Geoderma 385 (2021) 114875

Contents lists available at ScienceDirect

Geoderma

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

Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction

Depin Ou [a], Kun Tan [b,a,*], Jian Lai [c], Xiuping Jia [d], Xue Wang [b], Yu Chen [a], Jie Li [e]

[a] MNR Key Laboratory for Land Environment and Disaster Monitoring, China University of Mining and Technology, Xuzhou 221116, China
[b] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
[c] Shanghai Institute of Satellite Engineering, Shanghai 200240, China
[d] School of Engineering and Information Technology, The University of New South Wales, Canberra, ACT 2600, Australia
[e] Nantong Academy of Intelligent Sensing, Nantong 226000, China

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Hyperspectral imagery
Soil organic matter
Heavy metals
Semi-supervised deep neural network regression
Spatial transportation and aggregation
VNIR-SWIR spectroscopy

ABSTRACT

A number of algorithms have been developed for soil organic matter (SOM) or soil heavy metal detection in airborne hyperspectral imagery with high spatial and spectral resolutions. However, to achieve improved land management, the problems of the inconsistent features and low accuracy still need to be solved. In this paper, we propose a novel regression model to estimate the concentrations of SOM, arsenic (As), and chromium (Cr) in soil. Firstly, a hyperspectral unmixing technique is utilized to extract the bare soil pixels. We then combine the absorption depth feature after continuum removal, the original absorption feature, the band ratio feature, and the first-order differential feature, to form a set of features for parameter inversion. To solve the over-fitting problem caused by the small number of samples and the weak expression problem, the semi-supervised deep neural network regression (Semi-DNNR) model is introduced. The experimental were conducted using several datasets collected by HyMap, which is an airborne hyperspectral imaging sensor in VNIR-SWIR spectral range in Yitong county, Jilin province, China. The proposed Semi-DNNR model shows a good performance in this study, with the prediction R_p^2 values for SOM, As, and Cr being 0.71, 0.82, and 0.63, respectively. After the spatial distribution map of the soil components of the study area was overlaid with the stream network, which was obtained from the digital elevation model (DEM). It was found that snowmelt, the melting of frozen soil, and surface rainfall can transport SOM to low-lying areas. A similar phenomenon was also observed for As, due to SOM adsorption and dissolved organic matter (DOM) complexation. A comparison of the proposed method with both feature selection methods (competitive adaptive reweighted sampling (CARS), genetic algorithm (GA)) and regression methods (partial least squares regression (PLSR), support vector regression (SVR)) shows that the proposed feature selection method is more robust than the CARS and GA methods. The proposed Semi-DNNR model was found to be at least 18.80% higher in prediction accuracy for As than the SVR or PLSR methods, at least 25.71% higher for Cr, and at least 19.73% higher for SOM.

1. Introduction

Mining activities bring various minerals from underground to the Earth's surface. Without good mineral processing management, this can lead to heavy metal pollution in the soil and water in the surrounding areas. Pollution by heavy metals such as cadmium (Cd), nickel (Ni), copper (Cu), arsenic (As), mercury (Hg), chromium (Cr), lead (Pb), zinc (Zn), and manganese (Mn) can threaten human life and health (Malm, 1998; Stamatis et al., 2001). At present, the monitoring of soil organic matter (SOM) and heavy metal content is mainly carried out by field sampling, which is followed by laboratory chemical analysis (Mirsal, 2008). Although this approach can obtain high-precision results, it is time-consuming and laborious, and may incur a considerable cost in high-density sampling. Hyperspectral remote sensing can provide images with high spectral and spatial resolutions, and is now being widely used for soil composition monitoring and mapping (Chabrillat et al., 2019).

SOM is one of the most important parameters of soil, and it has a significant negative influence on soil spectral reflectance. Several studies have explored the spectral response bands for SOM. For example,

* Corresponding author at: Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China.
E-mail address: tankuncu@gmail.com (K. Tan).

https://doi.org/10.1016/j.geoderma.2020.114875
Received 12 May 2020; Received in revised form 21 November 2020; Accepted 23 November 2020
Available online 16 December 2020

Weidong et al. (2002) reported that visible bands could better measure the SOM content than near-infrared bands; Galvão and Vitorello (1998) found that SOM's absorption wavelengths are at 550–700 nm; and Liu et al. (2009) observed that the good spectral response bands of black SOM are between 620 and 810 nm, with a maximal spectral response at 710 nm. Furthermore, Fichot et al. (2015) used the partial least squares regression (PLSR) method to detect the concentration of soil organic carbon with visible to near-infrared hyperspectral remote sensing imagery. They showed that the feature band range of 480–700 nm has significant predictive capabilities. Therefore, in summary, the prominent spectral signature of SOM is in the range of 550–810 nm.

Most previous studies have investigated the heavy metal(loid)s concentration based on the soil reflectance spectroscopy obtained in the laboratory. For example, Tan et al. (2018) collected laboratory spectral data of soil from coal mining areas. They investigated As, Cr, Hg, and Pb contamination via the competitive adaptive reweighted sampling/partial least squares/support vector machine (CARS-PLS-SVM) method, and the results showed that the Cr prediction performance of CARS-PLS-SVM was superior to that of wavelet transform PLS (WT-PLS), synergy interval PLS (siPLS), and the original CARS-PLS model. Gannouni et al. (2012) used the PLSR method to investigate the heavy metal pollution from mine waste in Jalta and Bougrine in northern Tunisia. They found that the ratio of the 610/500 nm range is positively correlated with Pb, Zn, and Mn, while Ni and Cr have a strong correlation at 980 nm. Meanwhile, the important wavelengths in soil spectra for soil As prediction have been reported to be near 480, 600, 670, 810, 1980, 2050, and 2290 nm (Shi et al., 2016). It is also evident that As has absorption features around 428 nm (because of the influence of Fe oxides), 1290–1310 nm (because of O—H and C—H bonds), and 2250–2450 nm (related to the C—H bonds in SOM) (Ben-Dor et al., 1997; Chakraborty et al., 2017a, 2017b). However, the feature bands of the soil components obtained by these spectral pre-processing methods are inconsistent when applied to different data sets, and it may well be even harder to achieve consistency with airborne data (Gholizadeh et al., 2015). For example, Choe et al. (2008) investigated heavy metal contamination in river sediments of the Rodalquilar gold mining area of Spain. After preprocessing the airborne hyperspectral data, it was found that the absorption depth after continuum removal at 500 nm ($Depth_{500nm}$), the ratio of 610–500 nm ($R_{610500nm}$), the ratio of 1344–778 nm ($R_{1344778nm}$), the absorption depth after continuum removal at 2200 nm ($Depth_{2200nm}$), the asymmetry of the absorption feature at 2200 nm ($Asym_{2200nm}$), and the absorption area at 2200 nm ($Area_{2200nm}$) have a strong correlation with Pb, Zn and As. However, the model accuracy and the consistency of the feature bands vary with the different heavy metal measurement methods and different field sampling areas (Shi et al., 2014; Wang et al., 2018). Hence, data-driven methods are needed to select the optimal features.

The reflectance at each band describes the spectral characteristics of the materials sensed and can be treated as a feature. However, in different remote sensing applications, it is often necessary to use different processing methods to obtain distinguishable features to separate better the classes addressed. The spectral reflectance values in the 350–2500 nm wavelength range are highly collinear. Furthermore, using a full spectrum or selecting a part of the spectrum without proper guidelines will often lead to redundant or irrelevant information in the regression (Zou et al., 2010). Therefore, many methods have been developed to explore more effective inherent features, contrasted to techniques that are used a priori techniques or model-based techniques. In general, feature selection methodologies are divided into three types, referred to as filter (Khan et al., 2017), wrapper (Granitto et al., 2006; Leardi, 2000), and embedded methods (Ou et al., 2019). Embedded methods are widely used for feature selection in soil spectroscopy, including LASSO regularization (Kukreja et al., 2006), fuzzy rule-based methods (Tsakiridis et al., 2019) and sparse SVRs methods (Tsakiridis et al., 2020a). For example, Henderson et al. (1989) proposed an optimal feature correlation algorithm developed from the shape dominancy concept of Karhunen-Loeve; Coleman et al. (1991) used correlation, regression, and discriminant analyses for the spectral band selection; and Sarathjith et al. (2016) used an ordered predictor selection (OPS) approach for selecting the optimum number of spectral variables, to improve the regression performance. Besides, many state-of-the-art spectral band selection methods have also been developed, including the modified stepwise principal component analysis (MSPCA) approach proposed by Csillag et al. (1993), variable selection with the "variable importance in projection" (VIP) method developed by Cécillon et al. (2008), the genetic algorithm (GA)-based method of Leardi (2000), and the competitive adaptive reweighted sampling (CARS) method developed by Li et al. (2009). However, the CARS and GA-based methods have no unique solution because of the Monte Carlo strategy and random numbers (Sarathjith et al., 2016). Moreover, for these methods, the selected features are all original spectral features, and their feature responses are often weak. Feature combination after pre-processing by different methods is one way to solve these problems.

The main regression algorithms used in hyperspectral estimation are PLSR (Farifteh et al., 2007), multivariable linear regression (MLR) (Kleinbaum et al., 2013), stepwise regression (SR) analysis (Thompson, 1995), geographically weighted regression (GWR) (Chi and Wang, 2017), support vector regression (SVR) (Smola and Schölkopf, 2004), random forest (RF) (Liaw and Wiener, 2002), and other regression algorithms (Khajehsharifi et al., 2017). For both soil component estimation based on laboratory hyperspectral data and airborne/satellite hyperspectral image data, most of the previous studies have used traditional statistical models. Some studies (Selige et al., 2006; Stevens et al., 2008) have shown that conventional statistical models can provide good performance for SOM. However, compared with laboratory hyperspectral data, hyperspectral images have the problems of: 1) noise corruption due to atmospheric effects; and 2) mixed pixels. It is therefore a challenge to build a model that can perform well with hyperspectral imagery. Furthermore, due to the limitation of the sample data and the low content of heavy metals in soil, the application of the traditional statistical methods in imaging spectroscopy is often not ideal.

Deep learning techniques have performed well in feature training, and have been widely used in hyperspectral image classification (Du et al., 2020; Wang et al., 2019). Recently, Some studies have successfully applied convolutional neural network (CNN) (Padarian et al., 2019b), long short-term memory networks(LSTM) (Singh and Kasana, 2019), transfer learning (Padarian et al., 2019a), and other deep learning techniques(Tsakiridis et al., 2020b) in hyperspectral soil properties prediction. Pyo et al. (2019) used a convolutional neural network (CNN) in the hyperspectral image to predict phycocyanin and chlorophyll-a in rivers, achieving R^2 values of 0.86 and 0.73, respectively. However, the retrieval of soil properties based on hyperspectral technology needs costly field sampling and soil analysis. Moreover, due to the small amount of sample data, directly using deep learning methods for feature extraction and regression is prone to over-fitting (Srivastava et al., 2014). Regression models cannot be reliably used to calibrate using small number of analyzed soil samples. On the other hand, there are a large number of "cheap" hyperspectral data that can be used. To solve this problem, semi-supervised learning is one of the techniques that can use a large number of unlabeled samples to improve the learning performance (Zhou, 2006). Several studies have addressed regression modeling using the semi-supervised learning, such as co-training (Zhou and Li, 2007b), semi-supervised least squares regression (Brefeld et al., 2006), semi-supervised regression based on SVM co-training (Lei and Wang, 2011) and manifold learning-based regression (Wang et al., 2006). Soil components are strongly correlated in the spatial domain (Tobler, 1970), and unlabeled samples around labeled samples can be collected as training data. Therefore, compared to laboratory spectroscopy, imaging spectroscopy has the great advantage of using a semi-supervised technique to select unlabeled samples.

Most of the recent hyperspectral estimation works have focused on accuracy improvement, and there has been a lack of spatial analysis of

the results. In this study, based on these considerations, we aimed to: 1) obtain more robust and reliable feature bands by the use of a new feature band combination approach; 2) build a semi-supervised deep neural network regression (Semi-DNNR) model; and 3) conduct a hydrological analysis to study the spatial adsorption and transportation of the soil components in the study area.

The rest of this paper is organized as follows. Section 2 describes the study area, datasets, and methods; Section 3 provides the results and analysis; Section 4 provides a discussion; and our conclusions are drawn in Section 5.

2. Datasets and methodology

2.1. Study area

Yitong County is located in the southern part of Jilin province, China. It belongs to the Songnen Plain, part of which is saline-alkali land, and its soil is acidic. The climate is a cold temperate monsoon climate, with an average temperature of −12.9 °C in winter and 21.6 °C in summer. The average annual precipitation is 632.3 mm, mainly concentrated in the summer, accounting for 68% of the annual total. The average annual snowfall days and snowpack days are 38 days and 64.3 days, respectively. The predominant wind direction is southeasterly. Mineral resources are abundant in the area, including more than 30 minerals, such as gold (Au), silver (Ag), Cu, and iron (Fe). Our study area (125.32°E–125.46°E, 43.22°N–43.33°N), covering about 139 km^2 (as shown in Fig. 1), contains two gold mines, a large amount of agricultural land, many villages, and a few industrial plants. The Yitong River crosses the entire study area from southeast to northwest. The area is gently undulating, and the average elevation is 305 m, with a minimum elevation of 215 m and a maximum elevation of 430 m. Agriculture plays a dominant role in the study area, with corn being the main crop, and rice is also grown.

