Environ Monit Assess (2019) 191: 446
https://doi.org/10.1007/s10661-019-7510-4

Random forest–based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data

Kun Tan · Weibo Ma · Fuyu Wu · Qian Du

Received: 31 October 2018 / Accepted: 30 April 2019 / Published online: 18 June 2019

Abstract Heavy metals in the agricultural soils of reclaimed mining areas can contaminate food and endanger human health. The objective of this study is to effectively estimate the concentrations of heavy metals, such as zinc, chromium, arsenic, and lead, using hyperspectral sensor data and the random forest (RF) algorithm in the study area of Xuzhou, China. The RF's built-in feature selection ability and modeling expressive ability in heavy metal estimation of soil were explored. After the preprocessing of the spectrum obtained by an ASD (analytical spectral device) field spectrometer, the random forest algorithm was carried out to establish the estimation model based on the correlation-selected features and the full-spectrum features respectively. Results of all the different processes were compared with classical approaches, such as partial least squares (PLS) regression and support vector machine (SVM). In all the experimental results, from the perspective of models, the best estimation model for Zn (R^2 = 0.9061; RMSE = 6.5008) is based on the full-spectrum data of continuum removal (CR) pretreatment, and the best models for Cr (R^2 = 0.9110; RMSE = 4.5683), As (R^2 = 0.9912; RMSE = 0.5327), and Pb (R^2 = 0.9756; RMSE = 1.1694) are all derived from the correlation-selected features. And these best models of these heavy metals are all established by the RF method. The experiments in this paper show that random forests can make full use of the input spectral data in the estimation of four kinds of heavy metals, and the obtained models are superior to those established by traditional methods.

Keywords Soil heavy metal concentration · Random forest · Hyperspectral estimation

K. Tan
Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

K. Tan
School of Geographic Sciences, East China Normal University, Shanghai 200241, China

K. Tan (✉) · W. Ma · F. Wu
Key Laboratory for Land Environment and Disaster Monitoring of NASG, China University of Mining and Technology, Xuzhou 221116, China
e-mail: tankuncu@gmail.com

W. Ma (✉)
Ministry of Ecology and Environment, Nanjing Institute of Environmental Sciences, Nanjing 210042, China
e-mail: weibo_ma@126.com

Q. Du
Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA

Introduction

With rapid urbanization in China, less land is available for agricultural use, resulting in some reclaimed land being used for crop production. However, for the areas that use coal gangue and fly ash as reclamation material, the effects of climate change, land cultivation, soil water evaporation, and microbial action have led to the release of toxic heavy

metals and erosion of agricultural soil. These heavy metals can also enter human food chain through consuming agricultural products, which threatens human health and safety (Dong, Yu, Bian, Zhao, & Cheng, 2011; Wei & Yang, 2010; Wei, Jiang, Li, & Mu, 2009; Li, Ma, van der Kuijp, Yuan, & Huang, 2014). Therefore, estimation of heavy metal content in soils of reclaimed areas is an important task (Song et al., 2012). In recent years, the application of hyperspectral remote sensing technology for this purpose has become a popular research topic (J. Wang et al., 2014; Rathod, Rossiter, Noomen, & Fd, 2013; Soriano-Disla, Janik, Rossel, Macdonald, & Mclaughlin, 2014; Shi, Chen, Liu, & Wu, 2014; Summers, Lewis, Ostendorf, & Chittleborough, 2009). Compared with the traditional field sampling and laboratory tests, hyperspectral (350 to 2500 nm) data offers superiorities in both time and cost savings.

Since 1997, researchers have adopted remote sensing spectroscopy to estimate the content of heavy metals in soils (Malley & Williams, 1997). The heavy metal content in soils is usually low, so it does not have obvious spectral characteristic (Wang et al., 2014; Wu et al., 2005). It is therefore necessary to preprocess the spectra, so that the weak information can be reflected. A number of preprocessing methods—the Savitzky-Golay (SG) (Madden, 1978; Gorry, 1990) smoothing, first derivative (FD) (Han & Rundquist, 1997; Aguerssif, Benamor, Kachbi, & Draa, 2008), second derivative (SD) (Balestrieri, Colonna, Giovane, Irace, & Servillo, 1978; Kosmas, Curi, Bryant, & Franzmeier, 1984), standard normal variate (SNV) (Fearn, Riccioli, Garrido-Varo, & Guerrero-Ginel, 2009; Summers et al., 2009; Kinoshita, Moebiusclune, Es, Hively, & Bilgilis, 2012), and continuum removal (CR) (Meer, 2000; Gomez, Lagacherie, & Coulouma, 2008; Ding, Wang, & Ge, 2010)—are widely used, and they can smooth the spectra, eliminate the signal error caused by the instrument itself, and suppress the noise in data acquisition, thereby enhancing weak spectral information related to heavy metals. The aforementioned preprocessing methods have been shown to be effective in practical applications (Rinnan, Berg, & Engelsen, 2009; Summers et al., 2009; Xu, Xie, & Fan, 2011; Kinoshita et al., 2012; Soriano-Disla et al., 2014; Asadzadeh & Roberto, 2016; Candolfi, Maesschalck, Jouan-Rimbaud, Hailey, & Massart, 1999; Cen, Bao, Huang, & He, 2006; Jamshidi, Minaei, Mohajerani, & Ghassemian, 2012). Most of the current studies are based on ground spectrum. Quantitative estimation of heavy metals using imaging spectroscopy has been explored (Choe et al., 2008), but the performance is limited due to rough spatial resolution.

In this paper, we are more concerned with learning algorithms deployed to extract related features. Shi et al. (2014) reviewed the literature on the estimation of nine heavy metals, in which the most widely used modeling strategy was partial least squares (PLS) regression. Meanwhile, the importance of artificial neural networks (ANNs) (Tan, Ye, Cao, & Du, 2014) and genetic algorithms (GAs) (J. Wang et al., 2014) has been increasingly emphasized. Jie (2012) predicted the cadmium content of soils with a support vector machine (SVM) method and achieved satisfying results. However, researchers are always looking for learning algorithms that can offer higher accuracy and better generalization performance.

