Journal of Hazardous Materials 382 (2020) 120987

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest

Kun Tan[a,b,c,*], Huimin Wang[c], Lihan Chen[c], Qian Du[d,**], Peijun Du[e,**], Cencen Pan[c]

[a] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
[b] School of Geographic Sciences, East China Normal University, Shanghai 200241, China
[c] Key Laboratory for Land Environment and Disaster Monitoring of NASG, China University of Mining and Technology, Xuzhou 221116, China
[d] Department of Electrical and Computer Engineering, Mississippi State University, MS 39762, USA
[e] Key Laboratory for Satellite Mapping Technology and Applications of NASG, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Editor: R. Teresa

Keywords:
Airborne hyperspectral remote sensing
Random forest
Soil heavy metal

ABSTRACT

Hyperspectral imaging, with the hundreds of bands and high spectral resolution, offers a promising approach for estimation of heavy metal concentration in agricultural soils. Using airborne imagery over a large-scale area for fast retrieval is of great importance for environmental monitoring and further decision support. However, few studies have focused on the estimation of soil heavy metal concentration by airborne hyperspectral imaging. In this study, we utilized the airborne hyperspectral data in LiuXin Mine of China obtained from HySpex VNIR-1600 and HySpex SWIR-384 sensor to establish the spectral-analysis-based model for retrieval of heavy metals concentration. Firstly, sixty soil samples were collected in situ, and their heavy metal concentrations (Cr, Cu, Pb) were determined by inductively coupled plasma-mass spectrometry analysis. Due to mixed pixels widespread in airborne hyperspectral images, spectral unmixing was conducted to obtain purer spectra of the soil and to improve the estimation accuracy. Ten of estimated models, including four different random forest models (RF)—standard random forest (SRF), regularized random forest (RRF), guided random forest (GRF), and guided regularized random forest (GRRF)—were introduced for hyperspectral estimated model in this paper. Compared with the estimation results, the best accuracy for Cr, Cu, and Pb is obtained by RF. It shows that RF can predict the three heavy metals better than other models in this area. For Cr, Cu, Pb, the best model of RF yields R_p^2 values of 0.75,0.68 and 0.74 respectively, and the values of $RMSE_p$ are 5.62, 8.24, and 2.81 (mg/kg), respectively. The experiments show the average estimated values are close to the truth condition and the high estimated values concentrated near several industries, valifating the effectiveness of the presented method.

1. Introduction

With the rapid development of industry and economy, soil heavy metal pollution has become an increasingly serious problem. The study of soil heavy metal pollution is of great importance to the biological world, human health, and the sustainable development of social resources (Wang et al., 2017). The traditional method for obtaining the soil heavy metal contamination is mainly implemented laboratory chemical analyses which is expensive and time-consuming. Moreover, the reagents of chemical analyses are harmful to environment and generated the re-contamination (Shi et al., 2016). The emergence of hyperspectral remote sensing technology has made the rapid monitoring of soil heavy metal pollution in large areas a reality (Choe et al., 2008; Pascucci et al., 2012; Shi et al., 2014). Using hyperspectral images of visible and infrared bands can allow us to retrieve the heavy metal concentration of soil (Bonifazi et al., 2018). Different from most soil metal concentration estimation conducted in lab environment, this research focuses on estimating the distribution trend using airborne images of large-scale areas. Using airborne remote sensing images is much more challenging than using in situ sample measurements, and we have to deal with mixed pixels due to relatively rough spatial resolution.

Feature selection is also required for hyperspectral data (Guyon and Elisseeff, 2003; Phuong et al., 2006), in order to simplify the model, shorten the running time, and improve the generalization of the model. There are three types of feature-variable selection methods: filter, wrapper, and embedding methods (Tuia et al., 2010). The filter methods use variable ranking techniques as the principle criterion for

* Corresponding author at: Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China.
** Corresponding authors.
E-mail addresses: tankuncu@gmail.com (K. Tan), du@ece.msstate.edu (Q. Du), dupjrs@126.com (P. Du).

https://doi.org/10.1016/j.jhazmat.2019.120987
Received 27 January 2019; Received in revised form 6 August 2019; Accepted 7 August 2019
Available online 17 August 2019
0304-3894/

variable selection. Unlike the filter methods, the wrapper methods take into account the correlation between variables (Kim et al., 2006). Recursive feature elimination (RFE) is a typical example of encapsulation. RFE is a backward feature-variable selection strategy that selects features by recursively considering smaller sets of features (Kowalski, 2009). The embedding methods incorporate feature selection as part of the training process. There are many feature selection algorithms that can be embedded in the model, including sparse regression regularization techniques (Kim and Kwon, 2010; Tibshirani, 2011; Zou and Hastie, 2005), the least absolute shrinkage and selection operator (LASSO) (Deng and Runger, 2013), ridge regression, elastic net (EN) (Ho, 1995), regularized trees (Breiman, 2001), and so on.

Prediction using a single decision tree results in a lower accuracy of retrieval. A random forest integrated approach can be a solution to this limitation. In the 1990s, the concept of random forest was proposed by Ho (Adam et al., 2014), and the model was extended by Breiman (Vincenzi et al., 2011) in 2001. To date, the study of random forest models in hyperspectral-based information retrieval has mainly focused on biomass estimation (Abdel-Rahman et al., 2013; Svetnik et al., 2004) and the estimation of physical and chemical properties of vegetation (Shi et al., 2017). Svetnik et al. (Qiu et al., 2016) used a random forest model to quantitatively describe the molecular structure of compounds to predict their biological activity, and the results obtained with six public data sets showed that the random forest model is superior to other three methods (i.e., decision tree, partial least squares (PLS), SVM), in the absence of parameter optimization. Abdel-Rahman et al. (Shi et al., 2017) used cross-validation to optimize the parameters of a random forest model and successfully estimated the leaf nitrogen concentration of sugarcane in EO-1 Hyperion hyperspectral data. Abdel-Rahman et al. (2013) successfully estimated the biomass of a wetland using backward feature removal and random forest. Random forest is also used to predict Cd (Huang et al., 2009) in soil and to identify and apportion heavy metal pollution sources in agricultural soils on a local scale (Peterson and Stow, 2003). Wang et al. (2015) compared the performance of several ensemble learning methods including RF 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. Hong et al. (Sun and Xia, 2017) used RF to establish the relationship between spectral data and two heavy metals (Pb and Zn).

Despite the excellent performance of the random forest models, the presence of mixed pixels in airborne hyperspectral images still limits the retrieval accuracy. Spectral unmixing as a spectrum processing technology is widely used in natural resource monitoring (Sun and Xia, 2017) and environmental monitoring (Smith et al., 2007; Li et al., 2015). However, spectral unmixing is often a supervised approach with known endmembers. In this research, we adopt the minimum volume simplex analysis (MVSA) algorithm proposed by Li et al. (Ye et al., 2019), which is a fast unsupervised linear unmixing method suitable to airborne hyperspectral data.

The use of airborne hyperspectral data, improved by geometric correction and radiation correction, unmixed by the MVSA method, and estimated the distribution trend of heavy metal concentration by the RF method, can provide the information needed for management and remediation of heavy metal contamination in agriculture soils. Remediation can be achieved through the use of biological remediation. Ye et al. (2017) used incorporation of biochar as raw material into co-composting with agricultural organic matter, and found the concentrations of available metals and arsenic in soil with great reduction in the treatment of biochar-blended composting. Several biological remediation methods were summarized for the remediation of co-contaminated soil with heavy metals and organic pollutants by Ye et al. (Ren and Huo, 2010). After obtaining the distribution trend of soil heavy metal, we can use these methods to remedy heavy metal in agriculture soils.

Illustrated in Fig. 1, the rest of this article is structured as follows. In Section 2, we introduce the image acquisition and processing (i.e., geometric correction, radiation correction), collection of soil samples and their heavy metal concentration measurement. This section also details the modeling methods and the MVSA spectral unmixing method. In Section 3, we provide the results of heavy metal distribution estimation from the images. Section 4 provides a summary and discussion.

2. Data and methods

2.1. Study area

The study area is located in the Liuxin coal mining area (117°13′ E, 34°37′ N), in the northwest of Xuzhou, Jiangsu province, China (Fig. 2). Brown soil, cinnamon soil, and alluvial soil are the common soil types in this study area. Almost 79% of China's electricity comes from coal-fired power plants, and Xuzhou is the main base of the power source for Jiangsu province. There is a common phenomenon in the area of land subsidence caused by the coal mining. The mining subsidence area has been reclaimed by two different methods—filling with gangue and filling with fly ash—and is currently being used for agricultural purposes. Both these fillers contain a large volume of heavy metals. Due to the impact on water cycle and wind action, the open dumps of mining waste material in the mining areas are rapidly weathering and diffusing to the surrounding areas, resulting in the contamination of crops and a negative effect on human health.

Xuzhou Jielong Packaging Paper Co., Ltd. (Fig. 3c1) is located in Sunzhuang, Liuxin Town, Tongshan County, Xuzhou City. The company was founded in 2009 and mainly produced cardboard and cartons. The printing process of cartons can lead to heavy metal pollution. Various chemical liquid, such as corrosive liquid and electroplating liquid, are used in the process of printing plate making. The liquid contains cadmium, copper, nickel, zinc, acid and other substances. If the liquid is discharged to rivers and lakes without effective treatment, it causes water and soil pollution. Heavy metals such as lead, chromium, cadmium and mercury in the printing process also pollute the environment (Ye et al., 2017). Liuxin Industrial Park (Fig. 3c2) has Xuzhou Jinguan Industrial Textile Products Co., Ltd. founded in 2007, Xuzhou Pengjia Packaging Co., Ltd. founded in 2001, and Xuzhou Laique Biotechnology Co., Ltd. founded in 2016. These companies mainly produce textiles, cartons, food packaging bags. The major heavy metals produced in the textile process are mercury, cadmium, copper, etc.