2.2. Datasets and field sampling

A total of nine airborne hyperspectral image strip datasets were acquired between April 18 and April 22, 2017, using a HyMap airborne imaging spectrometer. During this period, some of the farmland had been ploughed to a depth of around 20 cm. The pre-processed hyperspectral images were made up of 2734 rows, 2508 columns, and 135 bands over the 466–2470 nm wavelength range, with a 10–20-nm spectral resolution and a 4.5-m spatial resolution. Simultaneous field sampling was also undertaken, following the sample distribution shown in Fig. 1. The spatial distribution of the sampling points was based on the rule that a spatial grid was used for the segmentation, with each grid cell containing at least one sampling point, so that the collected soil samples could reflect the overall soil information of the study area. The two gold mining areas in this area were the key research objects, so we increased the density of the sampling points around the mining areas. A total of 400 g of soil from a depth of about 5 cm was collected from each sampling point. At each sampling point, a 1-m rectangle was formed, and soil samples were taken from the top, bottom, left, right, and center points, to form each soil sample. Simultaneously, high-precision coordinate information for each soil sample was acquired by real-time kinematic (RTK) positioning. Finally, 95 soil samples were obtained from the farmland in this study area.

2.3. Airborne hyperspectral image pre-processing and soil sample analysis

The original hyperspectral images were first radiometrically calibrated using the standard data obtained by an integrating sphere, converting the digital number (DN) values into radiometric values. A lookup table was constructed using a high-precision position and orientation system (POS) data and digital elevation model (DEM) data, to perform the strip-by-strip geometric correction. Ground control points were then used for precise geometric correction, to obtain orthophoto images. Accurate parameters were acquired, including the primary conditions of the atmosphere in the study area. The MODTRAN4 atmospheric radiation transmission model (Berk et al., 1999) was used to perform atmospheric correction of all the orthophoto images. Due to the inconsistency of the acquisition time and attitude between strips, there were noticeable radiation differences among the different strips. This error was corrected using the bidirectional reflectance distribution function (BRDF)-based photometric correction algorithm (Yu et al., 2017). Finally, the multi-strip airborne hyperspectral images were combined

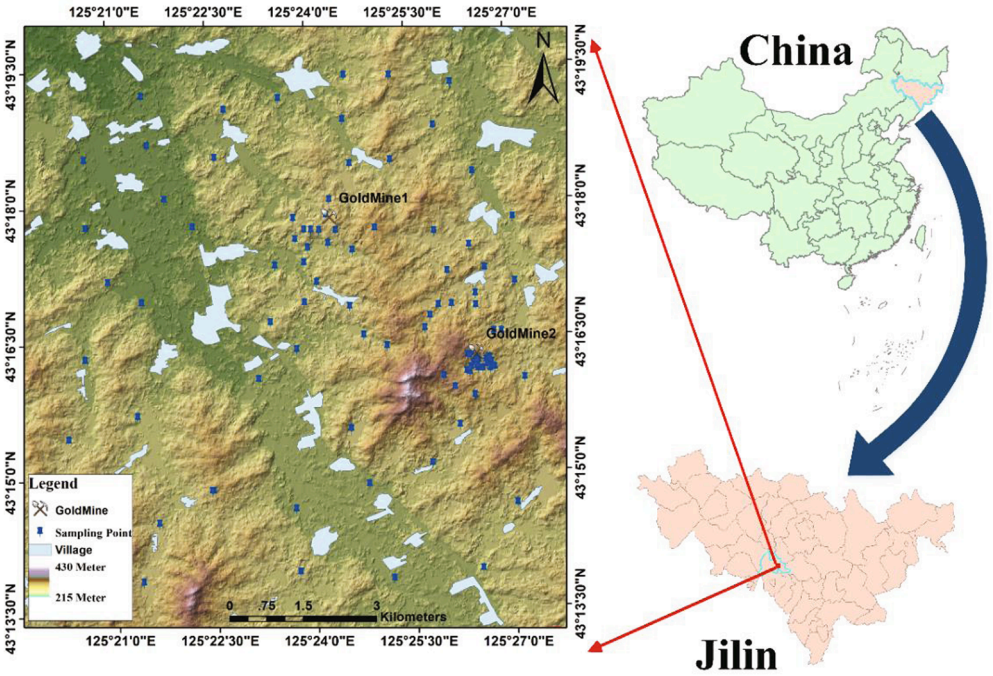

Fig. 1. Study area and field sampling points.

using the seamless mosaicking method to obtain a reflectance image of the whole study area. All the processing was undertaken in HyMap-C™ processing software. It is known that the absorption of water vapor results in low reflectance near 1400 nm and 1800 nm (Guanter et al., 2006). Therefore, the bands ranging over 1355–1514 nm and 1788–1996 nm were removed, and a total of 101 bands were finally retained in this airborne hyperspectral image. Fig. 2 shows the hyperspectral image after radiometric correction and atmospheric correction. Finally, 95 bare soil spectra were collected from the HyMap image based on the high-precision RTK coordinate information.

The 95 soil samples were treated by impurity removal, air drying and grinding, and 100-mesh screening. The concentrations of As and Cr for the 95 soil samples were then measured by inductively coupled plasma mass spectrometry (ICP-MS). Simultaneously, the SOM content for 93 samples (two samples were missing) was determined by the potassium dichromate volumetric method. Table 1 lists the statistical information for the SOM and heavy metal contents, wherein the units for SOM are g/kg and the units for the heavy metals are mg/kg. The Std value can reflect the dispersion of the soil samples. The Std values for As and Cr are high, indicating that the soil samples show a high degree of dispersion. The variation can be also reflected in the skewness and kurtosis values. The skewness and kurtosis values of SOM are relatively low, while the kurtosis value of As is up to 14 and the kurtosis value of Cr is up to 10. Table 2 shows that the Pearson correlation coefficient between SOM and the heavy metals is very low.

Table 1
Basic statistical information for the soil organic matter and heavy metals in the soil samples.

Statistic	SOM(g/kg)	As(mg/kg)	Cr(mg/kg)
Max	49.8369	419.9602	4617.558
Min	14.7571	6.3509	36.0396
Mean	30.5052	42.7515	399.4259
Std	6.3840	67.9779	864.1015
Skewness	0.1211	3.4761	3.3104
Kurtosis	0.6896	14.1855	10.7027

Table 2
Pearson correlation coefficient between soil organic matter and heavy metals.

	As	Cr	SOM
As	1		
Cr	−0.08019	1	
SOM	0.033935	0.102214	1

2.4. Methodology

2.4.1. Unmixing for the soil mask

This study was mainly focused on the pollution of the farmland soil in the study area. Therefore, it was necessary to extract the soil information before using the airborne hyperspectral data to perform the estimation task, and to build a mask to reduce the error caused by the seriously mixed pixels. The methods for generating a soil mask include classification-based and unmixing-based methods. For example, the combination of an ancillary spectral index (Nor-malized Burn Ratio 2, NBR2) and the normalized difference vegetation index (NDVI) have successfully applied in extracting the bare soil pixels (Demattê et al., 2018). In this study, NDVI and cellulose absorption index (CAI) threshold methods were also utilized to extract most of the bare soil pixels, but the soil-vegetation mixed boundary pixels and soil-building mixed pixels were often misclassified. We therefore used an unmixing method to obtain the soil abundance information. A superpixel can be defined as a group of spatial connected similar pixels in a local area (Ren and Malik, 2003). For the hyperspectral imagery, superpixels were first generated using the simple linear iterative clustering (SLIC) method proposed by Achanta et al. (2012). To reduce the time consumption, the vertex component analysis (VCA) (Nascimento and Dias, 2005) was used to extract the endmembers, and the fully constrained least squares (FCLS) (Heinz, 2001) method was applied to extract the soil abundance map. Because the soil from the village areas may have been affected by human activities, the village areas were masked manually. The boundary layers of the village were created using the ESRI® ArcGIS software. We then expanded it with a buffer zone of 10 m.

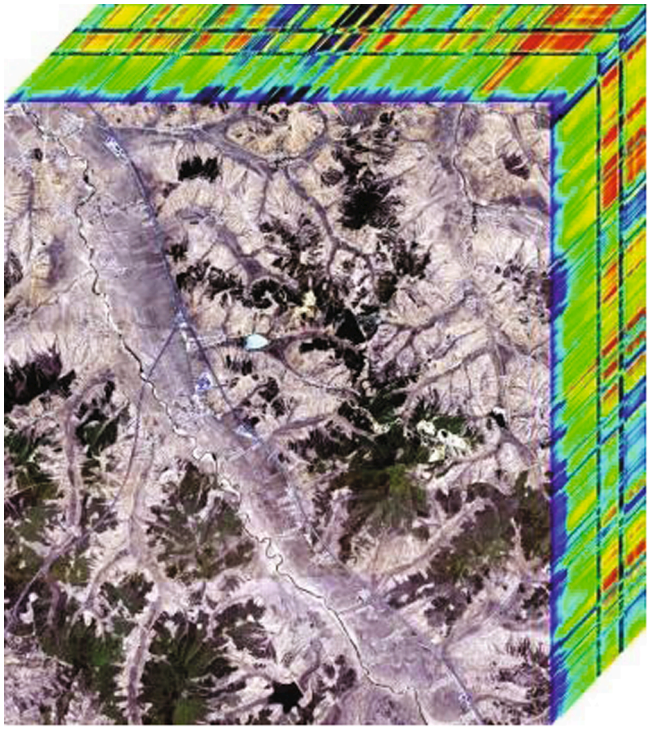

Fig. 2. Hyperspectral image after radiometric correction and atmospheric correction.

The field survey found that there were five types of ground objects in the study area: soil, vegetation, water, building, and roads. Therefore, the number of endmembers was set to five, but the parameters in the SLIC and VCA were set to the default. Since endmember extraction from the entire mosaicked hyperspectral image would require a large amount of memory, a subset of the image (size: $1142 \times 704 \times 101$) containing all the ground objects of the whole study area was used as the input data for the endmember extraction processing. Finally, based on the obtained endmembers, the abundance information of the whole image was calculated by FCLS, and then the bare soil pixels (obtained by applying a threshold of 0.7 in the soil abundance) were extracted.

2.4.2. Band selection and feature extraction

Soil heavy metals are often weakly correlated or uncorrelated with the original airborne hyperspectral data, so that it is difficult to find a good model using the original hyperspectral signal (Rinnan et al., 2009). The original spectra can be pre-processed by the first-order differential (Amigo et al., 2015), band ratio (Groves and Bajcsy, 2003), and continuum removal (Rezaei et al., 2008) methods, which can reduce the influence of noise and other interfering factors and highlight the feature information, to a certain extent. For example, Gholizadeh et al. (2015) used various spectral pre-processing methods to process visible and near-infrared spectra in the laboratory. They found that the first-order differential had the best predictive ability for heavy metals such as Cu, Mn, Pb, and Zn, while multiplicative scatter correction (MSC) and standard normal variate (SNV) pre-processing showed a weak predictive ability. Some recent studies (Tsakiridis et al., 2020b, 2019) have demonstrated that combining spectral pre-processing techniques

performed better than training with a single pre-processing technique. Therefore, we comprehensively applied various spectral pre-processing methods to find the optimal combination of features. The combination rules were as follows.

1) Calculate the absorption depth after continuum removal processing, and obtain the correlation between each band's depth and the soil components. The band depths with the optimal correlation are then selected. The continuum is removed by dividing it into the actual spectrum for each pixel in the image.

$$S_{cr} = S/C \quad (1)$$

$$D_{cr} = 1 - S_{cr} \quad (2)$$

where S_{cr} is the continuum-removed spectrum, S is the original spectrum, C is the continuum curve, and D_{cr} is the absorption depth after continuum removal.

2) Calculate all the band ratios between each pair of bands (in this study, the total number of band ratios was 10,201 (101 × 101 pairs)). Obtain the correlation with the soil components, and retain the highest band ratio for each band combination. The optimal band ratio is selected by combining the correlation and the frequency of the band occurrence. The optimal original band is also selected at the same time.
3) Select the optimal bands after first-order differential processing.
4) Combine all the above features (continuum removal feature, band ratio feature, original band feature, first-order derivative feature). The feature combinations tend to cause data redundancy, due to the high similarity of the various features, so it should remove those duplicated features.

2.4.3. The semi-supervised deep neural network regression model

Deep learning techniques, especially CNNs, have shown reliable performances in hyperspectral image processing. Compared with a deep neural network (DNN) model, the CNN model utilizes convolutional layers to capture the features, which consumes more time (Du et al., 2020). Therefore, we utilized a DNN as the regressor to learn the deep features. The DNN used in this study was set to six layers and one output, as shown in Fig. 3. Batch normalization was used for normalization before each linear layer, to avoid the DNN from updating in the wrong direction due to noisy samples. According to the characteristics of the regression analysis and the hyperspectral estimation task, a rectified linear unit (ReLU) activation function in the DNN is more suitable. To prevent or slow down the over-fitting phenomenon, we added a dropout function to the network, with a dropout rate of 0.5.

