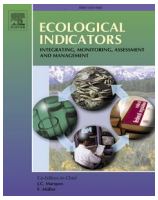
Ecological Indicators 133 (2021) 108400

Contents lists available at ScienceDirect

Ecological Indicators

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



Estimating chromium concentration in arable soil based on the optimal principal components by hyperspectral data

Fei Guo [a,b,c,*], Zhen Xu [d,*], Honghong Ma [a,b,c], Xiujin Liu [a,b,c], Shiqi Tang [a,b,c], Zheng Yang [a,b,c], Li Zhang [a,b,c], Fei Liu [a,b,c], Min Peng [a,b,c], Kuo Li [a,b,c]

[a] Institute of Geophysical & Geochemical Exploration, Chinese Academy of Geological Sciences, Langfang 065000, China
[b] Key Laboratory of Geochemical Cycling of Carbon and Mercury in the Earth's Critical Zone, Chinese Academy of Geological Sciences, Langfang 065000, China
[c] Geochemical Research Center of Soil Quality, China Geological Survey, Langfang 065000, China
[d] Department of Electronic and Information Engineering, Shantou University, Shantou 515063, China

ARTICLE INFO

Keywords:
Hyperspectral remote sensing technology
Soil chromium
Optimal principal components
Partial least squares regression (PLSR)
Gradient boosting decision tree (GBDT)

ABSTRACT

The heavy metal pollution in arable soil poses a significant threat to human health. Thus, it is of great significance to investigate the contamination of heavy metal elements in the soil. As the soil polluted by heavy metal is sensitive to spectral reflectance, thus the hyperspectral remote sensing technology could be a valuable tool for retrieving heavy metal components in the soil. This study, taking chromium (Cr) concentration as an example, proposes an optimal model for estimating heavy metal components in the soil by comprehensively taking account of the spectral pretreatment, dimensionality reduction with optimal parameters, and hyperspectral model. To this end, both the linear model, i.e., partial least squares regression (PLSR), and the nonlinear model, i.e., the gradient boosting decision tree (GBDT), are applied in this study. It is found in the study area, the Savitzky-Golay (SG) method can be regarded as an excellent spectral pretreatment for the hyperspectral data regardless of the applied model. By contrast, the dimensionality reduction in terms of the Principal Component Analysis (PCA) is closely related to hyperspectral model: the optimal principal components (PCs) in the estimation of Cr concentration are the first 9 PCs for the GBDT (nonlinear model), while that for the PLSR (linear model) become the first 8 PCs. Moreover, the examination of hyperspectral model shows the GBDT model has slightly better performance than the PLSR model for the Cr concentration estimation under most conditions. Finally, when the spectral pretreatment, dimensionality reduction, and hyperspectral model are fully considered, the best retrieval model for the Cr concentration in the study area is the SG-PCA-GBDT model. Numeric measures of model accuracy show the proposed model has a determination coefficient of 0.80 and a residual prediction deviation of 2.04, which provides a potentially new method for estimating Cr concentration in the polluted soil.

1. Introduction

Arable soil is the resource that humans depend on for survival, playing a vital role in agricultural development (Hong et al., 2019c). However, the heavy metal pollution in arable soil has become a severe problem in recent decades due to the rapid development of industrialization and urbanization (Chen et al., 2015; Sun et al., 2018). Nowadays, heavy metal pollution originates mainly from fertilizer, pesticides, and mining activities (Wei and Yang, 2010). Many elements can result in the heavy metal pollution of arable soil, wherein chromium (Cr), as a trace element, shows some peculiar characteristic features to the arable soil pollution. Cr is an essential trace element for the growth of organisms (Bhattacharya et al., 2016). Nevertheless, it would pollute the soil and water as long as its concentration exceeds a certain level. Generally, the Cr in the soil exists primarily in the form of Cr(III) and Cr(VI); the Cr(VI), which has higher toxicity and mobility, could pose a threat to the natural environment and human life (Chovanec et al., 2012; Kimbrough et al., 1999; Yalçın Tepe, 2014). Therefore, it is critical to investigate Cr concentration in arable soil.

The conventional methods for investigating heavy metal pollution in the soil include on-site sampling and laboratory analysis, which is time-consuming and expensive (Cheng et al., 2019; Leenaers et al., 1990). As a result, it has become an urgent problem for determining the concentration and spatial distribution of the heavy metals in the soil. It is

* Correspondence author at: Department of Electronic and Information Engineering, Shantou University, Shantou 515063, China.
E-mail addresses: flynn1991@sina.cn (F. Guo), xuzhen@stu.edu.cn (Z. Xu).

https://doi.org/10.1016/j.ecolind.2021.108400
Received 1 September 2021; Received in revised form 16 November 2021; Accepted 17 November 2021
Available online 19 November 2021

accepted that heavy metals can modify the soil's reflection characteristics (Sungur et al., 2014); in other words, the concentration of heavy metal in the polluted soil is sensitive to reflectance, especially those in visible and near-infrared bands. The visible and near-infrared reflectance (VNIR) hyperspectral spectroscopy has advantages of high spectral resolution, continuous band, and rapid acquiring of spectral information, making it well suited for studying the heavy metal concentration. As a result, the VNIR hyperspectral spectroscopy has become one of the powerful methods for quickly identifying the soil heavy metal pollution status. At present, a series of studies addressing the estimation of the heavy metal concentration using spectral reflectance have been conducted in recent years. (Horta et al., 2015; Shi et al., 2014; Tsai and Philpot, 1998).

The estimation of heavy metal using VNIR hyperspectral spectroscopy is affected by many factors, such as spectral preprocessing, spectral dimension reduction, and applied models (Lu et al., 2019). Among those factors, spectral preprocessing is an essential step to reduce (or eliminate) the signal noises for enhancing the spectral features of interest (Gholizadeh et al., 2018). Besides, it could extract crucial spectral information from the overlapping mixed peaks for spectral preprocessing (Chen et al., 2015; Douglas et al., 2018; Hong et al., 2019b). Accordingly, the spectral preprocessing is capable of improving the performance of the hyperspectral model, as proved in numerous studies (Nawar and Mouazen, 2018; Yin et al., 2014). At present, a large number of spectral preprocessing techniques have been proposed, e.g., the spectral derivative and related transformations, signal smoothing, light scattering corrections, etc. (Chakraborty et al., 2017; Chen et al., 2015; Kemper and Sommer, 2002; Liu et al., 2018). However, different spectral preprocessing techniques show disparate performances in enhancing spectral features and improving estimation accuracies for the hyperspectral models under different conditions. Interestingly, the physical and geometrical characteristics of the soil also play a critical role in determining the spectral preprocessing methods (Zhang et al., 2018). As a result, when considering the inversion of soil element at the different study areas, the heterogeneities of soil in both physical and geometrical characteristics makes it impossible to guarantee the robustness of the hyperspectral model even when applying the same spectral pretreatment. Even though, it might be feasible to yield an appropriate spectral pretreatment for small regions with similar physical and geometrical characteristics to enhance the spectral features of the hyperspectral data. Hence, the first aim of this study is choosing a suitable spectral pretreatment method for the arable soil in the study area. Such an aim can be achieved by comparing the accuracies of the hyperspectral model under different spectral pretreatment methods.

The other problem for estimating the heavy metal concentration using hyperspectral data is the problem of the "curse of dimensionality" (Shi et al., 2014; Wang et al., 2018). There are thousands of channels in the hyperspectral data, and most of those channel data are irrelative to the element of interest. Accordingly, the hyperspectral model performance is potentially affected by those unrelated variables that consist of most data in the spectrum. Several studies suggested that the data volume can be significantly reduced by selecting predictor variables and extracting feature parameters (Shi et al., 2014; Xie et al., 2015); thereafter, the hyperspectral model performance could be effectively promoted. At present, the variable selection methods have been explored in several studies, e.g., the Genetic Algorithm (GA), Successive Projection Algorithm (SPA), and Principal Component Analysis (PCA) (Cheng et al., 2019; Leardi and Lupiáñez González, 1998; Shi et al., 2014). The PCA is widely used in the dimensionality reduction and spectral features extraction of the hyperspectral data (Maduranga et al., 2020; Shi et al., 2017; Wu et al., 2005). The reason behind this is that the PCA is capable of extracting information from high-dimensional data by compressing those datasets into various principal components (PCs) (Mishra et al., 2017; Wu et al., 2007). On the other hand, it can retain most of the spectral information in the datasets for the PCs. However, few studies have simultaneously taken account of the impacts of PC numbers on the dimensionality reduction of hyperspectral data and estimation performance of the hyperspectral model. Such an analysis is performed in the study area to select the optimal PCs for minimizing the data volume on the foundation of keeping the performance of the hyperspectral model.

