Environmental Pollution 291 (2021) 118128

Contents lists available at ScienceDirect

Environmental Pollution

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

Diagnosis of cadmium contamination in urban and suburban soils using visible-to-near-infrared spectroscopy☆

Yongsheng Hong [a,b], Yiyun Chen [a,*], Ruili Shen [c], Songchao Chen [d], Gang Xu [e], Hang Cheng [a], Long Guo [f], Zushuai Wei [g], Jian Yang [g], Yaolin Liu [a], Zhou Shi [d], Abdul M. Mouazen [b]

[a] School of Resource and Environmental Sciences, Wuhan University, Wuhan, 430079, China
[b] Department of Environment, Ghent University, Coupure Links 653, 9000, Gent, Belgium
[c] Hubei Academy of Environmental Sciences, Wuhan, 430072, China
[d] Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, 310058, China
[e] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, 430079, China
[f] College of Resources and Environment, Huazhong Agricultural University, Wuhan, 430070, China
[g] South China Institute of Environmental Sciences, Ministry of Ecology and Environment, Guangzhou, 510530, China

ARTICLE INFO

Keywords:
Urban and suburban soil Cd contamination
Visible-to-near-infrared spectroscopy
Boruta algorithm
Synthetic minority over-sampling technique
Machine learning

ABSTRACT

Previous studies have mostly focused on using visible-to-near-infrared spectral technique to quantitatively estimate soil cadmium (Cd) content, whereas little attention has been paid to identifying soil Cd contamination from a perspective of spectral classification. Here, we developed a framework to compare the potential of two spectral transformations (i.e., raw reflectance and continuum removal [CR]), three optimization strategies (i.e., full-spectrum, Boruta feature selection, and synthetic minority over-sampling technique [SMOTE]), and three classification algorithms (i.e., partial least squares discriminant analysis, random forest [RF], and support vector machine) for diagnosing soil Cd contamination. A total of 536 soil samples were collected from urban and suburban areas located in Wuhan City, China. Specifically, Boruta and SMOTE strategies were aimed at selecting the most informative predictors and obtaining balanced training datasets, respectively. Results indicated that soils contaminated by Cd induced decrease in spectral reflectance magnitude. Classification models developed after Boruta and SMOTE strategies out-performed to those from full-spectrum. A diagnose model combining CR preprocessing, SMOTE strategy, and RF algorithm achieved the highest validation accuracy for soil Cd (Kappa = 0.74). This study provides a theoretical reference for rapid identification of and monitoring of soil Cd contamination in urban and suburban areas.

1. Introduction

Soil contamination with potentially toxic elements (PTEs) has been recognized as a global concern (Hou et al., 2017; Liu et al., 2013; McBratney et al., 2014; Zhang et al., 2019c; Zhao et al., 2015). In recent decades, the rapid industrialization and urbanization exert a great impact on the urban soil characteristics, which result in emissions of a large amount of pollutants, inevitably affecting the human health and urban ecosystems (Cheng et al., 2019; Hong et al., 2020; Li et al., 2018; Yuan et al., 2020). Urban soils mainly exist in parks, gardens, and other green spaces in the city, which are generally the repositories of pollutants. Urban residents have frequent and direct contact with the soils in these places (Li et al., 2018; Luo et al., 2012). Suburban soils are the spatial transition zones connecting urban and rural areas, and are vital in safeguarding food security and balancing local surrounding ecological systems (Hong et al., 2019; Wu et al., 2020). As the provincial capital of Hubei Province, Wuhan City has a large population density and developed industry. Since the beginning of the 21st century, Wuhan has experienced a rapid process of population growth, industrial development, and urbanization (Zhang et al., 2018). Soil cadmium (Cd) can pose a serious threat to human health, as it is a possible risk factor to human by causing lung cancer and other chronic diseases (Poggio et al., 2009; Proctor et al., 2006). Therefore, accurate diagnosing of suspected contaminated samples above the alert limit (i.e., a threshold) in urban

☆ This paper has been recommended for acceptance by Dr. Yong Sik Ok.
* Corresponding author.
E-mail address: chenyy@whu.edu.cn (Y. Chen).

https://doi.org/10.1016/j.envpol.2021.118128
Received 6 February 2021; Received in revised form 11 August 2021; Accepted 5 September 2021
Available online 9 September 2021

and suburban soils is necessary for risk assessment.

Commonly, the measurement of soil Cd concentration requires field sampling campaign, followed by laboratory chemical analysis. This approach faces challenges of being labor-intensive, time-consuming, environmentally unfriendly, and high expertise demanding, which do not allow meeting the needs for rapid measurement and high density sampling for soil Cd (Jia et al., 2021; Lassalle et al., 2020; Meng et al., 2020; Tan et al., 2020). Proximal soil sensing techniques like visible-to-near-infrared (Vis–NIR) spectroscopy for estimating PTEs have attracted the attention of researchers, mainly because of the advantages of its less sample preparation and rapid characterization (Cheng et al., 2019; Gholizadeh et al., 2018; Lassalle et al., 2020; Nawar et al., 2019; Ng et al., 2020; Sawut et al., 2018; Shi et al., 2014; Shi et al., 2017; Sun and Zhang, 2017; Todorova et al., 2014; Wang et al., 2018; Wang et al., 2014; Zhang et al., 2019b). There are various studies adopting Vis–NIR technique to estimate soil Cd, with sampling sites covering suburban soils, agricultural soils, mining regions, and sewage irrigation areas (Fig. 1). However, very few studies have focused on urban soils. Besides, the goal of all these studies was to quantitatively estimate soil Cd content. For practical applications concerning the qualitative discrimination of soil Cd contamination related to human health risk assessment and environmental ecosystem management, directly diagnosing of soil Cd contamination from Vis–NIR spectral data may be more efficient than quantifying soil Cd. Here's an interesting question: why don't we link Vis–NIR signals directly to soil Cd contamination so as to address the practical concerns? Therefore, by adopting the threshold of risk alarm, the quantitative estimation of soil Cd is converted into the multivariate classification for diagnosing soil Cd contamination.

Soil Vis–NIR spectral data having hundreds or thousands of wavelength variables are relatively of weak and broad absorption bands, mainly because of overtones and combinations of fundamental vibration occurring in the mid-infrared spectral region (Stenberg et al., 2010; Viscarra Rossel and Behrens, 2010; Viscarra Rossel et al., 2016). Many challenges arise during the application of the Vis–NIR spectral data to identify soil Cd contamination, such as the data high-dimensionality and the selection of most robust classification model (Shi et al., 2017). Linear classification models, such as partial least squares discriminant analysis (PLSDA), are often used in spectral modeling, due to their simple structure and easy interpretability (Yu et al., 2018). Nonparametric machine learning techniques, such as random forest (RF) and support vector machine (SVM), are versatile in modeling complicated and nonlinear relationships with high spectral dimensionality (Almeida et al., 2019; Nawar and Mouazen, 2017; Ng et al., 2020; Ng et al., 2019; Viscarra Rossel and Behrens, 2010). In addition to identifying the best classification algorithm, an equally important challenge in spectral analysis is to select the most informative spectral subset and get rid of redundant wavebands (Shi et al., 2017). The common practice of the use of the whole spectrum to train a model would be relatively complex, and potentially produce inefficient model interpretations (Raj et al., 2018; Shi et al., 2014; Vohland et al., 2014). Moreover, some spectral variables may contain irrelevant and even noisy information, which may distort the true relationship between soil Cd and predictors from Vis–NIR spectra. To overcome these shortcomings of the full-spectrum analyses, variable selection approaches should be explored. Among existing algorithms, the Boruta feature selection has the advantages of simplifying model structure, maximizing model performance, and facilitating model interpretation (Kursa et al., 2010; Kursa and Rudnicki, 2010; Prasad et al., 2019).

With the wide expansion of data availability in soil science, the problem of learning from imbalanced data with skewed distribution is a relatively new challenge that should be well explored. The imbalanced learning problem is related to the model accuracy of learning problems in the presence of underrepresented data and class distribution skews (He and Garcia, 2009). In soil data classification (for either digital soil mapping or spectral classification), the model accuracy is usually dependent on the number of classes and the frequency distribution of soil observations, which are the results of the environmental complexity of soil forming factor that affect soil property spatially (McBratney et al., 2003; Sharififar et al., 2019a; Sharififar et al., 2019b; Xie and Li, 2018). Thus, one critical issue that affects the model classification performance of soil contamination is the imbalanced number of samples among different classes. Several common machine learning (ML) algorithms consider balanced-based training dataset, whose all specific classes are broadly and equally represented. This pattern would result in a bias in predictions towards the majority classes with large sample size and the misclassification of or ignoring of the minority classes having a small number of samples (Chawla et al., 2002; Sharififar et al., 2019b). Many soil contamination datasets in real-world applications are imbalanced, especially for data containing local contamination hot spots. In the field of ML, this issue is called imbalanced class problem. Although imbalanced classification is recognized as a modeling problem in ML, this subject has not been well explored in soil spectral classification, particularly for diagnosing soil contamination (e.g., Cd). To address the issue of the imbalanced classification, the synthetic minority over-sampling technique (SMOTE) could be successfully used, however, it has not been utilized so far for Vis–NIR diagnose of soil Cd.