Xuzhou Gucheng Copper Industry Co., Ltd. (Fig. 3c3) was founded in 2000. It manufactures copper rods, copper wires, cables and wires. A large amount of pollutants including acid, Zn^{2+}, Cu^{2+}, Pb^{2+}, Cd^{2+}, Ni^{2+}, As^{3+}, and Co^{2+} are discharged during the copper smelting process (Ren and Huo, 2010). Xuzhou Tonghe Glass Product Factory (Fig. 3c4) was established in 2002 to produce various glass bottles. Abandoned coal mine are shown in Fig. 3c5. Solid wastes, such as coal gangue and coal slime, are produced in the process of coal mining and processing. The coal dust produced during the transportation process increased the concentration of particulates and the particles eventually precipitated on the surface of the soil. After long-term weathering and leaching, solid waste and particulate matter can cause heavy metals (Cd, Cu, Ni, Zn, Cr, Pb, Hg, As) pollution in soil and groundwater environment. The metal particulates emitted from iron and steel forging factory (Fig. 3c6) may contain heavy metals (zinc, cadmium, lead, nickel and chromium), depending on the grade of steel produced and the raw materials used.

Shitun coal mine (Fig. 3c7) was mined in 1992 and its mine production capacity was 180 thousand tons/year. Shitun Brick and Tile Factory (Fig. 3c8) was established in 1991. Clay, shale and coal gangue are generally used as raw materials when firing bricks and tiles. Coal gangue is mostly transported from the nearby Shitun Coal Mine, and its long-term accumulation causes pollution to the soil.

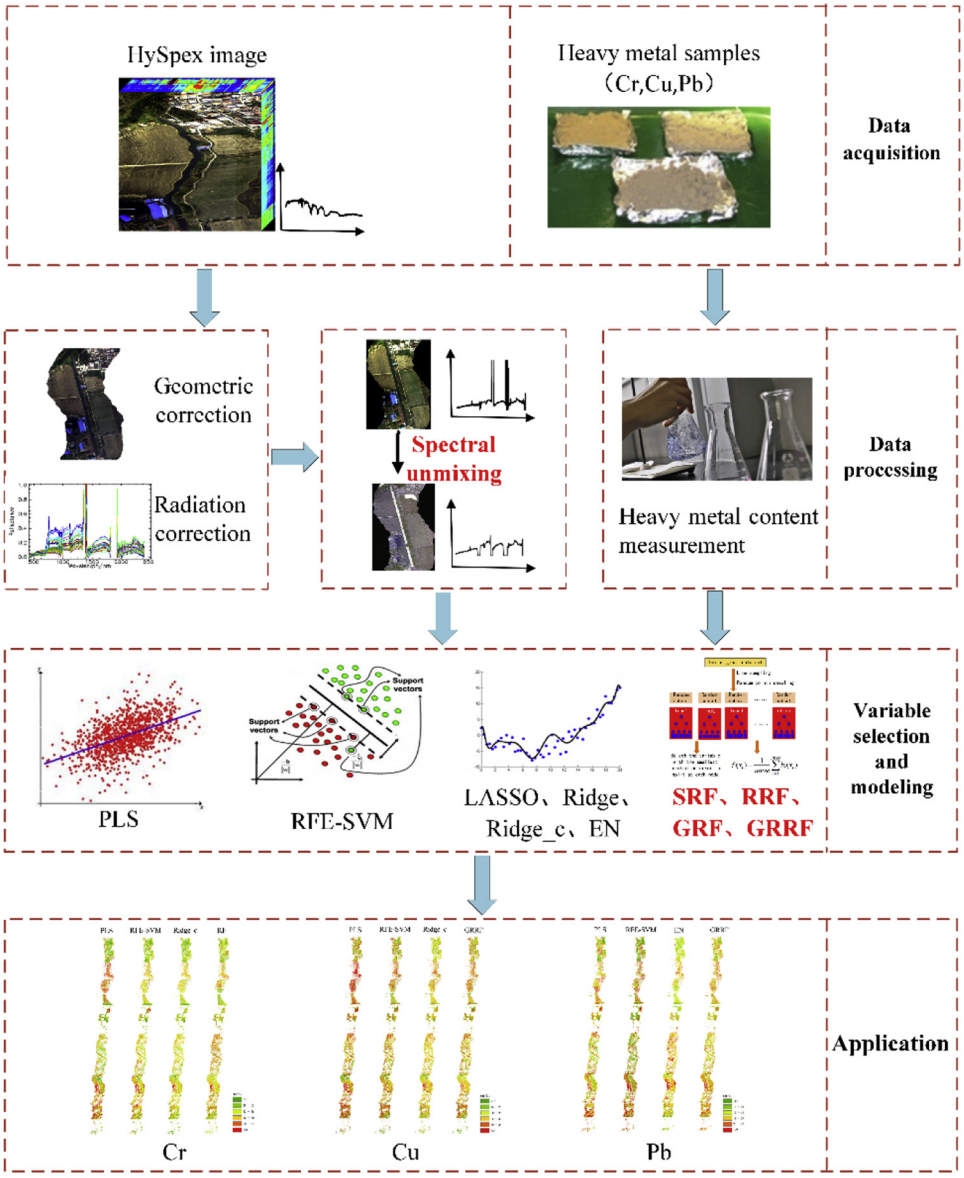

Fig. 1. The framework of soil heavy metal concentration estimation using airborne hyperspectral imagery.

2.2. Image acquisition and preprocessing

The airborne hyperspectral data were obtained by NEO imaging spectrometers produced in Norway. The HySpex VNIR-1600 and HySpex SWIR-384 hyperspectral cameras obtain visible and near-infrared (VNIR) data and short-wave infrared (SWIR) data, respectively. The main parameters of hyperspectral data are presented in Table 1.

It was necessary to perform geometric correction and radiation correction on the original images. The geometric correction involved spatial coordinate system conversion. The Hyspex VNIR-1600 data was resampled to the same spatial resolution (0.73 m) as the Hyspex SWIR-384 data. The two images have done the geometric correction using ground control points, and the RMSE is 0.54 m. The new image with spectral range of 400–2500 nm is generated after image registration and stitching. The geometric correction images are shown in Fig. 4.

The radiation correction consisted of system radiation correction and atmospheric radiation correction. HySpex RAD radiation calibration software was used, which is included in the HySpex imaging spectrometer system for system radiation correction. Converting a unit of data to a unit of radiance value through the radiation correction, the unit is W/nm•sr•m^2.In order to further eliminate the influence of atmospheric conditions, the radiance value was converted into real ground reflectance. The atmospheric correction model used was the MODerate resolution atmospheric TRANsmission (MODTRAN) atmospheric correction model. Setting different elevations in MODTRAN, can eliminate the influence of terrain and elevation. The spectral signatures of soil samples after atmospheric correction are presented in Fig. 5.

2.3. Sample acquisition

The time for collecting soil samples was November 8, 2014.We had investigated this study area before soil sampling. It was divided into the different sampling units based on the soil characteristics and the topographic maps of the sampling area. "S"-shaped sampling strategy was adopted in farmland, which could effectively avoid sampling errors caused by tillage and fertilization. Topsoil was the main distribution layer of the wheat roots, and was also the farming layer of agricultural production. Meanwhile, soil pollutants were mainly concentrated on the surface layer because these industries discharge the contaminant on the surface. Airborne hyperspectral imaging could only obtain the spectrum of topsoil. Therefore, about one kilogram topsoil was

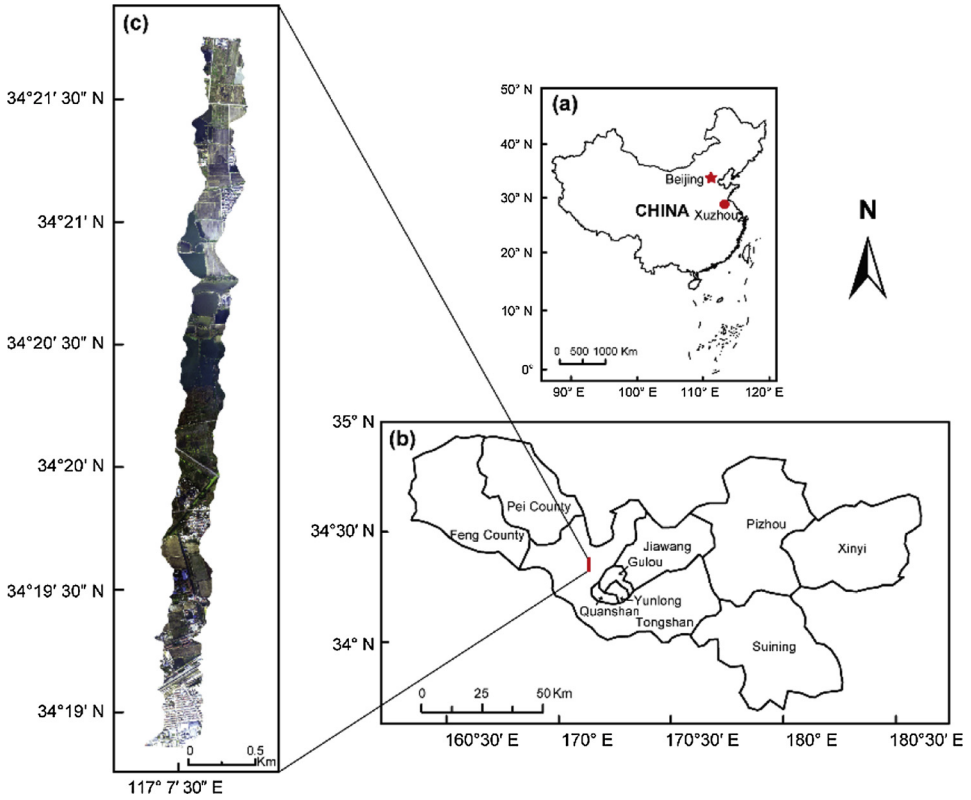

Fig. 2. Map showing the study area (c) at Xuzhou (b), Jiangsu Province, China (a). The area around the image includes several factories and coal mines. The following is an analysis of how industrial activities affect the heavy metals in agricultural soil.

collected at a depth of 0–20 cm surface.