Furthermore, it is vital to choose suitable hyperspectral models for the inversion of element concentration (Chen et al., 2015). Plenty of hyperspectral models were proposed to estimate the heavy metals concentration. Those hyperspectral models can be divided into two families; one is the linear models, e.g., the multiple linear regression (MLR), principal components regression (PCR), and partial least squares regression (PLSR). Another group of models takes the nonlinear forms, such as the gradient boosting decision tree (GBDT), support vector machine (SVM), random forest (RF), and artificial neural network (ANN) (Kemper and Sommer, 2002; Shen et al., 2019; Wu et al., 2005). At present, the linear PLSR model, together with appropriate dimensionality reduction methods, is widely used for investing the soil composition from the spectral reflectance data (Cheng et al., 2019; Viscarra Rossel et al., 2006). For example, Sun et al. (Sun and Zhang, 2017) built the GA-PLSR model for estimating Zn concentration in organic matter and clay minerals based on spectral data. Cheng et al. (Cheng et al., 2019) compared the estimation accuracies of PLSR models for different heavy metals under various spectral pretreatments. However, it is acknowledged that the linear model confronts some constraints to deal with the nonlinearity or randomness problems. Thereupon, the nonlinear hyperspectral models are utilized to solve those problems: Shen et al. applied various nonlinear hyperspectral models to select the optimum model for retrieving Cuprum (Cu) concentration in Daye under three spectral transformation methods(Shen et al., 2019). Wei et al. used three hyperspectral models, namely the PLSR, Radial Basis Function Neural Network (RBFNN) and Shuffled Frog Leaping Algorithm optimization of the RBFNN (SFLA-RBFNN), to establish characteristic wavelengths for estimating Arsenic (As) concentration in Daye (Wei et al., 2020). However, in that study, there is no optimal strategy that relating spectral preprocessing and dimensionality reduction to the selection those characteristic wavelengths; a large amount of spectral information had not been effectively used. Thus, it is meaningful and practical to explore the optimum hyperspectral models after spectral preprocessing and dimensionality reduction. In this regard, one can yield the optimization inversion accuracy using the least amount of spectral data volume possible with the best data quality, which is highly desirable for the hyperspectral inversion. However, there are few works addressing such problems in the hyperspectral technology.

To solve all aforementioned problems, in this study, various spectral pretreatments were performed to process the original hyperspectral data. Then, the PCA was applied to the original hyperspectral data and the data after various spectral pretreatments for the sake of reducing data dimensionality but preserving critical information. Following, the linear model (i.e., PLSR) and nonlinear model (i.e., the GBDT) are performed to retrieve the Cr concentration in soil from the hyperspectral data. Finally, the performances of both models with various spectral pretreatments and PCs are examined for choosing suitable models for the retrieval of Cr concentration in the study area. This study aims at solving three problems, namely: (i) choose a suitable spectral pretreatment for estimating Cr from polluted soil by comparing the performance of different spectral pretreatment methods for the study area; (ii) select the optimal PCs for minimizing the data volume under the premise of keeping the accuracies of hyperspectral models. Such a selection is achieved by comparing the impact of different PCs on the estimation accuracies of the hyperspectral model; (iii) propose an optimal model that comprehensively combines spectral pretreatment, dimensionality reduction and hyperspectral model to retrieve the Cr concentration in the soil from hyperspectral data of the study area.

2. Materials and methods

2.1. Study area and soil sample

Fig. 1 presents the location of the study area, which is in Daye District, the southeast of Hubei Province, China (114°30 ' ~ 115°30′ E, 29°30 ' ~ 30°20′ N). The study area is located in the middle and lower reaches of the Yangtze River. This region has a typical continental monsoon climate with distinct seasons, sufficient sunshine, and abundant rainfall. The terrain of the study area is high in the south and low in the north, while it is flat in the east and west (Shen et al., 2019); it is a hilly landform with an altitude ranging from 120 to 200 m. This region is rich in mineral resources, thus has many mines of large or medium sizes. Consequently, mining and smelting are the primary sources of heavy metal pollution to the surrounding soil.

In September 2019, around the mine regions, it collected 56 surface soil samples from the surrounding arable land, in which the primary type of soil was paddy soil and red soil. One topsoil sample consists of three soil sub-samples that were collected by bamboo shovels, the initial mass of which was heavier than 1000 g. Afterward, all soil samples were dried in the shade; then, they were ground and sieved through a 10-mesh nylon sieve (with a pore diameter of 2 mm) to remove plant residues, rocks, and large debris (Li et al., 2021). In the end, using the method of four distribution, the samples were divided into two portions for the indoor spectrum test and chemical analysis laboratory, respectively. Soil Cr concentration was measured using the determination of 22 metal elements-inductively coupled plasma optical emission spectrometry.

2.2. Spectral measurements and pretreatments

ASD FieldSpec4 spectroradiometer (Analytical Spectral Device, Inc., USA), which covers spectral bands ranging from 350 to 2500 nm, was used to obtain soil spectral data. The interval and resolution of the spectrometer in different bands were as follows: 1.4 nm sampling interval at the range of 350 ~ 1100 nm and 2 nm sampling interval at the range of 1000 ~ 2500 nm. After the spectral resample, a total number of 2,151 bands with a sampling interval of 1 nm were output. Spectral measurement was performed in the darkroom with a 50 W halogen lamp as a light source, which was mounted at an angle of 15°and 50 cm away from the sample without any shadows shade (Shen et al., 2019; Sun and Zhang, 2017). An optical probe was installed approximately 7 cm above the sample surface. To ensure the measurement's accuracy, the spectroradiometers were switched on 30 min earlier before the measurement. Also, the spectroradiometers were optimized using a standardized white $BaSO_4$ panel before the first scan. Ten spectral curves were successively collected from each soil sample uniformly tiled in a petri dish. Finally, wavelengths in the range of 350 ~ 399 nm and 2450 ~ 2500 nm were removed due to the relatively low signal-to-noise ratio (SNR) (Cheng et al., 2019).

Now, the spectral pretreatments can be carried out to the hyperspectral data. In this study, three kinds of spectral pretreatment methods, namely the Savitzky-Golay (SG), orthogonal signal correction (OSC), and first derivative (FD), were applied so as to eliminate noise interference and to improve the inversion accuracy. Besides, the spectral pretreatments might be helpful for exploring characteristic wavelengths or other featured parameters in a more accurate manner. Presented in Fig. 2, the original spectral reflectance and those after spectral pretreatments are plotted as a function of wavelength, from which we observe that there are three absorption valleys in the curvatures (especially in the curvatures after spectral pretreatments) near 1400, 1900, and 2200 nm due to the absorption characteristics of soil clay minerals (Kariuki and Van, 2003; Zhang et al., 2019b).

2.3. Principal component analysis

The PCA is widely used for analyzing datasets and reducing data volume (Abid et al., 2018; Viscarra Rossel et al., 2006). The core principle of PCA is to reduce the dimensionality of a dataset that consists of a large number of interrelated variables but to retain the largest variance and information in a dataset as much as possible (Bolcárová and Kološta, 2015). For high-dimensional spectral data, several sets of orthogonal vectors can be found in the measurement space through a series of matrix changes. The spectral data with multicollinearity variables are turned into a new set of irrelevant variables, a linear combination of the yielded original independent variables (Ağca, 2015; Hong et al., 2019a;

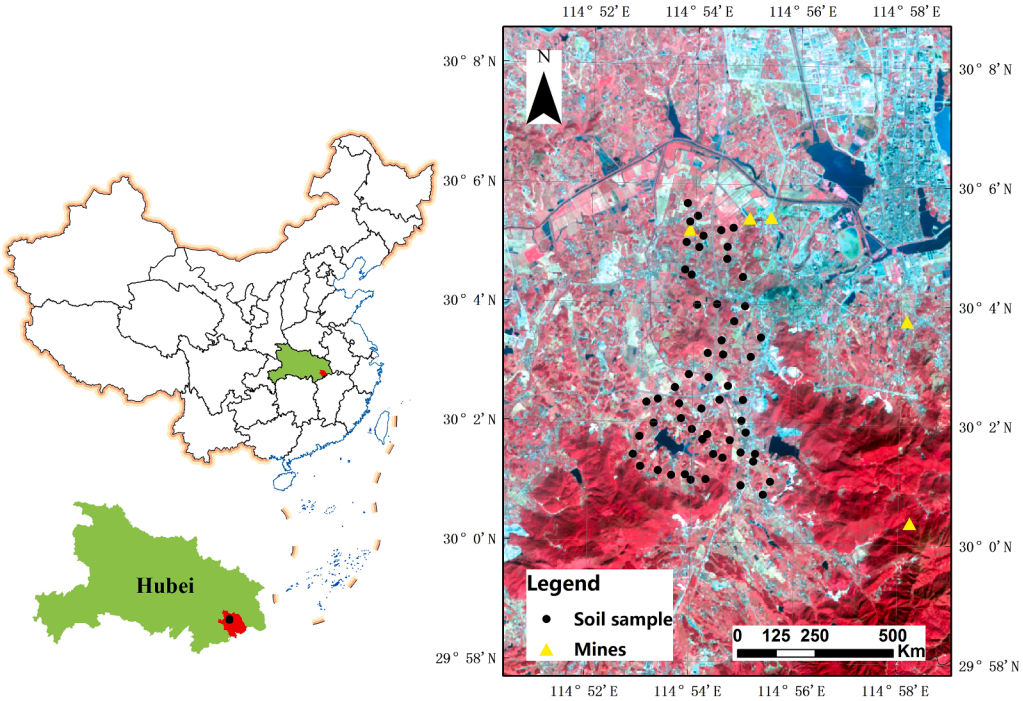


Fig. 1. Study area and sampling points.