Given the importance of assessing soil Cd contamination, the aim of

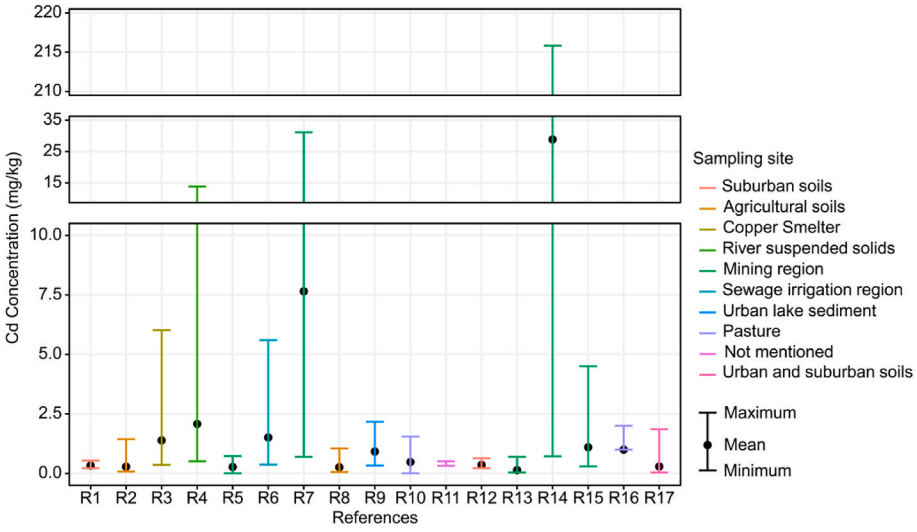

Fig. 1. Review summarizing previous studies reporting soil Cd concentration (including minimum, mean, and maximum value) estimated by Vis–NIR spectra. R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, and R17 refer to Wu et al. (2007), Song et al. (2012), Xie et al. (2012), Song et al. (2013), Gholizadeh et al. (2015), Chen et al. (2015), Rathod et al. (2016), St. Luce et al. (2017), Jiang et al. (2018), Stafford et al. (2018), Liu et al. (2018b), Cheng et al. (2019), Hou et al. (2019), Zhang et al. (2019b), Zhang et al. (2019a), Lamine et al. (2019), and this study, respectively. These references are classified according to the type of sampling site.

this study was to explore the best integrated modeling approach of Vis–NIR spectra to diagnose Cd in urban and suburban soils. The following objectives were thought: (1) analyze and understand the effect of soil Cd contamination on raw reflectance (RR) and continuum removal (CR) spectra; (2) compare the predictive potentials of three optimization strategies (i.e., full-spectrum, Boruta selection, and SMOTE) coupled with three ML classification algorithms (i.e., PLSDA, RF, and SVM) in diagnosing soil Cd contamination; and (3) determine the important wavelengths and spectral mechanism for soil Cd contamination diagnosis.

2. Materials and methods

2.1. Study area and sample collection

The study area, covering about 8569.15 km^2, is located in Wuhan City (the capital of Hubei province, China) (Fig. 2). This area is under a subtropical humid monsoon climate, and is characterized by four distinct seasons, with the average annual temperature and precipitation of 15.8–17.5 °C and 1150–1450 mm, respectively. Most of the rainfall occurs mainly from April to October. Rivers, lakes, and ponds are extensively scattered throughout the city, accounting for 26.1 % of the entire area. Soil parent materials are dominated by Quaternary clay, and river and lake sediments.

The prevailing agricultural products in suburban area contain fruits (including watermelons, grapes, peaches, and strawberries), vegetables, oilseed rape, wheat, and soybeans. Due to the increasing demands for foods in urban areas, agricultural production in the suburban region is generally intensive. According to interviews with local farmers and other available information, fertilizers and pesticides are intensively used in agricultural practices to improve the yield and quality of agricultural produces. In addition, there are some chemical industries, and iron and steel industries scattering in urban area.

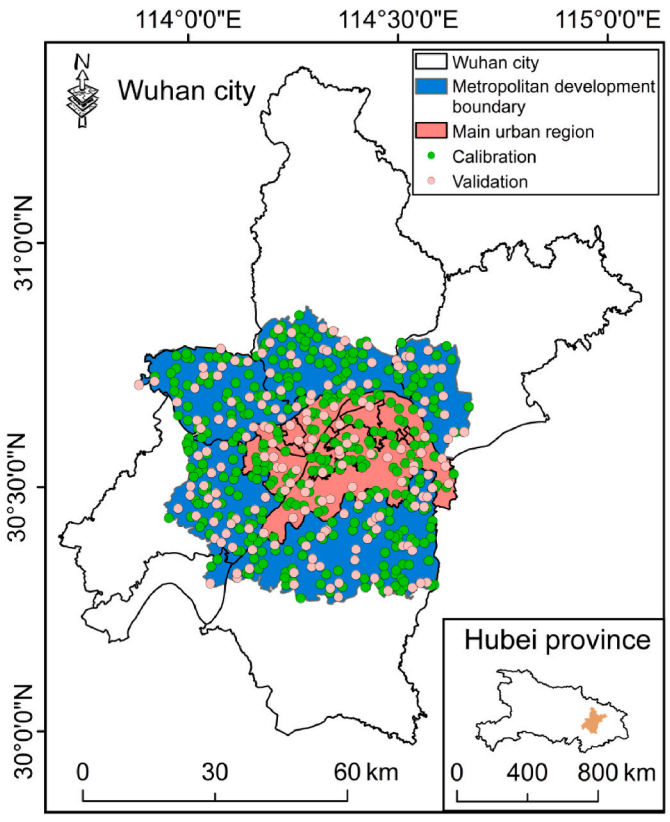

Fig. 2. Location of sampling points within the study area in the urban and suburban of Wuhan city, Hubei Province, China.

Depending on the specific land use type in urban and suburban areas, we adopt a random sampling scheme to collect 536 soil samples in 2012, 2013, and 2014 (Fig. 2). The sampling sites covered a wide range of land uses, including cultivated land, transportation land, grassland, and garden land. The geographical coordinates of all sampling sites were recorded with a hand-held global positioning system. Each sample consisted of 5 sub-samples, collected at a depth of 0–20 cm using a wooden shovel. During sampling, it was necessary to remove weeds, roots, gravel and other materials. Samples were packed into zip-lock plastic bags to avoid cross-contamination of the samples, and later brought back to the laboratory. Soil samples were air-dried in the laboratory, ground by an agate mortar, and then passed through 2 mm sieve before spectral measurement and laboratory chemical analysis to determine Cd and other key soil properties using methods detailed in the following section.

2.2. Chemical analysis and contamination assessment

To determine soil Cd and soil organic matter (SOM), iron (Fe), and pH, 536 soil samples were further ground and sieved through a 0.15 mm sieve. The following chemical analyses strictly followed the relevant standards of China's Technical Specifications for Soil Environmental Monitoring (Agricultural Chemistry Committee of China, 1983; CNMEE, 2018). For determining Cd concentration, the samples were first digested by HNO_3 and $HClO_4$, and then measured by an inductively coupled plasma mass spectrometry (Cheng et al., 2019). The SOM concentrations were measured by wet oxidation at 180 °C, following the potassium dichromate method (Cheng et al., 2019). The Fe contents were determined by a power X−ray fluorescence spectrometry (Cheng et al., 2019). Soil pH values were measured by an electronic digital pH meter with a water-to-soil ratio of 2.5:1 (Bao, 2005). For quality assurance and quality control, reagent blanks, analytical duplicates, and standard reference materials were utilized during the experiments (Qu et al., 2018).

Using a risk screening value of 0.30 mg/kg in China as the threshold value, the measured soil Cd content was classified into two categories, and further coded into binary 0 or 1 to indicate each individual as uncontaminated or contaminated sample, respectively (CNMEE, 2018).