The visible and near-infrared spectra of the soil samples under field conditions were measured with an ASD (Analytical Spectral Devices) field spectrometer (350–2500 nm) while collecting soil samples.

A total of 60 soil samples were collected according to the grid sampling method. The soil samples were sealed, marked, and brought back to the laboratory. In the laboratory, some sundries in the soil samples, such as stones, leaves, and roots, were removed. The soil samples were dried and ground and passed through a 120-mesh nylon sieve.

After drying, grinding and passing, soil samples were added to pre-cleaned digestion flask. A solution of HNO_3 and HCl with a ratio of 1:1 was poured into the samples. Then the soil samples were heated using an oven. The samples were removed from the oven and left to cool down to room temperature. After cooling down, the samples were diluted using pure deionized water and were placed on the hot plate until evaporated near to a dry state. The samples were then left to cool down and diluted using deionized water. At last, after filtering, the heavy metal concentrations of the soil samples were measured by inductively coupled plasma-mass spectrometry (ICP-MS).

From the preliminary analysis, Cr, Cu, and Pb were used in the experiments. Table 2 shows the max, min, mean, standard deviation (Std.), and coefficient of variation (CV) of the heavy metal concentrations. It can be seen that the coefficient of variation of Cr and Pb is about 0.2, which indicates that the data distribution is concentrated and suitable to statistical analysis. The CV of Cu is slightly larger than that of Cr and Pb, indicating the concentration of Cu is more dispersed than that of the other two metals.

Sixty soil samples were ranked from low to high values according to the heavy metal concentration in the soil. The sixty ranked soil samples were divided into 20 groups, and then two samples in each group were selected as training samples and one was used as the verification sample. It is better to balance the standard deviation and coefficient of variation for both data sets.

2.4. Hyperspectral feature selection using random forest

Feature selection can reduce the computation time, improve the prediction performance, and gain a better understanding of the data in machine learning or pattern recognition applications. The PLS algorithm performs multiple linear regression, while principal component analysis (PCA) and canonical correlation analysis (CCA) use statistical frameworks. The embedding method is also widely used, such as RFE-SVM, the regularization methods (LASSO, Ridge, EN, Ridge_c), and the random forest methods (i.e., standard random forest (SRF), regularized random forest (RRF), guided random forest (GRF), guided regularized random forest (GRRF). The main idea is to incorporate the variable selection as part of the training process. The regularization methods involve adding additional constraints or penalties to the existing model (loss function) to prevent over-fitting and improve the generalization ability. LASSO and Ridge use the L1 norm and L2 norm as penalty terms, respectively. LASSO is a good variable selection method due to the L1 regularization making the learned model very sparse. Ridge is appropriate for the understanding of the data because the corresponding coefficient is often non-zero. In order to simplify the combination of variables in ridge regression, the Ridge_c method based on the importance of the variables of ridge regression was proposed. EN combines the two normalization methods of the L1 norm and L2 norm.

A variety of effective variable selection strategies based on random forest have been presented. In this study, we constrain the information gain and the importance scores of the variables. RRF applies the regularized information gain and GRF applies the importance score to guide the random forest. GRRF applies the two strategies to make the selected variables more relevant and less redundant.

2.4.1. Standard random forest

The SRF regression algorithm is a bagging method based on classification and regression tree (CART) analysis. It employs recursive partitioning to divide the data into many homogenous subsets called

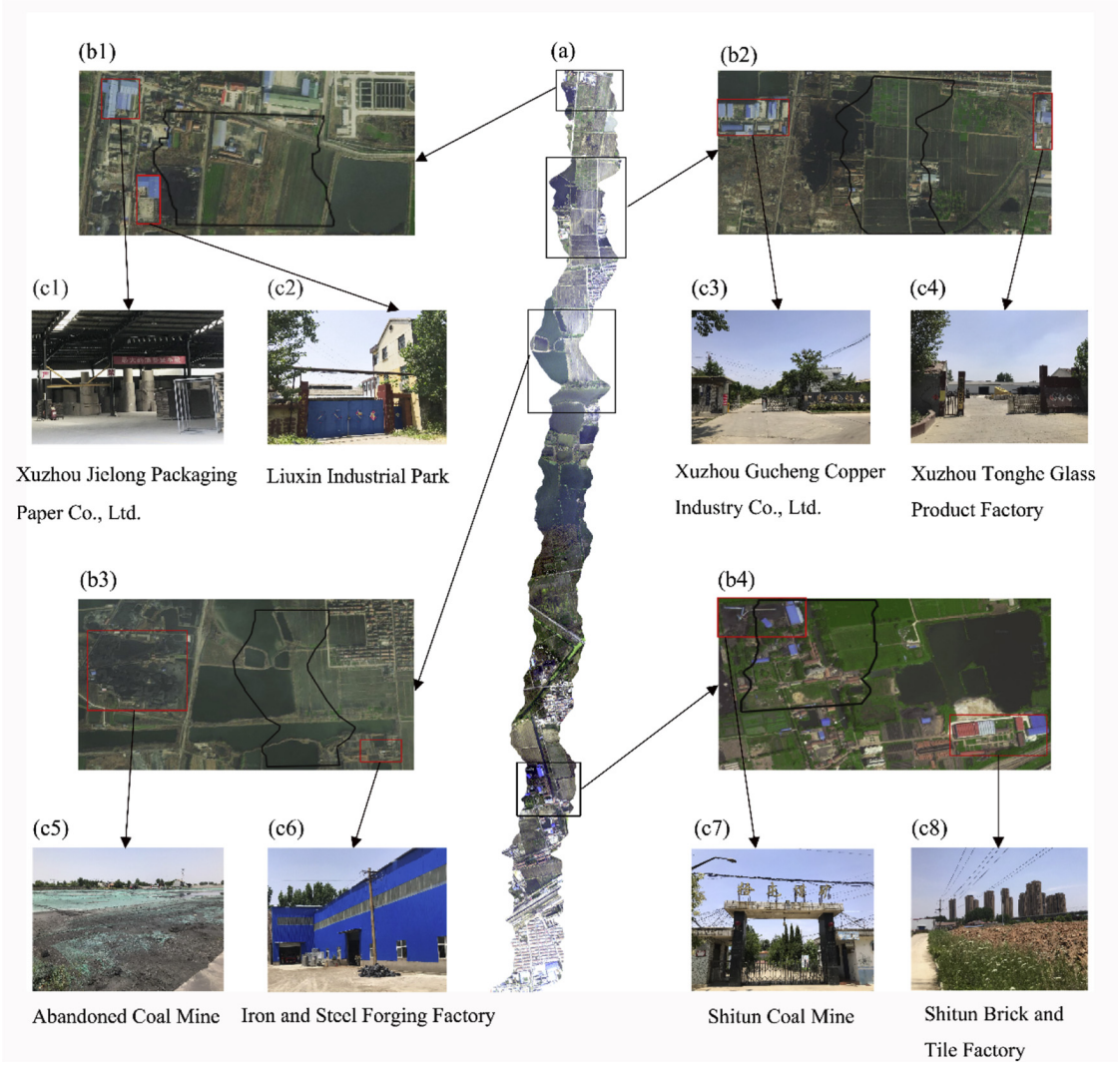

Fig. 3. A schematic diagram of the influence factors near the image. b1, b2, b3 and b4 are the four regions in the image and their surrounding areas. The black border indicates the location of the image in this area. The red border indicates the factory and the coal mine. c1, c2, c3, c4, c5, c6, c7 and c8 are photos taken on the spot (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 1
Main parameters of Hyspex VNIR-1600 and Hyspex SWIR-384 data.