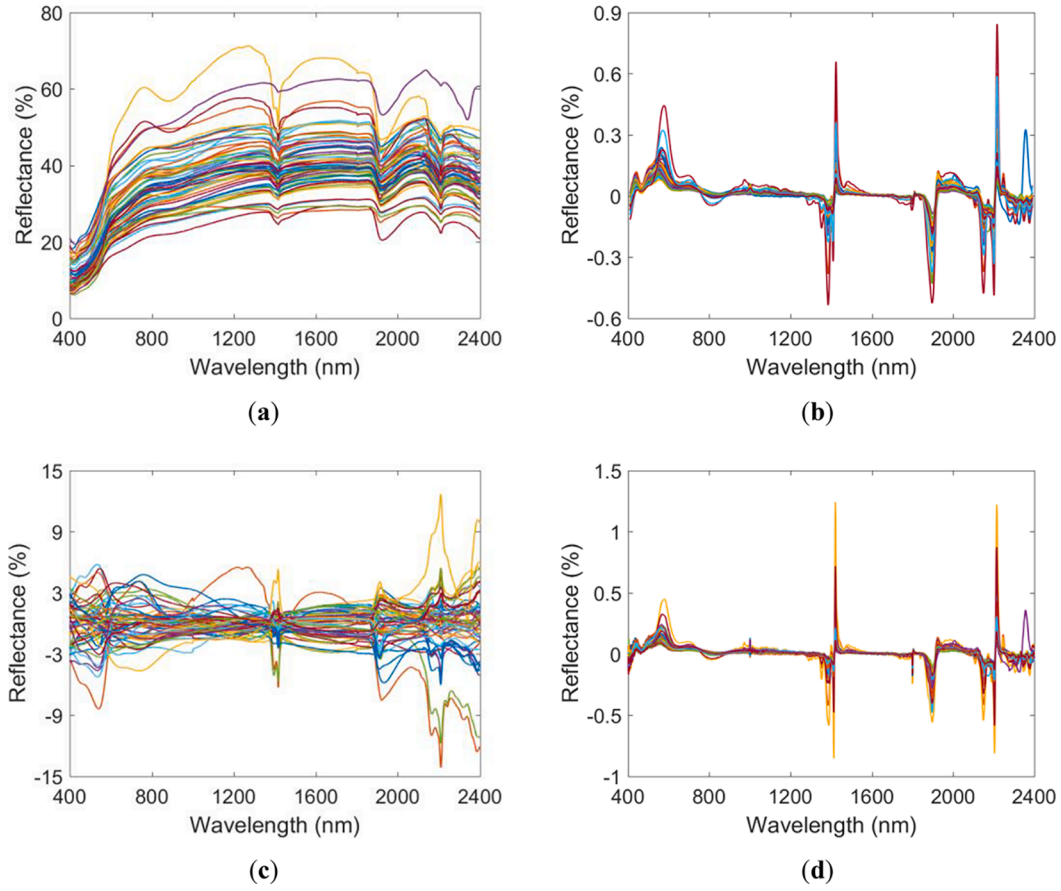


Fig. 2. The original spectral reflectance and the associated three transformations: (a) Original; (b) Savitzky-Golay; (c) OSC; (d) First derivative.

Mishra et al., 2017). It is noteworthy that the first few PCs can retain most of the variations present in all of the original variables. So, the number of retained PCs depends on the percentage of the cumulative variance of that part in the total variance, which would affect the estimation accuracy of the Cr concentration using the hyperspectral model. To analyze the effect of PC numbers, the PCA is performed on the whole dataset, including the original hyperspectral data and those after spectral pretreatment.

2.4. Hyperspectral models and accuracy validation

2.4.1. Hyperspectral models

In this study, the linear hyperspectral model (i.e., PLSR) and nonlinear model (i.e., GBDT) (Brown, 2007) were applied for estimating Cr concentration in soil. The following flowchart in Fig. 3 demonstrates the spectral pretreatments, PCs and hyperspectral models used in this study.

The PLSR is one of the most prevalent linear models proposed by Wold et al. in 1983; such a model is proposed to deal with the issue of spectral data with intense noise or high collinearity, or with the data with variables that significantly exceeds the number of available samples (Boulet et al., 2013; Hong et al., 2019a; Rossel and Behrens, 2010). The method used various principles (including multiple regression, canonical correlation analysis, and PCA) to project a set of variables into a low-dimensional space for the sake of eliminating noise and reducing dimensionality. In recent decades, the PLSR has been widely used to estimate soil heavy metal concentration based on hyperspectral data due to its advantage of providing direct relation between soil attributes and spectral vectors (Geladi and Kowalski, 1986).

Gradient boosting is a method based on the principle of training the newly added weak learners in accordance with the negative gradient information of the loss function of the current model; then combining

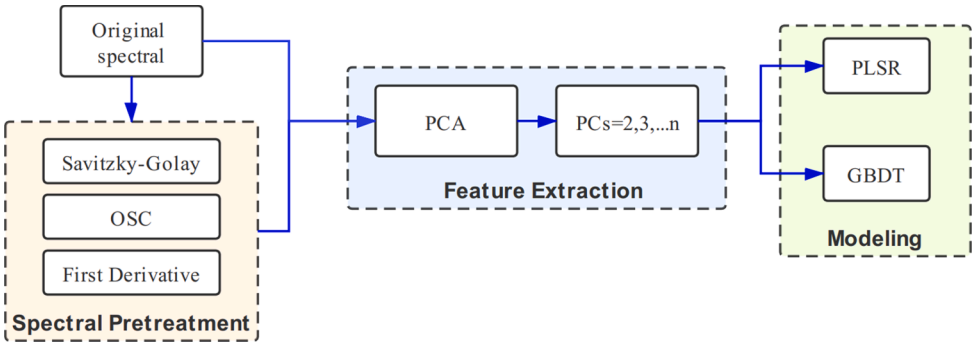


Fig. 3. Flowchart showing the methods of this study.

the trained weak learners into the existing model in a cumulative form (Nawar and Mouazen, 2017; Wei et al., 2019). The GB builds the model in a stepwise manner allowing the use of any differentiable loss function. The most typical basic learner in the GB algorithm is the decision tree (DT), which is thereafter signified as GBDT. The core concept of GBDT is that each calculation is done by a basic DT learner; in the process, it does not produce an independent prediction result but construct a series of DT. In other words, the single DT learner reflects only a part of the observation results, while the final result is the accumulation of all DT components based on a particular weight function. The GBDT model and regression are expressed as follows:

1) Model initialization:

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i, \gamma) \tag{1}$$

2) For m = 1 to M:

a) Calculate the negative gradient:

$$\overline{y}_i = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)} i = 1, 2 \cdots N \tag{2}$$

b) By minimizing the square error, $\overline{y}_i$ is fitted with the basic learner $h_m(x)$:

$$\{R_{jm}\}_1^J = \arg\min_{\{R_{jm}\}_1^J} \sum_{i=1}^{N} \left[\overline{y}_i - h_m\left(x_i; \{R_{jm}; b_{jm}\}_1^J\right)\right]^2 \tag{3}$$

c) Determine the step size γ_{jm} to minimize L

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}}^{N} L[y_i, f_{m-1}(x_i) + \gamma] \tag{4}$$

3) Calculate the negative gradient:

$$f_m(x) = f_{m-1}(x) + \sum_{i=1}^{N} \gamma_{jm} I\left(x \in R_{jm}\right) \tag{5}$$

2.4.2. Accuracy validation

To evaluate the performance of the hyperspectral model, three indexes are adopted in this study. They are determination coefficient (signified as R^2), Root-Mean-Square Error (RMSE), and Residual Prediction Deviation (RPD) (Saeys et al., 2005; Wang et al., 2014); their expressions take the forms of:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2} \tag{7}$$

$$RPD = \frac{S.D}{RMSE} \tag{8}$$

where n is the number of samples; y_i and $\widehat{y}_i$ are measured and predicted data in the validation set, and $\overline{y}_i$ is the mean of samples; S.D is the standard deviation.

A robust hyperspectral model has high R^2 and RPD but a low RMSE. Generally, the R^2 is nearer to 1.0, and the results are more accurate, while the lower the RMSE is, the more precise the results are. Regarding the RPD, it can be classified into three conditions, namely: an excellent prediction is identified by an RPD value exceeding 2.0; high and low possibilities to be distinguished respectively corresponds to RPD values of 1.4 to 2; and unsuccessful prediction has an RPD value below 1.4 (Chang et al., 2001; Sawut et al., 2018).

3. Results

3.1. Statistic analysis of Cr concentration in soil

In this study, all 54 samples were divided into 38 calibration samples and 18 validation samples. The statistical descriptions of Cr, including the minimum, maximum, mean, SD, and coefficient of variation (CV), were summarized in Table 1. From which we observe that the mean Cr concentration of the whole dataset was 65.24 mg/kg, and the maximal value can be up to 116.91 mg/kg. According to China soil element background values reported by China national environmental monitoring centre, the natural background value for the soil Cr is 86 mg/kg. Accordingly, the maximal Cr concentration is 46.1% larger beyond the nature value, indicating the soil in the study area is severely polluted by the Cr concentration. Besides, Table 1 also shows that the distribution of the Cr concentration is inhomogeneous, some region is little affected by the Cr concentration. At the same time, the Cr pollution might be significant in the other areas. The coefficients of variation (CV), which are all over 0.36, also implied that the spatial distribution of Cr concentration was inhomogeneous, suggesting the point source pollution might exist in the study area. Thus, it is urgently needed to establish an accurate inversion model for retrieving the Cr concentration so as to investigate the Cr pollution in the study area through hyperspectral remote sensing from a macro perspective.