2.3. Spectral measurement and preprocessing

All the ground and sieved samples had been placed in black petri dishes before they were scanned in a dark room to record the laboratory Vis–NIR spectral reflectance, using an ASD spectrometer that covers the spectral range from 350 to 2500 nm, with a final spectral output interval of 1 nm (Analytical Spectral Devices, Boulder, CO, USA). With a 45° light incident angle, a 50 W halogen lamp positioned 30 cm away from the soil sample was used as the illumination. The fiber optic sensor was mounted vertically 12 cm above the sample surface. The spectrometer was calibrated using a standard spectral panel that was repeated every 10 samples. Each sample was recorded 10 times, and the collected spectra were then averaged in one representative spectrum.

For spectral preprocessing, we only retained the spectral domain of 400–2400 nm to discard the noisy bands. A second order Savitzky–Golay algorithm with a window size of 11 was used to smooth the spectral data (Savitzky and Golay, 1964). To minimize the spectral multicollinearity, we resampled the spectral data to an interval of 10 nm, thus resulting in 201 wavebands in total. Spectral reflectance processed by the above steps was denoted as RR. The CR preprocessing was used to isolate the additional spectral peaks that are not easy to be observed from original spectra (Clark and Roush, 1984). Savitzky–Golay smoothing and CR processing were implemented in R with *prospectr* package (R Core Team, 2017; Stevens and Ramirez–Lopez, 2014).

2.4. Calibration and validation subsets

We used conditioned Latin hypercube sampling to divide the entire dataset ($N = 536$) into calibration (67 % of the data, $N = 360$) and validation (33 % of the data, $N = 176$) parts (Minasny and McBratney, 2006; Ramirez-Lopez et al., 2014). The calibration set aimed at identifying the spectral pattern of different classes by fitting the classification models, whereas the validation set was utilized for the evaluation of model performance.

2.5. Boruta feature selection

Boruta selection algorithm was used to reduce the data redundancy for minimizing the model complexity and to identify the most suitable predictor subset for modeling (Kursa et al., 2010; Kursa and Rudnicki, 2010; Prasad et al., 2019). Both RR and CR transformations were subjected to Boruta selection, which was computed in R software with *Boruta* package (Kursa and Rudnicki, 2010).

2.6. Synthetic minority over-sampling technique

Since samples in the minority classes are easily misclassified when ML algorithms are directly used to handle dataset with skewed and imbalanced sample distribution, new soil classes are generated using the SMOTE algorithm to obtain balanced (e.g., equal number of samples per class) class observations (Chawla et al., 2002; Sharififar et al., 2019a; Xie and Li, 2018). The SMOTE technique runs oversampling interpolation by introducing synthetic examples in the spectral space, joining any or all of the k-nearest neighbors (Chawla et al., 2002). Data treatment was performed in R software with the *DMwR* package. The *perc.over* and *perc.under* values in the SMOTE algorithm were set to ensure that, after application of the SMOTE, the number of sample in the minority class should be same to that in the majority class. In our case, the values of *perc.over* and *perc.under* were defined as 200 and 150, respectively. According to Xie and Li (2018), the parameter *k* that controls how many of the nearest neighbor samples are used to generate new examples was set to 5. For the original calibration dataset, two different types of new SMOTE-processed calibration sets were generated for RR and CR spectral transformations, respectively.

2.7. Model establishment

Three classification algorithms were considered (Table 1), encompassing three different types of approaches: (1) linear modeling (i.e., PLSDA); (2) tree-based modeling (i.e., RF); (3) kernel-based modeling (i.e., SVM). All these three methods were performed in R with *caret* package (Kuhn, 2008), which was cooperated with other packages also listed in Table 1. A brief introduction to each modeling approach is given below. For more detailed information, the reader is referred to the relevant literature provided.

The first modeling method was PLSDA, which is a parametric algorithm that can account for multivariate relationship between categorical response variable and spectral predictor (Wold et al., 2001). It converts high dimensional spectral data into some new latent variables, which are then used as new predictors for classification (Yu et al., 2018). The optimal number of components (i.e., *ncomp* parameter) was tuned with 10 repeats of 10-fold cross-validation.

The second modeling technique was RF, which is an integrated machine learning approach combining different predictions from multiple classification and regression trees in an average manner (Breiman, 2001). This method is widely used in the fields of remote sensing and spectral analysis because of its good predictive ability for high-dimensional spectral data (Belgiu and Drăguţ, 2016; Nawar and Mouazen, 2019). Moreover, it can effectively overcome problems associated with outliers, noisy samples, and model over-fitting. The output of RF technique can be interpreted by means of the variable importance, which provides a means for ranking the predictors influencing the response. The mean decrease in accuracy that is calculated from permuting out-of-bag data was specified as the measure type of variable importance. We used 10 repeats of 10-fold cross-validation method to tune the m_{try} parameter (i.e., number of randomly selected predictors) in the RF model.

The third modeling technique was SVM, which belongs to a non-parametric data mining algorithm, very popular in data regression and classification (Vapnik, 1999). With regard to the categorical variables, it divides the entire dataset into multiple classes by establishing hyperplanes in multidimensional feature space (Mountrakis et al., 2011). Following the principle of structural risk minimization, this algorithm is insensitive to over-fitting (Chen et al., 2019). The SVM modeling was carried out using *C*-classification and radial basis kernel "Gaussian" function. Within the framework of 10 repeats of 10-fold cross-validation, we used grid searching approach to optimize the parameters of *sigma* and *C* to achieve the optimal combination.

These three algorithms were applied following three different types of optimization strategies (i.e., full-spectrum, Boruta selection, and SMOTE) subjected to RR and CR spectral transformations for soil Cd binary contamination diagnosis, yielding 18 predictive models in total.

Table 1
Description of the three classification models considered in this study.

Type	Model	Abbr.	Parameter	R package
Linear	Partial least squares discriminant analysis	PLSDA	*ncomp*	*caret, pls*
Tree-based	Random forest	RF	*mtry*	*caret, randomForest*
Kernel-based	Support vector machines with radial basis function kernel	SVM	*C, sigma*	*caret, kernlab*

2.8. Accuracy assessment

Three metrics, including overall accuracy (OA), Kappa, and Matthews correlation coefficient (MCC), were used as evaluation indicators for assessing the predictive models (Boughorbel et al., 2017; Congalton, 1991). The OA indicates the proportion of total correctly diagnosed samples in the dataset as uncontaminated or contaminated. Kappa coefficient is used to measure the difference between observed agreement and accidental expected agreement (Cohen, 1960).

$$OA = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$Kappa = 1 - \frac{1 - OA}{1 - P_e} \tag{2}$$

where *TP*, *TN*, *FP*, and *FN* stand for the numbers of true positive, true negative, false positive, and false negative, respectively. P_e is the proportion of units of chance agreement.

The MCC is a balanced metric measuring the classification performance in multi-class problems and unbalanced datasets (Matthews, 1975). In the case of two-class classification, MCC can be written as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

For binary variables, MCC can be regarded as a discrete index of Pearson correlation. Following the suggestions by Xie and Li (2018), we interpreted the predictive model as: the absolute value of MCC within 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1 denote very weak, weak, moderate, strong, and very strong agreements, correspondingly. These three indicators were computed in R with *rminer* package (Paulo, 2013).

3. Results

3.1. Descriptive statistics

The summary statistics for laboratory measured soil Cd, SOM, Fe, and pH are displayed in Table 2. The Cd concentration of the entire dataset varied from 0.04 to 1.86 (mg/kg), with the coefficient of variation (CV) of 77.82 %. The CV values of SOM and Fe were 75.25 % and 22.12 %, respectively. The pH ranged from 4.28 to 9.17, with a mean value of 7.24. Following the classification levels of CV reported from Wilding (1985), values of CV < 15 %, 15 % < CV < 35 %, and CV > 35 % correspond to low, moderate, and high data variability, respectively. Thus, the 536 samples used in the study were categorized as high variability, indicating the high soil Cd spatial variation within the study area. According to the risk screening limit for soil Cd, approximately 33.96 % of soil samples in the entire dataset (i.e., 182 out of 536 samples) were contaminated (Table 2). These contaminated soil samples may be affected by anthropogenic activities and natural factors' associated changes in the landscape, e.g., erosion, flooding, and the like. Depending on the risk screening limit of soil Cd, 236 and 124 samples in the calibration dataset, and 118 and 58 samples in the validation dataset, were identified as uncontaminated and contaminated samples, respectively (Table 2).

To compare with the Cd concentration from other studies, we summarized the literature using spectral technique to estimate soil Cd sourced from various types of sampling sites (Fig. 1). Overall, our study presented lower mean value of soil Cd concentration, as compared to most other studies. In addition, compared to the two articles using suburban soils for Cd estimation of Wu et al. (2007) and Cheng et al. (2019), the averaged Cd concentration in this paper was also lower. However, the maximum value in our dataset was 1.32 and 1.22 (mg/kg) higher than that from Wu et al. (2007) and Cheng et al. (2019), respectively. These results further reflect that there are local hot spots of Cd contamination source in the study area.