Parameters	Hyspex VNIR-1600 data	Hyspex SWIR-384 data
Data of acquisition	2014-11-08	2014-11-08
Spectral range	400–1000 nm	1000–2500 nm
Channels	160	288
Spectral bandwidth	3.6nm	5.4nm
Spectral sampling interval	3.7nm	5.45 nm
Flight altitude	1km	1km
Ground sampling distance	0.19m	0.73m
Peak SNR	> 200	> 1100

regression trees (*ntree*), and then averages the results of all the trees. Each tree is independently grown to its maximum size based on a bootstrap sample from the training data set (approximately 70%), without any pruning (that is, without stopping the selection of the input variables at each node). In each tree, the SRF randomly selects a subset of variables (*mtry*) to determine the split at each node. The Gini coefficient $Gini(v)$ at node v can be computed as

$$Gini(v) = \sum_{j=1}^{p} \hat{p}_c^v (1 - \hat{p}_c^v) \tag{1}$$

where $\hat{p}_c^v$ is the observed value of the jth variable at node v. The X_i Gini information gain at split node v, i.e., $Gain(X_i, v)$, is the impurity difference between node v and the child node of node v, which is updated as:

$$Gain(X_i, v) = Gini(X_i, v) - w_L Gini(X_i, v^L) - w_R Gini(X_i, v^R) \tag{2}$$

where v^L and v^R are the left and right subnodes at node v, respectively, and w_L and w_R are the ratio of the characteristic variables to the left and right subnodes, respectively. At each node, the *mtry*($mtry \approx \sqrt{p}$)variables are randomly selected in the p variables, and the characteristic variables of the maximum information gain are finally obtained for the splitting of node v. The importance of the variable X_i is calculated as:

$$imp_i = \frac{1}{ntree} \sum_{v \in S_{Xi}} Gain(X_i, v) \tag{3}$$

where S_{X_i} is a collection of nodes that are split into X_i in a random forest of *ntree* trees. The importance score is used to evaluate the contribution of the characteristic variables for the prediction.

The random forest process simply includes four steps: 1) the original sample bootstrap sampling; 2) random selection of the *mtry* characteristics to establish the decision tree; 3) repeat the above two steps *ntree* times, that is, the formation of *ntree* decision trees making up the random forest; 4) for the new data, the predicted average of all the

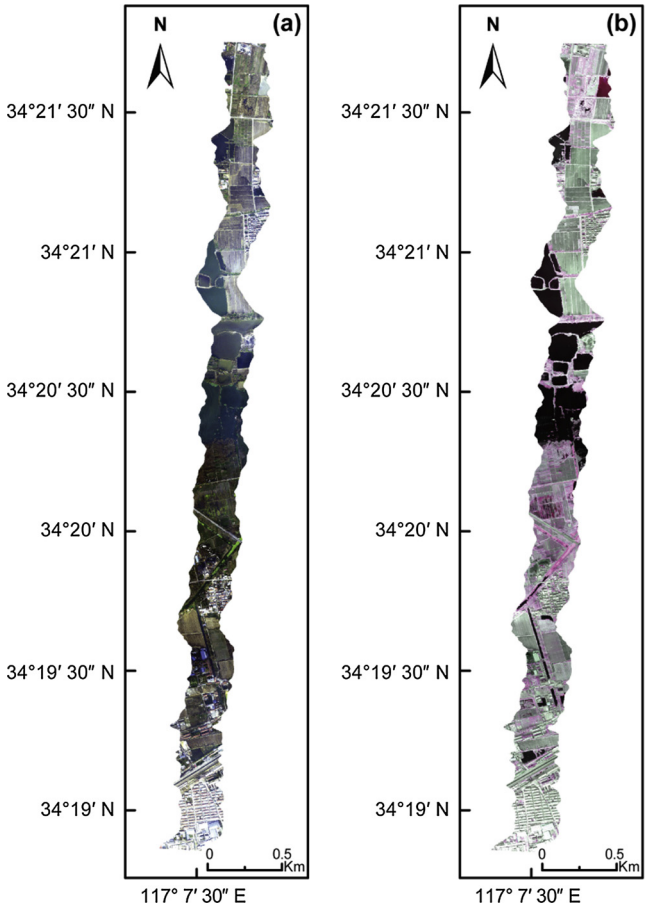

Fig. 4. Geometric correction images (a: VNIR; b: SWIR).

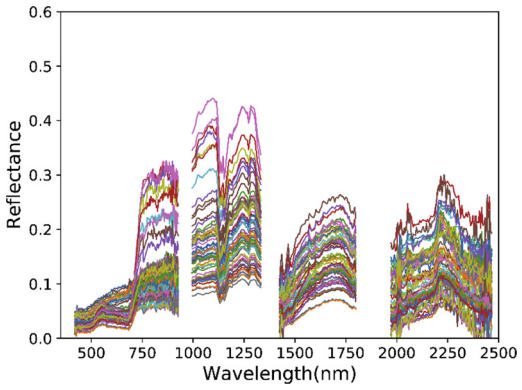

Fig. 5. Airborne reflectance curves of the soil samples after atmospheric correction.

Table 2
Statistical results of the soil heavy metal concentration.

Element	Max (mg/kg)	Min (mg/kg)	Mean (mg/kg)	Std. (mg/kg)	CV (%)
Cr	81.34	25.03	50.60	10.90	0.22
Cu	60.03	4.68	24.19	11.98	0.50
Pb	24.93	10.32	15.77	3.75	0.24

decision trees is predicted. The principle of node splitting is used to minimize the prediction error.

2.4.2. Regularized random forest

The features extracted from the decision tree model may be redundant, and the regularization tree method can solve this problem well. When the information gain of the next set of variable subsets is not very different from the previous set, the regularization tree does not repeatedly select new features to avoid feature redundancy. Each node recursively performs feature-variable splitting. Random forest uses the regularization strategy for the tree to form a regularized random forest (RRF), thus selecting a subset of the characteristic of the compression. The main difference from the original random forest is the application of regularized information gains as:

$$Gain_R(X_i, v) = \begin{Bmatrix} \lambda Gain(X_i, v)\, i \notin F \\ Gain(X_i, v)\, i \in F \end{Bmatrix} \tag{4}$$

where F is a set of feature indices for the splitting of the previous node, which is an empty set at the root node of the first tree. It not only suppresses the feature subset of the current tree splitting, but also suppresses the previously established tree. $\lambda \in (0,1]$ is the penalty coefficient; when $i \notin F$, the coefficient is used to divide the ith feature of node v. The smaller the value of λ, the greater the penalty. The RRF uses the regularized information gain $Gain_R(X_i, v)$ at each node, adding the index of the new variable to the set F when the feature variable adds enough predictive information for the existing variable. The difference between RRF and SRF is that RRF uses $Gain_R(X_i, v)$ to select the splitting feature. We calculate all the variables in $Gain_R(X_i, v)$ that belong to F and do not belong to the *mtry* variable in F. All the variables must increase the gain after the penalty to enter the set.

2.4.3. Guided random forest

Guided random forest (GRF) is a variable selection method based on random forest. It involves the use of the importance score in random forest to guide the random forest. $Gain(X_i)$ represents the Gini information gain for the feature X_i to be split at the tree node. The central idea of GRF is to weight $Gain(X_i)$ using the importance score in RF:

$$Gain_G(X_i) = \lambda_i Gain(X_i) \tag{5}$$

where λ_i is calculated as:

$$\lambda_i = 1 - \gamma + \gamma \frac{imp_i}{imp^*} \tag{6}$$

Here, imp_i is the importance score of variable X_i in the random forest, imp^* is the maximum importance score, $\frac{imp_i}{imp^*}$ is the normalized importance score, and $\gamma \in [0,1]$ controls the weight of the importance score in the SRF. It can be seen that features with smaller importance scores are more heavily penalized, and the penalty increases as γ becomes greater (GRF becomes SRF when $\gamma = 0$). In this research, the maximum penalty (i.e., $\gamma = 1$) is used, in order to select a small number of features in GRF. So $Gain_G(X_i)$ becomes:

$$Gain_G(X_i) = \frac{imp_i}{imp^*} Gain(X_i) \tag{7}$$

2.4.4. Guided regularized random forest

Guided regularized random forest (GRRF) involves the use of the importance score to guide the RRF, thus realizing the feature-variable selection process. The standardized importance score is defined as:

$$imp_i' = \frac{imp_i}{\max_{j=1}^{p} imp_j} \tag{8}$$

Unlike RRF, which imposes a unique penalty parameter on all the features, GRRF assigns a penalty factor to each feature as

$$Gain_R(X_i, v) = \begin{Bmatrix} \lambda_i \cdot Gain(X_i, v)\, i \notin F \\ Gain(X_i, v)\, i \in F \end{Bmatrix} \tag{9}$$

where $\lambda_i \in (0,1]$ is the coefficient for X_i, $i \in \{1, \ldots p\}$ and is calculated based on the importance score of X_i from ordinary random forest as

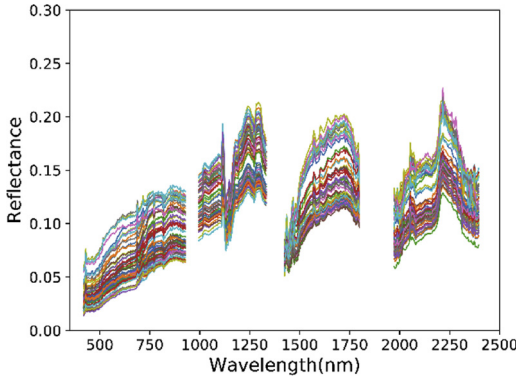

Fig. 6. Airborne reflectance curves of the soil samples after unmixing.

$$\lambda_i = (1-\gamma)\lambda_0 + \gamma \times imp_i' \tag{10}$$

where $\lambda_0 \in (0,1]$ is the base coefficient to control the degree of regularization, and $\gamma \in [0,1]$ is the importance coefficient to control the weight of the importance score after normalization. Note that the RRF is a special case of the GRRF with $\gamma = 0$. In general, γ and λ_0 together affect the size of the feature subset. To reduce the number of parameters of GRRF, λ_0 is fixed to be 1 and γ is considered as the only parameter for GRRF. With $\lambda_0 = 1$, we have:

$$\lambda_i = (1-\gamma) + \gamma \times imp_i' = 1 - \gamma(1 - imp_i') \tag{11}$$

A larger γ leads to a smaller λ_i, thus a larger penalty on $Gain(X_i, v)$ when X_i has not been used in the nodes prior to node v. Consequently, γ is essentially the degree of regularization.

