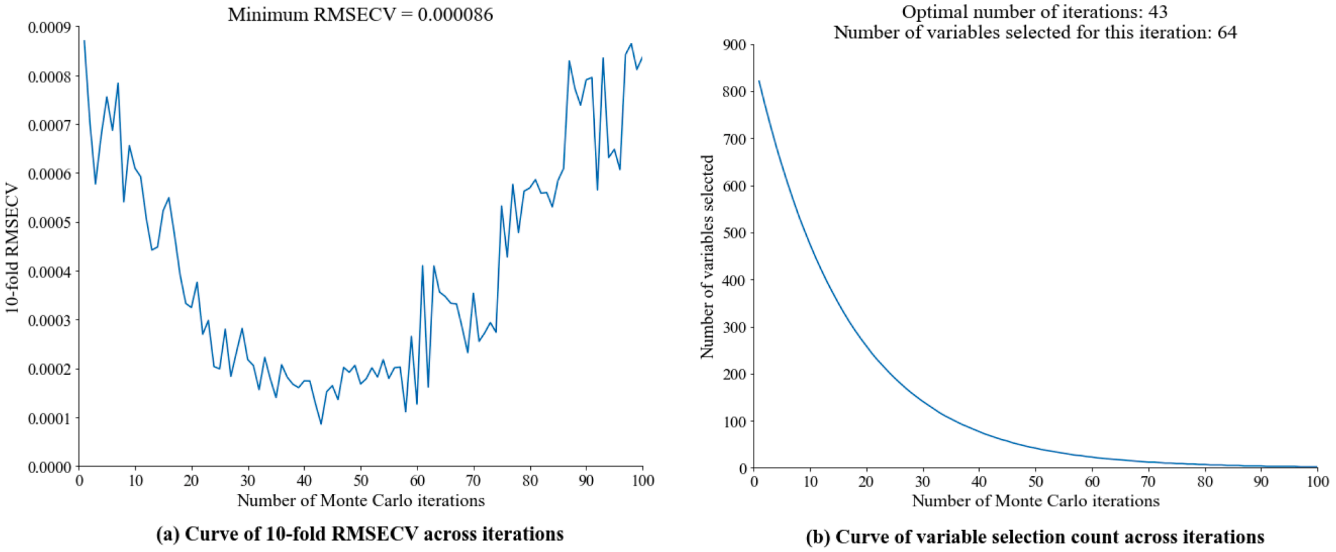
3. Results

3.1. Statistics toward soil heavy metal concentrations

The classification accuracy of GF-5AHSI imagery reached 98.826 %, with a corresponding Kappa coefficient of 0.982. Samples were extracted from bare land, cropland, and built-up areas based on the classification image, with sample counts of 74, 74, and 76, respectively. The concentration statistical results in different areas are described in Table 2. The mean concentration in each area exceeds the natural background levels of soil heavy metals in A-layer in Shaanxi Province. Moreover, the Cd content range has exceeded the standard of GB 15618-2018 (Ministry of Ecology and Environment of the People’s Republic of China, 2018), suggesting the presence of certain pollutants. Based on skewness and kurtosis statistics, it can be observed that the data distribution of the built-up area shows a strong right-skewed and relatively concentrated phenomenon, which is speculated to be the enrichment of heavy metals caused by human activities (Joanes and Gill, 1998; Westfall, 2014). The coefficient of variation revealed that Cd in the bare soil, Cu in the soil of the built-up area, and Hg in soils across the three distinct areas all exhibited strong variability ($CV > 65\%$) (Hu et al., 2008), indicating that human activities have markedly affected the soil heavy metal spread pattern within the region, consistent with the characteristics discussed regarding skewness and kurtosis. The degree of human influence on Pb and Cu in various areas was as follows: *CV*(Built-up Area) > *CV*(Bare Land Area) > *CV*(Cropland Area). The consequence of human activities on Cd in various areas followed this order: *CV*(Bare Land Area) > *CV*(Cropland Area) > *CV*(Built-up Area), which was opposite to that of Hg.

3.2. Calibration of GF-5 AHSI imagery with the DS algorithm

Evidently, taking cropland areas as an instance, the reflectance spectra of the GF-5 AHSI image (Fig. 3(b)) exhibited lower reflectance and were more coarse than the laboratory-measured spectrum (Fig. 3 (a)). However, distinct similarities in terms of shape, slope, and peak positions have been observed between these two kinds of spectra. The curves of cosine similarity θ in different areas, which vary with the size of the spectral transfer subset, are shown in Fig. 4. It was observed that the spectral subset sizes used to determine the DS transfer matrix are Bare Land Area $m_{51} \times p_{277}$, Cropland Area $m_{58} \times p_{277}$, and Built-up Area $m_{64} \times p_{277}$, respectively. It was evident that the reflectivity of the corrected image spectra (Fig. 3(c)) has been improved and exhibited



(a) Curve of 10-fold RMSECV across iterations

(b) Curve of variable selection count across iterations

Fig. 6. Process of CARS feature selection.