3.2. Spectral reflectance characterization

The mean spectra along with spectral standard deviations of contaminated and uncontaminated samples were plotted for RR and CR spectra in Fig. 3a and b, respectively. The RR spectra of the contaminated and uncontaminated samples presented similar spectral trends and shapes, but different absorption depths and reflectance intensity. The contaminated samples tended to have lower reflectance magnitude than the uncontaminated samples. This may be caused by the darker hue of the contaminated sample, resulting in higher absorption and lower reflection of the emitted light, and thus lower reflectance intensities observed (Horta et al., 2015; Nawar et al., 2019; Shi et al., 2014). However, it is not easy to visualize the contamination degree using the RR spectra. The preprocessed spectra after CR transformation exhibited pronounced absorption valleys around 450, 1430, and 1940 nm. Besides, the contaminated and uncontaminated samples could be visually distinguished at wavelengths approximately within 440–540, 1380–1550, and 1910–1990 nm. The CR preprocessing highlights the differences between contaminated and uncontaminated samples in different spectral regions and effectively constructs a new feature space.

We also plotted the mean raw spectral curves of SOM and Fe at different content classes (Fig. 3). Overall, spectral curves of these two soil properties followed a similar pattern: the reflectance decreased with increasing concentration. This trend may be attributable to the increasing light absorptions of the blue color associated with increasing SOM, and the red color associated with Fe oxides, resulting in a decrease in the intensity of the spectral curves (Nocita et al., 2014; Stenberg et al., 2010; Viscarra Rossel et al., 2016).

3.3. Spectral mechanism of soil Cd diagnoses

The correlation coefficients among the four analyzed soil properties in the calibration dataset (Fig. S1) demonstrated that Cd was substantially and positively correlated with SOM and pH, with correlation values of 0.52 ($P < 0.001$) and 0.33 ($P < 0.001$), respectively. There was a positive correlation between pH and Fe ($r = 0.19$, $P < 0.001$). In addition, we also analyzed the RR spectral correlations with soil Cd and the other three soil properties (Fig. S2). Among SOM, Fe, and pH, the correlation pattern of soil Cd to reflectance spectra was quite similar to that of SOM and pH, which corroborates the correlation analysis in Fig. S1. Since soil pH has no direct spectral response, the potential mechanism to explain the spectral correlation of soil Cd should be assigned to its surrogated correlation with SOM, through which an accurate spectral diagnosis model for soil Cd element can be developed.

3.4. Boruta selection and SMOTE

The results regarding Boruta selection carried out using RR and CR spectra are shown in Fig. 4a and b, respectively. A total of 21 wavebands were selected from RR spectra as significant variables, and were primarily located within 570–660, 1150–1160, 1940–1950, and 2220–2230 nm spectral bands. For the CR spectra, a total of 32 spectral predictors were selected mainly around 430–560, 1370–1460, 1910, 1950–1970, 2180–2220, and 2270–2320 nm spectral ranges. The wavelength distribution of the selected subsets for CR spectra in Fig. 4b compares well with that of the absorption features of CR spectra in Fig. 3b.

Following the working procedure of SMOTE algorithm defined in Section 2.6, with respect to RR spectra in the calibration set, two groups of resampled but balanced datasets, both with 372 samples, were generated for uncontaminated and contaminated samples, respectively. Likewise, two sets of balanced datasets with equal sample size (i.e., both with 372 samples) were also achieved for CR spectra.

3.5. Multivariate modeling

We numbered each diagnosis model to better distinguish them

Table 2
Summary statistics of soil Cd, soil organic matter (SOM), iron (Fe), and pH.

Variable	Sample set	N[a]	Min	Max	Mean	Median	CV (%)[b]	Background value[c]	Threshold[d]	Contamination ratio (%)[e]
Cd (mg/kg)	Entire	536	0.04	1.86	0.29	0.23	77.82	0.17	0.30	33.96
	Calibration	360	0.04	1.63	0.29	0.23	74.83	0.17	0.30	34.44
	Validation	176	0.05	1.86	0.29	0.23	83.77	0.17	0.30	32.95
SOM (%)	Entire	536	0.10	16.12	2.24	1.92	75.25			
Fe (%)	Entire	536	1.04	9.38	5.69	5.58	22.12			
pH	Entire	536	4.28	9.17	7.24	7.64	15.41			

[a] Sample number.
[b] Coefficient of variation.
[c] Soil Cd background value in Hubei province (Cheng et al., 2019).
[d] Risk screening value for soil Cd in China (CNMEE, 2018).
[e] Percentage of contaminated samples (the risk screening value is set as threshold).

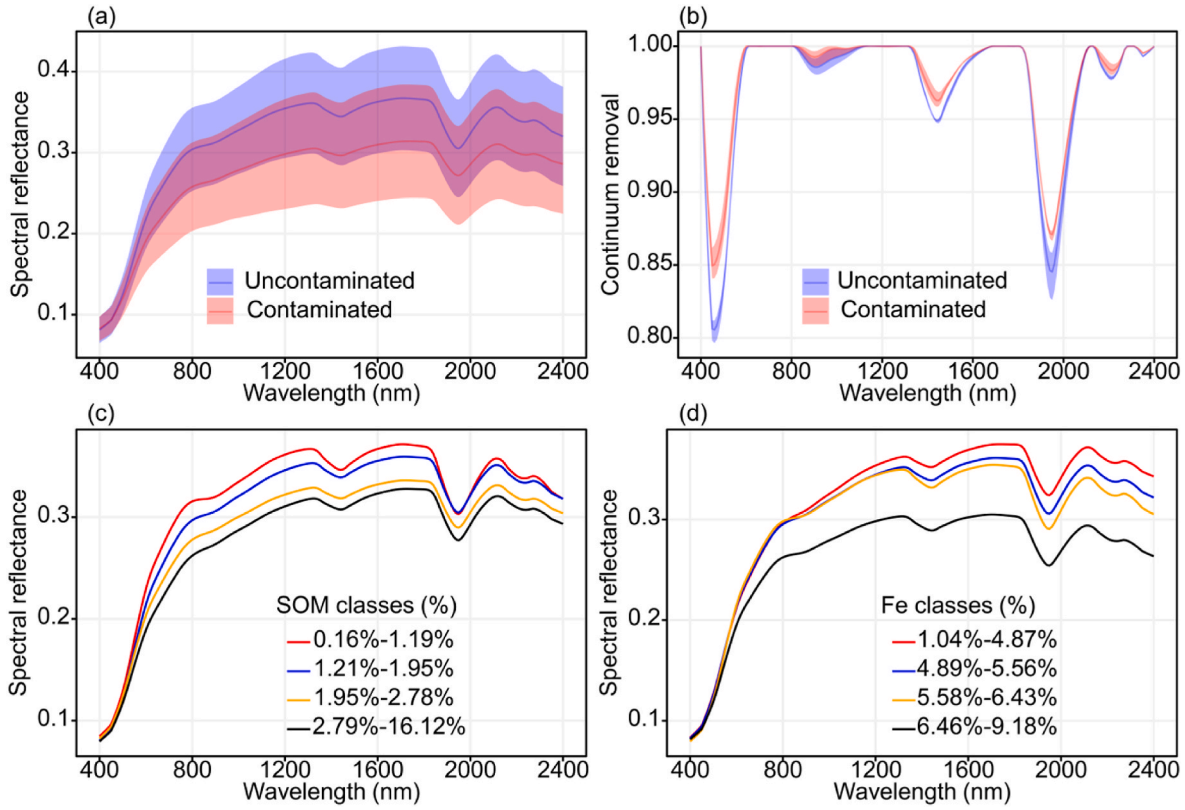

Fig. 3. Calibration dataset: mean RR (a) and CR processed (b) spectra of uncontaminated and contaminated soils, and mean reflectance spectra (c and d) grouped by different SOM and Fe classes that are uniformly divided based on the concentration of all samples from small to large. The colored regions in the first two subplots indicate the corresponding spectral standard deviations.