3.2. Relationship between the number of PCs and the cumulative contribution rate

PCA was used to reduce the dimensionality of Cr's original spectral and transformation spectral in arable soil. The relationship between the PCs and the cumulative contribution rate of the four kinds of spectral was shown in Fig. 4. When the cumulative contribution rate reached 99.99%, the number of different PCs in each data set was counted and taken as the input variable of the models. For the original spectral, 12 PCs were extracted by PCA when it explains 99.99% of the total variance. However, the number of PCs for other transformed spectral that corresponds to 99.99% of the total variance were 53 PCs of the SG, 19 PCs of the OSC, and 55 PCs of the FD, respectively. It was not that the more PCs that were preserved, the better the inversion performance of the models.

In the modeling process, if a few numbers of PCs were used, it could not fully reflect the changes in the spectral information of the unknown samples, and the accuracy of the models would be reduced, that is, insufficient fitting. If a large number of PCs were used to build the models, some PCs representing noise would be added to the models, which would reduce the prediction performance of the models, that is, overfitting.

3.3. Estimation accuracy based on different modeling methods

In this study, all of the input variables were extracted via dimensionality reduction of PCA to several kinds of spectral as a model of the independent variable (X). The Cr concentration in soil was the

Table 1
Statistic information for soil Cr concentration ($mg \cdot kg^{-1}$) in the study area.

Soil Cr ($mg \cdot kg^{-1}$)	Number	Min	Max	Mean	SD	CV
Calibration set	38	10.55	116.91	65.45	26.13	0.40
Validation set	18	17.36	101.53	64.79	23.80	0.37
Whole dataset	56	10.55	116.91	65.24	25.19	0.39

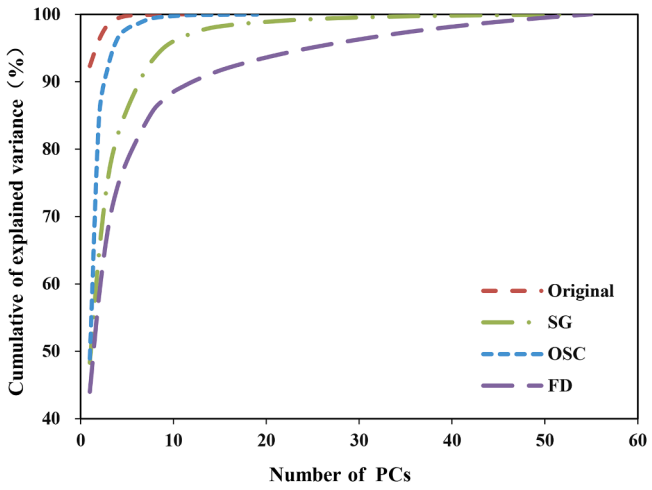


Fig. 4. The relationship between the number of PCs and the cumulative of explained variance.

dependent variable (Y). The PLSR model and the GBDT model are adopted to verify the influence of different PCs on the estimation performance for Cr concentration. Principally, an optimal model with the optimal number of PCs was proposed by comparing the various spectral transformations based on the dimensionality reduction bands of PCA.

3.3.1. Modeling of PLSR

The PLSR was employed to the original spectral data and data after spectral transformations for estimating Cr concentration. In this study, the contributions of different PC numbers (2, 3, …, n) are also probed into for the sake of exploring the optimal PC numbers used for dimensionality reduction. By comparing the effect of PLSR models under different PC numbers, using the evaluation indexes of the models R^2 and RPD, the validation results of the PLSR models were shown in Fig. 5. The experiment was repeated ten times, and the average values were regarded as the final results.

The 11 Ori-PLSR, 52 SG-PLSR, 18 OSC-PLSR, and 54 FD-PLSR models were built to estimate Cr concentration. In PLSR models, for the original spectral, the prediction accuracy of the models was the worst when the number of PCs was 2. It indicated that the model was insufficient fitting. With the number of PCs increasing, the R^2 of the PLSR models gradually improved. The R^2 and RPD of the PLSR models reached up to 0.61 and 1.64, respectively, when the number of PCs increased to 11. The sequence of the maximum R^2 and RPD for the different spectral transformations from largest to smallest was $SG_{max} > FD_{max} > Ori_{max} > OSC_{max}$. The spectral transformation of SG model had a better estimation performance with the maximum $R^2 = 0.78$ and RPD = 1.97 of the model, and its validation accuracy could be classified as a good prediction. Besides, the maximum R^2 and RPD of the FD model were respectively 0.76 and 1.92; therefore, the model also had a good prediction performance. However, the maximum R^2 and RPD of the OSC model were respectively 0.53 and 1.50, which shows a worse estimating performance.

3.3.2. Modeling of GBDT

Combined with the original spectral and spectral transformations, the GBDT models were used as a nonlinear model to estimate the Cr concentration in arable soil with a set of feature variables after dimensionality reduction of PCA. Thus it produces 11 Ori-GBDT, 52 SG-GBDT, 18 OSC-GBDT, and 54 FD-GBDT models, whose validation results are shown in Fig. 6. The arrangement of the maximum R^2 and RPD for the different spectral pretreatment from largest to smallest in GBDT models was $SG_{max} > OSC_{max} > Ori_{max} > FD_{max}$. The spectral transformation of SG in GBDT still had a prominent estimation performance with the maximum $R^2 = 0.80$ and RPD = 2.04 of the model, and its validation accuracy could be classified as an excellent prediction. The minimum R^2 of the SG-GBDT was only 0.14, and the model had an unsuccessful prediction. In comparison, the maximum R^2 and RPD of FD-GBDT models was respectively 0.25 and 1.15, suggesting the model also had an unsuccessful prediction. The estimation performance of the FD-GBDT models with the change of the number of PCs was presented as

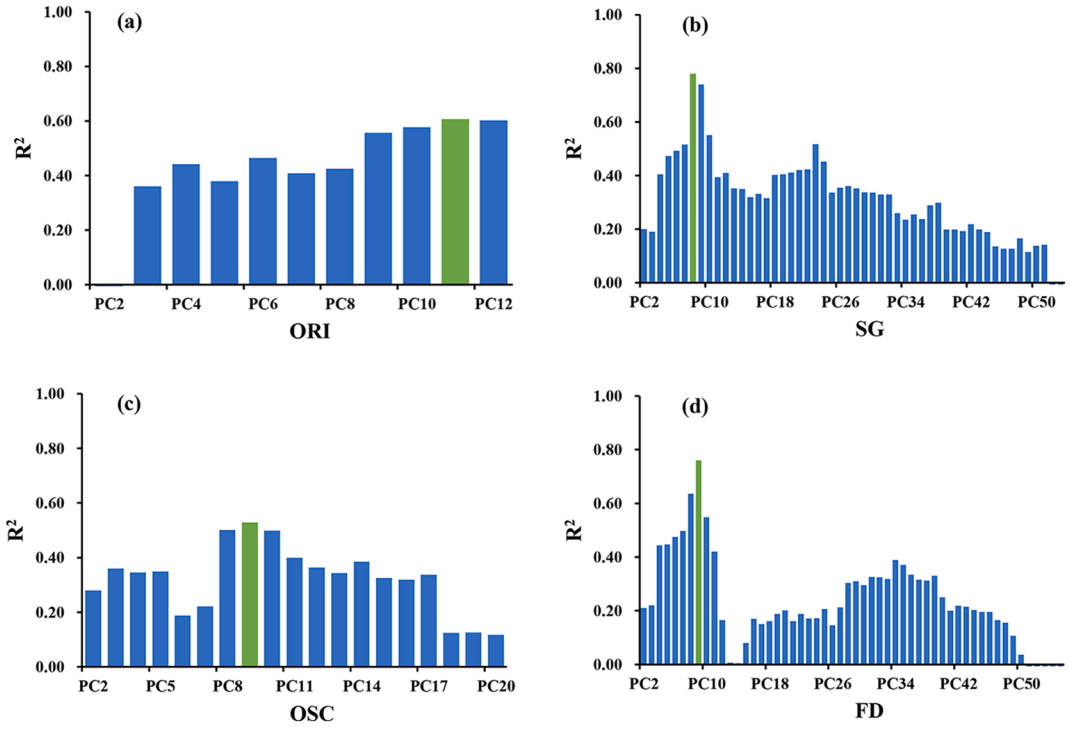


Fig. 5. The validation results of the number of PCs of the original spectral and spectral transformations in PLSR models: (a) Original; (b) Savitzky-Golay; (c) OSC; (d) First derivative.