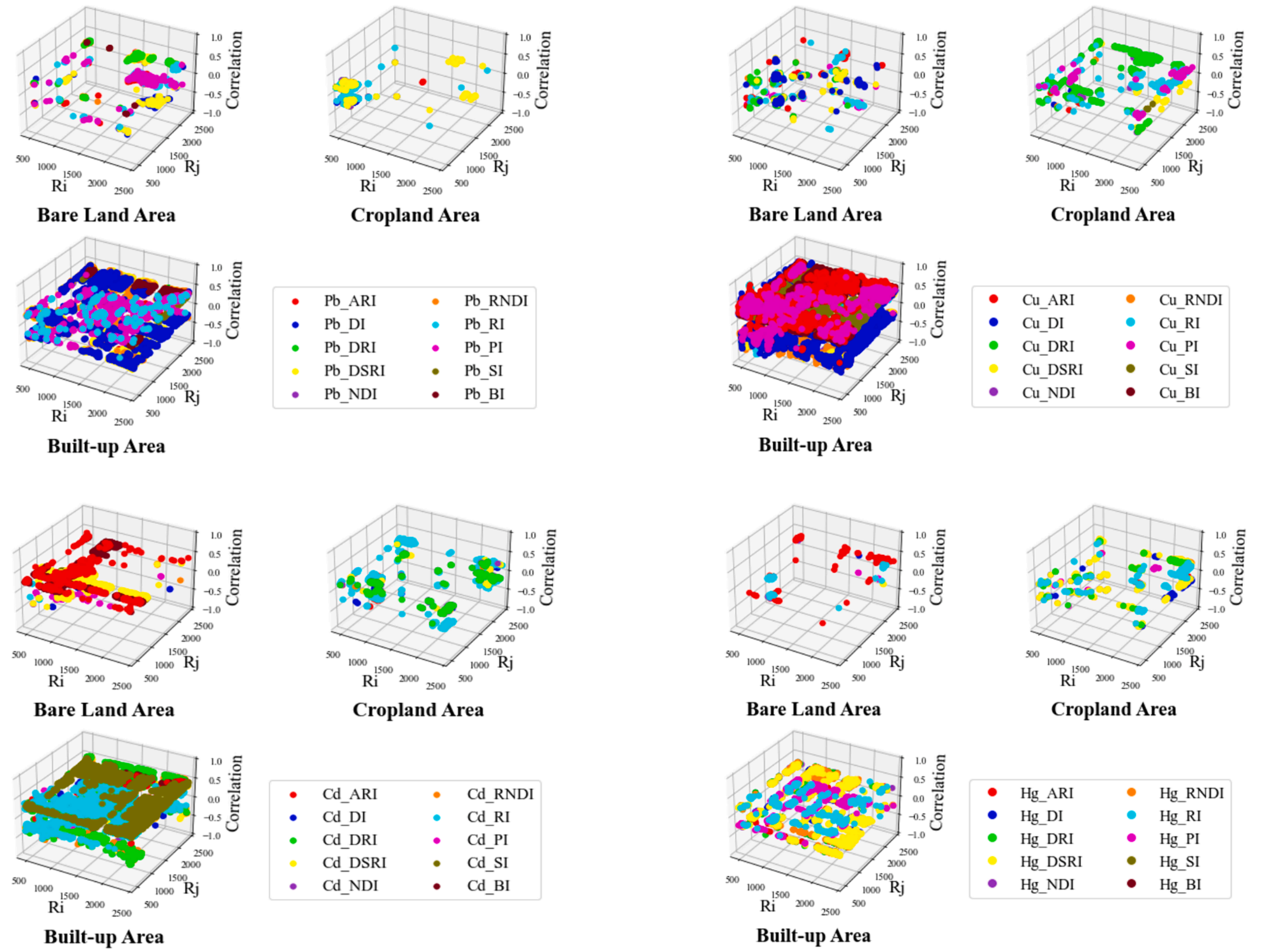


Fig. 7. Distributions of response characteristic bands of *2D-SSI* for soil heavy metals.

smoother characteristics than the original image spectra (Fig. 3(b)). Nine-point Savitzky-Golay convolution smoothing was verified and employed. The results demonstrated that the corrected smoothed spectra (Fig. 3(d)), not only preserved the detailed spectral characters at 550–700 nm and 1900–2000 nm, but also reduced the noise at 1000–1750 nm and 2250–2500 nm. The spectra of bare land areas and built-up areas have also shown similar patterns to those in cropland areas. Subsequent analysis in this study will be based on the corrected smoothed images.

3.3. Spectral feature selection

The distribution of significantly correlated bands between heavy metals and the DS-corrected GF-5 AHSI spectral bands is shown in Fig. 5 (a). Similar significant correlation bands were found for Cd and Cu in built-up areas, which were primarily distributed at 472–535 nm, 1450–1797 nm, and 1991–2463 nm with a low correlation of $|r| < 0.5$. It can be observed that DS-corrected GF-5 AHSI image spectral bands showed significant correlation with the concentration-dependent variables in cropland areas. However, only Cd in bare land areas has exhibited a significant correlation with DS-corrected GF-5 AHSI image spectral bands. Moreover, there are no significantly correlated bands between Hg in the built-up area and DS-corrected GF-5 AHSI image spectral bands. Consequently, it is challenging to attain accurate concentration estimation in various areas based on the significantly correlated bands of DS-corrected GF-5 AHSI image spectral. The maximum correlation coefficients between heavy metal content and *2D-SSI* across different areas were compared in Fig. 5(b), Fig. 5(c), and Fig. 5(d). Clearly, except for *SI* and *BI*, the maximum correlations between the constructed *2D-SSI* and heavy metal contents exceed those of the DS-corrected GF-5 AHSI image spectral bands and heavy metal contents. *DI*, *DRI*, *NDI*, *RI*, *RNDI* and *PI* have exhibited higher maximum correlations with Cu compared to other metals in bare land areas, while in cropland areas, except for *PI*, these *2D-SSI* indices show the highest correlations with Cd. The order of maximum correlation between heavy metal contents and *2D-SSI* in built-up areas was Cu > Cd > Pb > Hg.

The significant correlated *2D-SSI* variables selected were subjected to CARS feature selection. CARS feature screening contains four steps: Firstly, the MC sampling method was employed to randomly divide the samples into modeling and testing sets with the ratio of 8:2. Secondly, the ratio of variable retention was determined by EDF. Thirdly, ARS was adopted to remove variables forcefully. And finally iterate this process cyclically and calculate the 10-fold RMSECV. The CARS iteration was set to 100 times, and the 10-fold RMSECV was acquired. The CARS feature selection details can be seen from an instance exhibited in Fig. 6. The variable subset that matched the minimal RMSECV was then recognized as the final feature set. The positions of the response bands for each *2D-SSI* and their correlation coefficients are illustrated in Fig. 7. It is indicated that the distribution of response bands for *2D-SSI* of heavy metals was most dense and extensive in built-up areas, while it was relatively

Table 3
Regression outcomes of PLSR, RF, SVM, and BPNN (unit: mg.kg^{-1}).