The training samples and test samples were divided by a ratio of 2:1. With the training samples being very few in number, it can easily lead to the over-fitting problem in DNN training. Tobler's first law of geography clearly states that "everything is related to everything else, but near things are more related to each other" (Tobler, 1970). Hence, the semi-supervised deep neural network regression model (Semi-DNNR) constructed based on this law is more suitable for labeling the unlabeled samples, as well as obtaining a better training model. Fig. 4 shows the semi-supervised model framework, which is similar to Zhou's co-training style semi-supervised regression (COREG) algorithm (Zhou and Li, 2007a). However, this framework adds spatial similarity to make it more consistent with airborne hyperspectral remote sensing estimation. In the Semi-DNNR architecture, only a few parameters need to be adjusted. Firstly, all the trainable parameters in the network are for the three DNN regression models. The parameter tuning of DNN architecture is essential since there are only a few training data in the first iteration. Secondly, only a small number of the parameters exist in the proposed Semi-DNNR architecture, including threshold E, threshold *Delta* value, and the precision threshold. The Semi-DNNR algorithm pseudo-code is shown in Table 3, and the detailed rules are as follows.

2.4.3.1. Training set and test set. The training set and test set are divided before being input into the semi-supervised regression model. The division rule is: firstly, the SOM or heavy metal values are sorted according to the concentration gradient, and then three adjacent samples are taken as a group. For each group, two samples are selected randomly as the training set, and the one remaining sample is used as the test set, which can ensure the uniform distribution of the training samples.

2.4.3.2. Model. Model_1 and Model_2 are designed for sample augmentation, and their role is to ensure the validation of the additional labeled samples. If the predicted values obtained by Model_1 and Model_2 for the same sample are consistent, the reliability of this sample and its pseudo values is higher. Model_3 is set to further determine whether the selected pseudo-labeled samples are valid. Accuracy comparison is a more effective method, so the accuracy of Model_3 can be compared with the accuracy of Model_1 or Model_2. In this work, the accuracy of Model_1 is used as a reference for comparison. Hence, Model_1 and Model_3 are identical in both network architecture and hidden layer parameters. For each training operation, we use the k-fold cross-validation method to obtain the accuracy of each model, with k set to 10. The training sets are randomly divided into a training set (90%) and a validation set (10%). Finally, the mean value of accuracy across all folds is taken as the final accuracy of the model.

2.4.3.3. Candidate set. Due to the large amount of hyperspectral image data, where unlabeled samples account for the vast majority, it is difficult to guarantee the reliability of the pseudo-labels in the unlabeled samples. Therefore, the eight-neighborhood samples of all the samples are selected as candidate sets. Because their spatial and spectral characteristics are closer to the true labeled samples, they are more effective

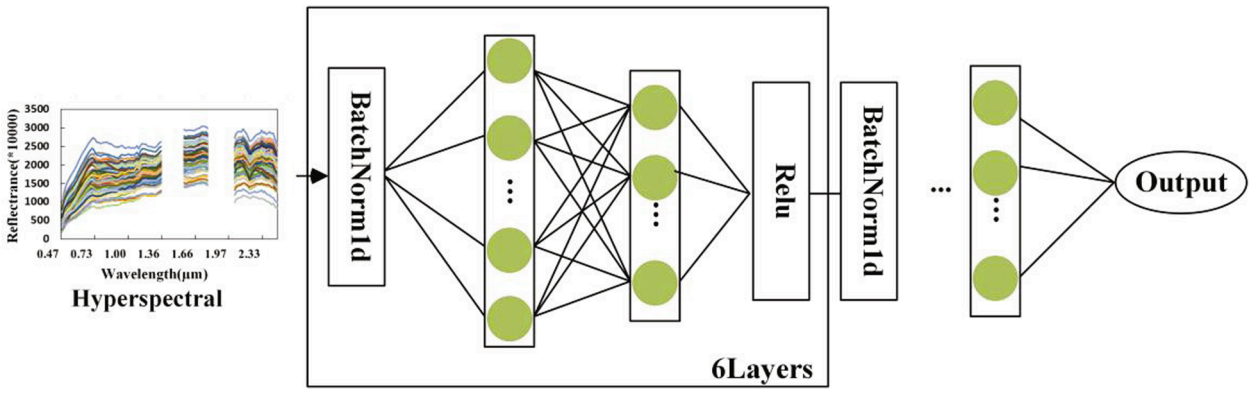

Fig. 3. The DNN regression model.

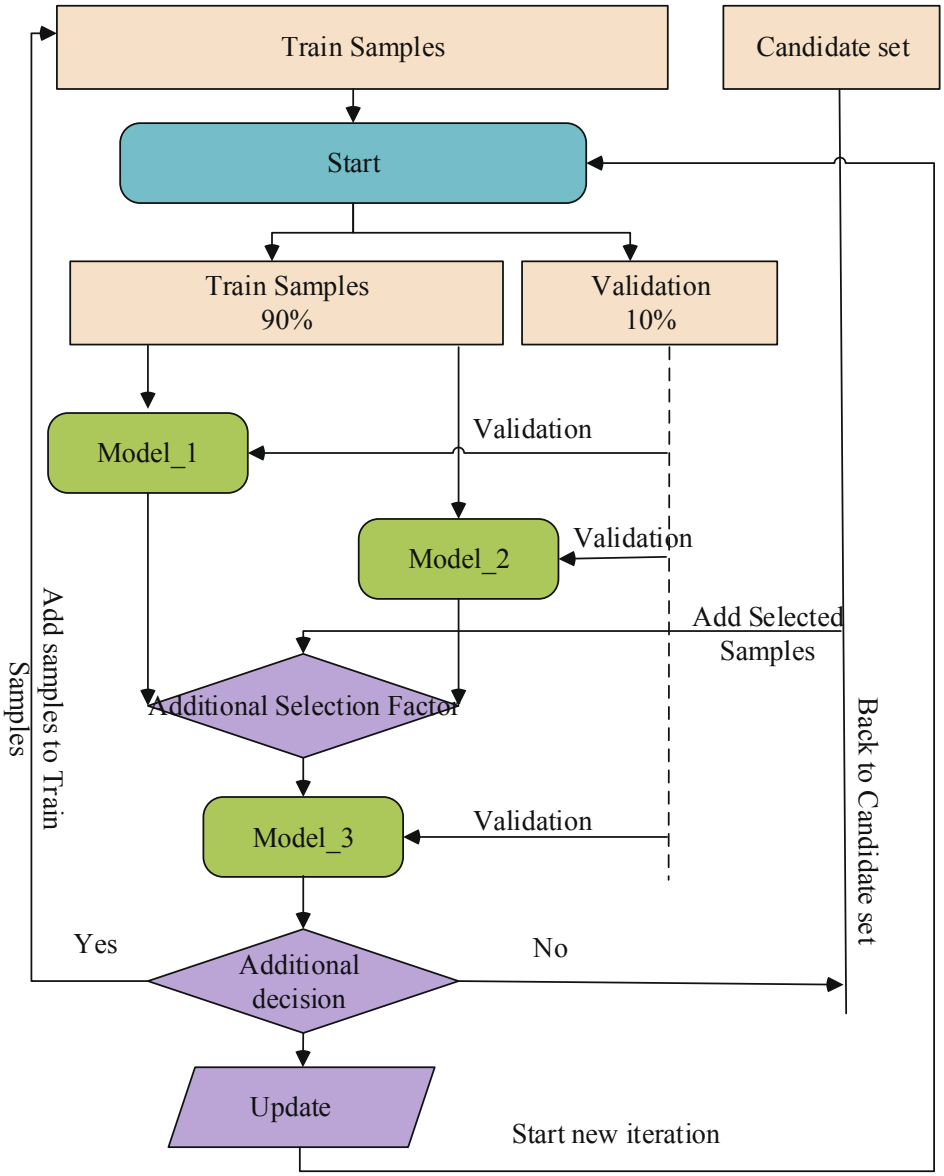

Fig. 4. The Semi-DNN framework.

in judging the accuracy of the pseudo-labels.

2.4.3.4. Additional selection factors. How to add reliable pseudo-labeled samples as training samples for the model training is very important. Firstly, we compare the candidate sets predicted by Model_1 and Model_2 in the previous step. If the difference between the values predicted by Model_1 and Model_2 for the same position is smaller than a threshold E, the pseudo-labeled sample is considered to be reliable.

$$|V_{m1} - V_{m2}| \leq E \tag{3}$$

where V_{m1} is the value predicted by Model_1, and V_{m2} is the value predicted by Model_2.

The principle of kriging interpolation states that the concentration of a sample component should be related to its adjacent spatial samples (Kleijnen, 2009). Therefore, taking into account the spatial distribution of the soil components, the SOM or heavy metal content of the samples should not be significantly different from that of the adjacent spatial samples, and should show spatial continuity. Consequently, the spectral angle, the distance, and the predictive difference between the unlabeled samples and their adjacent real labeled samples are used as the reliability criteria for the pseudo-labeled samples. The spectral angle is defined as follows.

$$\mathrm{SA}(X^*, X_i) = \cos^{-1}\left(\frac{(X^*)^{\mathrm{T}} X_i}{((X^*)^{\mathrm{T}} X^*)^{1/2} (X_i^T X_i)^{1/2}}\right) \tag{4}$$

where X^* represents the unlabeled samples and X_i is the nearest sample of X^*. The smaller the value of $\mathrm{SA}(X^*, X_i)$, the higher the similarity between the unlabeled sample and the nearest true labeled sample.

As regards the distance between samples, we use the Euclidean distance, which is defined as:

$$dist(X^*, X_i) = \sqrt{|X^* - X_i|^2} \tag{5}$$

where $dist(X^*, X_i)$ is the Euclidean distance between unlabeled sample X^* and its nearest labeled sample X_i. The smaller the distance, the closer the estimation result is to the real true value.

The definition of the difference between the unlabeled sample and its nearest labeled sample is:

Table 3
Pseudo-code of the Semi-DNNR algorithm.

ALGORITHM: Semi-DNNR

Input: labeled training set L, labeled validation set T, candidate sample set U,
Deep regression networks DNN_1, DNN_2, DNN_3,
Maximum number of iterations N,
Sample selection threshold *Delta*,
Deep regression network prediction threshold E
True value of the nearest labeled sample V
Process:
Repeat for N rounds:
$h_1 \leftarrow DNN_1(L); h_2 \leftarrow DNN_2(L)$
For each $X^* \in U$ **do**
$\widehat{y}_{1i} \leftarrow h_1(X^*); \widehat{y}_{2i} \leftarrow h_2(X^*)$
$\mathrm{SA}(X^*) = \mathrm{SA}(X^*, X_i) = \cos^{-1}\left(\frac{(X^*)^{\mathrm{T}} X_i}{((X^*)^{\mathrm{T}} X^*)^{1/2} (X_i^{\mathrm{T}} X_i)^{1/2}}\right)$
$dist(X^*, X_i) = \sqrt{|X^* - X_i|^2}$
$D_{rms}(X^*) = D_{rms} = \sqrt{\frac{1}{2}\left(\left(\widehat{y}_{1i} - V\right)^2 + \left(\widehat{y}_{2i} - V\right)^2\right)}$
if $|\widehat{y}_{1i} - \widehat{y}_{2i}| \le E$ **then**
$Delta(X^*) = \mathrm{SA}(X^*) * dist(X^*) * \frac{D_{rms}(X^*)}{V}$
else continue
if $Delta(X^*) \le Delta$ **then**
$\widehat{y}_i^{'} = (\widehat{y}_{1i} + \widehat{y}_{2i})/2, U^{'} \leftarrow \widehat{y}_i^{'}$
else pass
End of For
$L^{'} \leftarrow U^{'} \cup L$
$h_{1a} \leftarrow DNN_3(L^{'})$
if $R(h_{1a}(T)) > R(h_1(T)) \& R(h_{1a}(T)) > R(h_2(T)) \& R(h_{1a}(T)) > (\mathrm{R(last\ precision)} + 2\%)$
then $L \leftarrow U^{'} \cup L$
else pass
Update $(h_1(L), h_2(L))$
end of Repeat
Output:regressor h_1, h_2, h_{1a}

$$D_{rms} = \sqrt{\frac{(V_{m1} - V)^2 + (V_{m2} - V)^2}{2}} \tag{6}$$

where V_{m1} represents the values predicted by Model_1, V_{m2} represents the values predicted by Model_2, and V is the true value of the nearest labeled sample. The smaller the value of D_{rms}, the smaller the difference between the predicted values of the two models. Taking into account all of the above considerations, the *Delta* value is defined as the criterion for selecting additional samples:

$$Delta = \mathrm{SA} * dist * \frac{D_{rms}}{V} \tag{7}$$

where the smaller the spectral angle (SA) and the distance*dist*, the closer the estimation results are considered to be to the true value.$\frac{D_{rms}}{V}$ considers the ratio of the difference, which is more reasonable than directly using D_{rms}. Because the samples with large differences in distribution cause the additional samples to tend to the sample median, this makes the additional samples unable to represent the true conditions of most of the samples, resulting in over-fitting. Finally, a threshold can be used to limit the number of additional samples, preventing unreliable samples being added to the training set, thereby avoiding the problem of the low precision of the training model. The labeled values for all the unlabeled samples are the average of the two predicted values.