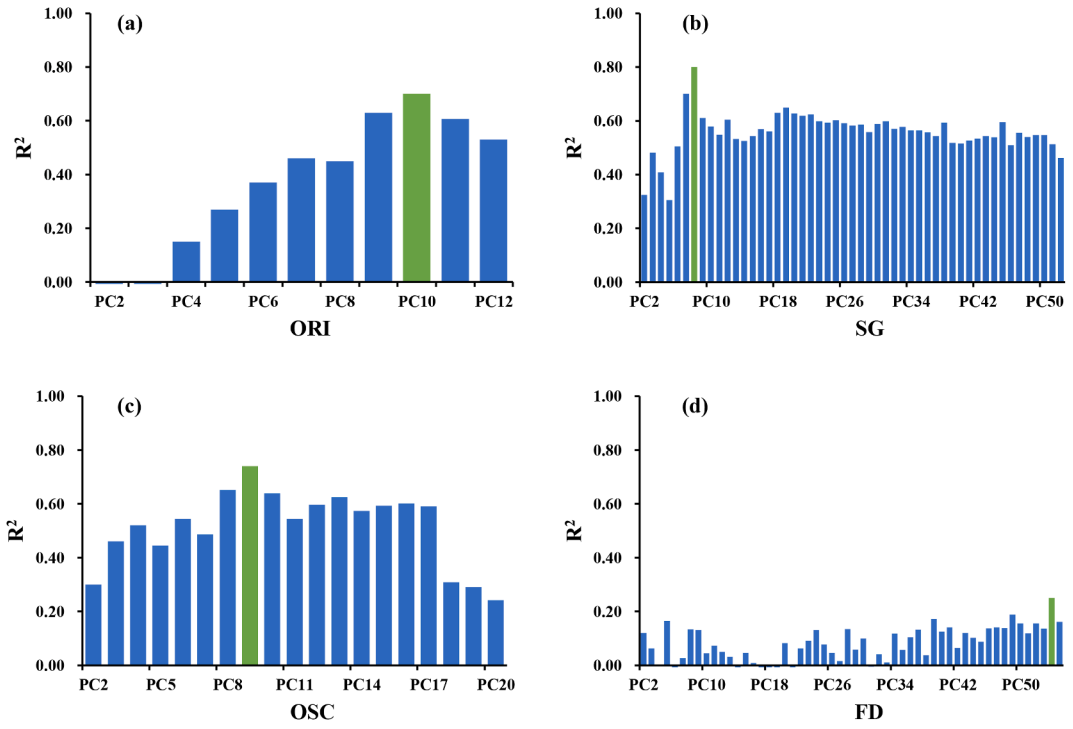


Fig. 6. The validation results of the number of PCs of the original spectral and spectral transformations in PLSR models: (a) Original; (b) Savitzky-Golay; (c) OSC; (d) First derivative.

overfitting. However, the OSC maximum R^2 and RPD of the GBDT were 0.74 and 1.80, respectively. The original GBDT model's maximum R^2 and RPD were 0.75 and 1.75, which showed that both models could distinguish between high and low values.

3.4. The optimal PCs for Cr prediction

The methods of PLSR and GBDT were employed to estimate Cr concentration with the different number of PCs based on the original spectral and three kinds of spectral transformations designed to compare and analyze the optimal estimation performance of the model, and the validation results were displayed in Table 2. In PLSR, the R^2 ranged from −0.02 (PCs = 2) to 0.61(PCs = 12) of original reflectance, however, when R^2 reached its maximum at PCs of 11 ($R^2 = 0.61$ and RPD = 1.64). For the spectral transformations of SG, OSC and FD, the maximum R^2 values were respectively observed from PCs of 8 ($R^2 = 0.78$ and RPD = 1.97), PCs of 9 ($R^2 = 0.53$ and RPD = 1.50), and PCs of 9 ($R^2 = 0.76$ and RPD = 1.92). It indicated that the spectral transformations using SG and FD could improve the inversion capability of the models related to Cr concentration in arable soil, especially the SG method.

In contrast to PLSR, the GBDT modeling performed better than PLSR, and the prediction accuracy had a certain extent of improvement, mainly in Ori-GBDT, SG-GBDT, and OSC-GBDT. The maximum R^2 values varied from 0.61 (PLSR) to 0.75 (GBDT) for the original spectral. Besides, the maximum R^2 values of GBDT models reached their maximum at PCs of 10 in original spectral. Similar to the PLSR trend, the model still had the best estimation performance with the SG transformation. The model of SG-GBDT provided the most successful prediction and outperformed all other tested estimation models. The maximum R^2 value of SG-GBDT was 0.80 (PCs = 9) higher than SG-PLSR 0.78 (PCs = 8) with the RPD = 2.04 > 2.0 indicating its excellent prediction. Through the OSC transformation, the GBDT model maximum R^2 value ($R^2 = 0.74$ and RPD = 1.80) had been greatly improved compared with PLSR ($R^2 = 0.53$ and RPD = 1.50), while the maximum R^2 values were both observed from PCs of 9. However, FD-GBDT is not suitable for estimating Cr concentration due to its maximum R^2 value only 0.25 and maximum RPD = 1.15.

Our results suggested that GBDT models performed better than those of the PLSR models when the input variables were extracted by PCA. Comparison of the R^2 values and RPD for all models with different spectral transformations demonstrated that the models developed with feature variables extracted by PCA using the SG transformation were superior to the other models. The optimal number of PCs in the SG-PCA-GBDT model was 9.

Table 2
Statistic information for soil Cr concentration (mg·kg^{-1}) in the study area.

Modeling	Transform	PC	R^2	RMSE (mg·kg^{-1})	RPD
PLSR	ORI	PC11	0.61	14.51	1.64
	SG	**PC8**	**0.78**	**12.10**	**1.97**
	OSC	PC9	0.53	15.89	1.50
	FD	PC9	0.76	12.42	1.92
GBDT	ORI	PC10	0.75	13.63	1.75
	SG	**PC9**	**0.80**	**11.65**	**2.04**
	OSC	PC9	0.74	13.19	1.80
	FD	PC54	0.25	20.69	1.15

3.5. Spatial distribution of soil Cr concentrations

In this paper, Kriging, one of the typical algorithms in the geostatistics, is used to describe the spatial distribution and variation of Cr concentration (Sun and Zhang, 2017; Wackernagel, 1995). As shown in Fig. 7, ordinary kriging was applied to simulate and map the spatial distribution of Cr concentration in the study area to show the potential polluted region. In Fig. 6, both the measured Cr concentration and estimated Cr concentration based on SG-PLSR with PC number of 8 and that in accordance with the SG-GBDT with PC number of 9 are shown. As can be seen, the spatial distribution of estimated Cr concentrations has approximately similar geographical trends with respect to that of the measured ones. The main difference between the measured Cr concentration and the estimated one based on SG-PLSR was the underestimation of Cr concentration at the interval of 24.5–44.4 mg kg^{-1}, appearing

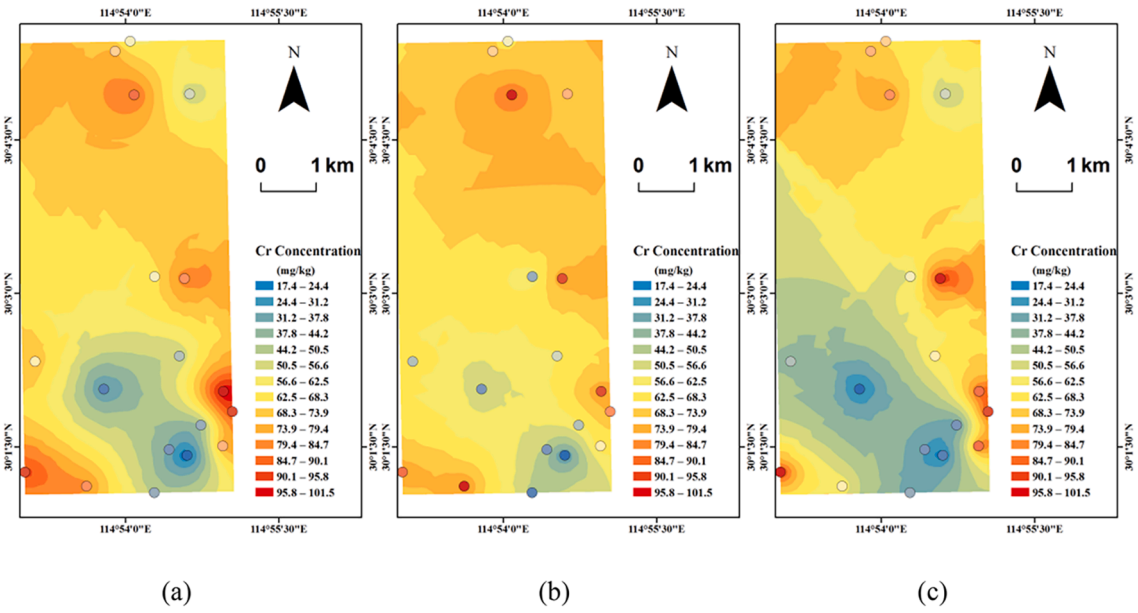


Fig. 7. The study area of soil Cr concentration spatial distribution map: (a) the field measured values; (b) the predicted values by Model SG-PLSR; (c) the predicted values by Model SG-GBDT.