Soil heavy metals	Spectral model	R^2_C	$RMSE_C$	RPD_C	R^2_V	$RMSE_V$	RPD_V
Bare Land Area_Pb	RNDI_PLSR	0.738	5.548	1.925	0.544	6.665	1.741
	DSRI_RF	0.682	8.455	1.827	0.428	9.143	1.470
	DSRI_SVM	**0.803**	**5.961**	**2.051**	**0.619**	**9.841**	**1.809**
	DI_BPNN	0.592	7.026	1.797	0.355	9.320	1.246
Cropland Area_Pb	**PI_PLSR**	**0.839**	**3.409**	**2.237**	**0.499**	**7.313**	**1.653**
	ARI_RF	0.729	5.089	1.919	0.417	8.629	1.423
	DI_SVM	0.496	7.254	1.641	0.253	8.756	1.157
	NDI_BPNN	0.666	2.004	1.817	0.439	2.601	1.335
	SG_PLSR	0.743	5.657	1.938	0.443	7.156	1.524
	SG_RF	0.655	6.245	1.817	0.379	9.254	1.302
	SG_SVM	0.384	8.895	1.322	0.224	9.550	1.134
	SG_BPNN	0.538	3.334	1.739	0.316	4.513	1.210
Built-up Area_Pb	RNDI_PLSR	0.969	2.389	2.376	0.811	3.013	2.088
	DSRI_RF	0.500	12.461	1.659	0.450	15.492	1.525
	RI_SVM	**0.939**	**7.734**	**2.324**	**0.871**	**8.478**	**2.248**
	RNDI_BPNN	0.764	1.691	1.963	0.639	2.089	1.816
	SG_PLSR	0.560	13.551	1.783	0.507	14.486	1.701
	SG_RF	0.425	16.388	1.457	0.380	16.517	1.317
	SG_SVM	0.767	10.620	1.967	0.603	11.107	1.803
	SG_BPNN	0.489	4.859	1.627	0.449	5.929	1.522
Bare Land Area_Cu	**RI_PLSR**	**0.912**	**2.686**	**2.315**	**0.788**	**5.157**	**1.971**
	NDI_RF	0.834	2.248	2.252	0.621	3.036	1.816
	NDI_SVM	0.793	2.482	1.986	0.512	2.599	1.674
	DI_BPNN	0.512	3.990	1.680	0.432	4.306	1.327
Cropland Area_Cu	DSRI_PLSR	0.559	1.502	1.769	0.363	2.568	1.268
	ARI_RF	**0.778**	**2.039**	**1.968**	**0.461**	**3.619**	**1.553**
	DRI_SVM	0.418	4.290	1.426	0.313	4.979	1.207
	DI_BPNN	0.452	3.960	1.541	0.355	4.080	1.239
	SG_PLSR	0.420	3.621	1.432	0.309	4.691	1.202
	SG_RF	0.639	3.309	1.817	0.382	3.896	1.312
	SG_SVM	0.287	4.746	1.186	0.216	4.979	1.131
	SG_BPNN	0.372	4.024	1.286	0.274	4.967	1.178
Built-up Area_Cu	RI_PLSR	0.999	1.133	2.433	0.673	2.344	1.818
	RI_RF	0.837	2.222	2.238	0.430	3.744	1.470
	DRI_SVM	0.838	2.296	2.239	0.433	2.674	1.328
	RI_BPNN	**1.000**	**0.068**	**2.433**	**0.883**	**0.202**	**2.272**
	SG_PLSR	0.819	1.582	2.132	0.546	3.382	1.744
	SG_RF	0.650	3.620	1.818	0.367	4.826	1.279
	SG_SVM	0.672	2.510	1.780	0.408	4.543	1.306
	SG_BPNN	0.897	1.325	2.299	0.696	1.898	1.819
Bare Land Area_Cd	**ARI_PLSR**	**0.984**	**0.031**	**2.426**	**0.709**	**0.262**	**1.851**
	DRI_RF	0.959	0.042	2.351	0.673	0.366	1.818
	NDI_SVM	0.665	0.082	1.816	0.503	0.086	1.665
	DSRI_BPNN	0.808	0.195	2.083	0.610	0.277	1.802
	SG_PLSR	0.404	0.073	1.308	0.351	0.410	1.262
	SG_RF	0.368	0.097	1.302	0.314	0.438	1.208
	SG_SVM	0.139	0.405	1.078	0.072	0.439	1.038
	SG_BPNN	0.215	0.319	1.129	0.204	0.445	1.121
Cropland Area_Cd	BI_PLSR	0.879	0.231	2.254	0.367	0.570	1.298
	ARI_RF	**0.778**	**0.127**	**1.965**	**0.589**	**0.130**	**1.788**
	RI_SVM	0.358	0.130	1.248	0.304	0.209	1.199
	ARI_BPNN	0.586	1.736	1.785	0.452	1.997	1.531
	SG_PLSR	0.448	0.330	1.536	0.341	0.578	1.249
	SG_RF	0.637	0.189	1.809	0.523	0.208	1.698
	SG_SVM	0.321	0.147	1.214	0.225	0.293	1.135
	SG_BPNN	0.501	1.975	1.618	0.393	2.375	1.376
Built-up Area_Cd	**DSRI_PLSR**	**0.963**	**0.071**	**2.376**	**0.834**	**0.225**	**2.252**
	DSRI_RF	0.546	0.094	1.748	0.443	0.096	1.524
	ARI_SVM	0.998	0.007	2.427	0.775	0.050	1.964
	RI_BPNN	0.954	0.030	2.327	0.790	0.064	1.979
	SG_PLSR	0.898	0.120	2.255	0.693	0.267	1.836
	SG_RF	0.511	0.114	1.679	0.321	0.142	1.214
	SG_SVM	0.776	0.059	1.965	0.577	0.086	1.777
	SG_BPNN	0.822	0.046	2.138	0.616	0.069	1.804
Bare Land Area_Hg	ARI_PLSR	0.625	0.084	1.816	0.414	0.112	1.419
	ARI_RF	0.704	0.023	1.838	0.415	0.034	1.420
	DRI_SVM	0.328	0.025	1.220	0.252	0.042	1.156
	ARI_BPNN	**0.789**	**0.023**	**2.177**	**0.418**	**0.038**	**1.311**

(*continued on next page*)

Table 3 (*continued*)

Soil heavy metals	Spectral model	R_C^2	$RMSE_C$	RPD_C	R_V^2	$RMSE_V$	RPD_V
Cropland Area_Hg	DRI_PLSR	0.514	0.090	1.712	0.389	0.149	1.368
	DI_RF	**0.821**	**0.019**	**2.164**	**0.614**	**0.028**	**1.804**
	ARI_SVM	0.689	0.023	1.832	0.355	0.034	1.246
	DSRI_BPNN	0.685	0.021	1.830	0.415	0.034	1.420
	SG_PLSR	0.342	0.188	1.233	0.291	0.360	1.189
	SG_RF	0.589	0.036	1.805	0.475	0.041	1.596
	SG_SVM	0.293	0.348	1.191	0.266	0.489	1.168
	SG_BPNN	0.424	0.044	1.453	0.345	0.062	1.214
Built-up Area_Hg	**DRI_PLSR**	**0.959**	**0.033**	**2.327**	**0.907**	**0.068**	**2.310**
	RNDI_RF	0.682	0.047	1.828	0.377	0.080	1.317
	ARI_SVM	0.887	0.033	2.275	0.484	0.045	1.622
	ARI_BPNN	0.740	0.034	1.963	0.674	0.069	1.820

sparse in bare land and cropland areas. The response features of various *2D-SSI* for Pb in bare land areas were formed by the band combination of 1450–2440 nm and 2000–2400 nm. *DSRI* and *RI* provided more features for Pb in cropland areas, with the main distribution range of 403–416 nm and 493–698 nm, while only a few features were distributed in the far-infrared bands. *DI* has indicated the most features for Pb in the built-up area, which almost covers the feature space between 400–2463 nm and 1450–2395 nm. *DI* and *DSRI* have provided the most features for Cu in bare land areas, while *DRI* offered the most extensive features for Cu in cropland areas, which were mainly distributed in the feature space between 724–1789 nm and 2244–2463 nm. The observation was aligned with the findings of Zhang et al. (2022). *BI* has the most selected features for Cu in the built-up area, followed by *DI*, *RNDI*, *NDI*, and *SI*. *DSRI* and *ARI* offered relatively abundant features for Cd in bare land areas, and these features were mainly concentrated in the spaces, between 425–1999 nm and 1991–2446 nm, as well as 420–805 nm and 1569–2454 nm. *DRI* and *RI* exhibit relatively most characteristics of Cd in cropland areas. The most abundant feature for Cd corresponded to *SI*, mainly distributed in the feature spaces within the spectral ranges of 403–1241 nm and 1300–2463 nm, as well as 1300–1789 nm and 1982–2463 nm, primarily associated with the adsorption effect of organic compounds containing Cd (Shi et al., 2014). Features selected for Hg in bare land and cropland areas were relatively few and scattered, whereas *PI* and *DSRI* corresponded to the greatest number of characteristics for Hg in built-up areas. The subsequent spectral model inversion and analysis were both based on the spectral features presented in Fig. 7. The number of features in the variable set selected by CARS accounts for less than 13 % of the total variables, which dramatically reduces the computational complexity of subsequent modeling. Nonetheless, an assessment of its effectiveness is still required.