The random forest (RF) algorithm (L. Breiman, 2001) is one of representative ensemble learning methods. In recent years, it has been studied in many remote sensing fields for various applications, i.e., classification (Pal, 2005; Ham, Chen, Crawford, & Ghosh, 2005), ship recognition (Huang, 2015), estimation of wheat biomass (Wang, Zhou, Zhu, Dong, & Guo, 2016), vegetation mapping (Feng, Liu, & Gong, 2015), etc. (Feng, Liu, & Gong, 2010; Belgiu & Drăguț, 2016). However, there have been very few studies on the application of RF to spectral analysis of soils (Belgiu & Drăguț, 2016). Wang, Xie, and Li (2015) compared the performance of several ensemble learning methods to estimate the heavy metal content of agricultural soil, but the number of variables used to build the model was very small, and data processing could not reflect the advantages of the RF algorithm. Rodriguez-Galiano, Sanchez-Castillo, Chica-Olmo, and Chica-Rivas (2015) used several machine learning methods, including ANN, RF, SVM, and regression tree, to map mineral prospectivity, and reported that the RF performed better than other methods. However, this was mainly in classification and assessment, rather than quantitative regression of heavy metals. By the use of random sampling, the RF algorithm can make full use of key information hidden in high-dimensional data (Hastie, Tibshirani, & Friedman, 2009). Therefore, in this paper, we explore the learning ability of RF with hyperspectral data and its application in the estimation of soil heavy metals. Meanwhile, the built-in feature extraction capability of RF is also tested. Specifically, models for estimating the soil content of four heavy metals—zinc (Zn); chromium (Cr); arsenic (As); and lead (Pb)—are

built by analyzing the spectral characteristics of a reclaimed mining area. Two kinds of data—selected bands and full-spectrum data—are analyzed by the RF, and the performance is compared with PLS regression and SVM.

This paper is organized as follows. The data and inversion algorithm are introduced in the "Materials and methods" section, and then the analysis of results and discussions are provided in the "Results" and "Discussion" sections, respectively. Final conclusions are drawn in the "Conclusions" section.

Materials and methods

The design and implementation of the framework for heavy metal estimation are described in Fig. 1. The main body can be divided into two components: model establishment and model evaluation. The establishment process of the model includes data preprocessing, feature extraction, and algorithm optimization. The evaluation process begins with applying the same preprocessing and feature extraction methods used in the model establishment process to the test data. The performance is then evaluated by comparison with true concentration values.

Data set description

Study area

The study area is located in a mining zone located 20 km north of Xuzhou, Jiangsu province, China. The area features a warm temperate semi-humid monsoonal climate, with annual average rainfall of 800–930 mm, and annual average temperature of 13.8 °C. The study area consists of three regions: reclaimed soil filled with coal gangue (site A), reclaimed in 1995 (34° 25′ 24 N–117° 08′ 26 E); reclaimed soil filled with fly ash (site B), reclaimed in 1999 (34° 23′ 43 N–117° 07′ 29 E); and a control site (site C, 34° 23′ 43 N, 117° 06′ 50 E) (Fig. 2).

Field soil sampling

Based on the area and shape, ten sampling points were selected from each region, following a snake-shaped pattern. Samples were collected from each sampling point. According to the principle of sampling of mixed soil samples and the collection of special soil samples, the sampling unit, sample number, and sampling section are determined. In the coal gangue filling site A area, B area, and fly ash filling site control site C, ten samples are collected in each district, and each sample is divided into 0 cm, 20 cm, 40 cm, interface layer (different 35–55 cm), or 60 cm (control site). Then, each of the four sides of soil sampling is mixed together to form a sample.

All 30 samples were kept in packages and tagged. After removing the sundries, such as roots, leaves, and stones, each sample was divided into two parts. One part was sent to a chemistry lab where soil heavy metal concentrations were measured, and the other was sent to the darkroom, where soil spectral signatures were measured.

Heavy metal measurement

The true values of Zn, Cr, As, and Pb concentrations in the reclaimed soil were measured in the laboratory. The basic statistics of the results of heavy metal measurements are shown in Table 1, including minimum (Min), maximum (Max), mean, and standard deviation (Std.). From the statistical values, the maximum concentrations are very close to the risk index values, so this study is indeed necessary and meaningful. Meanwhile, the standard deviation of As is relatively low, which may be caused by the difference in heavy metal concentrations of the three different sampling sites.

Laboratory optical measurements

The soil samples were scanned in the lab environment using ASD field spectrometer with a halogen lamp of 50 watts at an angle of 15°. The spectrometer acquired a continuous spectrum ranging from 350–1000 nm (at 1.4-nm intervals) and 1000–2500 nm (at 2-nm intervals). Each soil sample was scanned ten times, and the average spectrum was calculated for the follow-up study. After removing the abnormal spectra, the average spectrum was taken as the final reflectance spectrum of the soil sample. The spectra of 30 soil samples used in the experiments are shown in Fig. 3.

Spectral data preprocessing

The spectral reflectance data also contained other irrelevant information and noise. Therefore, it was necessary to implement some basic preprocessing steps to remove irrelevant information and noise. The common preprocessing methods in spectral analysis were applied, i.e.,