The models are evaluated with the coefficient of determination for calibration (R_c^2), the root-mean-square error of calibration ($RMSE_C$), the mean relative error of calibration (MREc), the coefficient of determination for prediction (R_p^2), the root-mean-square error of prediction ($RMSE_P$), and the mean relative error of prediction (MREp), which are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{13}$$

$$MRE = \frac{\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}}{n} \tag{14}$$

where n is the number of samples, y_i is the ith measured value, $\hat{y}_i$ is the ith predicted value, and $\bar{y}$ is the average of the measured values. In general, a large value of R^2 combined with small values of *RMSE* and *MRE* means better prediction.

2.5. Spectral unmixing

The presence of mixed pixels in airborne hyperspectral images would limit the retrieval accuracy. Some pixels which contain both vegetation and soil would have a serious effect on the accuracy of retrieval. The soil spectrum extracted from the pixel is improper while the vegetation in the pixel influences the whole pixel spectrum. However, such impact can be alleviated by spectral unmixing.

MVSA approaches hyperspectral unmixing by fitting a minimum volume simplex to the hyperspectral data, constraining the abundance fractions to belong to the probability simplex. The vertex component analysis (VCA) algorithm (Ahmed et al., 2008) is an efficient method based on the pure pixel hypothesis. Therefore, the initial of the MVSA algorithm is realized by the VCA algorithm.

Fig. 6 shows the airborne spectra of the soil samples after spectral unmixing, where the bands affected by water vapor have been removed. The spectra of the pixels after unmixing are smaller than the original spectra. The original spectra are between 0 and 0.45 and the spectra after unmixing are between 0 and 0.25.

Table 3
Regression results of PLS, RFE-SVM, LASSO, Ridge, EN, Ridge_c, SRF, RRF, GRF, GRRF for field spectra.

Element	Method	Rc2	RMSEc	MREc	Rp2	RMSEp	MREp	No. of variables
Cr	PLS	0.3665	8.5259	0.1380	0.1487	14.4045	0.2352	4
	RFE-SVM	0.9560	2.2597	0.0223	0.3500	14.8095	0.2493	233
	LASSO	0.4833	9.1266	0.1364	0.2675	19.6251	0.3028	0
	Ridge	0.3080	9.7283	0.1481	0.2640	16.5195	0.2677	394
	EN	0.4831	9.4766	0.1405	0.2677	18.3581	0.2896	0
	Ridge_c	**0.3704**	**9.0496**	**0.1388**	**0.5804**	**19.4263**	**0.3027**	**155**
	SRF	0.8542	5.1676	0.0789	0.3326	15.4499	0.2570	394
	RRF	**0.8682**	**5.2712**	**0.0871**	**0.4113**	**15.0728**	**0.2473**	**6**
	GRF	0.8536	5.1275	0.0790	0.3661	15.2697	0.2527	137
	GRRF	0.8320	5.3159	0.0815	0.3489	14.8008	0.2463	9
Cu	PLS	0.8606	5.8718	0.2843	0.0376	11.8562	0.8750	394
	RFE-SVM	**0.9652**	**2.6939**	**0.0454**	**0.2302**	**11.2012**	**0.8166**	**45**
	LASSO	0.2647	11.1132	0.5612	0.1597	12.6780	0.7015	0
	Ridge	0.2793	10.7163	0.5329	0.1426	12.5894	0.6841	394
	EN	0.2690	11.1419	0.5634	0.1912	12.6041	0.7036	0
	Ridge_c	0.2813	10.5979	0.5231	0.1405	12.8322	0.6715	300
	SRF	0.8606	5.8718	0.2843	0.0376	11.8562	0.8750	394
	RRF	**0.8434**	**6.4974**	**0.3331**	**0.4146**	**13.1818**	**0.9933**	**5**
	GRF	0.8491	5.9185	0.2923	0.0955	13.1024	1.0014	127
	GRRF	0.8002	6.7166	0.3363	0.3988	13.0845	1.0068	5
Pb	PLS	0.2262	3.2407	0.1763	0.2278	5.3203	0.2310	4
	RFE-SVM	0.2111	3.4022	0.1514	0.0833	4.9307	0.2883	6
	LASSO	0.4664	2.8426	0.1532	0.3690	7.1831	0.3192	2
	Ridge	**0.2280**	**3.5291**	**0.1918**	**0.5807**	**4.9440**	**0.2805**	**394**
	EN	0.4725	2.8318	0.1529	0.3749	7.6277	0.3703	3
	Ridge_c	0.2667	3.4488	0.1884	0.5618	4.9192	0.2675	295
	SRF	0.9040	1.7360	0.0922	0.2380	5.0296	0.2765	394
	RRF	**0.8400**	**1.9587**	**0.1056**	**0.5710**	**4.9364**	**0.2719**	**7**
	GRF	0.8950	1.7891	0.0957	0.3904	4.9612	0.2822	136
	GRRF	0.8606	1.9344	0.1061	0.5345	4.8534	0.2779	7

Table 4
Regression results of PLS, RFE-SVM, LASSO, Ridge, EN, Ridge_c, SRF, RRF, GRF, GRRF using the original pixels.

Element	Method	R_c^2	RMSEc	MREc	R_p^2	RMSEp	MREp	No. of variables
Cr	PLS	**0.6333**	**6.531**	**0.1032**	**0.2337**	**10.3331**	**0.1627**	**6**
	RFE-SVM	0.0851	10.5156	0.1559	0.1782	10.1048	0.1162	30
	LASSO	0.0629	10.6793	0.1649	0.2225	10.4021	0.1264	0
	Ridge	0.0988	10.3320	0.1594	0.2049	9.5907	0.1176	375
	EN	0.0000	10.7852	0.1662	0.0000	10.6154	0.1295	0
	Ridge_c	0.0894	10.3536	0.1612	0.2140	9.5708	0.1175	156
	SRF	0.9231	5.2718	0.0836	0.3018	9.0063	0.1280	375
	RRF	**0.8932**	**5.2224**	**0.0819**	**0.4164**	**8.4908**	**0.1160**	**13**
	GRF	0.9235	5.2652	0.0796	0.3511	8.8422	0.1266	134
	GRRF	0.8814	5.4282	0.0841	0.4040	8.6346	0.1137	11
Cu	PLS	0.4503	8.7784	0.4318	0.0690	11.8575	0.6234	5
	RFE-SVM	**0.9973**	**1.0505**	**0.0341**	**0.1486**	**10.9203**	**0.4876**	**33**
	LASSO	0.0872	11.4220	0.5851	0.0549	11.4452	0.5661	0
	Ridge	0.0646	11.5620	0.5932	0.0503	11.4866	0.5699	375
	EN	0.0877	11.4249	0.5854	0.0527	11.4509	0.5664	0
	Ridge_c	0.0713	11.4435	0.5796	0.0511	11.4100	0.5681	215
	SRF	0.9198	4.9589	0.2674	0.1649	10.6759	0.5726	375
	RRF	**0.9248**	**5.1768**	**0.2743**	**0.2358**	**10.2737**	**0.5379**	**15**
	GRF	0.9155	5.0480	0.2738	0.1685	10.6558	0.5724	150
	GRRF	0.9120	5.0161	0.2507	0.1330	10.9086	0.5390	16
Pb	PLS	0.5853	2.3284	0.1226	0.0110	4.1451	0.2007	6
	RFE-SVM	**0.9925**	**0.3935**	**0.0116**	**0.0693**	**3.7798**	**0.1821**	**18**
	LASSO	0.0510	3.5870	0.1965	0.0001	3.8206	0.1928	0
	Ridge	0.0463	3.5560	0.1979	0.0007	3.8265	0.1973	375
	EN	0.0510	3.5882	0.1964	0.0001	3.8207	0.1927	0
	Ridge_c	0.0298	3.5840	0.1963	0.0355	3.7814	0.1923	2
	SRF	0.8857	1.6904	0.0926	0.0508	3.8335	0.1900	375
	RRF	**0.9061**	**1.7289**	**0.0960**	**0.3773**	**3.1457**	**0.1508**	**8**
	GRF	0.8988	1.6966	0.0936	0.1106	3.6494	0.1802	135
	GRRF	0.9200	1.8004	0.1013	0.2242	3.4381	0.1735	7

Table 5
Regression results of PLS, RFE-SVM, LASSO, Ridge, EN, Ridge_c, SRF, RRF, GRF, GRRF using unmixed pixels.