3.4. Modeling evaluation

The *2D-SSI* and calibrated SG feature variables were separately introduced into the PLSR, RF, SVM, and BPNN regression prediction models. Statistics information of the calibration SG spectral model and the optimal *2D-SSI* model for different heavy metals within various areas is shown in Table 3. It can be observed that, under the same predictive model criteria, the R_C^2 and R_V^2 of the optimal *2D-SSI* models for each heavy metal in the same area are both greater than the corresponding calibrated SG spectral models, demonstrating the potential prospects of the *2D-SSI* model for predicting heavy metal content. For each heavy metal, the range of increase in R_V^2 when comparing the optimal *2D-SSI* model to the corresponding calibrated SG model was Pb $[0.074, 0.409]$, Cu $[0.080, 0.187]$, Cd $[0.035, 0.593]$, and Hg $[0.172, 0.396]$. In particular, compared to the corresponding calibration SG models, the optimal *2D-SSI* models for Pb BPNN and Hg RF in cropland areas show increases of 0.128 and 0.232 in R_C^2, and corresponding increases of 0.123 and 0.139 in R_V^2. For Cd in bare land areas, the optimal *2D-SSI* model compared to the corresponding calibrated SG model showed an increase in R_C^2 ranging from 0.526 to 0.593 and in R_V^2 ranging from 0.358 to 0.431. The ranking of the occurrence frequency of the optimal models corresponding to *2D-SSI* was as follows: *ARI*(12) > *DSRI*(8) > *RI*(7) > *DRI*(6) > *DI*(5) > *RNDI*(4) = *NDI*(4) > *PI*(1) = *B*1, indicating the feasibility of estimating heavy metal content with *2D-SSI*, except for *SI*. Models based on *DSRI*(3) and *RNDI*(3) have shown good performance on the training set for Pb, achieving R_C^2 values in the range of $[0.500, 0.970]$, while *RI*(4) exhibited strong generalization performance for Cu prediction with R_V^2 in the range of $[0.430, 0.883]$. For the optimal models of Cd and Hg, *ARI* appeared most frequently, with 4 and 6 occurrences, respectively, demonstrating the good applicability of *ARI* for predicting trace elements Cd and Hg. The R_C^2 values of the Pb models based on different optimal *2D-SSI* features all reached 0.5; however, only the PLSR model was found to be relatively superior to other models in estimating Pb concentration, with its R_V^2 values also reaching 0.5. The training accuracy R_C^2 of Cu models for bare land and built-up areas both achieved 0.512, surpassing that of cropland areas. Additionally, the R_V^2 values of predictive models for Cu in bare land areas also achieved an accuracy of 0.512, except for the BPNN model. The PLSR and BPNN models based on *RI* for Cu inversion, as well as the SVM model based on *ARI* for Cd prediction, which the R_C^2 all achieved 1 in built-up areas. Although over-fitting was present, the corresponding validation accuracy R_V^2 values were all greater than 0.65, demonstrating optimistic model predictive ability. For Cd prediction, except for the SVM model in cropland areas, the training accuracy range of the other models was $[0.546, 1]$. However, only models in bare land areas have demonstrated outstanding Cd prediction capability on the validation set, with R_V^2 in the range of $[0.503, 0.709]$. Except for the RF model, Cd inversion models in built-up areas were all achieved training set accuracy of $R_C^2 = 0.954$ and validation set accuracy of $R_V^2 = 0.775$, demonstrating better predictive capability compared to those in bare land areas. The training set performance R_C^2 for Hg in cropland and built-up areas fell within the range of $[0.514, 0.959]$, while only the PLSR and BPNN models in built-up areas performed well on the validation set, with R_V^2 of 0.67 and 0.91, respectively. Accuracy of the optimal indicator models for heavy metals in various areas of soil is illustrated in Fig. 8 and the distribution of corresponding optimal band combinations is listed in Table 4. It can be observed that for the same heavy metal in different sample areas, there exists a certain degree of overlapping in the optimalband combinations of *2D-SSI*, thereby highlighting the response characteristic bands of the heavy metals. Prediction models performed better in built-up areas than those in bare land and cropland areas. The prediction performance of optimal models in bare land areas, aside from Hg, was superior to that of cropland areas.

3.5. Mapping of heavy metal concentrations

The concentration distribution of soil heavy metal in bare land and cropland areas was mapped with the optimal *2D-SSI* prediction models

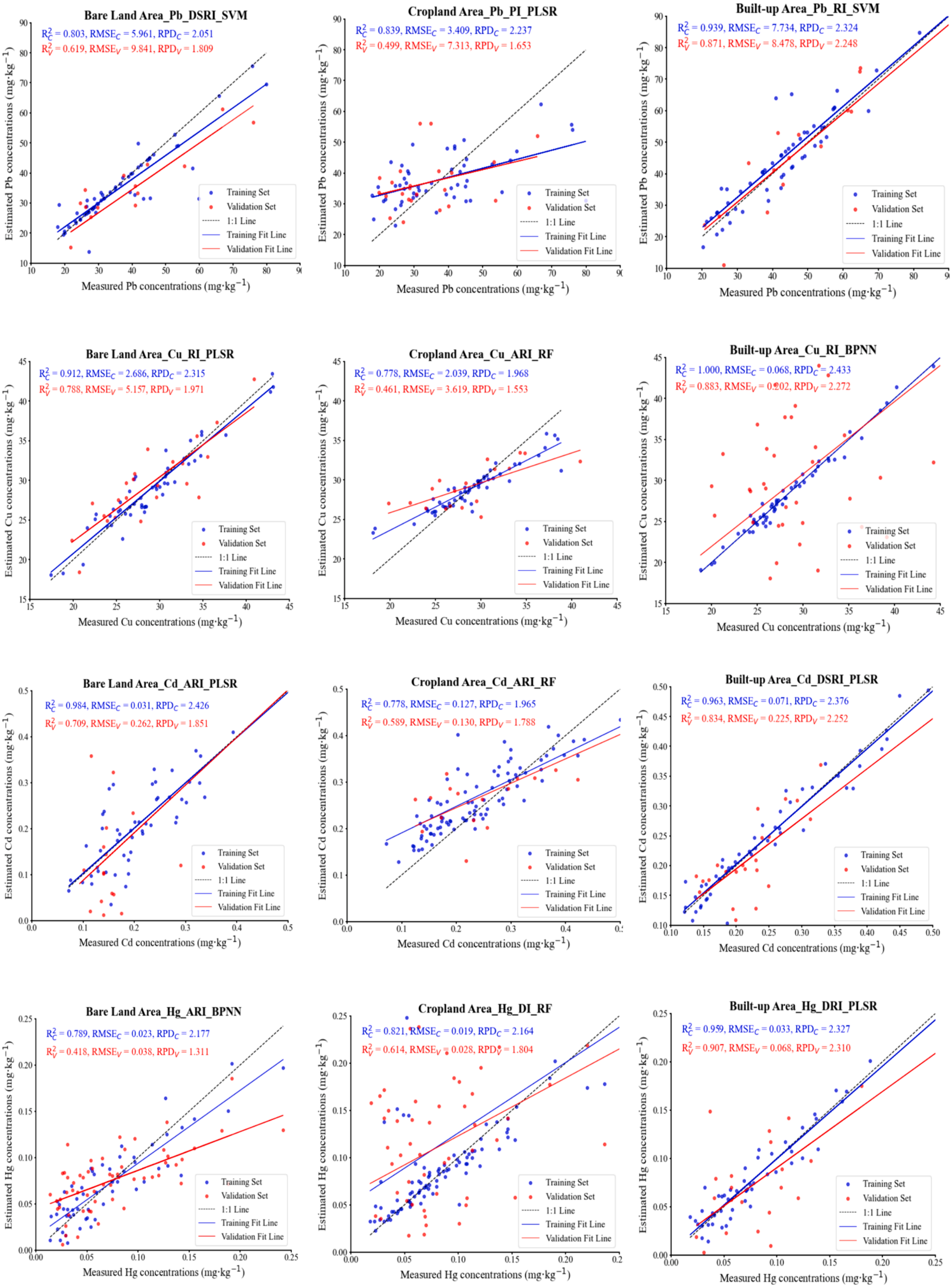


Fig. 8. Model effect of the optimal inversion from different areas.