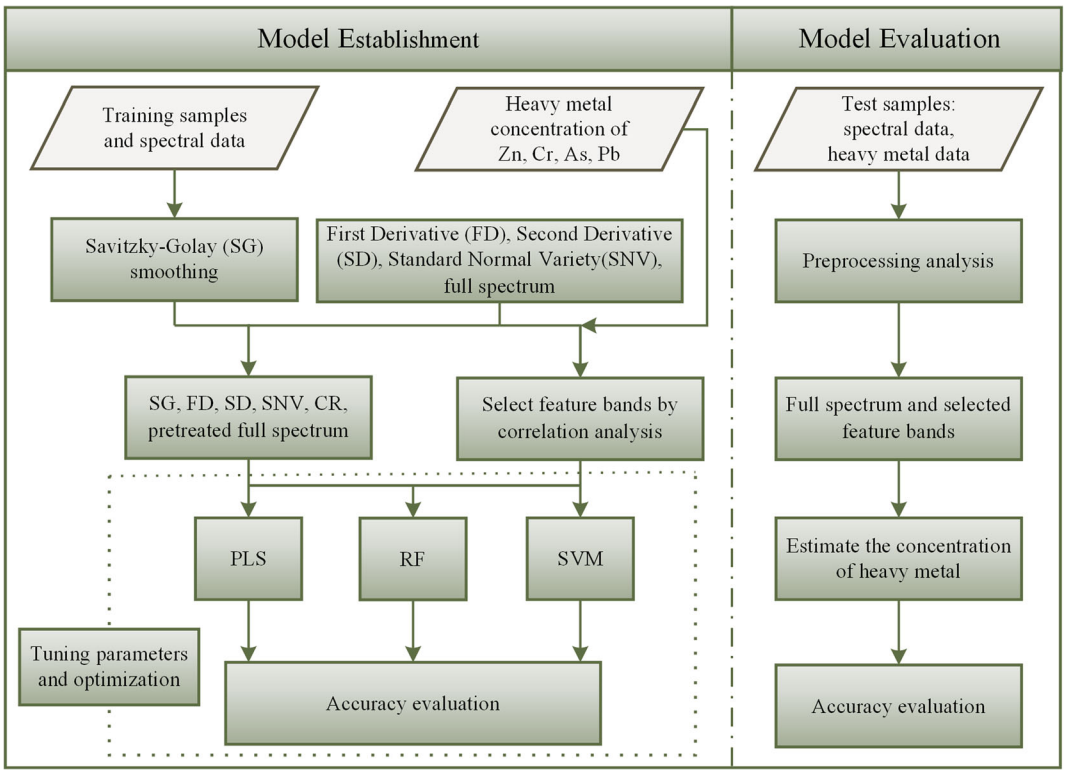

Fig. 1 The proposed framework for heavy metal estimation

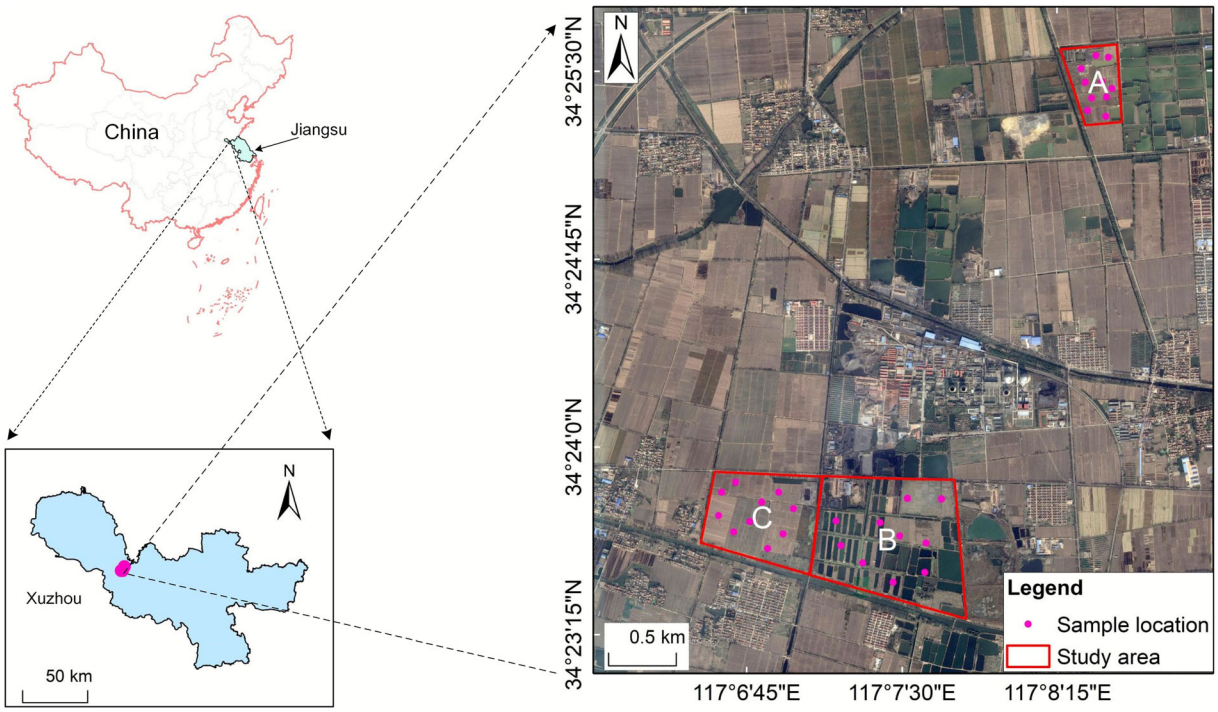

Fig. 2 Study sites A, B, and C near the city of Xuzhou, Jiangsu province, China. The red dots indicate the position of sampling points

Table 1 Statistics of the measured heavy metal concentrations

Metal	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	Std.
Zn	30.47	95.2	62.48	19.85
Cr	65.85	117.76	94.99	16.09
As	0.55	9.61	3.19	3.27
Pb	9.64	33.53	23.11	7.81

the Savitzky-Golay (SG) (Savitzky & Golay, 1964) smoothing, first derivative (FD), second derivative (SD), standard normal variate (SNV), continuum removal (CR), etc.

All the samples were first smoothed, and then two different strategies were implemented. On the one hand, the smoothed data separately obtained by FD, SD, SNV, and CR were treated as independent variables and input to the algorithms (i.e., PLS, SVM, and RF etc.) and to automatically start the feature extraction. The models were then tuned through cross-validation and other means. On the other hand, the correlation was analyzed between each band of the preprocessing results and the corresponding heavy metal samples. The bands with high correlation were then selected as the model input variables, which is a commonly used method in the literature. Table 2 lists the number of bands selected for various heavy metals after different preprocessing steps. In the process of selecting feature bands by correlation analysis, the results may vary due to the fact that the process may be subjective. This is one of the reasons that we use the RF for automatic feature selection in this study.

Methods

The random forest algorithm

The RF algorithm was proposed by Breiman (2001) and is a kind of predictive modeling algorithm based on classification and regression tree (CART) (L. I. Breiman, Friedman, Olshen, & Stone, 1984) and the bagging (Breiman, 1996) learning strategy. In bagging, a decision tree is generated from all of the properties each time, while in RF, it is randomly generated from a fixed-size subset of all the attributes, resulting in a reduced computational cost. Specifically, by the bootstrap (Efron & Tibshirani, 1993) resampling technique, random sampling is repeated K times to generate a fixed number of subset training samples (in general, the subset sample size is two-thirds of the training samples) from all the samples (here, K is the number of trees in the forest). Meanwhile, for each sample, only a fixed number of sub-attributes are selected. Each randomly selected subsample with its corresponding sub-attributes can then be used to generate a classification tree or regression tree, and all the trees make up the forest. Finally, the results are obtained according to the scores of the class voting from all the trees (certain algorithms can be implemented to determine the average of each tree, mostly for regression trees) (Hastie et al., 2009). The trained forest $\hat{F}_{RF}^{K}(x)$ with K trees can be expressed as

$$\hat{F}_{RF}^{K}(x) = \frac{1}{K}\sum_{K=1}^{K} T(x_s) \quad (1)$$

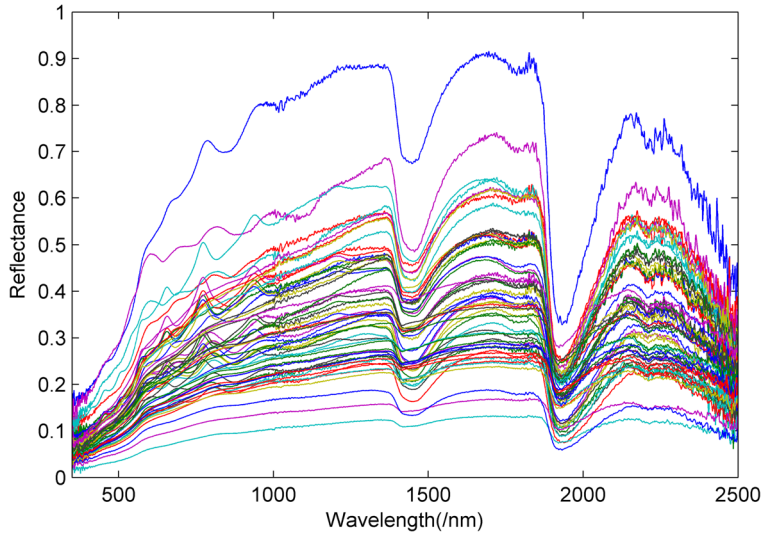

Fig. 3 Spectra of 30 soil samples from the three sampling sites

Table 2 The number of bands selected for heavy metals after different preprocessing steps