2.4.3.5. Additional decision. After the sample addition is performed under the above rules, the added samples still cannot be guaranteed to be reliable. Therefore, it is necessary to use the same regression as Model_1 for a new round of model training. If the accuracy after sample addition is higher than that of all the previous models, and its precision is 2% (this threshold is set according to the actual situation) higher than the highest precision of all the previous rounds, the selected samples are considered to be effective, and they are added to the training sample set; otherwise, they are returned to the candidate sample set.

2.4.3.6. Updating. Given the difference between regression analysis and classification, and due to the addition of pseudo-labeled samples in the model training process, pseudo-labeled sample updating is needed to ensure they have high precision. When the validation accuracy of the trained model is higher than that of the previous training round, the added pseudo-labeled samples in the training set should be updated. Depending on the characteristics of the deep learning network, this operation is also applicable if no suitable samples are added, allowing the model to be fine-tuned for self-training.

The semi-supervised regression method aims to obtain the optimal regression model, so the model with the highest accuracy is the final prediction model. This is because, in semi-supervised regression training, these three models can self-train to fine-tune the parameters and obtain better accuracy. A model with the same network structure can use the same parameters (In each iteration, the previous optimal parameters are used as the initial parameter settings). One reason for this is to ensure the accuracy of the model, using the added pseudo-labeled samples to fine-tune the network, and the other reason is to reduce the training time.

The Adam optimizer (Kingma and Ba, 2014) is used as the optimizer in the constructed DNN regression network, and its optimal step size is set to 2e − 4. According to the characteristics of regression analysis, the loss function is evaluated by the mean square error (MSE), which is consistent with the final accuracy evaluation in the estimation results. Multiple experiments showed that the optimal parameter settings for the different soil component estimation models in semi-supervised DNN regression are different, as shown in Table 4. A larger neuron number will lead to more massive training time, and can easily cause over-fitting due to the few samples. In comparison, a smaller neuron number will mean that it is more difficult to obtain a good result, so the parameter of the model in Table 4 is the final number of neurons. For the setting of the neuron number in each layer of the DNN network, the auto-coding idea is adopted. According to the regression problem, the DNN is set to encoding style. It can be seen from Table 1 that the Std of Cr is large, which means that it shows a high degree of spatial dispersion. It is therefore difficult to obtain a better training model if some abnormal samples are added. Thus, according to the normal distribution characteristics of the Cr samples, the Cr samples less than or equal to 300 mg/kg were divided into a training set and test set. Sample removal was not performed for As and SOM. Taking into account the dispersion of the different components, parameter *Delta* is inconsistent. The *Delta* value can eliminate most of the pseudo-labeled samples that do not meet the requirements, but there will still be a certain number of pseudo-labeled samples. Therefore, the maximum number of added samples was set to 100 in each sample adding operation, so as to avoid disturbance to the

Table 4
The parameters of Semi-DNNR.

Parameter	Component		
	As	SOM	Cr
Model_1/Model_3	input, 120, 80, 60, 40, 40, 20, output		
Model_2	input, 120, 90, 80, 60, 40, 20, output		
Training samples	63	62	50
Test samples	32	31	25
Filter values	≤420 mg/Kg (All)	≤50 g/Kg (All)	≤300 mg/Kg
Delta	0.3	0.02	0.15
E	40		
Batch size	8		
Maximum additional samples	100 per iteration		

accuracy by adding too many abnormal pseudo-labeled samples.

2.5. Model evaluation and analysis method

2.5.1. Estimation model evaluation method

The Semi-DNN method was implemented in the PyTorch (Paszke et al., 2017) deep learning framework, and the other methods were implemented in Python. The evaluation indices for the modeling are as follows.R_c^2, *RMSEc*, *MAEc*, *Biasc*, and *RPIQc* are the evaluation indices for the training set, while R_p^2, *RMSEp*, *MAEp*, *Biasp*, and *RPIQp* are the evaluation indices for the testing set. The optimal number of hidden layers was established through experiments, where the hidden layer nodes were set to 40. All the experiments were carried out on a personal computer with an Intel(R) Core(TM) i7-7700 CPU, 16 GB of RAM.

1) Coefficient of determination, R^2: Used to indicate the correlation between the predicted value and the real value. The closer the value is to 1, the better the prediction.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

2) Root-mean-square error, *RMSE*: The standard deviation of the residuals (prediction errors). It is a measure of how spread out these residuals are. The smaller the value, the higher the accuracy, and the better the prediction.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{9}$$

3) Mean absolute error, *MAE*:Average value of the difference between the predicted value and the real value. The smaller the value, the higher the accuracy, and the better the prediction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{10}$$

4) Bias: Average difference between the predicted value and the real value.

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{11}$$

5) Ratio of performance to inter-quartile distance, *RPIQ*: A larger *RPIQ* value indicates improved model performance.

$$RPIQ = \frac{Q3 - Q1}{RMSE} \tag{12}$$

where$\hat{y}_i$ is the predicted value of the sample and$y_i$is the real value of the sample. Q1 is the value below which we can find 25% of the samples; Q3 is the value below which we find 75% of the samples.

2.5.2. Hydrological analysis

Earth surface information is useful in many areas where it is important to understand the impact of water flow in the area, such as regional planning, agriculture, and forestry. Currently, surface stream networks are usually generated by the use of a DEM. In order to analyze the accumulation and transportation behavior of SOM and heavy metals in the study area, ASTER GDEM 2 data (Tachikawa et al., 2011) with a resolution of 30 m (released by the Ministry of Economy, Trade, and Industry (METI) of Japan and the United States National Aeronautics and Space Administration (NASA)) were used to generate a stream network. The stream network was obtained by pit-filling, a flow direction grid, flow accumulation, and threshold processing. The DEM hydrological analysis method used was a common method: pre-processing using a pit-filling algorithm (Jenson and Domingue, 1988) and flow direction calculation using the D-8 algorithm (O'Callaghan and Mark, 1984). All the processing was completed in ArcGIS software. The main formula for the hydrological analysis is shown in the following equation:

$$Max\_Drop = \frac{CZ\_Value}{D} * 100 \tag{13}$$

where *Max_Drop* is the maximum descent value, *CZ_Value* is the change value between each pixel and its eight neighbors, and *D* is calculated between the cell centers. If the cell size is 1, the *D* between two orthogonal cells is 1, and the *D* between two diagonal cells is 1.414 (the square root of 2).

3. Results and analysis

3.1. Feature band analysis

Original hyperspectral data often show a weak correlation with soil components (Rezaei et al., 2008). The results after continuum removal and first-order differential pre-processing are shown in Fig. 5. It can be seen that Fig. 5a (the SOM correlation after continuum removal pre-processing) and Fig. 5d (the SOM correlation after first-order differential pre-processing) are consistent in their peaks, and both show high correlation around 0.60 μm and 2.20 μm. However, the SOM after first-order differential pre-processing shows a good correlation around 0.89 μm, reaching a prediction R_p^2 value of 0.55. Hence, the first-order differential pre-processing in SOM is more effective at acquiring feature bands. Fig. 5b and e are the As correlation after continuum removal and first-order differential pre-processing, respectively. There are local peaks at 1.07 μm, 1.66 μm, and 2.24 μm, but with low correlation, and the feature bands obtained by the continuum removal and first-order differential pre-processing methods are not the same. Fig. 5c and f are the Cr results, which show a lower correlation than As, mainly around 0.62 μm.

It can also be seen from Fig. 5 that it is difficult to find stable feature bands for As and Cr after continuum removal and first-order differential pre-processing. Fig. 6 shows the SOM, As, and Cr correlation diagrams and frequency diagrams obtained by the band ratio method. It can be seen that the overall correlation performance for SOM reaches a prediction R_p^2 value of 0.6, while that for As and Cr reaches 0.4, which shows that the band ratio method is more efficient than the continuum removal and first-order differential methods. From Fig. 6a, it can be seen that the correlations between SOM and reflectance are similar for many wavelengths. The reason for this is that the frequency of the highest correlations is concentrated in several bands (0.57 μm, 0.67 μm, 0.75 μm, 0.89 μm, and 2.22 μm), which can be seen in Fig. 6d and g. In addition, the difference between adjacent bands is small, so the highest correlations at adjacent bands will be similar. From Fig. 6g, although the highest SOM correlation in band ratio has a frequency at 0.57 μm, its correlation is about 0.45, which is relatively low, so the feature band of SOM at 0.57 μm can be eliminated. At 0.89 μm, the correlation reaches the highest, at a prediction R_p^2 value of 0.65, which is consistent with the first-order differential pre-processing. There are also bands with a relatively high correlation at 0.67 μm, 0.75 μm, and 2.2 μm. According to these results, it can be considered that the feature bands of SOM are found at 0.67 μm, 0.75 μm, 0.89 μm, and 2.2 μm. For As, there is a higher frequency at 0.46–0.62 μm, but the average correlation is 0.25, so this is

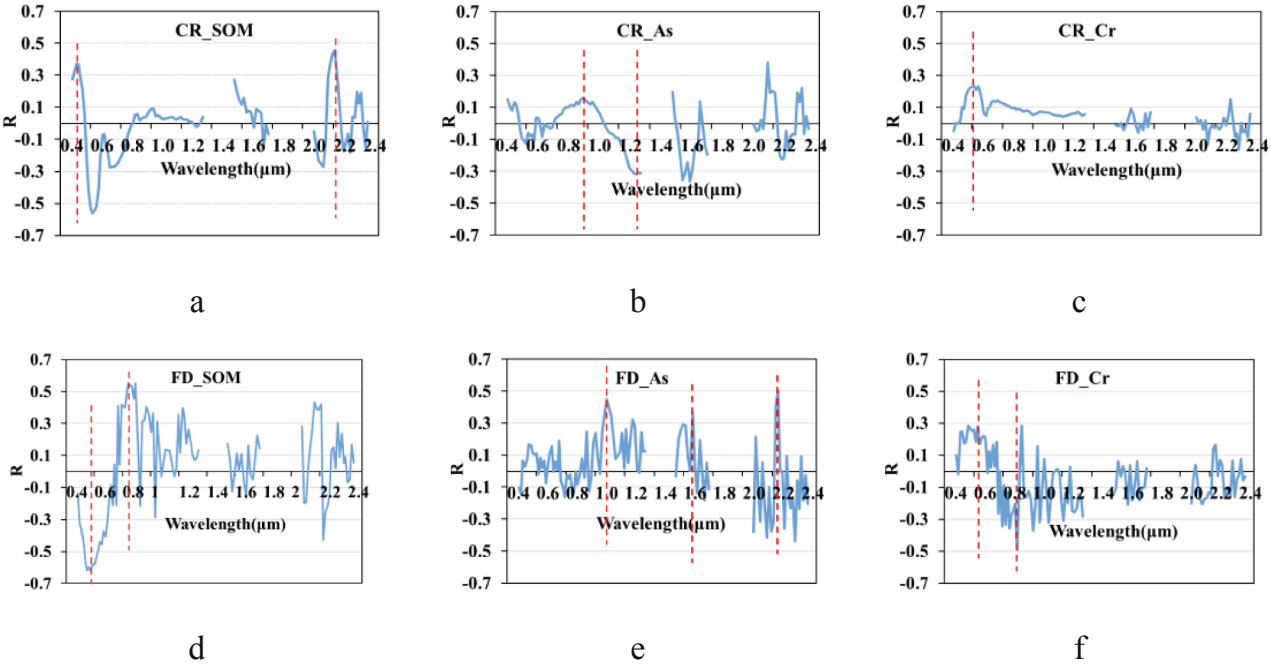

Fig. 5. Continuum removal and first-order differential pre-processing correlation analysis. (a) SOM correlation after continuum removal pre-processing. (b) As correlation after continuum removal pre-processing. (c) Cr correlation after continuum removal pre-processing. (d) SOM correlation after first-order differential pre-processing. (e) As correlation after first-order differential pre-processing. (f) Cr correlation after first-order differential pre-processing.

not considered. The highest correlation is mainly found in the vicinity of 1.0 μm, near 2.22 μm, and around 2.35 μm, with the highest correlation of 0.50 seen at 2.22 μm. As regards Cr, the best result is at 2.22 μm, and there are also corresponding features at 0.98 μm, 1.18 μm, and 2.40 μm. In summary, the feature bands of SOM, As, and Cr show good performance at 2.20 μm, but differ in the other feature bands. SOM mainly focuses on the infrared and near-infrared bands; As is dispersed, but is concentrated in the short-wave infrared band; and Cr is concentrated between the near-infrared to short-wave infrared region, and its most important feature band is around 2.20 μm.