Table 4
Distributions of the optimal band combinations corresponding to the best models for soil heavy metals.

Soil Heavy Metals		R_i	R_j	Spectral index
	Pb	1450–2440 nm	2000–2400 nm	DSRI
	Cu	429–476 nm \1460–2185 nm	818–1258 nm \2008–2395 nm	RI
	Cd	609–1342 nm \1545–2454 nm	420–1772 nm \472–981 nm	ARI
	Hg	1452–2463 nm	2130–2463 nm	ARI
	Pb	476–519 nm	1460–1511 nm	PI
	Cu	1047–1106 nm	733–775 nm	ARI
	Cd	844–916 nm	754–810 nm	ARI
	Hg	423–455 nm \2345–2454 nm	429–489 nm \2210–2345 nm	DI
Built-up Area	Pb	403–1199 nm \1460–2463 nm	407–1679 nm \2041–2437 nm	RI
	Cu	403–1780 nm \2016–2463 nm	416–1325 nm \711–1789 nm	RI
	Cd	822–1342 nm \1486–2454 nm	750–1325 nm \1199–2463 nm	DSRI
	Hg	429–540 nm \2404–2446 nm	2016–2463 nm \741–788 nm	DRI

founded upon corrected GF-5 AHSI imagery (Fig. 9). Concentrations of Pb and Cu were evidently higher than those of Cd and Hg. The content distribution of Pb was relatively high and exhibited significant spatial heterogeneity, while Cu was at lower concentration levels with a relatively uniform distribution. The eastern portion was primarily home to the high concentration dispersion of Cd, with the exception of a few isolated high values in the western portion. Regarding Hg, the high concentration zone was primarily found in the western portion, in spite of local high values in the northeast. According to the soil heavy metal background values in Shaanxi Province, the spatial proportion of heavy metals exceeding background values ranked as follows: Cu (98.208 %) > Cd (95.265 %) > Pb (65.914 %) > Hg (21.969 %). The concentration of Hg was found to be within the normal range with regard to the GB15618-2018 standard. The areas where Pb and Cu exceeded the standard limits accounted for 0.01 % of the entire study area, while those surpassing the Cd limit represented 6.861 %. Pairwise correlation analysis was performed on the areas with heavy metal content exceeding background values. The findings depicted in Table 5 implied the similarities and homogeneity in the distribution occurrence across these areas.

4. Discussion

4.1. The reliability of the spectral feature

The DS image spectral calibration has improved overall spectral accuracy by reducing the impact of external environmental interference on image spectra, a fact confirmed by earlier investigations (Ji et al., 2015a; Ji et al., 2015b). The calibrated spectra of GF-5 imagery (Fig. 3 (c)) exhibited greater consistency with the pertinent laboratory spectra (Fig. 3(a)), thereby providing substantial support for ongoing large-scale heavy metal concentration prediction (Zou et al., 2020). Furthermore, the features expanded by the constructed *2D-SSI* were refined and differentiated. It has been confirmed that soil organic matter influences the wavelengths of features for Cu and Cd mentioned above (Zhang et al., 2022). Additionally, the extracted spectral data from various land-use/land-cover areas were plotted and fitted to soil lines to assess the quality of image spectra (Fig. 10). The results revealed that the spectral data from the bare land area best represented the soil line relationship corresponding to bare soil ($R^2 = 0.675$), followed by the spectral data from the built-up area ($R^2 = 0.632$) and cropland area ($R^2 = 0.581$). It suggested that the soil spectra from the three types of areas, influenced by mixed pixels, all contain information originating from sources other than bare soil. Due to interference from vegetation and other factors, the spectral data from cropland areas contain less information related to bare soil. However, the spectral data from all three areas closely resemble the spectral characteristics of bare soil with $R^2 \geq 0.581$.

Table 5
Spatial correlation (r) of distribution of heavy metal content beyond background value.

Element	Pb	Cu	Cd	Hg
Pb	1	0.717	0.683	0.299
Cu		1	0.965	0.381
Cd			1	0.385
Hg				1

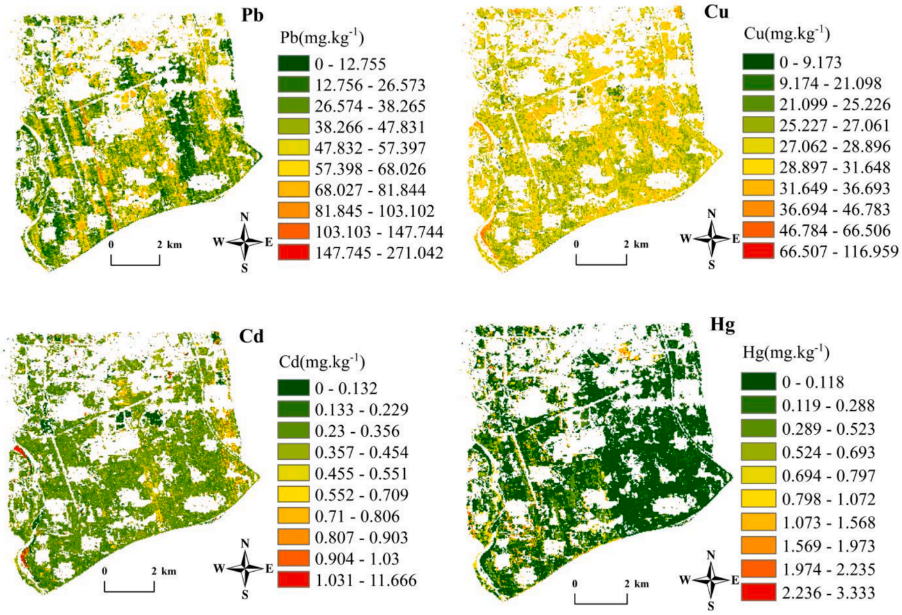


Fig. 9. Spatial distribution of soil heavy metal concentrations.