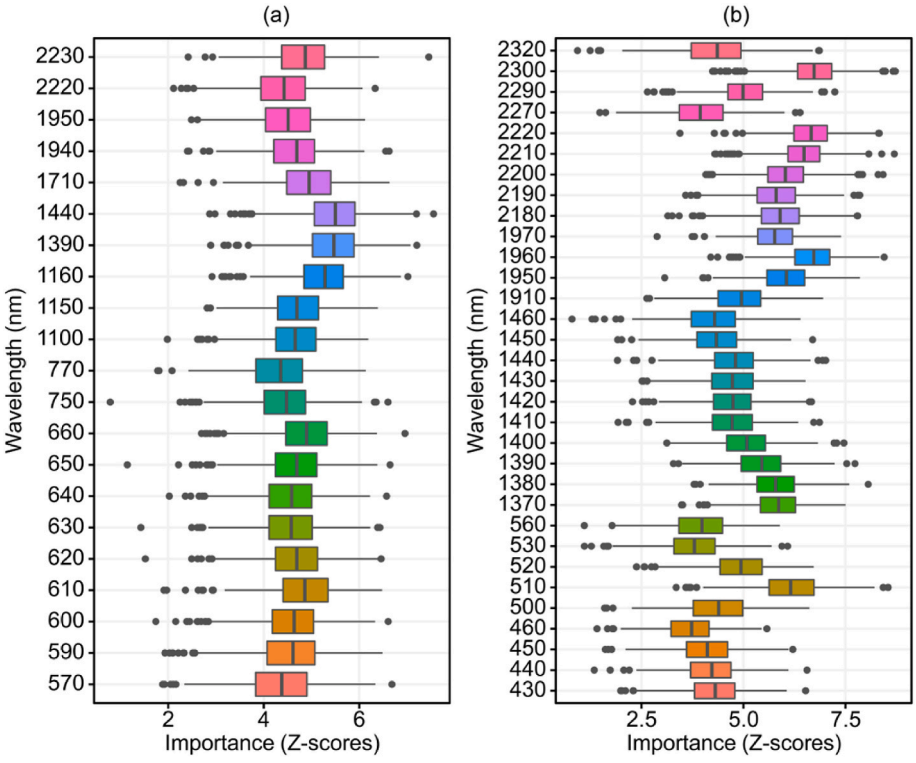

Fig. 4. Informative wavelengths selected by Boruta approach for RR (a) and CR (b) spectra in the calibration dataset ($N = 360$).

Table 3
Classification performance for soil Cd diagnosis using Vis–NIR spectra subjected to different optimization strategies, spectral transformations, and modeling algorithms.

Optimization strategy	Spectral transformation[b]	Modeling algorithm	Model[c]	Validation		
				OA[d]	Kappa	MCC[e]
Full-spectrum	RR	PLSDA	Model 1	0.75	0.43	0.58
		RF	Model 2	0.77	0.48	0.60
		SVM	Model 3	0.77	0.46	0.58
	CR	PLSDA	Model 4	0.77	0.46	0.58
		RF	Model 5	0.80	0.54	0.63
		SVM	Model 6	0.82	0.57	0.65
Boruta	RR	PLSDA	Model 7	0.82	0.59	0.66
		RF	Model 8	0.84	0.63	0.69
		SVM	Model 9	0.83	0.62	0.69
	CR	PLSDA	Model 10	0.85	0.65	0.70
		RF	Model 11	0.87	0.71	0.74
		SVM	Model 12	0.85	0.66	0.71
SMOTE[a]	RR	PLSDA	Model 13	0.84	0.64	0.69
		RF	Model 14	0.85	0.68	0.72
		SVM	Model 15	0.86	0.66	0.71
	CR	PLSDA	Model 16	0.85	0.67	0.72
		RF	Model 17	0.88	0.74	0.77
		SVM	Model 18	0.88	0.72	0.76

[a] Synthetic minority over-sampling technique.
[b] RR: raw reflectance; CR: continuum removal.
[c] Models are numbered depending on the optimization strategy, spectral transformation, and modeling algorithm used.
[d] Overall accuracy.
[e] Matthews correlation coefficient.

(Table 3). Overall, the classification performance for soil Cd depended on the optimization strategy, spectral preprocessing, and modeling algorithm applied (Table 3 and Fig. 5). For full-spectrum analysis, model generated using CR spectra combined with SVM algorithm provided the highest validation classification performance (i.e., classified as strong agreement), with OA, Kappa, and MCC being of 0.82, 0.57, and 0.65, respectively. Concerning the selection of significant wavebands using the Boruta algorithm, the optimal model was developed via CR spectra with RF approach, presenting the highest validation MCC value of 0.74 (i.e., classified as strong agreement). In comparison with the full-spectrum and the Boruta wavelength selection techniques, SMOTE resulted in consistent increase in the classification performance of the validation set, irrespective of the spectral transformation and modeling algorithm used. This improved classification accuracy highlight SMOTE as the most successful modeling strategy for stable enhancement of the spectral classification features related to soil Cd. Amongst all the models investigated, the combination of SMOTE, CR spectra with RF modeling algorithm (i.e., Model 17) resulted in the highest validation results in terms of OA, Kappa, and MCC values of 0.88, 0.74, and 0.77, respectively, a diagnosis that was classified as strong agreement. Although better classification performance is obtained by SMOTE over Boruta selection method in diagnosing soil Cd contamination (Table 3), the improvement in validation classification performance of Models 13–18 relative to Models 7–12 is limited. For instance, in terms of the best diagnosis models in Boruta selection and SMOTE strategies, the MCC value of Model 17 was only 0.03 higher than that of Model 11.

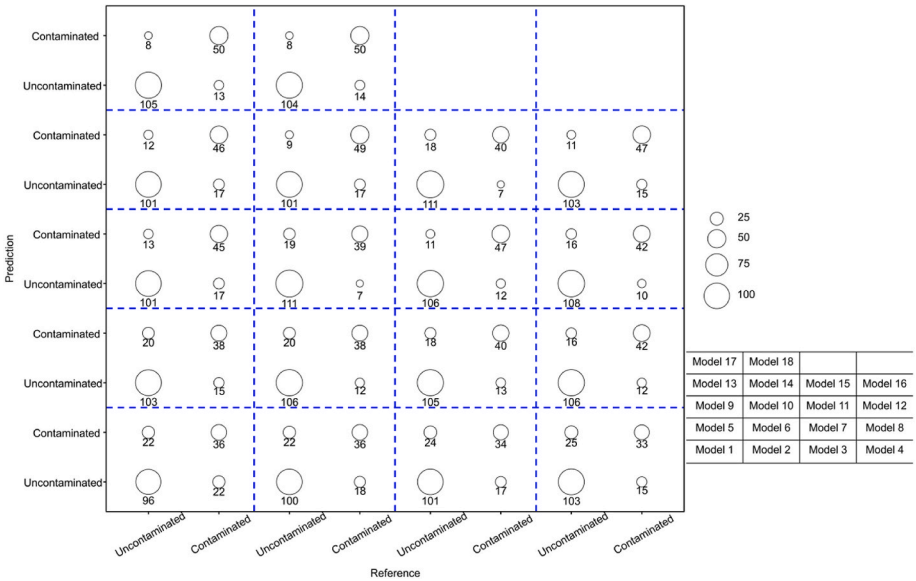

Fig. 5. Confusion matrix for soil Cd diagnosis for 18 different Vis–NIR classification models (shown for the validation dataset). Number of samples for each class is identified below the corresponding circle.

3.6. Important wavelengths for soil Cd diagnosis

Since Model 17 achieved the best classification accuracy among all 18 models explored, we used this model to interpret the important spectral bands related to soil Cd (Fig. S3). In addition, to compare the performance of the Model 17 with the corresponding full-spectrum model, the important wavelengths of Model 5 are also shown in Fig. S3. Both Models 5 and 17 showed overall similar variable importance trends across the entire spectral region, with large importance values occurring primarily within 410–610, 850–890, 1200–1220, 1350–1440, and 2150–2230 nm spectral bands, which are well in line with the wavelength positions of the Boruta selected variables on CR spectra (also shown in Fig. S3 and Fig. 4b).

4. Discussion

Although the mean value of soil Cd concentration in the entire dataset (i.e., 536 samples) was lower than that of most studies in Figs. 1 and 33.96 % of the samples were identified as contaminated with varying degrees of Cd concentrations. Elevated Cd content within the study area may be related to the rapid urbanization and industrialization of Wuhan City, and long-term agricultural cultivation nearby urban–rural transition zones. Commonly, the contamination sources of soil Cd mainly include: petrochemicals, galvanization, agricultural fertilizers, lubricating oil and tires, and coal combustion (Liu et al., 2018a). Wuhan City has a large population with well-developed road networks (Zhang et al., 2018), causing wearing of tires, as well as increased use of lubricating oil and brakes. Additionally, there were some petrochemical, steel production, galvanization manufacturing, and other industrial enterprises scattered in different locations of the city (Wu et al., 2020). These plants produce Cd emissions during the production process and consequently contaminate the surrounding soils through atmospheric deposition. In agricultural practices, due to the shortage of labor resources, the farmers in the suburban areas around Wuhan City tend to increase the phosphate fertilizer application in order to increase the crop production. When soil phosphorus fertilizer is low, increase of the phosphorus fertilizer input to improve soil effective phosphorus level is necessary; but when soil reaches phosphorus rich-level, further increase in effective phosphorus would aggravate the threat of the loss of soil phosphorus in the farmland. Excessive usage of phosphate fertilizer may lead to an increase in soil Cd concentration, because Cd is the inherent impurity in phosphate rock (Lv and Wang, 2019). Therefore, it is necessary to establish a rapid and efficient approach to diagnose the soil Cd contamination, especially for some areas that pose risks to the public health.