in the northeast of the study area. Interestingly, a comparison of Fig. 7 (a) to (c) showed that some differences also exist between the measured value and estimated one based on SG-GBDT, but the variation in the spatial distribution is lower than that in Fig. 7(b). Moreover, from the distribution of the Cr concentration, it can be concluded that the Cr pollution in the middle part of the study area was lower than the China

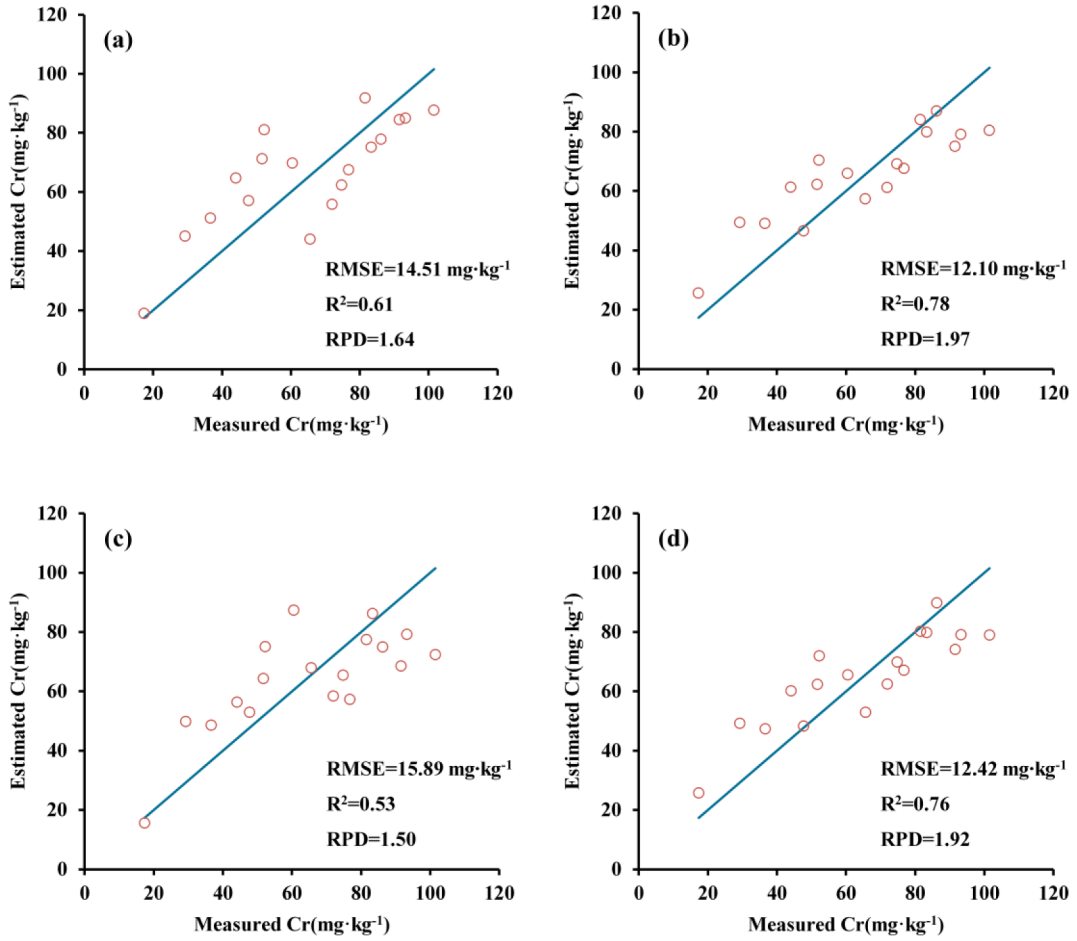


Fig. 8. Scatter plots of optimal number of PCs to estimate Cr concentration using the original spectral and spectral transformations in PLSR models: (a) Original; (b) Savitzky-Golay; (c) OSC; (d) First derivative.

soil elements background values. In contrast, in the north, southeast, and southwest parts of the study area, the enrichment trend of Cr concentration was exceeded the China Soil Elements Background Values, indicating the soil is polluted by the Cr concentration in those regions. Consequently, one might conclude that the polluted region is mainly located in the north, southeast, and southwest regions of the study area.

4. Discussion

Different spectral pretreatments could affect the spectral information and thus further impact the estimation performance of the hyperspectral model. The PCA was capable of achieving the dimensionality reduction for the multi-dimension dataset, while retaining the data information as much as possible (Maduranga et al., 2020; Mishra et al., 2017). The feature variables extracted by the PCA could be good input variables for the hyperspectral models. However, the performance of the hyperspectral model was not related to the number of PCs; in other words, the accuracy of the hyperspectral model did not increase with the increasing of PCs. It was recognized in this study that the models had the best performance only when the optimal PCs were applied.

In this study, the linear model, i.e., the PLSR, and the nonlinear model, i.e., the GBDT, were employed to estimate the Cr concentration in arable soil. The performance of those hyperspectral models could be evaluated by R^2. Fig. 8 presented the accuracy of the PLSR model with the optimal number of PCs for the estimation of the Cr concentration from the data of the original spectral and those based on the spectral transformations.

It can be identified from Fig. 8 that the R^2 of the PLSR model under different spectral transformations (including the original one) was within the scope ranging from 0.53 to 0.78. Such results were similar to the results yielded by the PLSR model in the other related studies, as presented in Table 3. Accordingly, the PLSR model, together with the optimal PCs, was capable of providing a potential new method to the retrieval of Cr concentration from the polluted arable soil in a more effective way with satisfactory accuracies.

However, the relationship between the Cr and the spectral of the arable soil is further complicated, which is generally randomness and nonlinearity. Thus, the spectral proprieties of the arable soil are difficult to explain in several bands. For a more rigorous analysis, a nonlinear model, i.e., the GBDT, was selected to estimate the Cr concentration, the accuracies of which are shown in Figue.9. Additionally, the R^2 and RPD were used to evaluate the performance of the GBDT model. As seen from Fig. 9 and Table 2, the performance of the GBDT model was overall slightly superior to the PLSR models when the original spectral and spectral pretreatments of SG and OSC were applied. Interestingly, in contrast to the PLSR model with FD spectral preprocessing, the GBDT model with the FD spectral preprocessing performs worse, the R^2 of which was only 0.25. Although the FD transformations had a range of advantages in reducing overlapping spectral bands, improving the spectral resolutions and sensitivities, and eliminating the interferences caused by other background factors (Hong et al., 2019a), yet it introduced some noises and reduced the spectral strength. This may be the reason why the FD was not applicable to estimate the Cr concentration with GBDT (Hong et al., 2019a). Simultaneously, it suggested that the data pretreatments played a significant role in the estimation performance of the hyperspectral model.

It was also found that the optimal estimation performance for the Cr estimation was obtained by the GBDT model combined with the data based on SG spectral pretreatment method. The SG also had the best performance in the PLSR with respect to the rest three spectral pretreatments. Besides, in view of the problem of information redundancy in hyperspectral data, most studies used the correlation coefficients or stepwise regression methods to filter out feature bands. Still, a large number of bands with an abundance of helpful information were discarded when such approaches were applied. It was of great significance to select latent parameters which could not only guarantee the main information concentration of the band but also reduce the input variables. One of the aims of this study was to check whether the feature variables screened by the PCA were suitable input variables to the models and whether the optimal number of PCs had the best prediction performance regarding the hyperspectral models. For this purpose, several spectral pretreatments, i.e., the SG, OSC, and FD, and two kinds of models, i.e., the PLSR and GBDT, were applied. We analyzed the estimation results in the PLSR (Fig. 8), and GBDT models (Fig. 9) used different PCs obtained by PCA to identify whether PCA could provide suitable input variables to the models. As shown in Fig. 8, all the PLSR models had a certain ability to estimate the Cr concentration in arable soil. Moreover, SG-PCA-PLSR exhibited the best performance for estimation when the first 8 PCs were applied. Similarly, the dimensionality reduction of PCA for the GBDT models also showed good performance for the original spectral, as well as the SG and OSC spectral pretreatment. The SG-PCA-GBDT showed the most excellent estimation performance with the highest precision by using the first 9 PCs.

Although the SG-PCA-GBDT model shows a good performance for the retrieval of soil Cr using the hyperspectral data, yet the corresponding processing is relatively complex. Besides, there are still some shortcomings that limit the further application of the SG-PCA-GBDT model. For example, the accuracy of the proposed model is affected by many factors, including the types of soil, land-use types, heavy metal element and its concentration, as well as the chemical state of the heavy metal. In addition, the spatial resolution and sample sizes of the hyperspectral data would also affect the hyperspectral model performance. Among the aforementioned factors, the land-use types (i.e., grassland, forest land, and cropland) and soil types (i.e., paddy soil, red soil, calcareous soil, and soil PH) could directly affect the spectral reflectance even without considering heavy metal element. Those characteristics, together with the impact of heavy metal concentration, make the mechanism of the hyperspectral model more complicated. Inversely, the corresponding study could significantly contribute to methodological development and to reduce uncertainty in the hyperspectral model, which would be further remediated in our future study.

Table 3
The estimation accuracy of Cr concentration in other studies and this study.