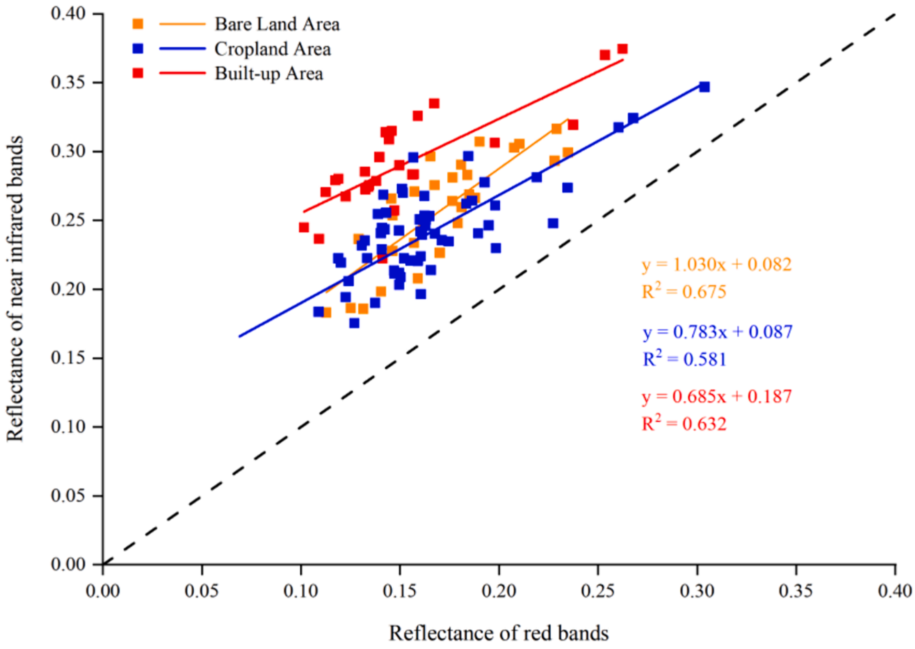


Fig. 10. Soil line analysis in different areas.

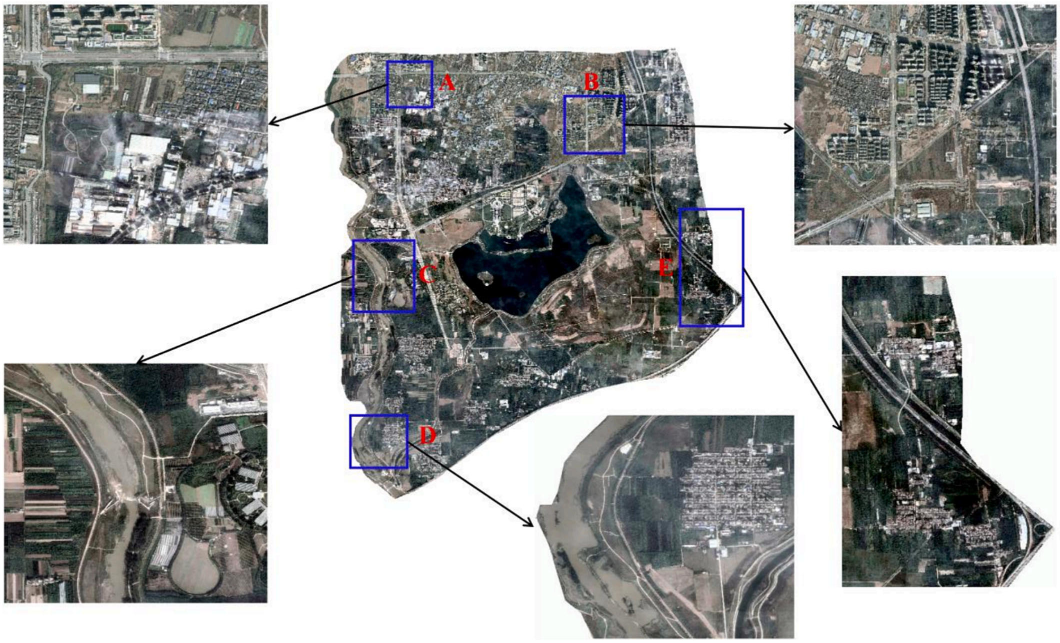


Fig. 11. Local enlarged view of interest regions with high concentrations of soil heavy metals.

Finally, features were plotted in the two-dimensional spectral feature space (Fig. 7) to facilitate a clearer analysis of the composition of the spectral response characteristic band. Pb, being a strictly regulated heavy metal, exhibited spectral features closer to the ultraviolet wavelength, making its features more prominent in this range. This explains why Pb has characteristic bands between 400 nm and 700 nm, while the spectral wavelength range of 2000 nm to 2463 nm represents features related to other heavy metals (Tan et al., 2021).

4.2. Spatial characteristics of heavy metal contamination

The concentration distributions have exceeded the background levels of soil heavy metals, showing local anomalies (Fig. 9). To our knowledge, the distribution pattern is influenced not only by natural factors such as topographical features, meteorological parameters, soil characteristics, and geochemical processes but also by anthropogenic variables as industrial emissions, agricultural fertilization, and construction operations (Arif et al., 2022). The correlation coefficients for the super-background spread pattern of Pb, Cu, and Cd are all greater than 0.68, indicating similar enrichment patterns for these metals. The interest regions with high concentrations were marked with blue rectangles and labeled with red letters A to E (Fig. 11) on the distribution map for further analysis and interpretation. Pb and Cd exhibited abnormal pollution in areas A and E, both of which were located at the intersection of traffic arteries and residential areas, and it was speculated that the accumulation was attributed to transportation and other human activities. Due to the demolition and reconstruction of abandoned houses in residential areas, as well as the random accumulation of construction waste and dust, Pb, Cd, and Hg accumulated around area B, resulting in concentrations exceeding the soil heavy metal background values of Shaanxi Province. Similar abnormal distributions of Cu and Cd pollution were observed in areas C and D, which are formed by the sedimentary belt on both sides of the river and the highway. This is speculated to be caused by factors such as vehicle exhaust emissions, traffic dust, and sedimentation of river transport. The distribution pattern of Cu and Cd above background values was correlated with $r = 0.965$, indicating a similarity and homogeneity in the aggregation of Cd and Cu. Overall, soil contamination within the study area was closely associated with human activities, with vehicle exhaust emissions, traffic dust, and the demolition and reconstruction of abandoned houses identified as the dominant contributing factors.

4.3. Uncertainty and prospect of the inversion strategy

Although we have proposed an soil heavy metal content inversion method guided by satellite images in accordance with spectral correction and data mining, the fact that spectral characteristics of soil may be influenced by factors such as dampness, structure, shade, and surface coarseness should not be ignored (Lin et al., 2022). On the other hand, the laboratory spectral data were acquired with a 30°angle halogen lamp as the light source, which is consistent with most previous studies, but differs from the solar altitude angle during the acquisition and observation of GF-5 satellite imagery. All of these factors are vital for optical calibration of remote sensing data, however, our consideration of these factors may not be comprehensive enough. The correlation between the DS-corrected spectra and dependent variables of concentration has all reached a significant level, and the correlation has been significantly improved compared to the original bands. However, any method based on hyperspectral images for estimating soil physical and chemical parameters requires careful analysis and validation specific to the conditions (Ge et al., 2022). SVM algorithm used for hyperspectral image classification is limited by the challenges of manual data annotation and highly correlated spectral features. Additionally, it is difficult to implement with large-scale training samples, and the classification results are prone to salt-and-pepper noise. Therefore, improved network intelligent classification approaches such as unified multiscale learning (UML) framework (Wang et al., 2022) and capsule-vectored neural network (CVNN) (Wang et al., 2023) are preferred for future research to address the issues of insufficient feature representation and poor classification performance with limited labeled samples inherent in traditional models. The performance evaluation of models claimed that the PLSR, RF, SVM, and BPNN models exhibited varying regional applicability for estimating soil heavy metal content. The use of a small-scale training dataset may bring about model over-fitting (Xin et al., 2020), as evidenced by the PLSR and BPNN models based on *RI* for Cu, as well as the SVM inversion model based on *ARI* for Cd in the built-up area. Therefore, quantitatively characterizing the association regarding image spectra and pure soil signal is crucial for further improving the precision and robustness of the model.