Due to the small amount samples of SOM, As, and Cr, a single index can easily lead to model training failure. Therefore, in this study, it was necessary to comprehensively combine the selected different components, including continuum removal features, band ratio features, original feature bands, and first-order differential features. This feature pre-processing serves as a guide and an assistance to the deep layers to learn discriminant features more effectively, leading to faster convergence of the deep neural network model. From Figs. 5 and 6, it can be seen that the features obtained by the band ratio method have a higher correlation coefficient than the continuum removal and first-order differential methods, which means that the features of band ratio were more suitable for model training. According to the rule described in Section 2.4.2, the final feature combination is shown in Table 5. It can be seen from this table that the features of SOM are mainly found at 0.57 μm, and around 0.67 μm, 0.75 μm, 0.89 μm, and 2.21 μm. These features are consistent with published literature(Galvão and Vitorello, 1998; Liu et al., 2009; Rossel and Behrens, 2010; Soriano-Disla et al., 2014), that the prominent spectral signature of SOM is in the range of 550–810 nm, and around 1100 nm and 2200–2400 nm, there are absorptions by the C–H, C-O, and C-N functional groups that dominate in organic matter. The features of As are all short-wave infrared, and are mainly found in the vicinity of 1.07 μm, 1.22 μm, 2.22 μm, and 2.35 μm. And the features of Cr are mainly found at 0.58 μm, and around 0.89 μm, 0.98 μm, 2.22 μm, and 2.40 μm.

3.2. Semi-supervised deep neural network regression model analysis

In the Semi-DNNR network, as the number of iterations increases, the regression accuracy of the test samples is gradually increased by adding pseudo-labeled samples or self-updating, as shown in Fig. 7. The figure indicates that the proposed Semi-DNNR network can effectively fine-tune the model to improve the prediction accuracy. For SOM, As, and Cr, additional samples can be effectively selected, and with the improvement of the model accuracy, the selected samples are re-predicted by self-updating, which allows the model to develop in a better direction.

Table 6 and Fig. 8 show the optimal model accuracy information and scatter plots of the prediction results for SOM, As, and Cr, respectively. From Table 6, the training model accuracies (R_c^2) for SOM, As, and Cr all reach 0.90 or more, which indicates that the training was sufficient. The prediction accuracy R_p^2 of the SOM model is 0.71, the accuracy of the As model is 0. 82, and the accuracy of the Cr model is 0.63. This represents an excellent result in the airborne hyperspectral soil composition estimation field. It also shows that the selected feature bands are representative and accurate, and that the model training ability of the Semi-DNNR network is good. The combination of feature bands and the Semi-DNNR network can complement each other in the hyperspectral estimation of soil composition, and can obtain results with reliable accuracy. It can be seen in Table 5 that the RMSEs of As and Cr are both higher than that of SOM, because of the high degree of dispersion. The highest concentration of As in the soil samples participating in the training is 419.96 mg/kg, and the minimum is 6.35 mg/kg. Meanwhile, the highest concentration of Cr is 297 mg/kg, and the minimum is 36.04 mg/kg. Fig. 8 clearly shows that the predictions of the three soil components are all around the prediction lines, indicating that the models show a good predictive performance. For SOM, the predictive ability for the training set is very strong. Although the predictive ability for the test set is relatively low, the prediction values of all the samples are concentrated around the prediction line, and the over-fitting phenomenon is not apparent. For As, it can be seen that most of the samples are concentrated at lower concentrations, so that there is a certain difference in the higher concentration samples, which is also the reason for

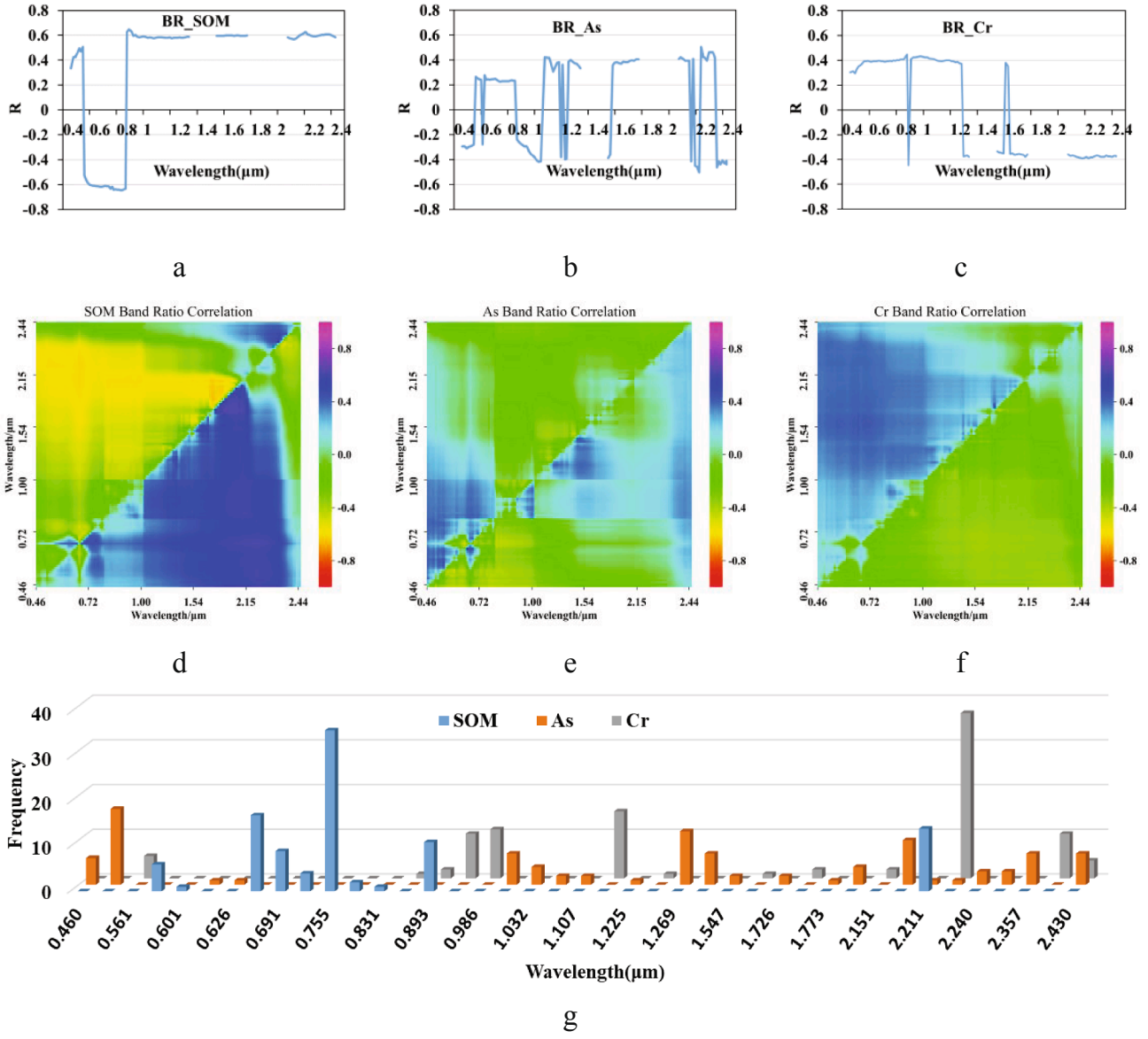

Fig. 6. The correlation after band ratio pre-processing and the frequency diagram. (a) SOM correlation after band ratio pre-processing. (b) As correlation after band ratio pre-processing. (c) Cr correlation after band ratio pre-processing. (d) All the correlations between band ratios and SOM. (e) All the correlations between band ratios and As. (f) All the correlations between band ratios and Cr. (g) Frequency diagram.

Table 5
Feature combinations.

	Features	Total number of features
SOM	$R_{0.89/0.83}$,$R_{0.84/0.89}$,$R_{0.78/0.89}$,$R_{0.91/0.78}$,$R_{2.21/0.69}$,$R_{0.67/2.21}$,$R_{2.37/0.67}$,$R_{0.94/0.75}$,$R_{1.77/0.73}$,$R_{0.54/0.57}$,$FD_{0.57}$,$FD_{0.84}$,$FD_{0.89}$,$FD_{2.21}$,$CR_{0.51}$,$CR_{0.60}$,$CR_{2.19}$	17
As	$R_{2.24/2.22}$,$R_{2.32/2.35}$,$R_{1.07/1.03}$,$FD_{1.07}$,$FD_{2.24}$,$FD_{2.35}$,$B_{1.01}$,$B_{1.03}$,$B_{1.07}$,$B_{1.10}$,$B_{1.22}$,$B_{1.26}$,$B_{1.54}$,$B_{1.72}$,$B_{1.77}$,$B_{2.09}$,$B_{2.19}$,$B_{2.21}$,$B_{2.22}$,$B_{2.24}$,$B_{2.30}$,$B_{2.35}$,$B_{2.43}$,$CR_{1.29}$,$CR_{1.64}$,$CR_{2.15}$,$CR_{2.37}$	27
Cr	$R_{0.88/0.89}$,$R_{0.98/2.22}$,$R_{1.18/2.15}$,$R_{0.58/2.40}$,$R_{2.40/0.56}$,$FD_{0.89}$,$FD_{1.00}$,$B_{0.56}$,$B_{0.88}$,$B_{0.89}$,$B_{0.97}$,$B_{0.98}$,$B_{1.18}$,$B_{1.25}$,$B_{1.61}$,$B_{1.75}$,$B_{2.15}$,$B_{2.22}$,$B_{2.40}$,$B_{2.43}$,$CR_{0.62}$,$CR_{2.28}$,$CR_{2.34}$	23

the higher RMSEp. As regards Cr, similar to the case of As, one issue is the small number of samples, and the other is that the sample concentration distribution is higher, resulting in low-accuracy samples in higher concentrations, but within acceptable limits.

3.3. Spatial distribution map analysis

When the estimation accuracy of the prediction model reaches 0.5 or above, the map generated via the hyperspectral imagery can reflect the actual situation of the whole research area. We used the prediction models for SOM, As, and Cr to produce spatial distribution maps, and overlaid the stream network generated by the DEM onto these maps to study the water sources and motion. Fig. 9 shows the spatial distribution maps for SOM, As, and Cr. It can be clearly seen from Fig. 9a and b that the spatial distribution of SOM and As shows a close correlation with the topography, and that a significant aggregation effect is apparent in the terrain depressions. The SOM content in the whole study area is high, showing a relatively uniform distribution, while As is at a lower level across the entire study area, and shows a strong correlation with the pollution caused by mining activities. The environment around the gold mine shows heavy As pollution, and crop growth is strongly inhibited. The As pollution also shows a state of movement and aggregation around the gold mining area; that is, the As component accumulates in the low-lying areas due to water movement. Cr shows a low spatial

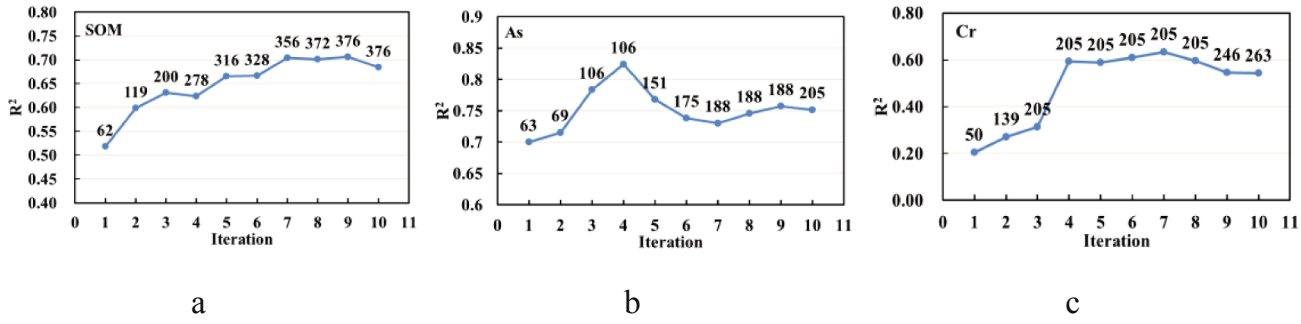

Fig. 7. The accuracy tendency chart in the testing set (the label information is the number of training samples). (a) SOM accuracy tendency chart. (b) As accuracy tendency chart. (c) Cr accuracy tendency chart.

Table 6
Optimal model accuracies for SOM, As, and Cr.

No.	Parameter	Training set					Test set				
		R_c^2	RMSEc	MAEc	Biasc	RPIQc	R_p^2	RMSEp	MAEp	Biasp	RPIQp
1	SOM (g/kg)	0.97	1.15	0.69	−0.16	6.60	0.71	3.52	2.59	−0.75	2.04
2	As (mg/kg)	0.93	15.06	9.40	2.52	2.09	0.82	34.74	24.05	−1.38	0.85
3	Cr (mg/kg)	0.99	3.70	2.70	−0.74	12.47	0.63	38.26	23.02	4.60	1.36

distribution density, with randomness. However, the Cr pollution around the gold mining area is relatively high, so that it can be concluded that the anthropogenic mining is the main source of Cr pollution, and there is basically no correlation with factors such as topography and water. One reason for this may be that some of the high-value samples had been removed during the training, resulting in insufficient expression in the highly polluted areas.