Although soil Cd is spectrally featureless with no direct spectral response in the Vis–NIR spectral region, its relation (if exist) to other spectrally active soil properties would support the successful estimation (Gholizadeh et al., 2018; Horta et al., 2015; Nawar et al., 2019; Shi et al., 2014; Stenberg et al., 2010). From Fig. S1, SOM was positively and more strongly correlated with soil Cd than Fe, suggesting that SOM is the major property to explain the mechanism of predicting soil Cd from the Vis–NIR reflectance spectra, as SOM has direct spectral responses in Vis–NIR spectroscopy. Furthermore, as soil becomes darker in color with increasing SOM, this corresponds to the significant bands for Cd in the visible range (Fig. 4). This result can be explained by the fact that different interaction forms of the metal complex and the decomposition of SOM result in the binding of soil Cd with SOM (Chen et al., 2020; Piccolo and Stevenson, 1982; Shi et al., 2014). However, this type of spectral estimation mechanism may vary depending on the study site. The finding in our case is comparatively consistent with those reported by Chen et al. (2015), Pandit et al. (2010), Song et al. (2012), and St. Luce et al. (2017), but not with Wu et al. (2007), who noticed the correlation with Fe is the major mechanism for explaining the successful estimation of soil Cd by the Vis–NIR reflectance spectra. These differences in prediction mechanism may be caused by diverse soil formation environments, such as terrain, soil parent materials, and anthropogenic management practices.

Selecting the most appropriate subset of spectral bands to diagnose the soil Cd contamination is a vital factor that influences the model classification performance (Table 3). The use of the Boruta algorithm for waveband variable selection reduced the number of spectral predictors from 201 to 21 and 32 for RR and CR spectra, respectively (Fig. 4). Besides, all the six resulted models (i.e., Models 7–12) have improved the validation accuracy (all were classified as strong diagnosis agreement), in comparison with their corresponding full-spectrum predictive models (i.e., Models 1–6), irrespective of spectral transformation and modeling algorithm. The relatively poor classification results developed based on the full-spectrum input may be because of its waveband predictors containing irrelevant and noisy information, which leads to poor results, although the high-resolution spectral data can fully reflect well-defined narrow spectral features (Castaldi et al., 2016; Raj et al., 2018; Vohland et al., 2014; Yang et al., 2012). The elaborated feature selection procedure followed in this study has not only resulted in improved prediction for Cd, but also produces parsimonious model structure for practical applications. Overall, the majority of selected variables by the Boruta selection algorithm on CR spectra in Fig. 4b approximately match the locations of possible absorption features of key soil attributes having direct spectral responses at 430–560 (related to goethite, ferric oxide, hematite, ferrihydrite, and organic matter), 1370–1460 (associated with OH stretch, kaolin doublet, and C═O), 1950–1970 (attributed to OH in the first overtone and C–OH and smectite), 2180–2220 (corresponded to SOM, illite, kaolinite, and Al–OH bend plus O–H stretch), and 2270–2320 nm (linked to Fe–OH, Mg–OH, C–H stretch fundamentals, and humic acid) (Coblinski et al., 2020; Knadel et al., 2013; Viscarra Rossel and Behrens, 2010). Moreover, the locations of these significant wavebands were comparable with those of relatively large importance value resulted from Models 5 and 17, identified in Fig. S3, confirming the efficiency of the Boruta selection approach. Indeed, direct diagnosing of whether samples are contaminated or not through reflectance spectra, mainly depends on how well a soil pollutant is correlated to the spectrally active soil properties like SOM and absorption features of molecules (Bray et al., 2009; Shi et al., 2017; Wang et al., 2014). The substantial correlation of soil Cd with SOM and to a less extent with Fe, and good match between Boruta selected variables from CR spectra and absorption features mentioned-above explain the good classification accuracy for soil Cd contamination diagnosis (Table 3).

A limitation to cause deterioration in data classification accuracy is the highly imbalanced dataset among different classes, under which statistical classifiers are often overpowered by the majority class (having the largest number of samples) and overlook the minority class (Chawla et al., 2002; Xie and Li, 2018). According to Table 2, the uncontaminated and contaminated samples accounted for 66.56 % and 34.44 % of the total observations in the calibration dataset, respectively. By using SMOTE algorithm to create balanced datasets, the validation classification accuracies in all cases (i.e., Models 13–18) were improved in comparison to full-spectrum analysis (Table 3), regardless of spectral transformation and modeling algorithm used. Previous researches have shown that the classification accuracy has been improved using resampling techniques in other research areas of soil studies. For instance, Sharififar et al. (2019a), Sharififar et al. (2019b), and Xie and Li (2018) all used oversampling technique to improve the classification performance for prediction or mapping of soil type classes.

With regard to the spectral transformation, the prediction models built with the CR spectra slightly outperformed those of the RR spectra (Table 3). The improvement of classification accuracy after CR preprocessing indicated CR to be as effective technique in highlighting key absorption features and minor peaks (difficult to detect in the RR) beneficial for soil Cd diagnose, as demonstrated in Fig. 3b (Clark and Roush, 1984). This method has also been successfully applied in other published studies to predict soil properties with competitive results (Lagacherie et al., 2008; Nawar et al., 2016; Vašát et al., 2014). As

shown in Table 3, the validation accuracy of PLSDA, RF, and SVM depended upon both the optimization method and spectral transformations adopted. In summary, the RF and SVM methods provided similar predictive ability in diagnosing soil Cd contamination, and both algorithms outperformed PLSDA. The outperformance might be attributed to the nonlinear relationship that might exist between soil Cd concentration and spectral data (Hong et al., 2019; Shi et al., 2017). Once advanced machine-learning methods, such as RF and SVM, are utilized, improved classification accuracy is expected as these algorithms can effectively deal with the complex and non-linear spectral behaviors (Dotto et al., 2018; Nawar and Mouazen, 2019; Ng et al., 2020).

Due to the complexity of soils, diversity of soil formation, and spatial heterogeneity, a single sensor to estimate or classify soil attributes is reported to underperform the sensor fusion approach (Horta et al., 2015; Wan et al., 2020; Xu et al., 2019). This necessitates innovative solutions to integrate and fuse multi-sensors or integrate modeling approaches (multi-models) to maximize accurate estimation of multiple soil properties. With the continuous advancements in proximal soil sensing technologies, a multiple-sensors data fusion approach is strongly advancing in soil analysis including soil contamination. Sensor fusion include the combinations of Vis–NIR, X-ray fluorescence, mid-infrared, and laser-induced breakdown spectroscopy, which have been successfully adopted with advanced modeling approach to estimate various soil properties (Gholizadeh et al., 2018; Horta et al., 2015; Nawar et al., 2019; Shi et al., 2014). More researches for assessing soil Cd are required by considering the joint use of multiple sensors in combination with advanced modeling approaches similar to the one adopted in the current work. Although the combined use of multiple sensors may decrease the measurement speed and increase the instrument costs, the improved model accuracy is worth expecting.

5. Conclusions

This study developed a framework based on Vis–NIR spectroscopy to explore the potentials of three different optimization strategies (i.e., full-spectrum, Boruta selection, and SMOTE) for diagnosing soil Cd contamination in urban and suburban soils, using three different classification methods, including PLSDA, SVM, and RF. Based on the results obtained, the following conclusions were made:

(1) Soils contaminated by Cd showed lower reflectance spectral magnitude, compared with uncontaminated soils. Spectra preprocessing by CR has highlighted significant spectral features related to soil Cd, useful for correct spectral classification into contaminated or uncontaminated samples.
(2) Both Boruta and SMOTE methods improved the classification accuracy for soil Cd, and the best accurate diagnosis was achieved by the combination of SMOTE applied on CR spectra with RF modeling (Kappa in the validation set = 0.74).
(3) The predictor wavebands for diagnosing soil Cd contamination were at 410–610, 850–890, 1200–1220, 1350–1440, and 2150–2230 nm, most of which are associated with well-defined soil absorption features.
(4) The correlations of soil Cd with the spectrally active soil attributes (i.e., SOM in particular and Fe to a less extent) were important for assessing the contamination of the spectrally featureless Cd element.