Element	Method	R_c^2	RMSEc	MREc	R_p^2	RMSEp	MREp	No. of variables
Cr	PLS	0.4960	7.6565	0.1372	0.3627	9.2370	0.1467	6
	RFE-SVM	**0.8520**	**4.5489**	**0.0675**	**0.7234**	**5.5090**	**0.0803**	**35**
	LASSO	0.1286	10.7234	0.1648	0.4483	10.5000	0.1275	0
	Ridge	0.2442	9.5526	0.1585	0.6351	7.9588	0.1033	375
	EN	0.1284	10.2261	0.1566	0.4495	9.3813	0.1117	0
	Ridge_c	0.2345	9.6102	0.1575	0.6484	8.0473	0.1037	355
	SRF	**0.9200**	**3.9849**	**0.0665**	**0.7460**	**5.6241**	**0.0836**	**375**
	RRF	0.9144	4.2060	0.0713	0.7250	5.8046	0.0924	13
	GRF	0.9138	4.2897	0.0713	0.7433	5.7199	0.0823	142
	GRRF	0.9178	4.5130	0.0745	0.6734	6.4840	0.0999	8
Cu	PLS	0.5473	7.7109	0.4423	0.5630	8.1702	0.3940	10
	RFE-SVM	**0.5582**	**7.6969**	**0.4299**	**0.5758**	**8.3345**	**0.3909**	**16**
	LASSO	0.2435	10.0307	0.5966	0.3999	9.8157	0.4380	0
	Ridge	0.2087	10.2628	0.5959	0.3754	10.0129	0.4347	375
	EN	0.2432	10.0444	0.5961	0.3961	9.8576	0.4403	0
	Ridge_c	0.2952	9.6762	0.5852	0.3998	9.6489	0.4304	42
	SRF	0.9343	4.8343	0.2643	0.2776	10.5051	0.5191	375
	RRF	0.9070	5.1337	0.2854	0.6383	8.3616	0.4018	9
	GRF	0.9295	4.7051	0.2517	0.3793	9.7912	0.4983	140
	GRRF	**0.8936**	**5.4407**	**0.3021**	**0.6826**	**8.2408**	**0.3742**	**8**
Pb	PLS	0.2631	3.1039	0.1694	0.2571	3.3256	0.1612	7
	RFE-SVM	**0.4455**	**2.7356**	**0.1128**	**0.4212**	**3.5216**	**0.2007**	**16**
	LASSO	0.3863	2.8849	0.1605	0.3892	3.0474	0.1639	0
	Ridge	0.2493	3.2241	0.1795	0.2413	3.2379	0.1764	16
	EN	0.3940	2.8667	0.1595	0.3991	3.0021	0.1577	0
	Ridge_c	0.2431	3.2366	0.1801	0.2317	3.2582	0.1777	248
	SRF	0.9143	1.5931	0.0860	0.6156	2.8899	0.1492	375
	RRF	0.8959	1.7219	0.0929	0.6493	2.8141	0.1490	8
	GRF	0.9269	1.5990	0.0870	0.6921	2.7727	0.1425	138
	GRRF	**0.8878**	**1.7368**	**0.0967**	**0.7375**	**2.8099**	**0.1417**	**9**

3. Results and discussion

3.1. The influence factors of retrieval accuracy

In general, there are four main factors that influence predication accuracy. Firstly, the acquisition process of airborne hyperspectral data would influence the accuracy of the predictions, such as flight conditions, flight route and instrument precision etc. Before the flight, the weather and flight conditions would be considered and a suitable time would be selected. Meanwhile, the instrument calibration should be

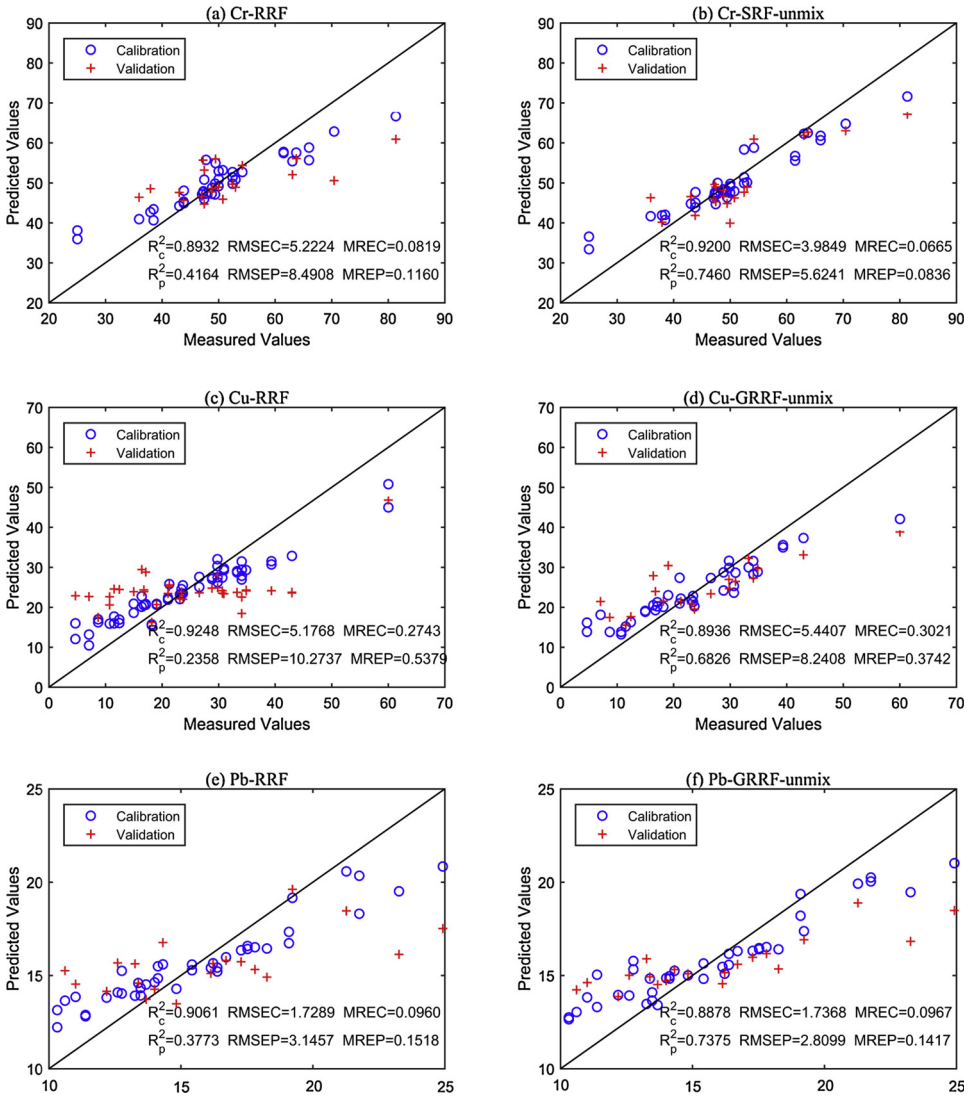

Fig. 7. Scatter plots of the measured against predicted concentrations of the best methods for Cr, Cu, Pb using the original (a, c, e) and unmixed spectra (b, d, f).

finished. Secondly, the quality of airborne imagery should be improved by spectral unmixing methods MVSA, geometric correction and radiation correction. Thirdly, the acquisition process of soil samples and the concentration of heavy metal also have the impact, such as sampling strategy and the operating process of the concentration measure of heavy metal. In this research, "S"-shaped sampling strategy is adopted in farmland and heavy metals are measured strictly in accordance with nationality standards to eliminate these effects. Last, the performance of models, including feature selection strategy, also have effect to the accuracy of the predictions.

3.2. Ten models regression analysis for field, original airborne and unmixed spectra

The heavy metals of Cr, Cu, and Pb were retrieved by PLS, RFE-SVM, LASSO, Ridge, EN, and Ridge_c, and our proposed methods. The field spectra, original airborne spectra and unmixed airborne spectra were used for the regression analysis, and the results are listed in Tables 3–5, respectively. The results of ten models using field spectra confirm that the random forest methods are more robust and can consistently offer excellent performance. The results of Cr and Pb are better than that of Cu. The random forest models perform slightly better than the others but cannot meet the retrieval requirements using original airborne spectra. The results of the retrieval unmixed pixels are significantly better than those of the original airborne spectra. The random forest models produce excellent results using the unmixed spectra for the retrieval analysis, and they also perform significantly better than other retrieval methods. The SRF model obtains the highest accuracy for Cr prediction (R_p^2 = 0.7460, $RMSE_P$ = 5.6241, MRE_P = 0.0836). The accuracy of Cr is better than that of Cu and Pb. The accuracy of the retrieval of Pb is highest (R_p^2 = 0.7375, $RMSE_P$ = 2.8099, MRE_P = 0.1417) using the GRRF method. The random forest models are more stable and extract about 10 characteristic variables, which has advantages for large-area applications. Random forest methods work better due to the variable selection strategy of the information gain and the importance score of the variable. GRRF of the random forest models has a better performance. GRRF involves using the importance score to guide the RRF, and assigns a penalty factor controlled by the importance score to each feature thus realizing the feature-variable selection process. RRF is a modification of a random forest that incorporates regularization into the tree growing algorithm. Specifically, RRF establishes a penalty for the use of a feature that was not previously used in a current tree construction. This penalty is proportional to the potential information gain from building a split on this feature, so that only features with significant information that is not redundant with respect to already built splits will be included in the model (Fig. 7).