To verify the reliability of the estimation results, we conducted a field study of the entire study area in April 2019. The red box region marked ABCD in Fig. 9c shows the key research region, where unusual phenomena were apparent. Fig. 9d shows field pictures from the research area. Image is from the national highway to the east of region A, where region A is the river confluence area. The field research showed that there often are trucks carrying stones and other goods on the road. The road conditions are poor and there is an obvious dust problem. Therefore, some of the pollution in region A comes from vehicle emissions and dust. Images ② and ③ are located in region B. Image ④ is in region D. Both regions B and D show higher SOM contents in the spatial distribution maps. It can be seen from Images ② and ④ that the soils in this region are of a black color, and there is a huge difference with the surrounding soil. In addition, both regions are obviously located in low-lying areas, where the water content is high. These findings indicate that the topography in this study area has a significant impact on the SOM distribution. Image ③ shows the black soil after drying, which shows white spots. From our on-site investigation, expert consultation, and the relevant literature (Wang et al., 2009), we found that part of the study area features saline-alkali land, which is rich in SOM, where the soil texture is sticky and the water and fertilizer retention performance is good. This is a clear demonstration of how accurate the predictions of the proposed method are. Image ⑤ is located in region C, which is a large abandoned pit after gold mine excavation, where there is no river recirculation. Since the whole of the research area has frozen soil, the water in this large pit is collected after snow and ice melt and the thawing of frozen soil. We also found that, in other parts of the study area, as long as the gullies were dug in the lower areas, water began to accumulate after a period of time. Therefore, ice and snowmelt and frozen soil thawing affect the water flow in the entire study area, and there are many signs of water flow. In the original hyperspectral imagery (Fig. 2a), it is clear that the blackened area is the lower topographical area. The main reason for the soil blackening is the transportation and agglomeration of soil components, such as SOM and As, due to the action of water motion. Image ⑥ shows a used pesticide bottle discarded at random. Image ⑦ shows the treatment of domestic garbage in the village, where the garbage is directly incinerated outside. Image ⑧ shows different farmers using different fertilizers and different corn seeds. We also found during the survey that farmyard manure is still used in some areas. This shows that there is no uniform farming pattern between the different cultivated plots, and the management is not strict. As a result, the pollution of the cultivated land is scattered, which is also consistent with the spatial distribution maps.

In summary, the following conclusions can be drawn: 1) the SOM, As, and Cr values obtained by the Semi-DNNR model can accurately reflect the distribution of the whole research area, which also indicates that the selected feature bands have high reliability; 2) the topographic changes in the study area have a significant impact on the transportation and agglomeration of SOM and As, mainly due to snowmelt, ice melt, and frozen soil thawing; and 3) gold mining activities are the main source of heavy metal pollution. Vehicle emissions and dust, garbage disposal in the villages, the direct discharge of domestic sewage, and the different farming methods also bring some pollution to the research area.

3.4. Transportation of SOM and as

Due to the influence of topographic changes, the flow and accumulation of surface water provide the transport capacity for SOM and As. Most interestingly, it can be seen in the spatial distribution maps for SOM and As that the concentration of As increases with the increase of SOM, but when As increases beyond a certain point, the SOM decreases. This phenomenon can be clearly seen in Fig. 10, which shows the spatial distribution maps for SOM and As. Many studies (Bauer and Blodau, 2006; Kalbitz and Wennrich, 1998; Mcarthur et al., 2004; Redman et al., 2002; Wang and Mulligan, 2006) have indicated that organic matter has an adsorption capacity for As, and the valence states of As(III) and As(V) can be converted to each other by combination with organic matter. In addition, the dissolved organic matter (DOM) and As can undergo complexation to form As-NOM complexes. Therefore, affected by the micro-topography, the movement and aggregation of water flow affect the transportation and aggregation of SOM. Due to the adsorption and complexation of As by SOM and DOM, respectively, As also shows a similar phenomenon to SOM. When As pollution is significantly

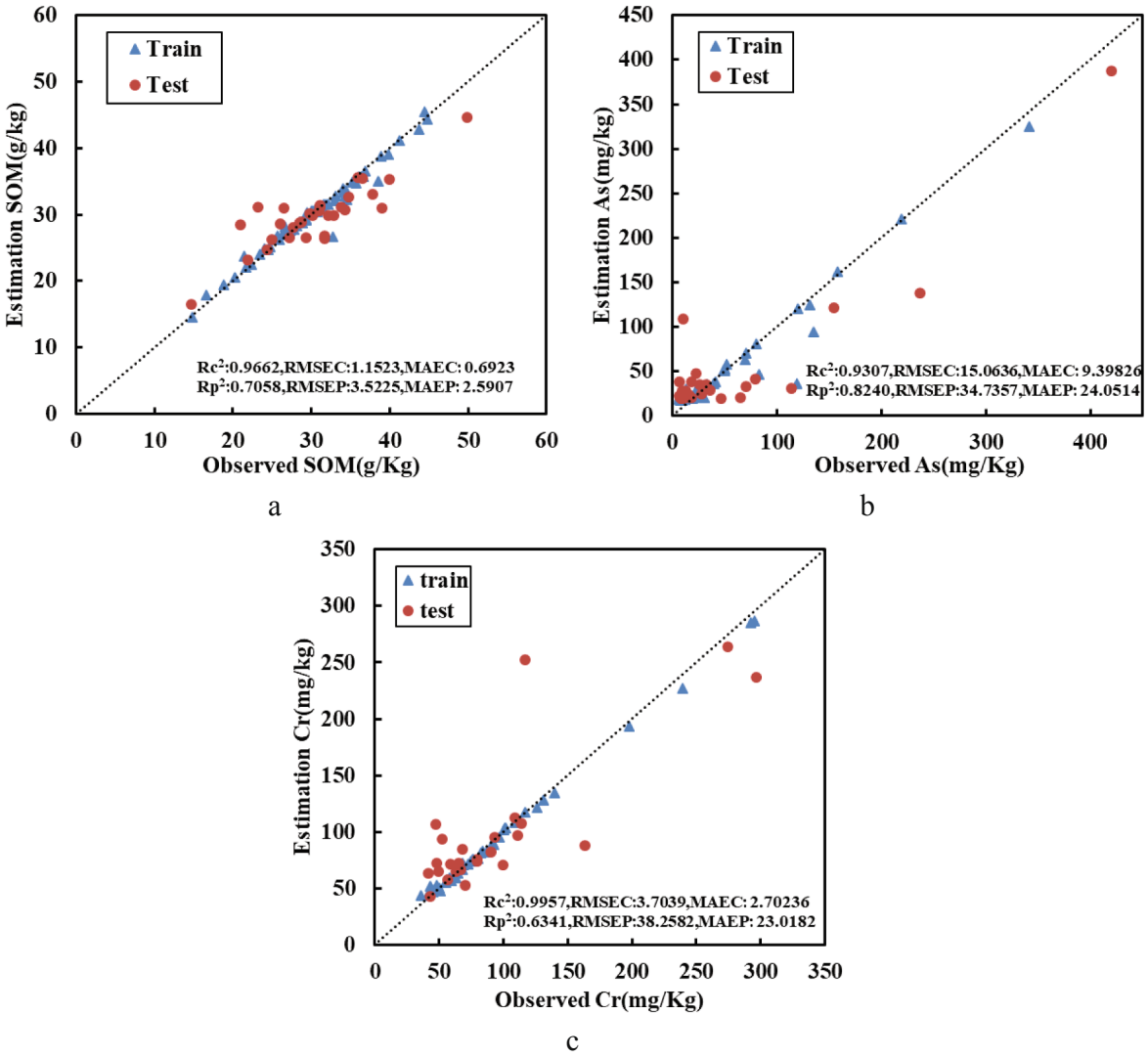

Fig. 8. Scatter plots of the prediction results for SOM, As, and Cr. (a) SOM. (b) As. (c) Cr.

increased due to external forces, it affects the crop growth, so that SOM remains at a low level. In this regard, these findings prove the reliability of this experiment, which provides us with a macro perspective to analyze the transportation of SOM and As.

In the spatial distribution maps of both SOM and As, it can be seen that the interaction between SOM and As is more apparent in low-lying areas and contaminated locations, which means that water is a direct factor. It can be seen from the figures that when the concentration of As is at a low level, SOM shows an upward trend, which indicates that SOM results in adsorption and complexation of As. However, when the As pollution is high, SOM remains at a low level, meaning that As has a great impact on crop growth. Although the correlation between SOM and As is not strong, it can provide a simple reference relationship for SOM and As.

In summary, the transportation and aggregation of As in the study area are mainly affected by the adsorption and complexation of SOM and DOM. Hyperspectral remote sensing estimation can provide us with a macroscopic view to analyze the transportation, adsorption, and complexation between SOM and As.

4. Discussion

In the Semi-DNN framework, threshold parameters *E* and *Delta* can be used to limit the number of additional samples. Both *E* and *Delta* are empirical values. $|V_{m1} - V_{m2}| \leq E$ means that the difference between the values predicted by Model_1 and Model_2 for the same position is smaller than a threshold *E*. If both Model _1 and Model _2 have a high prediction accuracy, *E* will close to 0. For As and Cr, which with high Std values, so *E* should be set as a large value; however, for SOM, parameter *E* has little effect on the results. Compared with E, the influence of *Delta* is more important, because parameter *E* is only used to remove very unreliable samples in the first step, and *Delta* is related to the number of selected samples. After sorting the additional samples according to the *Delta* value, the lower-ranked samples with low *Delta* value can be considered as unreliable samples. Therefore, the empirical *Delta* value should be determined case by case.

For the DNN model, the number of hidden layers affects the regression performance. Fig. 11 shows the SOM accuracy and time consumption performance for the DNN model with different hidden layers. Generally speaking, the accuracy of the regression is poor if there are too few layers. Although a larger number of layers results in a better regression accuracy, the time consumption is also higher. It can be seen from Fig. 11 that the optimal number of hidden layers is six, and using the six-layer network is computationally more time efficient than using the seven-layer network. Therefore, when the number of training samples is small, the number of hidden layers can be set to six, obtaining the

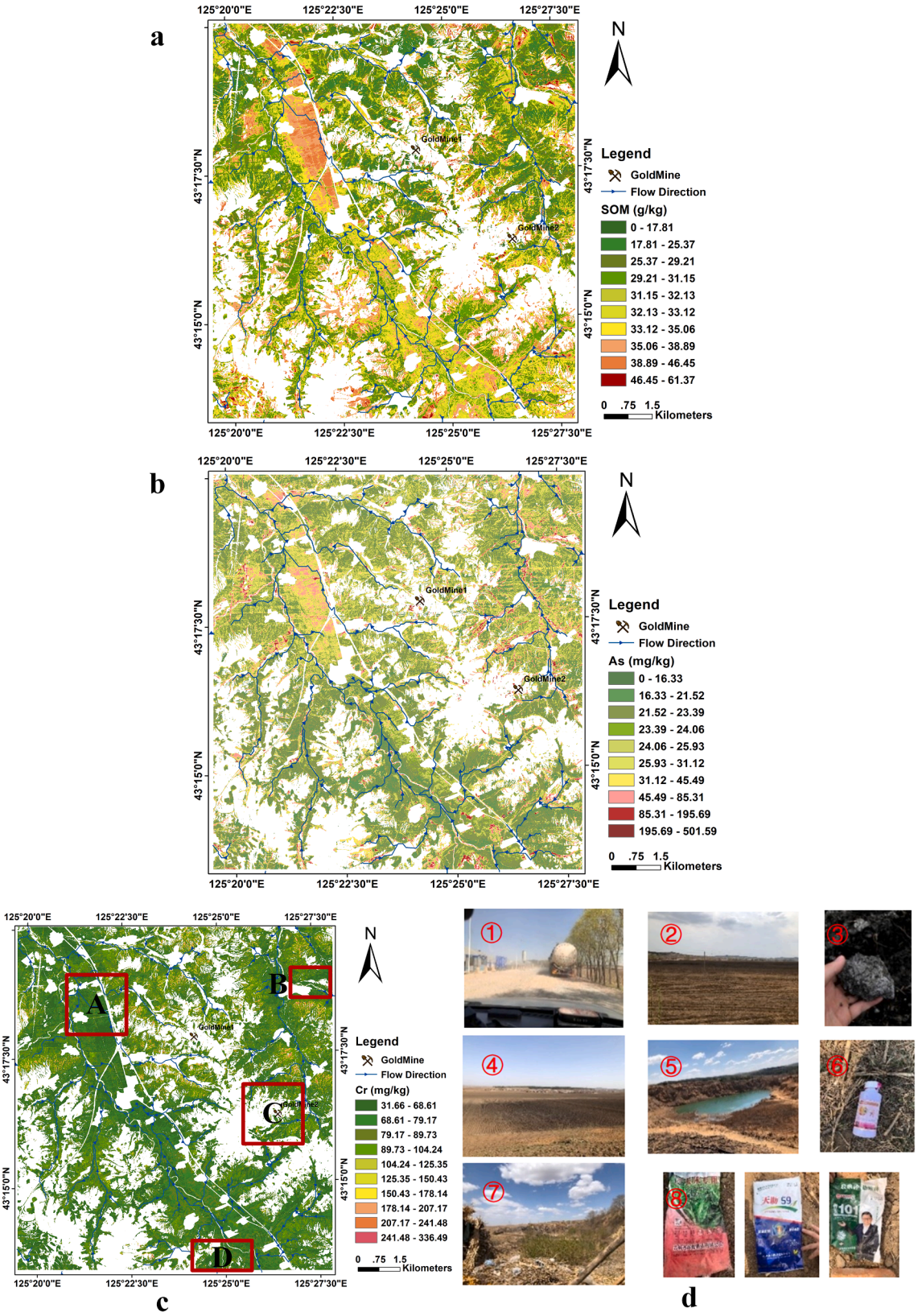

Fig. 9. Spatial distribution maps and field surveys. (a) SOM spatial distribution map. (b) As spatial distribution map. (c) Cr spatial distribution map. (d) Images from the field survey.