Pretreatment	Zn	Cr	As	Pb
CR	12	11	13	9
FD	11	5	7	6
SD	5	12	10	11
SNV	5	6	4	7
Total	33	34	34	33

where $T(x)$ is a single tree, x is all the training samples, and x_s is each tree's training sample data obtained with the bootstrap sampling method. Another parameter that is not noted in Eq. (1) is the number of sub-attributes selected from all the attributes with bootstrap sampling.

The most important advantage of RF with respect to other ensemble learning algorithms lies in the interpretation of results. RF results indicate the importance of variables. That is to say, the RF algorithm has a built-in feature extraction function (Archer & Kimes, 2008). This is an automatic mechanism for attribute selection. Whether the voting is correct or not, the out-of-bag (OOB) error also needs to be counted, in addition to the training of the model itself. Furthermore, the OOB error is an unbiased estimator with bootstrap sampling and a large number of trees, which is similar to the error obtained by cross-validation (Hastie et al., 2009).

RF is among the most popular machine learning methods due to relatively high accuracy, robustness, and ease of use. Two straightforward methods for feature selection can be adopted, i.e., mean decrease impurity and mean decrease accuracy (Archer & Kimes, 2008). We focus on the mean decrease impurity in this research. RF consists of a number of decision trees. Each node in a decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. The measure based on which the (locally) optimal condition is chosen is called impurity. For classification, it is typically either Gini impurity or information gain/entropy, and for regression trees, it is variance. Thus, when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree. For a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure. By calculating the importance of variables, rank is determined by sorting. The average decrease of classification accuracy and correct classification rate is computed before disturbance. For each decision tree, prediction error of out-of-pocket data is recorded by using the data outside the bag, and a new out-of-pocket data is formed for verification by random change of each predictor variable. For a prediction variable, the importance of the calculation is the mean of the difference between the predicted error before and after the transformation. (For more detailed discussion and experimental validation, please refer to Genuer, Poggi, and Tuleau-Malot (2010); Hapfelmeier and Ulm (2014); and Stańczyk (2015)).

The essence of the RF algorithm is an improvement of the decision tree algorithm, and it can handle a large number of input variables. Meanwhile, the tree structure can deal with missing sample attributes, and it can maintain accuracy with missing data (Carranza & Laborte, 2015). These properties of RF make it suitable for the processing of hyperspectral data. The characterization of high-dimensional hyperspectral data can be exploited by learning the attribute subset from a large number of trees. With more trees, we can not only improve the attribute subset to cover all of the property space, but also prevent overfitting.

In this study, the RF algorithm was implemented in R software using the Random Forest package (Liaw & Wiener, 2001). Two important parameters need to be specified: the number of input variables randomly chosen at each tree at the root node (mtry) and the number of trees (ntree) in the forest.

Partial least squares regression and support vector machine

In the field of hyperspectral inversion study, two of the algorithms that are commonly used are PLS and SVM. PLS regression was developed by Herman Wold in 1966 (Tenenhaus, Vinzi, Chatelin, & Lauro, 2005; Wold, Sjöström, & Eriksson, 2001). It is an improvement over multiple linear regression (MLR), and its main objective is to establish a linear regression model about independent and dependent variable matrices. By incorporating with the principal component analysis (PCA), PLS can process high-dimensional data and can partially remove the correlation among variables (Wold et al., 2001). SVM is a learning method based on the statistics of the minimum structure risk, and it can use the given error separation to optimize the separation hyperplane of training data (Balabin & Lomakina, 2011). It can solve either classification or regression problems (Drucker, Burges, Kaufman, Smola, & Vapnik, 1996). The use of kernel function makes it successful in nonlinear analysis. Its main advantage lies

Table 3 Model performance statistics for selected bands

Metal	Evaluation	PLS	SVM	RF
Zn	R^2	0.8730	0.8851	0.8929
	RMSE	6.9413	7.0360	6.7740
Cr	R^2	0.8681	0.9005	0.9110
	RMSE	5.9399	4.9732	4.5683
As	R^2	0.9431	0.9720	0.9912
	RMSE	1.3683	0.8760	0.5327
Pb	R^2	0.9071	0.9622	0.9756
	RMSE	2.6327	1.5919	1.1694

in its ability to deal with a small number of samples, nonlinear and high-dimensional problems, and local optima problem. In view of this, SVM is suitable for processing hyperspectral data. In this study, the PLS algorithm was implemented in MATLAB, and the SVM algorithm was run in R software using the package e1071 (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2015).

Model performance evaluation

With the test dataset, model performance was evaluated by the coefficient of determination (R^2) and the root-mean-square error (*RMSE*) expressed as

$$R^2 = \frac{\left[\sum_{i=1}^{N}\left(y_{ai}-\overline{y}_a\right)\sum_{i=1}^{N}\left(y_{mi}-\overline{y}_m\right)\right]^2}{\sum_{i=1}^{N}\left(y_{ai}-\overline{y}_a\right)^2\sum_{i=1}^{N}\left(y_{mi}-\overline{y}_m\right)} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_{ai}-y_{mi}\right)^2}{N-1}} \quad (3)$$

where y_{ai} and y_{mi} are the predicted and measured values, respectively, $\overline{y}_a$ and $\overline{y}_m$ are the average predicted and measured values, respectively, and N is the number of samples. R^2 is widely used to measure the goodness of model fitting, with values between 0 and 1. The closer the values is to 1, the higher the accuracy of model fitting. RMSE is the variance of the difference data to measure the "average error."

Results

All the 30 preprocessed samples of Zn, Cr, As, and Pb were divided evenly: 20 as training and ten as testing samples. In order to test the ability of the RF algorithm, the input data of each heavy metal were estimated from a comparison of the two strategies. One strategy was using selected bands from correlation analysis, and the other was using full-spectrum data without band selection. The 20 training samples were trained for each type of heavy metal with different algorithms, i.e., PLS, SVM, and RF. After parameter optimization, the 10 testing samples were fed into the optimized model, and the accuracy was evaluated. It should be noted that the results obtained by the RF algorithm were not the same, so the results of the RF algorithm for all the experiments in this paper are the average values obtained from fifteen runs, by setting a random number of seeds.