5. Conclusions

The study specifically produced an efficient approach for predicting soil heavy metal concentrations by constructing *2D-SSI* from the DS-corrected GF-5 imagery. The complete workflow and methods used in this study have been thoroughly described, and the principal findings are as follows:

(1) The correlation has been enhanced by the constructed *2D-SSI* feature variables, and the feature selection approach of the significant correlation method coupled with the CARS algorithm has been validated as straightforward and productive, which performed dimensional reduction and enhanced the interpretability of model features.

(2) The DS algorithm has proven to be practicable and dependable for GF-5 AHSI imagery calibration, thereby enhancing the estimation accuracy. Models based on the *2D-SSI* have demonstrated excellent performance with the R_V^2 intervals of $[0.253, 0.871]$, $[0.313, 0.883]$, $[0.304, 0.834]$, and $[0.252, 0.907]$, respectively, for Pb, Cu,Cd, and Hg. The distribution pattern is generally compatible with the findings of actual observations. This approach can effectively portray soil heavy metal concentrations across expansive spatial scales.

(3) Human activities such as vehicle exhaust emissions, demolition and reconstruction of abandoned houses, and dust from transportation were suspected as the predominant pollution sources. This information may serve as a reference for issuing warnings in polluted areas.

There were limitations in analyzing the factors influencing the spatial distribution of soil heavy metals. The research focused primarily on human factors, neglecting the impacts of geochemistry, crops, and even seasonal fluctuations in heavy metal transformation and transportation. Therefore, further investigation should delve into these factors for a more comprehensive understanding.

CRediT authorship contribution statement

Nannan Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Liangzhi Li:** Writing – review & editing, Supervision, Investigation. **Ling Han:** Resources, Project administration, Funding acquisition, Data curation. **Kyle Gao:** Writing – review & editing, Visualization. **Songjie Qu:** Writing – review & editing, Data curation. **Jonathan Li:** Writing – review & editing, Resources.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

The work was supported by the Key Research and Development Project of Shaanxi Province (Program No. 2022ZDLSF07-05) and the High-Resolution Special Project of ecological protection and high quality development industrialization in key areas of the upper and middle reaches of the Yellow River (Program No. 91-Y50G32-9001-22/23). Special thanks to the relevant staff for their work in data collection. The authors would like to thank the editors and reviewers for their valuable comments.

References

Arif, M., Qi, Y., Dong, Z., Wei, H., 2022. Rapid retrieval of cadmium and lead content from urban greenbelt zones using hyperspectral characteristic bands. J. Clean. Prod. 374, 133922 https://doi.org/10.1016/j.jclepro.2022.133922.

Bao, Y., Ustin, S., Meng, X., Zhang, X., Guan, H., Qi, B., Liu, H., 2021. A regional-scale hyperspectral prediction model of soil organic carbon considering geomorphic features. Geoderma 403, 115263. https://doi.org/10.1016/j.geoderma.2021.115263.

Bellinaso, H., Silvero, N.E.Q., Ruiz, L.F.C., Amorim, M.T.A., Rosin, N.A., Mendes, W.D.S., de Sousa, G.P.B., Sepulveda, L.M.A., de Queiroz, L.G., Nanni, M.R., Dematte, J.A.M., 2021. Clay content prediction using spectra data collected from the ground to space platforms in a smallholder tropical area. Geoderma 399, 115116. https://doi.org/10.1016/j.geoderma.2021.115116.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., 2015. Recent advances and emerging challenges of feature selection in the context of big data. Knowl.-Based Syst. 86, 33–45. https://doi.org/10.1016/j.knosys.2015.05.014.

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

Chen, T., Chang, Q., Clevers, J.G.P.W., Kooistra, L., 2015. Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy. Environ. Pollut. 206, 217–226. https://doi.org/10.1016/j.envpol.2015.07.009.

Cheng, H., Shen, R., Chen, Y., Wan, Q., Shi, T., Wang, J., Wan, Y., Hong, Y., Li, X., 2019. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. Geoderma 336, 59–67. https://doi.org/10.1016/j.geoderma.2018.08.010.

China Centre for Resources Satellite Data and Application, 2024. Land Satellite Observation Data Service Platform. https://data.cresda.cn/#/home. (accessed 30 March 2024).

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

Escadafal, R., 1989. Remote sensing of arid soil surface color with landsat thematic mapper. Adv. Space Res. 9 (1), 159–163. https://doi.org/10.1016/0273-1177(89)90481-X.

Ge, X., Wang, J., Ding, J., Cao, X., Zhang, Z., Liu, J., Li, X., 2019. Combining uav-based hyperspectral imagery and machine learning algorithms for soil moisture content monitoring. PeerJ 7, e6926.

Ge, X., Ding, J., Teng, D., Wang, J., Huo, T., Jin, X., Wang, J., He, B., Han, L., 2022. Updated soil salinity with fine spatial resolution and high accuracy: the synergy of sentinel-2 msi, environmental covariates and hybrid machine learning approaches. Catena (giessen). 212, 106054 https://doi.org/10.1016/j.catena.2022.106054.

Gitelson, A.A., Merzlyak, M.N., Chivkunova, O.B., 2001. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. Photochem. Photobiol. 74 (1), 38–45. https://doi.org/10.1562/0031-8655(2001)074%3C0038:OPANEO%3E2.0.CO;2.

Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., Zhang, H., Zeng, C., 2021. Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. Geoderma 398, 115118. https://doi.org/10.1016/j.geoderma.2021.115118.

Hu, W., Shao, M., Wang, Q., Reichardt, K., 2008. Spatial variability of soil hydraulic properties on a steep slope in the loess plateau of china. Sci. Agric. 65, 268–276. https://doi.org/10.1590/S0103-90162008000300007.

Husnizar H., Wilopo W., Yuliansyah A.T., Gadjah M.U., 2018. The prediction of heavy metals lead (pb) and zinc (zn) contents in soil using nirs technology and plsr regression method. J. Degrad. Min. Land Manage. 5(3): 1153-1159. 10.15243/jdmlm.2018.053.1153.

Ji, W., Viscarra Rossel, R.A., Shi, Z., 2015a. Accounting for the effects of water and the environment on proximally sensed vis–nir soil spectra and their calibrations. Eur. J. Soil Sci. 66 (3), 555–565. https://doi.org/10.1111/ejss.12239.

Ji, W., Viscarra Rossel, R.A., Shi, Z., 2015b. Improved estimates of organic carbon using proximally sensed vis-NIR spectra corrected by piecewise direct standardization. Eur. J. Soil Sci. 66 (4), 670–678. https://doi.org/10.1111/ejss.12271.

Jiang, Q., Liu, M., Wang, J., Liu, F., 2018. Feasibility of using visible and near-infrared reflectance spectroscopy to monitor heavy metal contaminants in urban lake sediment. Catena 162, 72–79. https://doi.org/10.1016/j.catena.2017.11.020.