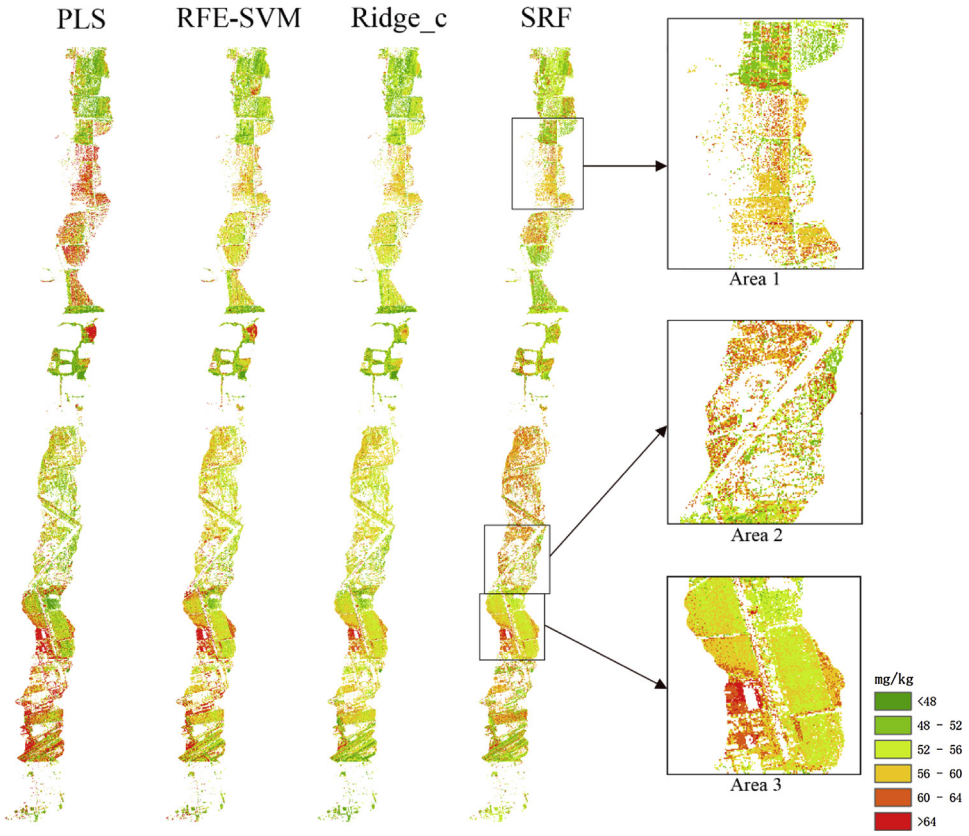

Fig. 8. Unmixed airborne image heavy metal (Cr) estimation map.

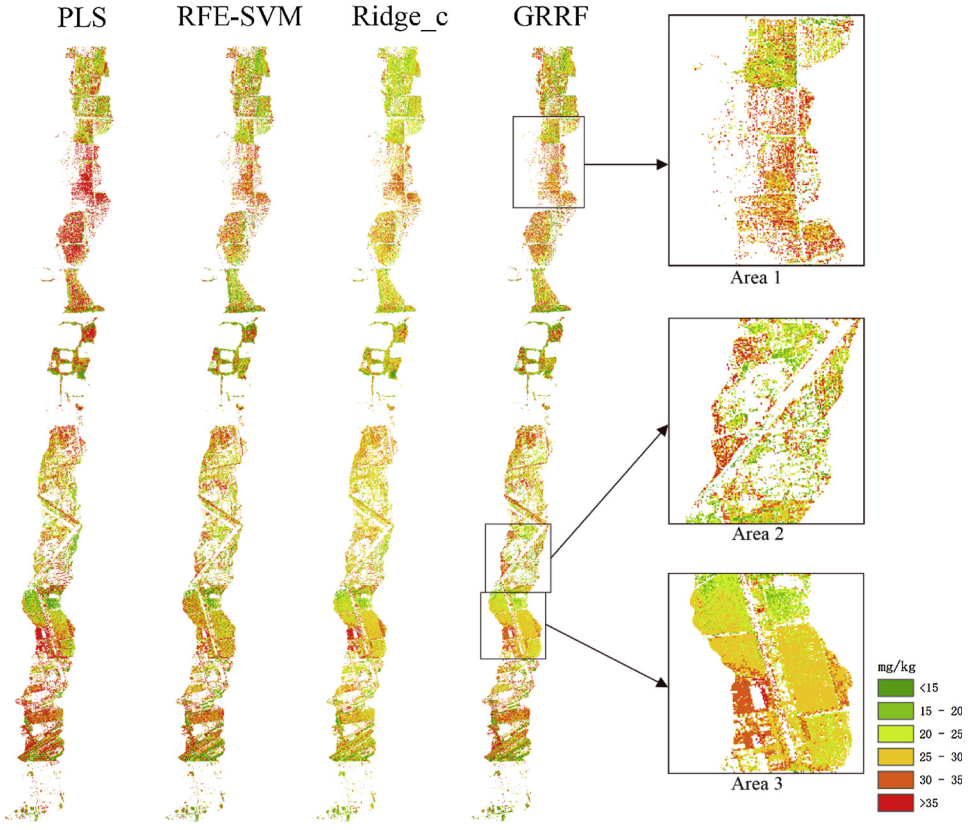

Fig. 9. Unmixed airborne image heavy metal (Cu) estimation map.

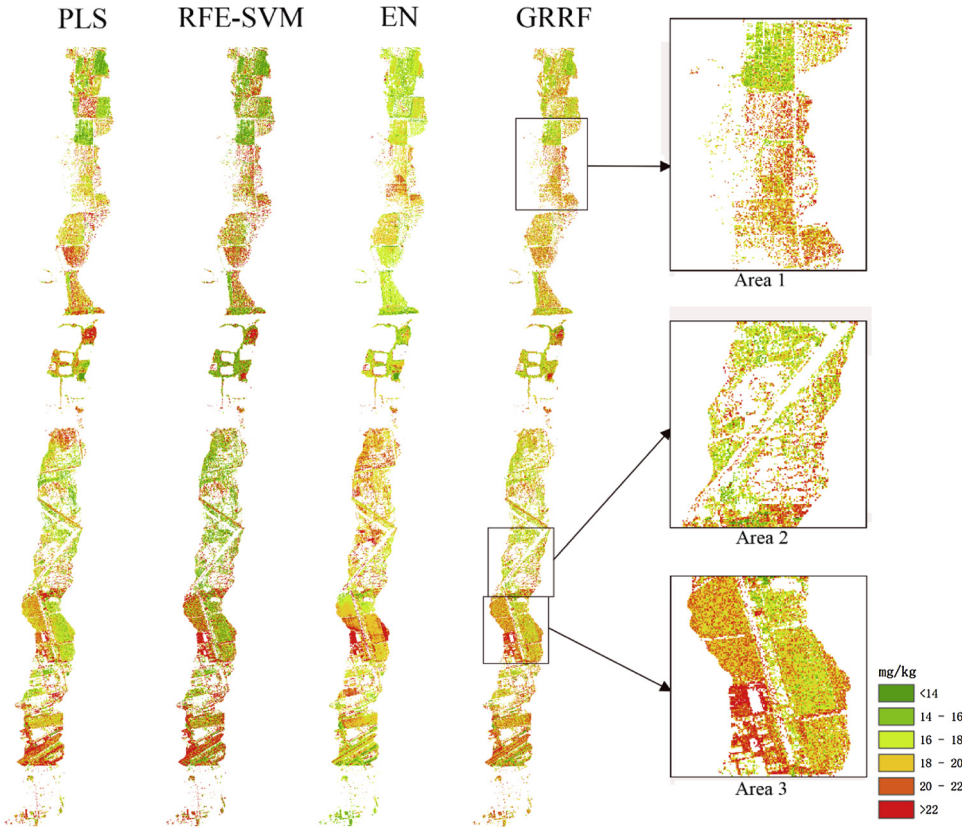

Fig. 10. Unmixed airborne image heavy metal (Pb) estimation map.

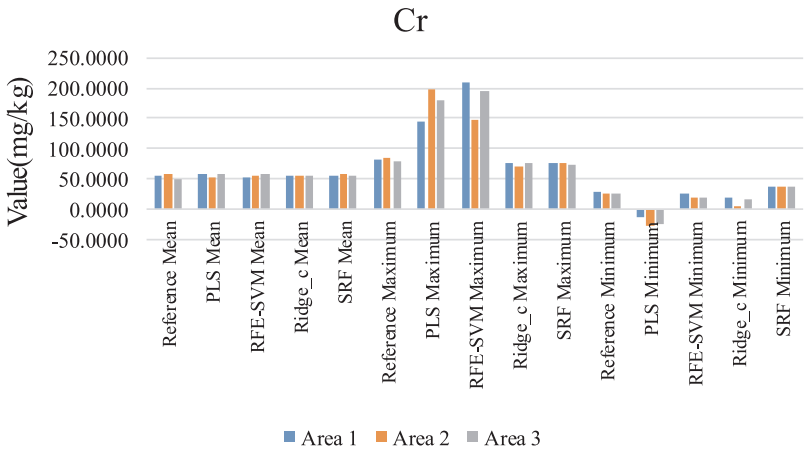

Fig. 11. Estimated and measured values of Cr concentration from the airborne image.

3.3. Heavy metal estimation of the airborne hyperspectral image

The best retrieval structures of the four models (linear: PLS; kernel function: RFE-SVM; regularization method; random forest) were extracted and applied to the heavy metal estimation of the airborne hyperspectral image. Figs. 8–10 show the distribution of three heavy metals (unit: mg/kg). The estimates of the random forest model are concentrated, with Cr between 40–70 mg/kg, Cu between 10–50 mg/kg and Pb between 5–40 mg/kg. From the estimated results of the random forest model, it can be seen that the values in Area1 and Area3 are very high. The effects of surrounding environment on soil heavy metals will be described in detail in the discussion section.

To assess the accuracy of the heavy metal estimation with the HySpex image, three areas with high estimated values were selected and compared with reference values. The results are shown in Figs. 11–13.

It can be seen from Fig. 11 that the PLS method is unstable for Cr estimation because its estimated maximum is far greater than the reference maximum, and the minimum estimate is unreasonably negative. The maximum estimate of the RFE-SVM method is higher than the reference maximum, and the estimated minimum is smaller than the reference minimum. The estimated mean, maximum of the Ridge_c model are close to the reference. The SRF model is stable, although the estimated minimum is slightly higher than the reference minimum.