Model analysis of selected bands

By means of correlation analysis, statistical characteristics of heavy metal concentrations can be obtained. The "Spectral data preprocessing" section describes a variety of pretreatments, including FD, SD, SNV, and CR, used to enhance spectral characteristics of heavy metals after smoothing. There were 33 or 34 feature bands selected for each heavy metal. The accuracy of different models is evaluated in Table 3, through the validation of the trained model using the test data. The parameter setting of SVM and RF are shown in Table 4. It can be

Table 4 The parameter setting of SVM and RF

Element	SVM	RF
Zn	type = nu-regression, kernel = sigmoid	ntree = 50, mtry = 6
Cr	type = nu-regression, kernel = sigmoid	ntree = 120, mtry = 3
As	type = nu-regression, kernel = radial	ntree = 200, mtry = 6
Pb	type = nu-regression, kernel = sigmoid	ntree = 500, mtry = 6

seen that the accuracy of the RF model is higher than the other two algorithms in the prediction of each type of heavy metals, according to both RMSE and R^2. Although the accuracy of the SVM model for Cr, As, and Pb is higher than PLS, it is still less accurate than the RF model. Furthermore, it can be seen that the RF model has a better generalization performance, and it can make full use of the information contained in the variables.

Model analysis of full-spectrum data

Due to high spectral resolution across 350–1000 nm (at 1.4-nm intervals) and 1000–2500 nm (at 2-nm intervals), there are many bands in full-spectrum data, which may interfere with the model established by the traditional linear regression method, and model stability may be poor. However, RF can overcome the problem of high spectral correlation due to the use of bootstrap sampling, and it can treat all the trees comprehensively. Therefore, the accuracy of the model using the full-spectrum data or the preprocessed full-spectrum data depends on the learning and generalization abilities of the algorithm itself.

The Savitzky-Golay-smoothed full-spectrum data

After smoothing, random noise in the spectrum can be reduced, and the influence of numerical jumps of adjacent bands is eliminated (Savitzky & Golay, 1964). Using the three algorithms and the SG-smoothed full-spectrum data, the accuracy of the test data is shown in Table 4. From the results, only PLS for Pb performs the best in both statistical indicators. For the prediction of Cr, R^2 (0.7595) from SVM is lower than that of PLS (0.7884), but the RMSE of SVM (7.7894) is lower than that of PLS (8.6438). Similar results can be found in the prediction of Zn. Overall, in the prediction of almost all the heavy metals, the RF models using the smoothed spectral data show a poor performance. This indicates that the ability of RF to find useful information from the original spectra of the heavy metals is limited.

First derivative–preprocessed full-spectrum data

Using FDs to smooth out measurement noise is regarded as an effective pretreatment method (Kinoshita et al., 2012; Asadzadeh & Roberto, 2016). As shown in Table 5, the RF performs the best in the prediction of all four heavy metals. In the prediction of Pb, As (mostly), and Zn, PLS does not perform as well as RF, but obtains better results than SVM. Meanwhile, for Cr, the accuracy of PLS is lower than that of SVM. It can be inferred that some weak spectral features can be enhanced after noise reduction through the FD preprocessing.

Second derivative–preprocessed full-spectrum data

SD is a further extension of FD. Table 6 shows the performance of using SD-preprocessed full-spectrum data. In the prediction of Zn, Cr, and Pb, the RF offers the highest accuracy, but SVM obtains a slightly higher accuracy for As. In comparison with Table 5, we can see that the accuracy of the models trained by RF from the SD-preprocessed data set are lower than those of the FD-preprocessed data. A further comparison shows that the prediction results of PLS and SVM for Zn using the SD-preprocessed data are also lower than those of the FD-preprocessed data. For Cr and Pb, only one metric

Table 4 Model performance statistics for the SG-smoothed full-spectrum data

Metal	Evaluation	PLS	SVM	RF
Zn	R^2	0.8030	0.8548	0.3471
	RMSE	10.4291	14.1878	15.9842
Cr	R^2	0.7884	0.7593	0.2853
	RMSE	8.6438	7.7894	15.2252
As	R^2	0.7461	0.5446	0.1126
	RMSE	3.3110	2.7695	3.3469
Pb	R^2	0.9530	0.9129	0.1538
	RMSE	2.2303	2.9949	7.8589

Table 5 Model performance statistics for the FD-preprocessed full-spectrum data

Metal	Evaluation	PLS	SVM	RF
Zn	R^2	0.8591	0.8557	0.8973
	RMSE	8.4920	9.2301	7.9850
Cr	R^2	0.7901	0.9013	0.9086
	RMSE	7.9061	7.0240	5.8085
As	R^2	0.8594	0.8583	0.9435
	RMSE	1.6860	2.1191	1.5950
Pb	R^2	0.8965	0.9074	0.9499
	RMSE	2.4269	2.4242	2.0025

Table 6 Model performance statistics for the SD-preprocessed full-spectrum data

Metal	Evaluation	PLS	SVM	RF
Zn	R^2	0.8155	0.8486	0.8788
	RMSE	9.4964	9.2478	8.7723
Cr	R^2	0.8334	0.8722	0.8903
	RMSE	7.3611	6.8716	6.8514
As	R^2	0.9440	0.9146	0.9143
	RMSE	1.5780	1.6067	1.7877
Pb	R^2	0.8979	0.8965	0.9346
	RMSE	2.4489	2.4217	2.5233

indicates a good validation accuracy, either RMSE or R^2. In contrast, using the data preprocessed by SD, with the SVM and PLS algorithms, the prediction accuracy of As is improved significantly. However, in general, for RF, SD preprocessing is not better than FD preprocessing, because SD may be more sensitive to noise than FD.

Standard normal variate–preprocessed full-spectrum data

SNV is a commonly used and effective method for spectrum scattering correction (Fearn et al., 2009). The model validation accuracy for the full-spectrum data after SNV scattering correction is shown in Table 7. Similar to the results in Table 5, RF yields the best results in the prediction of almost all four heavy metals. In the prediction of Cr and Zn, SVM performs as well as PLS, but, for Pb, SVM performs better than PLS. For As, the RMSE of SVM is lower than PLS's, and the R^2 of SVM is not as good as the R^2 of PLS.