Joanes, D.N., Gill, C.A., 1998. Comparing measures of sample skewness and kurtosis. J. r. Stat. Soc. Ser. D (the Statistician) 47 (1), 183–189. https://doi.org/10.1111/1467-9884.00122.

Li, Q., Huang, Y., Song, X., Zhang, J., Min, S., 2019. Moving window smoothing on the ensemble of competitive adaptive reweighted sampling algorithm. Spectrochim. Acta A Mol. Biomol. Spectrosc. 214, 129–138. https://doi.org/10.1016/j.saa.2019.02.023.

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 (1), 77–84. https://doi.org/10.1016/j.aca.2009.06.046.

Lin, N., Jiang, R., Li, G., Yang, Q., Li, D., Yang, X., 2022. Estimating the heavy metal contents in farmland soil from hyperspectral images based on stacked adaboost ensemble learning. Ecol. Ind. 143, 109330 https://doi.org/10.1016/j.ecolind.2022.109330.

Liu, Q., He, L., Guo, L., Wang, M., Deng, D., Lv, P., Wang, R., Jia, Z., Hu, Z., Wu, G., Shi, T., 2022. Digital mapping of soil organic carbon density using newly developed bare soil spectral indices and deep neural network. Catena 219, 106603. https://doi.org/10.1016/j.catena.2022.106603.

Meng, X., Bao, Y., Liu, J., Liu, H., Zhang, X., Zhang, Y., Wang, P., Tang, H., Kong, F., 2020. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. Int. J. Appl. Earth Obs. 89, 102111 https://doi.org/10.1016/j.jag.2020.102111.

Ministry of Ecology and Environment of the People's Republic of China, 2018. Chinese National Standard. https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/trhj/201807/t20180703_446029.shtml. (accessed 30 March 2024).

Nyarko, F., Tack, F.M.G., Mouazen, A.M., 2022. Potential of visible and near infrared spectroscopy coupled with machine learning for predicting soil metal concentrations at the regional scale. Sci. Total Environ. 841, 156582 https://doi.org/10.1016/j.scitotenv.2022.156582.

Odebiri, O., Odindi, J., Mutanga, O., 2021. Basic and deep learning models in remote sensing of soil organic carbon estimation: a brief review. Int. J. Appl. Earth Obs. 102, 102389 https://doi.org/10.1016/j.jag.2021.102389.

Qin, G., Niu, Z., Yu, J., Li, Z., Ma, J., Xiang, P., 2021. Soil heavy metal pollution and food safety in china: effects, sources and removing technology. Chemosphere 267, 129205. https://doi.org/10.1016/j.chemosphere.2020.129205.

Roujean, J., Breon, F., 1995. Estimating par absorbed by vegetation from bidirectional reflectance measurements. Remote Sens. Environ. 51 (3), 375–384. https://doi.org/10.1016/0034-4257(94)00114-3.

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

Sun, W., Zhang, X., 2017. Estimating soil zinc concentrations using reflectance spectroscopy. Int. J. Appl. Earth Obs. 58, 126–133. https://doi.org/10.1016/j.jag.2017.01.013.

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. Soil. Sediment. 18 (5), 2008–2022. https://doi.org/10.1007/s11368-018-1930-6.

Tan, K., Ma, W., Chen, L., Wang, H., Du, Q., Du, P., Yan, B., Liu, R., Li, H., 2021. Estimating the distribution trend of soil heavy metals in mining area from hymap airborne hyperspectral imagery based on ensemble learning. J. Hazard. Mater. 401, 123288 https://doi.org/10.1016/j.jhazmat.2020.123288.

Tao, C., Wang, Y., Cui, W., Zou, B., Zou, Z., Tu, Y., 2019. A transferable spectroscopic diagnosis model for predicting arsenic contamination in soil. Sci. Total Environ. 669, 964–972. https://doi.org/10.1016/j.scitotenv.2019.03.186.

Wang, X., An, S., Xu, Y., Hou, H., Chen, F., Yang, Y., Zhang, S., Liu, R., 2020. A back propagation neural network model optimized by mind evolutionary algorithm for estimating Cd, Cr, and Pb concentrations in soils using vis-NIR diffuse reflectance spectroscopy. Appl. Sci. 10 (1), 51. https://doi.org/10.3390/app10010051.

Wang, F., Gao, J., Zha, Y., 2018. Hyperspectral sensing of heavy metals in soil and vegetation; Feasibility and challenges. Isprs J. Photogramm. 136, 73–84. https://doi.org/10.1016/j.isprsjprs.2017.12.003.

Wang, X., Tan, K., Du, P., Pan, C., Ding, J., 2022. A unified multiscale learning framework for hyperspectral image classification. IEEE T Geosci Remote. 60, 1–19. https://doi.org/10.1109/TGRS.2022.3147198.

Wang, X., Tan, K., Du, P., Han, B., Ding, J., 2023. A capsule-vectored neural network for hyperspectral image classification. Knowl.-Based Syst. 268, 110482 https://doi.org/10.1016/j.knosys.2023.110482.

Wang, L., Wang, R., 2022. Determination of soil pH from vis-NIR spectroscopy by extreme learning machine and variable selection: a case study in lime concretion black soil. Spectrochim. Acta A Mol. Biomol. Spectrosc. 283, 121707 https://doi.org/10.1016/j.saa.2022.121707.

Westfall P.H., 2014. Kurtosis as peakedness, 1905 - 2014. R.i.p. Am. Stat. 68(3): 191-195. 10.1080/00031305.2014.917055.

Wilford, J., de Caritat, P., Bui, E., 2016. Predictive geochemical mapping using environmental correlation. Appl. Geochem. 66, 275–288. https://doi.org/10.1016/j.apgeochem.2015.08.012.

Xin, Z., Jun, S., Yan, T., Quansheng, C., Xiaohong, W., Yingying, H., 2020. A deep learning based regression method on hyperspectral data for rapid prediction of cadmium residue in lettuce leaves. Chemometr. Intell. Lab. 200, 103996 https://doi.org/10.1016/j.chemolab.2020.103996.

Xavier, G., Yoshua, B., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Yee W.T., Mike T. (Eds.). PMLR, pp. 249-256. https://api.semanticscholar.org/CorpusID:5575601.

Yang, N., Han, L., Liu, M., 2023. Inversion of soil heavy metals in metal tailings area based on different spectral transformation and modeling methods. Heliyon. 9 (9), e19782.

Ye, B., Tian, S., Cheng, Q., Ge, Y., 2020. Application of lithological mapping based on advanced hyperspectral imager (ahsi) imagery onboard gaofen-5 (gf-5) satellite. Remote Sens. 12 (23), 3990. https://doi.org/10.3390/rs12233990.

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

Zou, B., Tu, Y., Jiang, X., Tao, C., Zhou, M., Xiong, L., 2019. Estimation of cd content in soil using combined laboratory and field ds spectroscopy. Spectrosc. Spect. Anal. 39 (10), 3223–3231. https://doi.org/10.3964/j.issn.1000-0593(2019)10-3223-09.

Zou, B., Jiang, X., Feng, H., Tu, Y., Tao, C., 2020. Multisource spectral-integrated estimation of cadmium concentrations in soil using a direct standardization and spiking algorithm. Sci. Total Environ. 701, 134890 https://doi.org/10.1016/j.scitotenv.2019.134890.