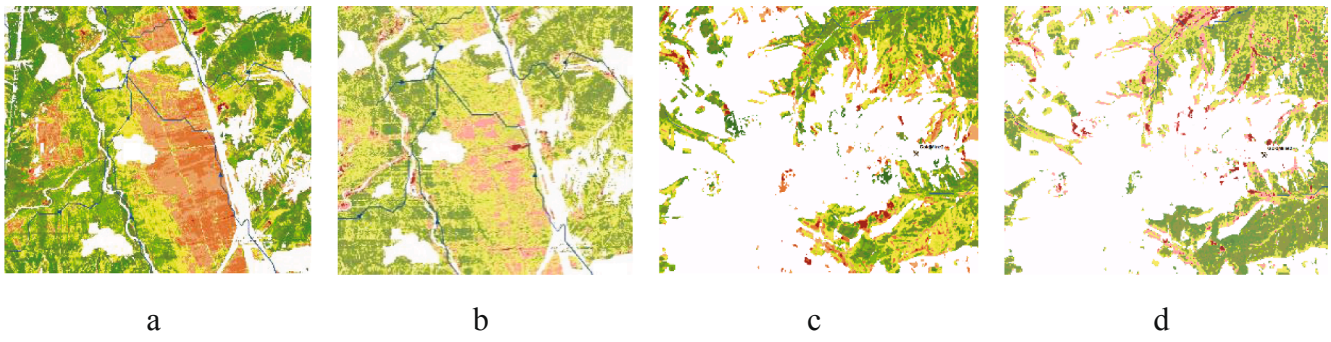

Fig. 10. Zoomed areas in region A and region C. (a) The SOM of the zoomed area in region A. (b) The As of the zoomed area in region A. (c) The SOM of the zoomed area in region C. (d) The As of the zoomed area in region C.

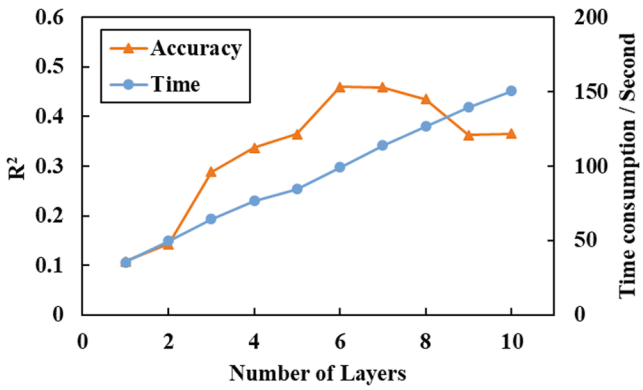

Fig. 11. The SOM accuracy and time consumption performance for the DNN model with different hidden layers.

best performance.

Some of the previous studies (Bajcsy and Groves, 2004; Underwood et al., 2003) have shown that the band ratio pre-processing method can obtain a better feature expression, to some extent, but the same band ratio combination may not exist for airborne hyperspectral imagery. Hence, it is necessary to adopt effective feature selection methods for different data. To verify the effectiveness of both the feature selection and the Semi-DNNR model, we used different methods for the comparative experiments. We compared the proposed feature selection method with the state-of-the-art methods of CARS (Li et al., 2009) and the GA-based method (Leardi, 2000). The regression methods were SVR and PLSR. The search space of the default hyperparameters of PLSR were defaulted, and its search space of components was [2, 20]. The kernel function applied in SVR was Radial Basis Function, (RBF). The search space of penalty parameter C was [2^-5, 2^20], while gamma was [2^-20, 2^20]. All the training and test sets for the same soil component were the same for all the methods tested. The results of the comparison between the proposed Semi-DNNR method and the feature selection methods (CARS, GA) and regression methods (PLSR, SVR) are shown in Table 7. From Table 7, it can be seen that the proposed feature selection method is more robust than the CARS, GA, and all bands without feature selection methods. When using the PLSR regression method, the proposed feature selection method obtains the highest prediction accuracy for SOM, As and Cr. When using the SVR regression method, the proposed feature selection method gets the highest prediction accuracy for SOM and As. Overall, the performance of SVR is better than that of PLSR, and the proposed Semi-DNNR method shows significant advantages in accuracy improvement. In addition, when using the Semi-DNNR method, all the training performance R_c^2 values for SOM, As, and Cr reach 0.9 or above, which is a better performance than the PLSR and SVR regressors. For the SOM prediction, the prediction accuracy $R_p^2$of the Semi-DNNR model is 19.73% higher than that of SVR (Proposed + SVR) and 23.65% higher than that of simple SVR with all features. For the As prediction, the $R_p^2$of the Semi-DNNR model is 18.80% higher than that of SVR (Proposed + SVR) and 22.86% higher than that of simple SVR with all features. For the Cr prediction, the$R_p^2$of the Semi-DNNR model is 25.71% higher than that of PLSR (Proposed + PLSR) and 27.34% higher than that of simple SVR with all features. Therefore, the feature selection method described in Section 2.4.2 combined with the proposed Semi-DNNR model shows the best prediction performance. Meanwhile, the multi-input model which combines spectral pre-processing techniques shows a better result than the single-input model by one pre-processing method.

A potential drawback of this methodology is that it is sensitive to the quality of the initial training data set. In the first iteration, actually, it is a DNN network, there are only a few training data. If the training is not good enough, the prediction of the candidate set samples will be low, which makes it challenging to select persuasive unlabeled samples. It should further be noted that the feature combination method should be more intelligent, which can be possible when a large number of samples are available.

5. Conclusion

In this study, hyperspectral remote sensing image data collection and field sampling in the black soil farmland of Northeast China were carried out to estimate the concentrations of SOM, As, and Cr in the soil. Furthermore, in this paper, we have innovatively provided the entire processing flow and analysis methods used in this study. Firstly, the soil mask information file was constructed by the use of an unmixing method. The feature bands were then obtained by combining multiple spectral pre-processing methods. The SOM, As, and Cr were found to have a strong response in the vicinity of 2.20 μm, which means that the –OH group plays an important role in soil composition estimation. However, these three components have different feature bands. The feature bands of SOM are mainly concentrated in the vicinity of 0.57 μm, 0.67 μm, 0.75 μm, and 0.89 μm. The feature bands of As are mainly concentrated in the vicinity of 1.07 μm and 1.22 μm. The feature bands of Cr are mainly found at 0.58 μm, and adjacent to 0.89 μm, 0.98 μm, and 2.40 μm. Finally, the novel semi-supervised deep neural network regression (Semi-DNNR) model was proposed. The semi-supervised idea is introduced into the DNN regression model to solve the training problem caused by the limited samples. The proposed Semi-DNNR method has a strong training ability, with the prediction R_p^2 accuracy for SOM, As, and Cr being 0.71, 0.82, and 0. 63, respectively. Our research found that mining activities are the main source of soil pollution in the mining area. The field survey also showed that the decentralized farmland management practices, the garbage disposal in the villages, the direct discharge of domestic sewage, and the road dust and emissions from trucks have also contributed to the soil contamination.

Table 7
Comparison of the proposed method with feature selection methods (CARS, GA) and regression methods (PLSR, SVR).

Parameter	Menthod	Total number of features	Training set					Testing set				
			R^2c	RMSEc	MAEc	Biasc	RPIQc	R^2p	RMSEp	MAEp	Biasp	RPIQp
SOM	ALL + PLSR	101	0.3774	4.9514	3.9776	0.0000	1.5569	0.3659	5.1719	4.0660	0.1787	1.4587
	CARS + PLSR	11	0.3713	4.9756	3.9045	0.0000	1.5493	0.3778	5.1230	3.9690	0.0942	1.4726
	GA + PLSR	14	0.4035	4.8466	3.7923	0.0000	1.5906	0.4269	4.9168	3.8399	−0.3441	1.5344
	Proposed + PLSR	17	0.3962	4.8761	3.9796	0.0000	1.5809	**0.4453**	4.8370	3.8316	−0.2115	1.5597
	ALL + SVR	101	0.4480	4.6622	3.3699	0.1124	1.6303	0.4693	4.7314	3.6534	−0.1904	1.5223
	CARS + SVR	11	0.5124	4.3819	3.1708	0.0832	1.7346	0.4962	4.6098	3.4818	−0.2599	1.5625
	GA + SVR	14	0.4222	4.7699	3.6352	0.1272	1.5935	0.4545	4.7969	3.8108	−0.5447	1.5015
	Proposed + SVR	17	0.5628	4.1494	2.7911	−0.0077	1.8318	**0.5085**	4.5533	3.7530	−0.5116	1.5819
	Semi-DNNR	17	0.9663	1.1523	0.6924	−0.1565	6.5959	**0.7058**	3.5225	2.5907	−0.7514	2.0447
As	ALL + PLSR	101	0.1777	51.8927	32.7583	0.0000	0.6136	0.2093	73.6362	39.0349	−13.1060	0.4370
	CARS + PLSR	18	0.1895	51.5177	32.9753	0.0000	0.6181	0.2386	72.2585	41.5309	−11.5137	0.4453
	GA + PLSR	30	0.1624	52.3723	31.1565	0.0000	0.6080	0.2079	73.7019	35.5044	−12.0820	0.4366
	Proposed + PLSR	27	0.3144	47.3849	29.3349	0.0000	0.6720	**0.4742**	60.0484	29.6957	−12.9565	0.5359
	ALL + SVR	101	0.9260	15.5631	3.7453	−1.7044	2.0221	0.5955	52.6688	34.0320	−8.2007	0.5582
	CARS + SVR	18	0.8437	22.6209	9.8378	−0.6180	1.3912	0.5833	53.4546	35.0492	−7.0263	0.5500
	GA + SVR	30	0.8719	20.4815	7.0760	−3.1479	1.5366	0.6125	51.5521	32.8717	−4.9379	0.5703
	Proposed + SVR	27	0.8228	24.0905	10.0493	−1.0739	1.3064	**0.6361**	49.9543	34.4013	−10.2171	0.5885
	Semi-DNNR	27	0.9307	15.0636	9.3982	2.5237	2.0892	**0.8241**	34.7357	24.0514	−1.3847	0.8464
Cr	ALL + PLSR	101	0.0855	54.1600	37.3403	0.0000	0.8577	0.2369	55.2548	39.6870	−4.1432	0.9747
	CARS + PLSR	19	0.0958	53.8568	36.8598	0.0000	0.8625	0.2186	55.9123	39.3700	−3.8362	0.9632
	GA + PLSR	10	0.1186	53.1723	36.7192	0.0000	0.8736	0.2550	54.5956	38.7819	−4.4654	0.9864
	Proposed + PLSR	23	0.1834	51.1796	34.6943	0.0000	0.9077	**0.3770**	49.9254	37.3864	1.5049	1.0787
	ALL + SVR	101	0.4729	41.1175	18.8514	−12.1206	1.1236	**0.3607**	50.5718	36.6290	−10.8638	1.0300
	CARS + SVR	19	0.5991	35.8609	17.8105	−5.7075	1.2883	0.2814	53.6166	37.0133	4.2673	0.9715
	GA + SVR	10	0.2100	50.3407	23.1166	−12.2114	0.9177	0.2139	56.0805	37.9056	−22.3851	0.9288
	Proposed + SVR	23	0.3175	46.7891	23.6013	−11.6017	0.9874	0.2118	56.1563	37.2824	−3.0128	0.9275
	Semi-DNNR	23	0.9957	3.7040	2.7024	−0.7420	12.4727	**0.6341**	38.2583	23.0182	4.5970	1.3615

ALL: all features.
Proposed: feature selection method described in Section 2.4.2.

Finally, and most importantly, the stream network generated by the DEM was successfully used to analyze the transportation and aggregation of SOM, As, and Cr. It was found that the distribution of As in the soil was mainly affected by the adsorption and complexation of SOM and DOM. There is micro-topography in the study area, so the water movement due to the melting of snow, the melting of frozen soil, and the flow of surface rainfall have carried and concentrated SOM and As, as we found high SOM and As aggregations in the low-lying areas. Furthermore, when As pollution reaches a certain level, it has a negative effect on crop growth, resulting in a decrease in SOM content. In conclusion, the method proposed in this paper was able to accurately describe the spatial distribution of SOM, As, and Cr in the study area. It could also provide macroscopic observations for the study of the adsorption and transportation between SOM and As.