Table 7 Model performance statistics for the SNV-preprocessed full-spectrum data

Metal	Evaluation	PLS	SVM	RF
Zn	R^2	0.8655	0.8140	0.8798
	RMSE	7.5063	9.5108	8.6126
Cr	R^2	0.8646	0.7906	0.8301
	RMSE	6.3827	7.4220	6.5935
As	R^2	0.9253	0.8653	0.9795
	RMSE	1.4210	1.3929	0.7417
Pb	R^2	0.9264	0.9499	0.9655
	RMSE	2.1203	1.8329	1.4300

Table 8 Model performance statistics for the CR-preprocessed full-spectrum data

Metal	Evaluation	PLS	SVM	RF
Zn	R^2	0.8816	0.8760	0.9061
	RMSE	6.5046	7.8462	6.5008
Cr	R^2	0.7903	0.8448	0.8870
	RMSE	7.5112	7.2607	5.3972
As	R^2	0.9432	0.8684	0.9772
	RMSE	1.1914	1.5383	1.1314
Pb	R^2	0.9025	0.8730	0.9596
	RMSE	2.7137	2.6970	1.4991

Continuum removal–smoothed full-spectrum data

The CR method can effectively highlight the absorption, reflection, and emission characteristics of a spectral curve (Gomez et al., 2008). As shown in Table 8, for the estimation of Zn and As, PLS obtains a higher accuracy than SVM, but does not perform as well as RF. For Cr, SVM overall performs better than PLS, but does not perform as well as RF. For Pb, the $R^2$of SVM is lower than for PLS, and the RMSE of SVM is better than the RMSE of PLS.

Discussion

In order to obtain more stable experimental results, all the experimental results here are the results of 15 sets of random seed numbers, while the other parameters are the same. That is, each experiment is the average of the results of 15 groups. For illustration purposes, we take Pb as an example in Fig. 4. The accuracies of the displayed models are not the average value, but very close to it.

As can be seen from Fig. 4, the distribution of heavy metal concentrations is not uniform because the data are collected in three different regions. Compared with the accuracy of the scatter plots, except for the smoothed spectrum, all the other spectra can obtain high-accuracy results with the RF algorithm. Although the data obtained by spectral band selection can establish a model with the highest accuracy, the accuracy of the model using the whole spectrum is very close to it. In previous study,

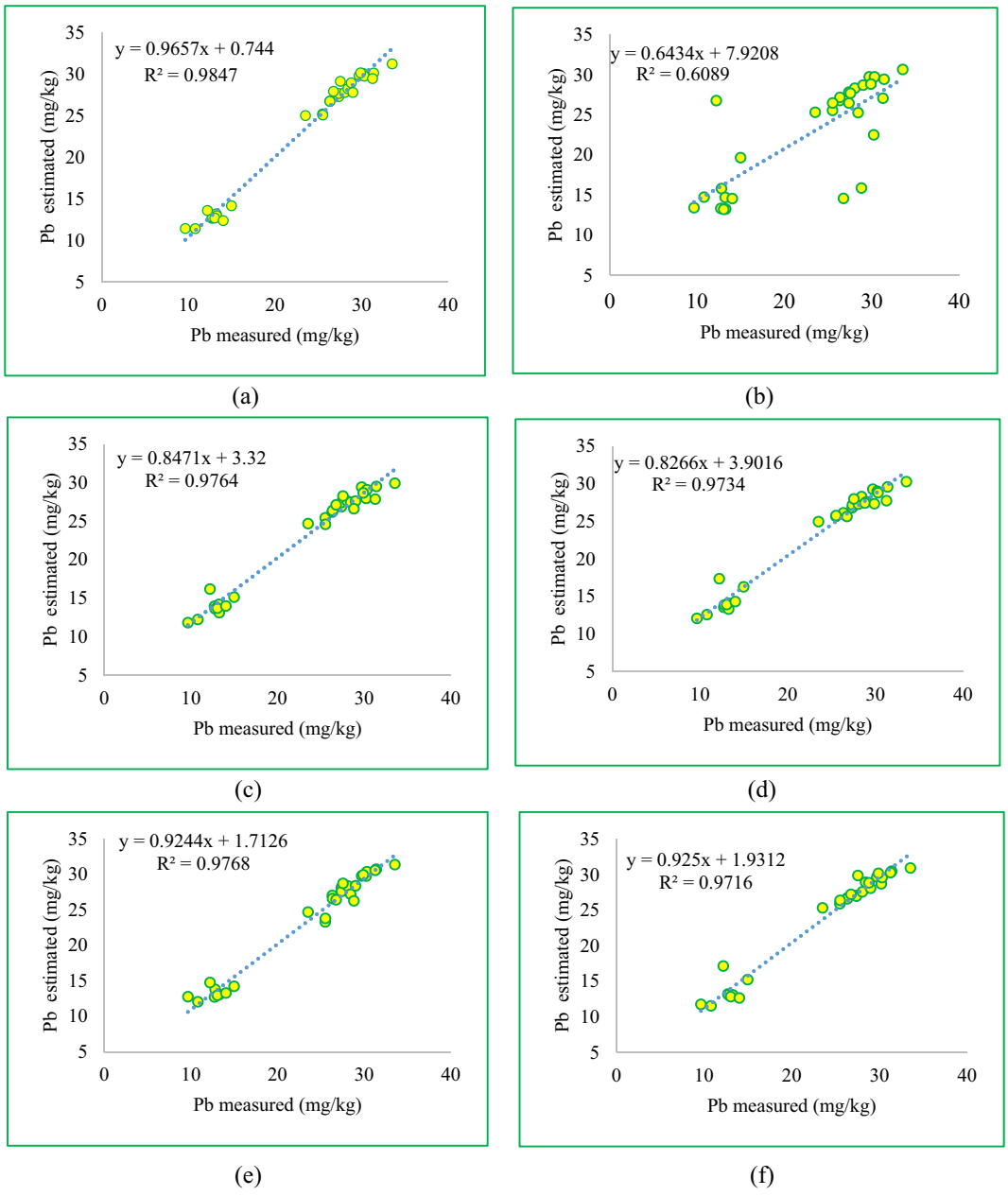

Fig. 4 Scatter plots of all the RF models for Pb. **a** Scatter plot of Pb with selected bands; **b** scatter plot of Pb with SG-preprocessed full-spectrum data; **c** scatter plot of Pb with FD-preprocessed full-spectrum data; **d** scatter plot of Pb with SD-preprocessed full-spectrum data; **e** scatter plot of Pb with SNV-preprocessed full-spectrum data; **f** scatter plot of Pb with CR-preprocessed full-spectrum data

the performance of FD and SD is inconclusive. That is to say, the performance of FD is more stable and reliable under the full-spectrum data at a particular spectral resolution (350–1000 nm at 1.4-nm intervals and 1000–2500 nm at 2-nm intervals).