This study provides a direct solution for the use of the Vis–NIR spectroscopy to diagnose soil Cd contamination in urban and suburban soils.

Author statement

Yongsheng Hong, Songchao Chen: Conceptualization, Methodology, Software, Validation, Visualization; Yongsheng Hong, Songchao Chen: Writing – original draft; Yiyun Chen, Long Guo, Yaolin Liu, Zhou Shi, Abdul M. Mouazen: Writing – review & editing; Yiyun Chen, Ruili Shen, Hang Cheng, Yi Liu: Investigation, Resources; Yiyun Chen, Gang Xu, Zushuai Wei, Jian Yang, Abdul M. Mouazen: Supervision, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Key R&D Program of China (2019YFC1803404) and the National Natural Science Foundation of China (grant number: 41771440). Authors also acknowledged the financial support received from the European Commission H2020-SFS-38-2018 call for proposals for SIEUSOIL project, under Grant Agreement No. 818346.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2021.118128.

References

Agricultural Chemistry Committee of China, 1983. Conventional Methods of Soil and Agricultural Chemistry Analysis (in Chinese). Science Press, Beijing, pp. 70–165.
Almeida, C.T.D., Galvão, L.S., Aragão, L.E.O.C., Ometto, J.P.H.B., Jacon, A.D., Pereira, F. R.D.S., Sato, L.Y., Lopes, A.P., Graça, P.M.L.A., Silva, C.V.D.J., Ferreira-Ferreira, J., Longo, M., 2019. Combining LiDAR and hyperspectral data for aboveground biomass modeling in the Brazilian Amazon using different regression algorithms. Remote Sens. Environ. 232, 111323.
Bao, S.D., 2005. Agricultural Chemistry Analysis. Agricultural Press, Beijing.
Belgiu, M., Drăguț, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS-J. Photogramm. Remote Sens. 114, 24–31.
Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS One 12 (6).
Bray, J.G.P., Viscarra Rossel, R.A., McBratney, A.B., 2009. Diagnostic screening of urban soil contaminants using diffuse reflectance spectroscopy. Aust. J. Soil Res. 47 (4), 433–442.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S., Casa, R., 2016. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. Remote Sens. Environ. 179, 54–65.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
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.
Chen, S., Li, S., Ma, W., Ji, W., Xu, D., Shi, Z., Zhang, G., 2019. Rapid determination of soil classes in soil profiles using vis-NIR spectroscopy and multiple objectives mixed support vector classification. Eur. J. Soil Sci. 70 (1), 42–53.
Chen, W., Peng, L., Hu, K., Zhang, Z., Peng, C., Teng, C., Zhou, K., 2020. Spectroscopic response of soil organic matter in mining area to Pb/Cd heavy metal interaction: a mirror of coherent structural variation. J. Hazard. Mater. 393, 122425.
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.
Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy - quantitative-analysis techniques for remote-sensing applications. J. Geophys. Res. 89 (NB7), 6329–6340.
CNMEE, 2018. Soil Environmental Quality – Risk Control Standard for Soil Contamination of Agricultural Land. Ministry of Ecology and Environment of China. GB 15618-2018. (in Chinese).
Coblinski, J.A., Giasson, É., Demattê, J.A.M., Dotto, A.C., Costa, J.J.F., Vašát, R., 2020. Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. Catena 189, 104485.
Cohen, J., 1960. A coefficient of agreement for Nominal scales[J]. Educ. Psychol. Meas. 20 (1), 37–46.
Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37 (1), 35–46.
Dotto, A.C., Dalmolin, R.S.D., ten Caten, A., Grunwald, S., 2018. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for

multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma 314, 262–274.
Gholizadeh, A., Borůvka, L., Vašát, R., Saberioon, M., Klement, A., Kratina, J., Tejnecký, V., Drábek, O., 2015. Estimation of potentially toxic elements contamination in anthropogenic soils on a Brown coal mining dumpsite by reflectance spectroscopy: a case study. PloS One 10 (2), e0117457.
Gholizadeh, A., Saberioon, M., Ben-Dor, E., Boruvka, L., 2018. Monitoring of selected soil contaminants using proximal and remote sensing techniques: background, state-of-the-art and future perspectives. Crit. Rev. Environ. Sci. Technol. 48 (3), 243–278.
He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284.
Hong, Y., Shen, R., Cheng, H., Chen, S., Chen, Y., Guo, L., He, J., Liu, Y., Yu, L., Liu, Y., 2019. Cadmium concentration estimation in peri-urban agricultural soils: using reflectance spectroscopy, soil auxiliary information, or a combination of both? Geoderma 354, 113875.
Hong, N., Guan, Y., Yang, B., Zhong, J., Zhu, P., Ok, Y.S., Hou, D., Tsang, D.C.W., Guan, Y., Liu, A., 2020. Quantitative source tracking of heavy metals contained in urban road deposited sediments. J. Hazard. Mater. 393, 122362.
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.
Hou, D., O'Connor, D., Nathanail, P., Tian, L., Ma, Y., 2017. Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: a critical review. Environ. Pollut. 231, 1188–1200.
Hou, L., Li, X., Li, F., 2019. Hyperspectral-based inversion of heavy metal content in the soil of coal mining areas. J. Environ. Qual. 48 (1), 57–63.
Jia, X., O'Connor, D., Shi, Z., Hou, D., 2021. VIRS based detection in combination with machine learning for mapping soil pollution. Environ. Pollut. 268, 115845.
Jiang, Q., Liu, M., Wang, J., Liu, F., 2018. Feasibility of using visible and near-infrared reflectance spectroscopy to monitor heavy metal contaminants in urban lake sediment. Catena 162, 72–79.
Knadel, M., Viscarra Rossel, R.A., Deng, F., Thomsen, A., Greve, M.H., 2013. Visible-near infrared spectra as a proxy for topsoil texture and Glacial boundaries. Soil Sci. Soc. Am. J. 77 (2), 568–579.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Software 28 (5), 1–26.
Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Software 36 (11), 1–13.
Kursa, M.B., Jankowski, A., Rudnicki, W.R., 2010. Boruta - a system for feature selection. Fundam. Inf. 101 (4), 271–286.
Lagacherie, P., Baret, F., Feret, J.-B., Madeira Netto, J., Robbez-Masson, J.M., 2008. Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. Remote Sens. Environ. 112 (3), 825–835.
Lamine, S., Petropoulos, G.P., Brewer, P.A., Bachari, N.-E.-I., Srivastava, P.K., Manevski, K., Kalaitzidis, C., Macklin, M.G., 2019. Heavy metal soil contamination detection using combined geochemistry and field spectroradiometry in the United Kingdom. Sensors 19 (4), 762.
Lassalle, G., Fabre, S., Credoz, A., Dubucq, D., Elger, A., 2020. Monitoring oil contamination in vegetated areas with optical remote sensing: a comprehensive review. J. Hazard. Mater. 393, 122427.
Li, G., Sun, G.X., Ren, Y., Luo, X.S., Zhu, Y.G., 2018. Urban soil and human health: a review. Eur. J. Soil Sci. 69 (1), 196–215.
Liu, Y.L., Wen, C., Liu, X.J., 2013. China's food security soiled by contamination. Science 339 (6126), 1382–1383.
Liu, J., Liu, Y.J., Liu, Y., Liu, Z., Zhang, A.N., 2018a. Quantitative contributions of the major sources of heavy metals in soils to ecosystem and human health risks: a case study of Yulin, China. Ecotoxicol. Environ. Saf. 164, 261–269.
Liu, J., Zhang, Y., Wang, H., Du, Y., 2018b. Study on the prediction of soil heavy metal elements content based on visible near-infrared spectroscopy. Spectroc. Acta Pt. A-Molec. Biomolec. Spectr. 199, 43–49.
St Luce, M., Ziadi, N., Gagnon, B., Karam, A., 2017. Visible near infrared reflectance spectroscopy prediction of soil heavy metal concentrations in paper mill biosolid- and liming by-product-amended agricultural soils. Geoderma 288, 23–36.
Luo, X.-s., Yu, S., Zhu, Y.-g., Li, X.-d., 2012. Trace metal contamination in urban soils of China. Sci. Total Environ. 421–422, 17–30.
Lv, J., Wang, Y., 2019. PMF receptor models and sequential Gaussian simulation to determine the quantitative sources and hazardous areas of potentially toxic elements in soils. Geoderma 353, 347–358.
Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme[J]. BBA-Protein Struct. M. 405 (2), 442–451.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1), 3–52.
McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213.
Meng, X., Bao, Y., Liu, J., Liu, H., Zhang, X., Zhang, Y., Wang, P., Tang, H., Kong, F., 2020. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. Int. J. Appl. Earth Obs. Geoinf. 89, 102111.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388.
Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. ISPRS-J. Photogramm. Remote Sens. 66 (3), 247–259.
Nawar, S., Mouazen, A.M., 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).
Nawar, S., Mouazen, A.M., 2019. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. Soil Tillage Res. 190, 120–127.
Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., Mouazen, A.M., 2016. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. Soil Tillage Res. 155, 510–522.
Nawar, S., Cipullo, S., Douglas, R.K., Coulon, F., Mouazen, A.M., 2019. The applicability of spectroscopy methods for estimating potentially toxic elements in soils: state-of-the-art and future trends. Appl. Spectrosc. Rev. 1–33.
Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251–267.
Ng, W., Minasny, B., McBratney, A., 2020. Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. Sci. Total Environ. 702, 134723.
Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347.
Pandit, C.M., Filippelli, G.M., Li, L., 2010. Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression. Int. J. Rem. Sens. 31 (15), 4111–4123.
Paulo, C., 2013. rminer: Simpler use of data mining methods (e.g. NN and SVM) in classification and regression. R package version 1.3. https://CRAN.R-project.org/package=rminer.
Piccolo, A., Stevenson, F.J., 1982. Infrared spectra of Cu2+ Pb2+ and Ca2+ complexes of soil humic substances. Geoderma 27 (3), 195–208.
Poggio, L., Vrščaj, B., Schulin, R., Hepperle, E., Ajmone Marsan, F., 2009. Metals pollution and human bioaccessibility of topsoils in Grugliasco (Italy). Environ. Pollut. 157 (2), 680–689.
Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2019. Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. Catena 177, 149–166.
Proctor, S.D., Dreher, K.L., Kelly, S.E., Russell, J.C., 2006. Hypersensitivity of prediabetic JCR : LA-cp rats to fine airborne combustion particle-induced direct and noradrenergic-mediated vascular contraction. Toxicol. Sci. 90 (2), 385–391.
Qu, M., Wang, Y., Huang, B., Zhao, Y., 2018. Source apportionment of soil heavy metals using robust absolute principal component scores-robust geographically weighted regression (RAPCS-RGWR) receptor model. Sci. Total Environ. 626, 203–210.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Lanzhou, China (URL). http://www.R-project.org/.
Raj, A., Chakraborty, S., Duda, B.M., Weindorf, D.C., Li, B., Roy, S., Sarathjith, M.C., Das, B.S., Paulette, L., 2018. Soil mapping via diffuse reflectance spectroscopy based on variable indicators: an ordered predictor selection approach. Geoderma 314, 146–159.
Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Demattê, J.A.M., Scholten, T., 2014. Sampling optimal calibration sets in soil infrared spectroscopy. Geoderma 226–227, 140–150.
Rathod, P.H., Müller, I., Van der Meer, F.D., de Smeth, B., 2016. Analysis of visible and near infrared spectral reflectance for assessing metals in soil. Environ. Monit. Assess. 188 (10), 558.
Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36 (8), 1627–1639.
Sawut, R., Kasim, N., Abliz, A., Hu, L., 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 Obs. Geoinf. 73, 14–25.
Sharififar, A., Sarmadian, F., Malone, B.P., Minasny, B., 2019a. Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma 350, 84–92.
Sharififar, A., Sarmadian, F., Minasny, B., 2019b. Mapping imbalanced soil classes using Markov chain random fields models treated with data resampling technique. Comput. Electron. Agric. 159, 110–118.
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.Z., Liu, H.Z., Chen, Y.Y., Fei, T., Wang, J.J., Wu, G.F., 2017. Spectroscopic diagnosis of arsenic contamination in agricultural soils. Sensors 17 (5).
Song, Y., Li, F., Yang, Z., Ayoko, G.A., Frost, R.L., Ji, J., 2012. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. Appl. Clay Sci. 64, 75–83.
Song, Y., Ji, J., Mao, C., Ayoko, G.A., Frost, R.L., Yang, Z., Yuan, X., 2013. The use of reflectance visible–NIR spectroscopy to predict seasonal change of trace metals in suspended solids of Changjiang River. Catena 109, 217–224.
Stafford, A.D., Kusumo, B.H., Jeyakumar, P., Hedley, M.J., Anderson, C.W.N., 2018. Cadmium in soils under pasture predicted by soil spectral reflectance on two dairy farms in New Zealand. Geoderma Reg 13, 26–34.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter five - visible and near infrared spectroscopy in soil science. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 163–215.
Stevens, A., Ramirez-Lopez, L., 2014. An introduction to the prospectr package. In: R Package Version 0.1, p. 3.
Sun, W., Zhang, X., 2017. Estimating soil zinc concentrations using reflectance spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 58, 126–133.
Tan, K., Wang, H., Chen, L., Du, Q., Du, P., Pan, C., 2020. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. J. Hazard. Mater. 382, 120987.