Sampling site	Number of samples	Content range ($mg \cdot kg^{-1}$)	Method	R^2	Studies
Suburban area	120	28.7–105.0	PLSR	0.76	(Wu et al., 2005)
River sediments	117	5.0–175.0	PLSR	0.66	(Moros et al., 2009)
Reclaimed	39	103.0–397.0	PLSR	0.75	(Zhang et al., 2019a)
Arable land	**58**	**10.6–116.9**	**PLSR**	**0.78**	**This study**
Arable land	**58**	**10.6–116.9**	**GBDT**	**0.80**	**This study**

5. Conclusions

In this study, we took the arable soil around the mining area in Daye, Hubei Province as the study area. After spectral preprocessing with various methods, the PCA was applied to reduce the data dimensionality, based on which the performances of the linear model (i.e., the PLSR) and the nonlinear model (i.e., the GBDT) for the estimation of the Cr concentration in arable soil were analyzed. The conclusions were as follows:

(1) The study area is polluted since the maximum Cr concentration of the study area was 46.1% larger beyond the national pollution threshold; besides, the distribution of the Cr concentration is

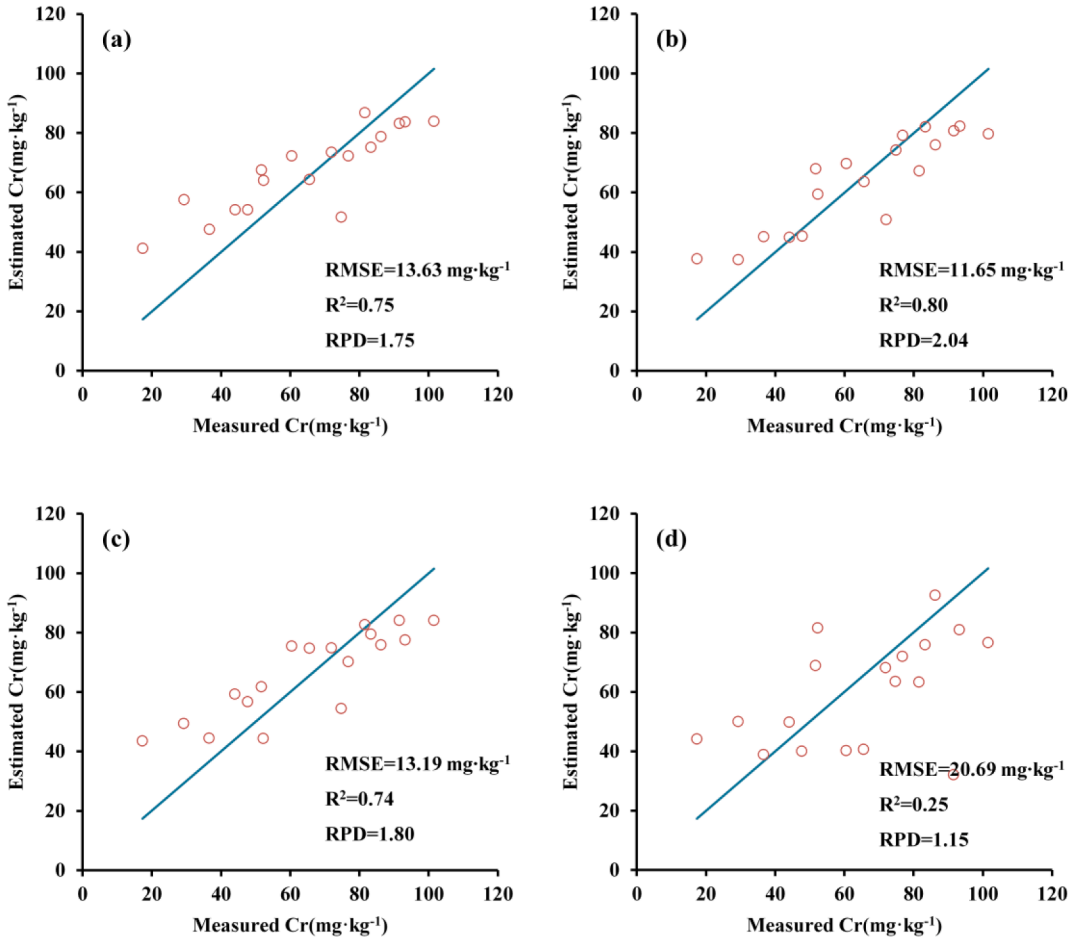


Fig. 9. Scatter plots of optimal number of PCs to estimate Cr concentration using the original spectral and spectral transformations in GBDT models: (a) Original; (b) Savitzky-Golay; (c) OSC; (d) First derivative.

inhomogeneous, and there might be point sources of pollution in the study area.

(2) The SG method had the best performance with respect to the rest spectral pretreatment methods for estimating Cr concentration in the study area, which could be regarded as a suitable pretreatment method to deal with the spectral data in the retrieval of Cr concentration.

(3) The PCA could effectively reduce the dimensionality of the hyperspectral data. The optimal PC numbers were related to the specified hyperspectral model. The best optimal PCs in the PCA were the first 9 PCs for the GBDT model, while that for PLSR is the first 8 PCs.

(4) The GBDT model shows better performance and robustness with respect to the PLSR model; in other words, the nonlinear hyperspectral model performs better than the linear hyperspectral model in this study. Further, once the spectral pretreatment and dimensionality reduction are comprehensively taken into account, the optimal estimation model for the Cr concentration was the SG-PCA-GBDT model, the R^2 and RPD values of which were 0.80 and 2.04, respectively.

In conclusion, the proposed method can effectively retrieve the Cr concentration from the polluted arable soil in the study area. However, it is noteworthy that the land-use types and soil types could potentially affect the performance of hyperspectral models, even the combination of spectral pretreatment, dimensionality reduction, and hyperspectral models. Such an interesting topic would be probed in our future study.

CRediT authorship contribution statement

Fei Guo: Conceptualization, Methodology, Writing – original draft, Formal analysis, Software. **Zhen Xu:** Formal analysis, Writing – review & editing, Software. **Honghong Ma:** Formal analysis, Validation. **Xiujin Liu:** Investigation. **Shiqi Tang:** Investigation. **Zheng Yang:** Formal analysis. **Li Zhang:** Formal analysis. **Fei Liu:** Resources. **Min Peng:** Supervision. **Kuo Li:** 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.

Acknowledgements

This work was supported in part by the Director Foundation of the Institute of Geophysical and Geochemical Exploration, Chinese Academy of Geological Sciences under Grant AS2019J02, in part by National Science Foundation of China under Grant 42101398, and in part by Shantou University (STU) Scientific Research Foundation for Talents under Grant NTF20023.

Appendix. Nomenclature in the paper

PLSR	Partial least squares regression
GBDT	Gradient boosting decision tree
PCA	Principal component analysis
PCs	Principal components
GA	Genetic algorithm
SPA	Successive projection algorithm
MLR	Multiple linear regression
PCR	Principal components regression
SVM	Support vector machine
RF	Random forest
ANN	Artificial neural network
RBFNN	Radial basis function neural network
SFLA-RBFNN	Shuffled frog leaping algorithm optimization of the RBFNN
SG	Savitzky-golay
OSC	Orthogonal signal correction
FD	First derivative
RMSE	Root-mean-square
RPD	Residual prediction deviation
CV	Coefficient of variation

References

Abid, A., Zhang, M.J., Bagaria, V.K., Zou, J., 2018. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nat. Commun. 9, 2134.

Ağca, N., 2015. Spatial distribution of heavy metal content in soils around an industrial area in Southern Turkey. Arabian J. Geosci. 8 (2), 1111–1123.

Bhattacharya, P.T., Misra, S.R., Hussain, M., 2016. Nutritional Aspects of Essential Trace Elements in Oral Health and Disease: An Extensive Review. Scientifica (Cairo) 2016, 1–12.

Bolcárová, P., Kološta, S., 2015. Assessment of sustainable development in the EU 27 using aggregated SD index. Ecol. Ind. 48, 699–705.

Boulet, J.-C., Bertrand, D., Mazerolles, G., Sabatier, R., Roger, J.-M., 2013. A family of regression methods derived from standard PLSR. Chemometr. Intell. Laboratory Syst. 120, 116–125.

Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140 (4), 444–453.

Chakraborty, S., Weindorf, D.C., Deb, S., Li, B., Paul, S., Choudhury, A., Ray, D.P., 2017. Rapid assessment of regional soil arsenic pollution risk via diffuse reflectance spectroscopy. Geoderma 289, 72–81.

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. Soil Sci. Soc. Am. J. 65 (2), 480–490.

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.

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.

Chovanec, P., Sparacino-Watkins, C., Zhang, N., Basu, P., Stolz, J., 2012. Microbial Reduction of Chromate in the Presence of Nitrate by Three Nitrate Respiring Organisms. Front. Microbiol. 3.

Douglas, R.K., Nawar, S., Cipullo, S., Alamar, M.C., Coulon, F., Mouazen, A.M., 2018. Evaluation of vis-NIR reflectance spectroscopy sensitivity to weathering for enhanced assessment of oil contaminated soils. Sci. Total Environ. 626, 1108–1120.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Analytica Chimica Acta 185, 1–17.