All the models based on the RF algorithm are summarized in Figs. 5 and 6. It can be seen that the accuracy for Cr, As, and Pb using the selected bands are the highest, and that of Zn is slightly lower than using the CR-processed full-spectrum data. Comparing FD and

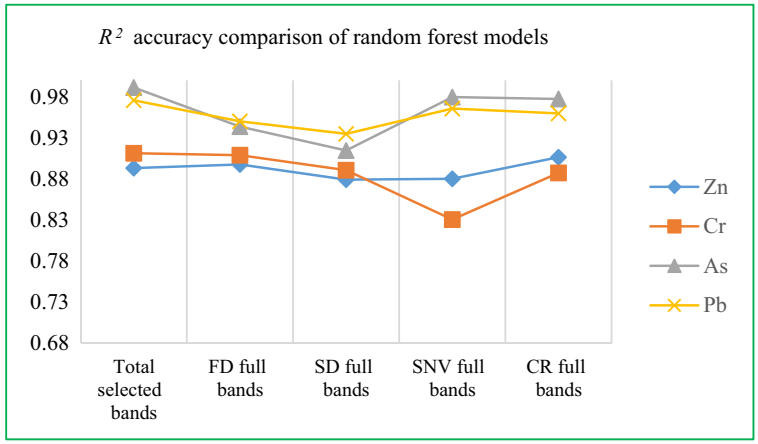

Fig. 5 R^2 accuracy comparison for the RF models

SD for the full-spectrum results, the results of FD are better. In all the full-spectrum models, using SNV-preprocessed data for As and Pb offers the best performance, and for Zn, using CR-processed is the best. For Cr, the R^2 value for FD is the highest, and the RMSE is the lowest for CR. In general, the FD and CR preprocessing can better enhance the spectral characteristics. It also shows that the feature selection method built in the RF can achieve the same effect as band selection, and this method is automatic. In addition, the RF owns specific bagging mechanism and random attribute selection mechanism, resulting in sample attribute disturbance and base learner diversity, and then ultimate performance can be improved by the final integration.

Taking all the experiments into consideration, we can see that the RF offers many advantages. By repeating the selection of feature subset, it forms many different trees. The deviation of each tree is small and the variance is large. Finally, the effect of the variance of the overall model is reduced by summing the trees. That is to say, the diversity of basic tree learners can achieve mutual compensation in ensemble learning, constructing a powerful learning algorithm. Meanwhile, due to repeated sampling, the learning space spanned by training samples can be extended to a certain level, which can make up for the problem of small sample set and avoid overfitting.

Conclusions

Few previous studies have directly used full-spectrum data to estimate the concentration of heavy metals, but, in this paper, an effective attempt is made, and the

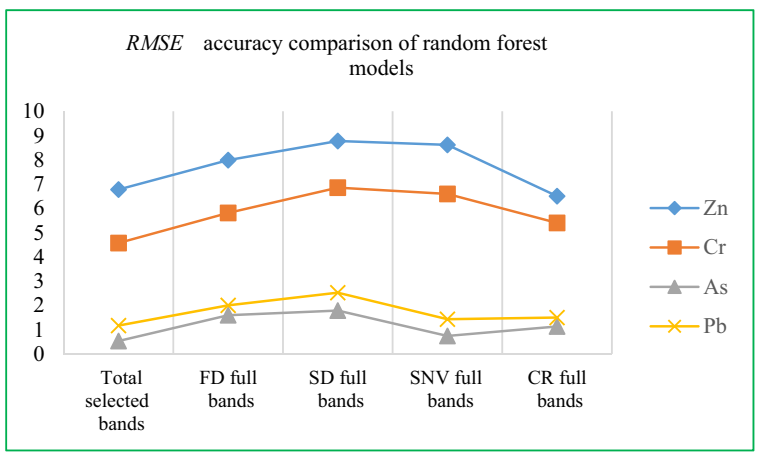

Fig. 6 RMSE accuracy comparison for the RF models

results show that RF can cope with full-spectrum data, providing an outstanding performance. In almost all the 20 experimental cases of heavy metal estimation, RF performs better than PLS and SVM, except using the SG-preprocessed full-spectrum data. The original smoothed spectra cannot provide effective feature information, and it is found that in all five preprocessing methods, FD and CR have a more stable ability for all four heavy metals.

Although the precision of the models using selected bands are the highest, the accuracy with full-spectrum data are very close. Feature selection in RF makes the following feature extraction more automatic. Considering the fact that selecting feature bands is often subjective and data-/application-dependent, RF offers a more convenient and efficient solution without careful band selection for hyperspectral applications.

Acknowledgments The authors would like to thank Professor Jihong Dong.

Funding information This research is supported in part by the Natural Science Foundation of China (No. 41871337, 41471356, 51874306), the Xuzhou Scientific Funds (KC16SS092), Key Laboratory for National Geographic Census and Monitoring, National Administration of Surveying, Mapping and Geoinformation 2018NGCM08 and Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Aguerssif, N., Benamor, M., Kachbi, M., & Draa, M. T. (2008). Simultaneous determination of Fe (III) and Al(III) by first-derivative spectrophotometry and partial least-squares (PLS-2) method – application to post-haemodialysis fluids. *Journal of Trace Elements in Medicine & Biology Organ of the Society for Minerals & Trace Elements, 22*(3), 175–182.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249–2260.

Asadzadeh, S., & Roberto, D. S. F., Carlos (2016). A review on spectral processing methods for geological remote sensing. *International Journal of Applied Earth Observation & Geoinformation, 47*, 69-90.

Balabin, R. M., & Lomakina, E. I. (2011). Support vector machine regression (SVR/LS-SVM)–an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Analyst, 136*(8), 1703–1712.

Balestrieri, C., Colonna, G., Giovane, A., Irace, G., & Servillo, L. (1978). Second-derivative spectroscopy of proteins. A method for the quantitative determination of aromatic amino acids in proteins. *European Journal of Biochemistry, 90*(3), 433–440.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *Isprs Journal of Photogrammetry & Remote Sensing, 114*, 24–31.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Breiman, L. I., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees (CART). *Encyclopedia of Ecology, 40*(3), 582–588.

Candolfi, A., Maesschalck, R. D., Jouan-Rimbaud, D., Hailey, P. A., & Massart, D. L. (1999). The influence of data preprocessing in the pattern recognition of excipients near-infrared spectra. *Journal of Pharmaceutical & Biomedical Analysis, 21*(1), 115–132.

Carranza, E. J. M., & Laborte, A. G. (2015). Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). *Computers & Geosciences, 74*, 60–70.

Cen, H., Bao, Y., Huang, M., & He, Y. (2006). *Comparison of data pre-processing in pattern recognition of milk powder Vis/NIR spectra*. Berlin Heidelberg: Springer.

Choe, E., Meer, F. V. D., Ruitenbeek, F. V., Werff, H. V. D., Smeth, B. D., & Kim, K. W. (2008). Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: a case study of the Rodalquilar mining area, SE Spain. *Remote Sensing of Environment, 112*(7), 3222–3233.