Todorova, M., Mouazen, A.M., Lange, H., Atanassova, S., 2014. Potential of near-infrared spectroscopy for measurement of heavy metals in soil as affected by calibration set size. Water, air. & Soil Pollution 225 (8), 2036.
Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Network. 10 (5), 988–999.
Vašát, R., Kodešová, R., Borůvka, L., Klement, A., Jakšík, O., Gholizadeh, A., 2014. Consideration of peak parameters derived from continuum-removed spectra to predict extractable nutrients in soils with visible and near-infrared diffuse reflectance spectroscopy (VNIR-DRS). Geoderma 232–234, 208–218.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1–2), 46–54.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230.
Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection. Geoderma 223–225, 88–96.
Wan, M., Hu, W., Qu, M., Li, W., Zhang, C., Kang, J., Hong, Y., Chen, Y., Huang, B., 2020. Rapid estimation of soil cation exchange capacity through sensor data fusion of portable XRF spectrometry and Vis-NIR spectroscopy. Geoderma 363, 114163.
Wang, J., Cui, L., Gao, W., Shi, T., Chen, Y., Gao, Y., 2014. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216, 1–9.
Wang, 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.
Wilding, L.P., 1985. Spatial variability: it's documentation, accommodation and implication to soil surveys, soil spatial variability. Workshop 166–194.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. 58 (2), 109–130.
Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., Ma, H., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71 (3), 918–926.
Wu, Z., Chen, Y., Han, Y., Ke, T., Liu, Y., 2020. Identifying the influencing factors controlling the spatial variation of heavy metals in suburban soil using spatial regression models. Sci. Total Environ. 717, 137212.
Xie, X.-L., Li, A.-B., 2018. Identification of soil profile classes using depth-weighted visible–near-infrared spectral reflectance. Geoderma 325, 90–101.
Xie, X.-L., Pan, X.-Z., Sun, B., 2012. Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a copper smelter. Pedosphere 22 (3), 351–366.
Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., Shi, Z., 2019. Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China. Eur. J. Soil Sci. 70 (1), 162–173.
Yang, H., Kuang, B., Mouazen, A.M., 2012. Quantitative analysis of soil nitrogen and carbon at a farm scale using visible and near infrared spectroscopy coupled with wavelength reduction. Eur. J. Soil Sci. 63 (3), 410–420.
Yu, K., Van Geel, M., Ceulemans, T., Geerts, W., Ramos, M.M., Serafim, C., Sousa, N., Castro, P.M.L., Kastendeuch, P., Najjar, G., Ameglio, T., Ngao, J., Saudreau, M., Honnay, O., Somers, B., 2018. Vegetation reflectance spectroscopy for biomonitoring of heavy metal pollution in urban soils. Environ. Pollut. 243, 1912–1922.
Yuan, Y., Cave, M., Xu, H., Zhang, C., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). J. Hazard. Mater. 393, 122377.
Zhang, Y., Liu, Y., Zhang, Y., Liu, Y., Zhang, G., Chen, Y., 2018. On the spatial relationship between ecosystem services and urbanization: a case study in Wuhan, China. Sci. Total Environ. 637–638, 780–790.
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. Spectroc. Acta Pt. A-Molec. Biomolec. Spectr. 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.
Zhang, Y., Hou, D., O'Connor, D., Shen, Z., Shi, P., Ok, Y.S., Tsang, D.C.W., Wen, Y., Luo, M., 2019c. Lead contamination in Chinese surface soils: source identification, spatial-temporal distribution and associated health risks. Crit. Rev. Environ. Sci. Technol. 49 (15), 1386–1423.
Zhao, F.J., Ma, Y.B., Zhu, Y.G., Tang, Z., McGrath, S.P., 2015. Soil contamination in China: current status and mitigation strategies. Environ. Sci. Technol. 49 (2), 750–759.