Declaration of Competing Interest

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

Acknowledgements

This research was supported in part by the Natural Science Foundation of China (No. 41871337) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence 34(11), 2274-2282.
Amigo, J.M., Babamoradi, H., Elcoroaristizabal, S., 2015. Hyperspectral image analysis. A tutorial. Anal. Chim. Acta 896 (X), 34–51.
Bajcsy, P., Groves, P., 2004. Methodology for hyperspectral band selection. Photogramm. Eng. Remote Sens. 70 (7), 793–802.
Bauer, M., Blodau, C., 2006. Mobilization of arsenic by dissolved organic matter from iron oxides, soils and sediments. Sci. Total Environ. 354 (2), 179–190.
Ben-Dor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. Remote Sens. Environ. 61 (1), 1–15.
Berk, A., Anderson, G.P., Bernstein, L.S., Acharya, P.K., Dothe, H., Matthew, M.W., Adler-Golden, S.M., Chetwynd Jr, J.H., Richtsmeier, S.C., Pukall, B., 1999. MODTRAN4 radiative transfer modeling for atmospheric correction, Optical spectroscopic techniques and instrumentation for atmospheric and space research III. Int. Soc. Opt. Photon. 348–353.
Brefeld, U., Gärtner, T., Scheffer, T., Wrobel, S., 2006. Efficient co-regularised least squares regression, International Conference on Machine Learning.
Cécillon, L., Cassagne, N., Czarnes, S., Gros, R., Brun, J.-J., 2008. Variable selection in near infrared spectra for the biological characterization of soil and earthworm casts. Soil Biol. Biochem. 40 (7), 1975–1979.
Chabrillat, S., Ben-Dor, E., Cierniewski, J., Gomez, C., Schmid, T., van Wesemael, B., 2019. Imaging spectroscopy for soil mapping and monitoring. Surv. Geophys. 40 (3), 361–399.
Chakraborty, S., Li, B., Deb, S., Paul, S., Weindorf, D.C., Das, B.S., 2017a. Predicting soil arsenic pools by visible near infrared diffuse reflectance spectroscopy. Geoderma 296, 30–37.
Chakraborty, S., Weindorf, D.C., Deb, S., Li, B., Paul, S., Choudhury, A., Ray, D.P., 2017b. Rapid assessment of regional soil arsenic pollution risk via diffuse reflectance spectroscopy. Geoderma 289, 72–81.
Chi, G., Wang, D., 2017. Small-area population forecasting: a geographically weighted regression approach. Springer, Cham.
Choe, E., van der Meer, F., van Ruitenbeek, F., van der Werff, H., de Smeth, B., Kim, K.-W., 2008. Mapping of heavy metal pollution in stream sediments using combined

geochemistry, field spectroscopy, and hyperspectral remote sensing: a case study of the Rodalquilar mining area, SE Spain. Remote Sens. Environ. 112 (7), 3222–3233.
Coleman, T., Agbu, P., Montgomery, O., Gao, T., Prasad, S., 1991. Spectral band selection for quantifying selected properties in highly weathered soils. Soil Sci. 151 (5), 355–361.
Csillag, F., Pásztor, L., Biehl, L.L., 1993. Spectral band selection for the characterization of salinity status of soils. Remote Sens. Environ. 43 (3), 231–242.
Demattê, J.A.M., Fongaro, C.T., Rizzo, R., Safanelli, J.L., 2018. Geospatial soil sensing system (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sens. Environ. 212, 161–175.
Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H., Liu, W., 2020. Advances of four machine learning methods for spatial data handling: a review. J. Geovisualiz. Spatial Anal. 4, 13.
Farifteh, J., Van der Meer, F., Atzberger, C., Carranza, E., 2007. Quantitative analysis of salt-affected soil reflectance spectra: a comparison of two adaptive methods (PLSR and ANN). Remote Sens. Environ. 110 (1), 59–78.
Fichot, C.G., Downing, B.D., Bergamaschi, B.A., Windham-Myers, L., Marvin-Dipasquale, M.C., Thompson, D.R., Gierach, M., 2015. High-resolution remote sensing of water quality in the San Francisco Bay-Delta Estuary. Environ. Sci. Technol. 50 (2), 573.
Galvão, L.S., Vitorello, Í., 1998. Variability of laboratory measured soil lines of soils from southeastern Brazil. Remote Sens. Environ. 63 (2), 166–181.
Gannouni, S., Rebai, N., Abdeljaoued, S., 2012. A spectroscopic approach to assess heavy metals. J. Geograp. Inform. Syst. 4, 242–253.
Gholizadeh, A., BoruVka, L., Saberioon, M.M., Kozák, J., Vašát, R., NěMeček, K., 2015. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. Soil & Water Research 10(4), 218-227.
Granitto, P.M., Furlanello, C., Biasioli, F., Gasperi, F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr. Intell. Labor. Syst. 83 (2), 83–90.
Groves, P., Bajcsy, P., 2003. Methodology for hyperspectral band and classification model selection, IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data.
Guanter, L., Richter, R., Moreno, J., 2006. Spectral calibration of hyperspectral imagery using atmospheric absorption features. Appl. Opt. 45 (10), 2360–2370.
Heinz, D.C., 2001. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. 39 (3), 529–545.
Henderson, T.L., Szilagyi, A., Baumgardner, M.F., Chen, C.C.T., Landgrebe, D.A., 1989. Spectral band selection for classification of soil organic matter content. Soil Sci. Soc. Am. J. 53 (6), 1778–1784.
Jenson, S.K., Domingue, J.O., 1988. Extracting topographic structure from digital elevation data for geographic information system analysis. Photogramm. Eng. Remote Sens. 54 (11), 1593–1600.
Kalbitz, K., Wennrich, R., 1998. Mobilization of heavy metals and arsenic in polluted wetland soils and its dependence on dissolved organic matter. Sci. Total Environ. 209 (1), 27–39.
Khajehsharifi, H., Eskandari, Z., Sareban, N., 2017. Using partial least squares and principal component regression in simultaneous spectrophotometric analysis of pyrimidine bases. Arabian Journal of Chemistry 10(S1), S1878535212001748.
Khan, T.M., Bailey, D.G., Khan, M.A.U., Kong, Y., 2017. Efficient hardware implementation for fingerprint image enhancement using anisotropic gaussian filter. IEEE Trans. Image Process. 26 (5), 2116–2126.
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kleijnen, J.P., 2009. Kriging metamodeling in simulation: a review. Eur. J. Oper. Res. 192 (3), 707–716.
Kleinbaum, D., Kupper, L., Nizam, A., Rosenberg, E., 2013. Applied Regression Analysis and Other Multivariable Methods. Nelson Education.
Kukreja, S.L., Löfberg, J., Brenner, M.J., 2006. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. IFAC Proceedings Volumes 39 (1), 814-819.
Leardi, R., 2000. Application of genetic algorithm–PLS for feature selection in spectral data sets. J. Chemom. 14 (5–6), 643–655.
Lei, M.A., Wang, X., 2011. Semi-supervised regression based on support vector machine co-training. Comput. Eng. Appl. 25 (2).
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.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18–22.
Liu, H., Zhang, Y., Zhang, B., 2009. Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China. Environ. Monit. Assess. 154 (1–4), 147.
Malm, O., 1998. Gold mining as a source of mercury exposure in the Brazilian Amazon. Environ. Res. 77 (2), 73–78.
Mcarthur, J.M., Banerjee, D.M., Hudson-Edwards, K.A., Mishra, R., Purohit, R., Ravenscroft, P., Cronin, A., Howarth, R.J., Chatterjee, A., Talukder, T., 2004. Natural organic matter in sedimentary basins and its relation to arsenic in anoxic ground water: the example of West Bengal and its worldwide implications. Appl. Geochem. 19 (8), 1255–1293.
Mirsal, I.A., 2008. Soil Pollution. Springer.
Nascimento, J.M., Dias, J.M., 2005. Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Trans. Geosci. Remote Sens. 43 (4), 898–910.
O'Callaghan, J.F., Mark, D.M., 1984. The extraction of drainage networks from digital elevation data. Comput. Vision Graph. Image Process. 28 (3), 323–344.
Ou, D., Tan, K., Du, Q., Zhu, J., Wang, X., Chen, Y., 2019. A novel Tri-Training technique for the semi-supervised classification of hyperspectral images based on regularized local discriminant embedding feature extraction. Remote Sensing 11 (6).
Padarian, J., Minasny, B., McBratney, A.B., 2019a. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma 340, 279–288.
Padarian, J., Minasny, B., McBratney, A.B., 2019b. Using deep learning to predict soil properties from regional spectral data. Geoderma Regional 16, e00198.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch. NIPS. Workshop.
Pyo, J., Duan, H., Baek, S., Kim, M.S., Jeon, T., Kwon, Y.S., Lee, H., Cho, K.H., 2019. A convolutional neural network regression for quantifying cyanobacteria using hyperspectral imagery. Remote Sens. Environ. 233, 111350.
Redman, A.D., Macalady, D.L., Dianne, A., 2002. Natural organic matter affects arsenic speciation and sorption onto hematite. Environ. Sci. Technol. 36 (13), 2889–2896.
Ren, X., Malik, J., 2003. Learning a classification model for segmentation, IEEE International Conference on Computer Vision. IEEE, pp. 10.
Rezaei, Y., Mobasheri, M.R., Zoej, M.J.V., 2008. Unsupervised information extraction using absorption line in Hyperion images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37, 383-388.
Rinnan, Å., van den Berg, F., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC, Trends Anal. Chem. 28 (10), 1201–1222.
Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1), 46–54.
Sarathjith, M., Das, B.S., Wani, S.P., Sahrawat, K.L., 2016. Variable indicators for optimum wavelength selection in diffuse reflectance spectroscopy of soils. Geoderma 267, 1–9.
Selige, T., Böhner, J., Schmidhalter, U., 2006. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. Geoderma 136 (1–2), 235–244.
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.
Shi, T., Wang, J., Chen, Y., Wu, G., 2016. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. Int. J. Appl. Earth Obs. Geoinf. 52, 95–103.
Singh, S., Kasana, S.S., 2019. Estimation of soil properties from the EU spectral library using long short-term memory networks. Geoderma Regional 18, e00233.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statist. Comput. 14 (3), 199–222.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M. J., 2014. The performance of Visible, Near-, and Mid-Infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49 (2), 139–186.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Stamatis, G., Voudouris, K., Karefilakis, F., 2001. Groundwater pollution by heavy metals in historical mining area of Lavrio, Attica, Greece. Water Air Soil Pollut. 128 (1–2), 61–83.
Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma 144 (1–2), 395–404.
Tachikawa, T., Kaku, M., Iwasaki, A., Gesch, D.B., Oimoen, M.J., Zhang, Z., Danielson, J. J., Krieger, T., Curtis, B., Haase, J., 2011. ASTER global digital elevation model version 2-summary of validation results, NASA.
Tan, K., Wang, H., Zhang, Q., Jia, X., 2018. An improved estimation model for soil heavy metal (loid) concentration retrieval in mining areas using reflectance spectroscopy. J. Soils Sediments 18 (5), 2008–2022.
Thompson, B., 1995. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Sage Publications Sage CA, Thousand Oaks, CA.
Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit Region. Econ. Geograp. 46 (sup1), 234–240.
Tsakiridis, N.L., Chadoulos, C.G., Theocharis, J.B., Ben-Dor, E., Zalidis, C.G., 2020a. A three-level multiple-kernel learning approach for soil spectral analysis. Neurocomputing 389, 27–41.
Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020b. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma 367, 114208.
Tsakiridis, N.L., Theocharis, J.B., Panagos, P., Zalidis, G.C., 2019. An evolutionary fuzzy rule-based system applied to the prediction of soil organic carbon from soil spectral libraries. Appl. Soft Comput. 81, 105504.
Underwood, E., Ustin, S., DiPietro, D., 2003. Mapping nonnative plants using hyperspectral imagery. Remote Sens. Environ. 86 (2), 150–161.
Wang, F., Gao, J., Zha, Y., 2018. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. ISPRS J. Photogramm. Remote Sens. 136, 73–84.
Wang, L., Seki, K., Miyazaki, T., Ishihama, Y., 2009. The causes of soil alkalinization in the songnen plain of northeast China. Paddy Water Environ, 7 (3), 259–270.
Wang, M., Hua, X.S., Song, Y., Dai, L.R., Zhang, H.J., 2006. Semi-supervised kernel regression, Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006). IEEE.
Wang, S., Mulligan, C.N., 2006. Effect of natural organic matter on arsenic release from soils and sediments into groundwater. Environ. Geochem. Health 28 (3), 197–214.

Wang, X., Tan, K., Du, Q., Chen, Y., Du, P., 2019. Caps-TripleGAN: GAN-Assisted CapsNet for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57 (9), 7232–7245.
Weidong, L., Baret, F., Xingfa, G., Qingxi, T., Lanfen, Z., Bing, Z., 2002. Relating soil surface moisture to reflectance. Remote Sens. Environ. 81 (2–3), 238–246.
Yu, J., Yan, B., Liu, W., Li, Y., He, P., 2017. Seamless mosaicking of multi-strip airborne hyperspectral images based on Hapke model, International Conference on Sensing and Imaging. Springer, pp. 285–292.
Zou, X., Jiewen, Z., Povey, M.J., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667 (1–2), 14–32.
Zhou, Z.-H., 2006. Learning with unlabeled data and its application to image retrieval. In: Yang, Q., Webb, G. (Eds.), PRICAI 2006: Trends in Artificial Intelligence. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 5–10.
Zhou, Z.-H., Li, M., 2007a. Semisupervised regression with cotraining-style algorithms. IEEE Trans. Knowl. Data Eng. 19 (11), 1479–1493.
Zhou, Z.H., Li, M., 2007b. Semi-Supervised regression with Co-Training. IEEE Trans. Knowl. Data Eng. 19 (11), 1479–1493.