Ding, L. X., Wang, Z. H., & Ge, H. L. (2010). Continuum removal based hyperspectral characteristic analysis of leaves of different tree species. *Journal of Zhejiang Forestry College, 27*(6), 809–814.

Dong, J., Yu, M., Bian, Z., Zhao, Y., & Cheng, W. (2011). The safety study of heavy metal pollution in wheat planted in reclaimed soil of mining areas in Xuzhou, China. *Environmental Earth Sciences, 66*(2), 673–682.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems, 28*(7), 779–784.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Fearn, T., Riccioli, C., Garrido-Varo, A., & Guerrero-Ginel, J. E. (2009). On the geometry of SNV and MSC. *Chemometrics & Intelligent Laboratory Systems, 96*(1), 22–26.

Feng, Q., Liu, J., & Gong, J. (2010). Retrieval of remote sensing images using color, texture and spectral features. *International Journal of Engineering Science & Technology, 7*(1), 1074–1094.

Feng, Q., Liu, J., & Gong, J. (2015). UAV remote sensing for urban vegetation mapping using random forest and texture analysis. *Remote Sensing, 7*(1), 1074–1094.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recogn. Lett., 31*(14), 2225–2236.

Gomez, C., Lagacherie, P., & Coulouma, G. (2008). Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma, 148*(2), 141–148.

Gorry, P. A. (1990). General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. *Analytical Chemistry, 62*(6), 570–573.

Ham, J., Chen, Y., Crawford, M. M., & Ghosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience & Remote Sensing, 43*(3), 492–501.

Han, L., & Rundquist, D. C. (1997). Comparison of NIR/RED ratio and first derivative of reflectance in estimating algal-chlorophyll concentration: a case study in a turbid reservoir. *Remote Sensing of Environment, 62*(3), 253–261.

Hapfelmeier, A., & Ulm, K. (2014). Variable selection by random forests using data with missing values. *Computational Statistics & Data Analysis, 80*(80), 129–139.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). The elements of statistical learning. *Springer, 167*(1), 192–192.

Huang, S. (*2015*). A remote sensing ship recognition using random forest. In *Proceedings of The fourth International Conference on Information Science and Cloud Computing (ISCC2015)* (pp. 18–19). *Guangzhou, China*: Sissa Medialab srl Partita.

Jamshidi, B., Minaei, S., Mohajerani, E., & Ghassemian, H. (2012). Reflectance Vis/NIR spectroscopy for nondestructive taste characterization of Valencia oranges. *Computers & Electronics in Agriculture, 85*(5), 64–69.

Jie, L. (2012). Hyperspectral remote sensing estimation model for cd concentration in rice using support vector machines. *Yingyong Kexue Xuebao/journal of Applied Sciences, 30*(1), 105–110.

Kinoshita, R., Moebiusclune, B. N., Es, H. M. V., Hively, W. D., & Bilgilis, A. V. (2012). Strategies for soil quality assessment using visible and near-infrared reflectance spectroscopy in a Western Kenya Chronosequence. *Soil Science Society of America Journal, 76*(76), 1776–1788.

Kosmas, C. S., Curi, N., Bryant, R. B., & Franzmeier, D. P. (1984). Characterization of iron oxide minerals by second-derivative visible spectroscopy. *Soil Science Society of America Journal, 48*(2), 401–405.

Li, Z., Ma, Z., van der Kuijp, T. J., Yuan, Z., & Huang, L. (2014). A review of soil heavy metal pollution from mines in China: pollution and health risk assessment. *Science of the Total Environment, 468-469*, 843–853.

Liaw, A., & Wiener, M. (2001). Classification and regression by random forest. *R News, 2*(3), 18–22.

Madden, H. H. (1978). Comments on the Savitzky-Golay convolution method for least-squares-fit smoothing and differentiation of digital data. *Analytical Chemistry, 50*(3), 1383–1386.

Malley, D., & Williams, P. (1997). Use of near-infrared reflectance spectroscopy in prediction of heavy metals in freshwater sediment by their association with organic matter. *Environmental science & technology, 31*(12), 3461–3467.

Meer, F. V. D. (2000). Spectral curve shape matching with a continuum removed CCSM algorithm. *International Journal of Remote Sensing, 21*(16), 3179–3185.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. 2015 e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien,. *R package version* (pp. 1.6-7).

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217–222.

Rathod, P. H., Rossiter, D. G., Noomen, M. F., & Fd, V. D. M. (2013). Proximal spectral sensing to monitor phytoremediation of metal-contaminated soils. *International Journal of Phytoremediation, 15*(15), 405–426.

Rinnan, Å., Berg, F. V. D., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry, 28*(10), 1201–1222.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews, 71*, 804–818.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639.

Shi, T., Chen, Y., Liu, Y., & Wu, G. (2014). Visible and near-infrared reflectance spectroscopy—an alternative for monitoring soil contamination by heavy metals. *Journal of Hazardous Materials, 265*(2), 166–176.

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. *Applied Clay Science, 64*(4), 75–83.

Soriano-Disla, J. M., Janik, L. J., Rossel, R. A. V., Macdonald, L. M., & Mclaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews, 49*(2), 139–186.

Stańczyk, U. (2015). Feature selection for data and pattern recognition. *Studies in Computational Intelligence, 584*, 1–7.

Summers, D., Lewis, M., Ostendorf, B., & Chittleborough, D. (2009). Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. *Ecological Indicators, 11*(1), 123–131.

Tan, K., Ye, Y., Cao, Q., & Du, P. (2014). Estimation of arsenic contamination in reclaimed agricultural soils using reflectance spectroscopy and ANFIS model. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 7*(6), 2540–2546.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48*(1), 159–205.

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*(4), 1–9.

Wang, Q., Xie, Z., & Li, F. (2015). Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. *Environmental Pollution, 206*, 227–235.

Wang, L. A., Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal, 4*(3), 212–219.

Wei, B., & Yang, L. (2010). A review of heavy metal contaminations in urban soils, urban road dusts and agricultural soils from China. *Microchemical Journal, 94*(2), 99–107.

Wei, B., Jiang, F., Li, X., & Mu, S. (2009). Spatial distribution and contamination assessment of heavy metals in urban road

dusts from Urumqi, NW China. *Microchemical Journal, 93*(2), 147–152.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics & Intelligent Laboratory Systems, 58*(2), 109–130.

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. *Applied Geochemistry, 20*(6), 1051–1059.

Xu, L., Xie, D., & Fan, F. (2011). Effects of pretreatment methods and bands selection on soil nutrient hyperspectral evaluation. *Procedia Environmental Sciences, 10*, 2420–2425.