Gholizadeh, A., Saberioon, M., Carmon, N., Boruvka, L., Ben-Dor, E., 2018. Examining the Performance of PARACUDA-II Data-Mining Engine versus Selected Techniques to Model Soil Carbon from Reflectance Spectra. Remote Sens. 10 (8), 1172. https://doi.org/10.3390/rs10081172.

Hong, Y., Chen, S., Liu, Y., Zhang, Y., Yu, L., Chen, Y., Liu, Y., Cheng, H., Liu, Y.i., 2019a. Combination of fractional order derivative and memory-based learning algorithm to improve the estimation accuracy of soil organic matter by visible and near-infrared spectroscopy. Catena 174, 104–116.

Hong, Y., Liu, Y., Chen, Y., Liu, Y., Yu, L., Liu, Y.i., Cheng, H., 2019b. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. Geoderma 337, 758–769.

Hong, Y., Shen, R., Cheng, H., Chen, S., Chen, Y., Guo, L., He, J., Liu, Y., Yu, L., Liu, Y.i., 2019c. Cadmium concentration estimation in peri-urban agricultural soils: Using reflectance spectroscopy, soil auxiliary information, or a combination of both? Geoderma 354, 113875. https://doi.org/10.1016/j.geoderma.2019.07.033.

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R., Pozza, L., 2015. Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: A prospective review. Geoderma. 241-242, 180–209.

Kariuki, P.C., Van, D., 2003. Determination of soil activity from optical spectroscopy. International Conference on Remote Sensing.

Kemper, T., Sommer, S., 2002. Estimate of Heavy Metal Contamination in Soils after a Mining Accident Using Reflectance Spectroscopy. Environ. Sci. Technol. 36 (12), 2742–2747.

Kimbrough, D.E., Cohen, Y., Winer, A.M., Creelman, L., Mabuni, C., 1999. A Critical Assessment of Chromium in the Environment. Crit. Rev. Environ. Sci. Technol. 29 (1), 1–46.

Leardi, R., Lupiáñez González, A., 1998. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemometr. Intell. Laboratory Syst. 41 (2), 195–207.

Leenaers, H., Okx, J.P., Burrough, P.A., 1990. Employing elevation data for efficient mapping of soil pollution on floodplains. Soil Use Manag. 6 (3), 105–114.

Li, C., Yang, Z., Yu, T., Hou, Q., Liu, X.u., Wang, J., Zhang, Q., Wu, T., 2021. Study on safe usage of agricultural land in karst and non-karst areas based on soil Cd and prediction of Cd in rice: A case study of Heng County, Guangxi. Ecotoxicol. Environ. Safety 208, 111505. https://doi.org/10.1016/j.ecoenv.2020.111505.

Liu, J., Zhang, Y., Wang, H., Du, Y., 2018. Study on the prediction of soil heavy metal elements content based on visible near-infrared spectroscopy. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 199, 43–49.

Lu, Q., Wang, S., Bai, X., Liu, F., Wang, M., Wang, J., Tian, S., 2019. Rapid inversion of heavy metal concentration in karst grain producing areas based on hyperspectral bands associated with soil components. Microchem. J. 148, 404–411.

Maduranga, U., Wijegunarathna, K., Weerasinghe, S., Perera, I., Wickramarachchi, A., 2020. Dimensionality Reduction for Cluster Identification in Metagenomics using Autoencoders.

Mishra, S.P., Sarkar, U., Taraphder, S., Datta, S., Swain, D.P., Saikhom, R., Panda, S., Laishram, M., 2017. Multivariate Statistical Data Analysis- Principal Component Analysis (PCA).

Moros, J., Vallejuelo, S.-O., Gredilla, A., Diego, A.d., Madariaga, J.M., Garrigues, S., Guardia, M.d.l., 2009. Use of Reflectance Infrared Spectroscopy for Monitoring the Metal Content of the Estuarine Sediments of the Nerbioi-Ibaizabal River (Metropolitan Bilbao, Bay of Biscay, Basque Country). Environ. Sci. Technol. 43 (24), 9314–9320.

Nawar, S., Mouazen, A.M., 2018. Optimal sample selection for measurement of soil organic carbon using online vis-NIR spectroscopy. Comput. Electron. Agric. 151, 469–477.

Nawar, S., Mouazen, A., 2017. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. Sensors 17 (10), 2428. https://doi.org/10.3390/s17102428.

Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1-2), 46–54.

Saeys, W., Mouazen, A.M., Ramon, H., 2005. Potential for Onsite and Online Analysis of Pig Manure using Visible and Near Infrared Reflectance Spectroscopy. Biosyst. Eng. 91 (4), 393–402.

Sawut, R., Kasim, N., Abliz, A., Hu, L.i., Yalkun, A., Maihemuti, B., Qingdong, S., 2018. Possibility of optimized indices for the assessment of heavy metal contents in soil around an open pit coal mine area. Int. J. Appl. Earth Observ. Geoinform. 73, 14–25.

Shen, Q., Xia, K.e., Zhang, S., Kong, C., Hu, Q., Yang, S., 2019. Hyperspectral indirect inversion of heavy-metal copper in reclaimed soil of iron ore area. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 222, 117191. https://doi.org/10.1016/j.saa.2019.117191.

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., Liu, H., Chen, Y., Fei, T., Wang, J., Wu, G., 2017. Spectroscopic Diagnosis of Arsenic Contamination in Agricultural Soils. Sensors 17 (5), 1036. https://doi.org/10.3390/s17051036.

Sun, W., Zhang, X., 2017. Estimating soil zinc concentrations using reflectance spectroscopy. Int. J. Appl. Earth Observ. Geoinform. 58, 126–133.

Sun, W., Zhang, X., Sun, X., Sun, Y., Cen, Y.i., 2018. Predicting nickel concentration in soil using reflectance spectroscopy associated with organic matter and clay minerals. Geoderma 327, 25–35.

Sungur, A., Soylak, M., Ozcan, H., 2014. Investigation of heavy metal mobility and availability by the BCR sequential extraction procedure: relationship between soil properties and heavy metals availability. Chem. Speciat. Bioavailab. 26 (4), 219–230.

Tepe, Y., A., 2014. Toxic Metals: Trace Metals – Chromium, Nickel, Copper, and Aluminum. Encyclopedia of Food Safety, 356-362.

Tsai, F., Philpot, W., 1998. Derivative analysis of hyperspectral data. Remote Sens. Environ. 66 (1), 41–51.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1-2), 59–75.

Wackernagel, H., 1995. Multivariate Geostatistics: An Introduction with Applications. Multivariate Geostatistics: An Introduction with Applications.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D.e., Simpson, M., McGowen, I., Sides, T., 2018. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. Ecol. Ind. 88, 425–438.

Wang, J., Cui, L., Gao, W., Shi, T., Chen, Y., Gao, Y., 2014. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216, 1–9.

Wei, B., Yang, L., 2010. A Review of Heavy Metal Contaminations in Urban Soils, Urban Road Dusts and Agricultural Soils From China. Microchem. J. 94 (2), 99–107.

Wei, L., Pu, H., Wang, Z., Yuan, Z., Yan, X., Cao, L., 2020. Estimation of Soil Arsenic Content with Hyperspectral Remote Sensing. Sensors 20 (14), 4056. https://doi.org/10.3390/s20144056.

Wei, L., Yuan, Z., Zhong, Y., Yang, L., Hu, X., Zhang, Y., 2019. An Improved Gradient Boosting Regression Tree Estimation Model for Soil Heavy Metal (Arsenic) Pollution Monitoring Using Hyperspectral Remote Sensing. Appl. Sci. 9 (9), 1943. https://doi.org/10.3390/app9091943.

Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., Ma, H., 2007. A Mechanism Study of Reflectance Spectroscopy for Investigating Heavy Metals in Soils. Soil Sci. Soc. Am. J. - SSSAJ 71.

Wu, Y., Chen, J., Wu, X., Tian, Q., Ji, J., Qin, Z., 2005. Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. Appl. Geochem. 20, 1051–1059.

Xie, H., Zhao, J., Wang, Q., Sui, Y., Wang, J., Yang, X., Zhang, X., Liang, C., 2015. Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis. Sci. Rep. 5, 10930.

Yin, G., Lijuan, C., Bing, L., Yanfang, Z., Tiezhu, S., 2014. Estimating Soil Organic Carbon Content with Visible–Near-Infrared (Vis-NIR) Spectroscopy. Appl. Spectrosc.

Zhang, B., Dai, D., Huang, J., Zhou, J., Gui, Q., Dai, F., 2018. Influence of physical and biological variability and solution methods in fruit and vegetable quality nondestructive inspection by using imaging and near-infrared spectroscopy techniques: A review. Crit. Rev. Food Sci. Nutr. 58, 2099–2118.

Zhang, S., Shen, Q., Nie, C., Huang, Y., Wang, J., Hu, Q., Ding, X., Zhou, Y., Chen, Y., 2019a. Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods. Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 211, 393–400.

Zhang, X., Sun, W., Cen, Y., Zhang, L., Wang, N., 2019b. Predicting cadmium concentration in soils using laboratory and field reflectance spectroscopy. Sci. Total Environ. 650, 321–334.