As can be seen from Fig. 12, PLS and RFE-SVM are unstable, producing an estimated maximum much larger than the reference maximum, and their estimates may be negative. Although the estimated mean and estimated maximum values of Ridge_c are close to the reference values, the estimated value can also be negative. Therefore, GRRF is the best predictor in the estimation of Cu concentration.

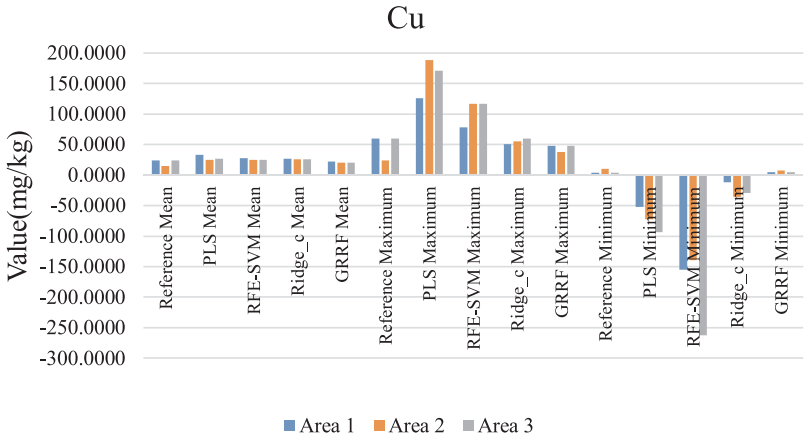

Fig. 12. Estimated and measured values of Cu concentration from the airborne image.

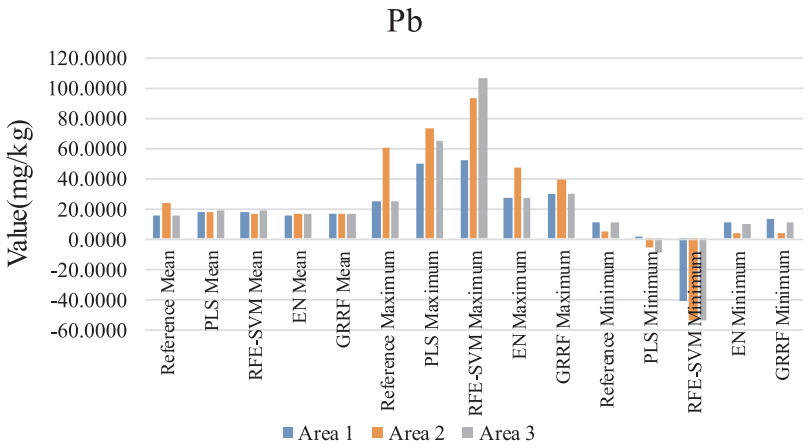

Fig. 13. Estimated and measured values of Pb concentration from the airborne image.

Fig. 13 is about the Pb concentration, which shows the estimated mean values of the four methods are close to the reference mean. The most stable method is still GRRF, because the maximum and minimum estimates using the PLS and RFE-SVM methods are far greater than the reference values and the minimum estimates are even negative.

Some minimum estimates using PLS, RFE-SVM and Ridge_c are negative. Due to the introduction of randomness of variable selection and feature selection, random forest has the ability to prevent overfitting and noise. In contrast, PLS, RFE-SVM and Ridge_c could appear overfitting, yielding negative estimated values. In conclusion, PLS, RFE-SVM and Ridge_c are not as stable as random forest.

The soil background values of three heavy metals Cr, Cu and Pb in Xuzhou are 55.5,12.61,16.3 mg/kg respectively (Nascimento and Dias, 2005). The concentration of Cr and Pb in a few areas is higher than the soil background value of Xuzhou. But in most areas, the concentration of Cu exceeded the background value.

After analyzing the estimation results with the industries around the study area, the copper industry has great influence on the concentration of Cu in the soil, which makes the Cu concentration in most areas higher than the soil background value of Xuzhou. The mining and transportation of coal and the stacking of solid waste cause heavy metal pollution in the surrounding soil. In particular, Shitun coal mine has a long working time, resulting in the concentration of the three heavy metals higher than other areas.

4. Conclusions

In this paper, the spatial distribution of heavy metal from airborne hyperspectral images based on spectral unmixing and random forest has been studied. The results show that MVSA can quickly provide the solution of unsupervised unmixing and using unmixed pixels can improve the retrieval accuracy. For Cr, with the original mixed pixels, the random forest models yielded R_p^2 values of 0.30–0.42, which are improved to 0.64–0.74 when using unmixing pixels. Comparing the retrieval results of various models, it was found that the existing methods, such as PLS, LASSO, Ridge, EN, and Ridge_c models, perform poorly, RFE-SVM performs slightly better, and the best model is random forest, which provides the most robust estimates. For the three metals using unmixing pixels, except the prediction results of SRF and GRF models slightly worse ($R_p^2 < 0.5$), the other random forest models reported R_p^2 values of 0.55–0.74. In a large area estimation of soil heavy metal concentration, the estimated values of the random forest models are the closest to the reference values. Therefore, the random forest model is a promising approach to predict the low heavy metal concentrations from airborne hyperspectral imagery. The estimated results are consistent with the real situation surveyed, which further validated the effectiveness of the method. The concentrations of Cr and Pb in a few areas are higher than the soil background value of Xuzhou. But in most areas, the concentrations of Cu are higher than background value. The copper industry has great influence on the concentration of Cu in the soil. The mining and transportation of coal and the stacking of solid waste cause heavy metal pollution in the surrounding soil. In particular, Shitun coal mine has a long working time, resulting in three heavy metals concentration higher than other areas. In this paper, few samples in the field and narrow image lead to limited coverage. Sediment contamination can also be predicted if there are the hyperspectral data and heavy metal concentration of sediment samples. In future research, more detailed plans will be made. The relationship between soil spectra

and chemical properties such as organic matter, carbon, phosphorus and potassium will be fully explored. Comprehensive utilization of spatial and spectral information will make the results more accurate.

Acknowledgment

The authors would like to thank Professors Lixin Wu, Jihong Dong and Xin Xiao in the experiments. This research was supported in part by the National National Natural Science Foundation of China (No. 41871337, 41471356).

References

Abdel-Rahman, E.M., Ahmed, F.B., Ismail, R., 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. Int. J. Remote Sens. 34, 712–728.

Adam, E., Mutanga, O., Abdel-Rahman, E.M., Ismail, R., 2014. Estimating standing biomass in papyrus Cyperus papyrus L. swamp: exploratory of in situ hyperspectral indices and random forest regression. Int. J. Remote Sens. 35, 693–714.

Ahmed, B.C., Bhadrinarayana, N.S., Anantharaman, N., Km, M.S.B., 2008. Heavy metal removal from copper smelting effluent using electrochemical cylindrical flow reactor. J. Hazard. Mater. 152, 71–78.

Bonifazi, G., Capobianco, G., Serranti, S., 2018. Asbestos containing materials detection and classification by the use of hyperspectral imaging. J. Hazard. Mater. 344, 981–993.

Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32.

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 Sens. Environ. 112, 3222–3233.

Deng, H., Runger, G., 2013. Gene selection with guided regularized random forest. Pattern Recognit. 46, 3483–3489.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Ho, T.K., 1995. Random decision forests. International Conference on Document Analysis and Recognition. IEEE Computer Society, Nancy, France, pp. 278.

Huang, J.F., Chen, D.Y., Cosh, M.H., 2009. Sub-pixel reflectance unmixing in estimating vegetation water content and dry biomass of corn and soybeans cropland using normalized difference water index (NDWI) from satellites. Int. J. Remote Sens. 30, 2075–2104.

Kim, K.I., Kwon, Y., 2010. Single-image super-resolution using sparse regression and natural image prior. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1127–1133.

Kim, Y., Kim, J., Kim, Y., 2006. Blockwise sparse regression. Stat. Sin. 16, 375–390.

Kowalski, M., 2009. Sparse regression using mixed norms. Appl. Comput. Harmon. Anal. 27, 303–324.

Li, J., Agathos, A., Zaharie, D., Bioucas-Dias, J.M., Plaza, A., Li, X., 2015. Minimum volume simplex analysis: a fast algorithm for linear hyperspectral unmixing. IEEE Trans. Geosci. Remote. Sens. 53, 5067–5082.

Nascimento, J.M.P., Dias, J.M.B., 2005. Vertex component analysis: a fast algorithm to unmix hyperspectral data. Ieee Trans. Geosci. Remote. Sens. 43, 898–910.

Pascucci, S., Belviso, C., Cavalli, R.M., Palombo, A., Pignatti, S., Santini, F., 2012. Using imaging spectroscopy to map red mud dust waste: the Podgorica Aluminum Complex case study. Remote Sens. Environ. 123, 139–154.

Peterson, S.H., Stow, D.A., 2003. Using multiple image endmember spectral mixture analysis to study chaparral regrowth in southern California. Int. J. Remote Sens. 24, 4481–4504.

Phuong, T.M., Lin, Z., Altman, R.B., 2006. Choosing SNPs using feature selection. J. Bioinform. Comput. Biol. 4, 241–257.

Qiu, L., Wang, K., Long, W., Wang, K., Hu, W., Amable, G.S., 2016. A comparative assessment of the influences of human impacts on soil Cd concentrations based on stepwise linear regression, classification and regression tree, and random forest models. PLoS One 11, e0151131.

Ren, X.S., Huo, L.J., 2010. Case study of life cycle assessment for corrugated board Box production technology. Packag. Eng. 31, 54–57.

Shi, T., Liu, H., Chen, Y., Wang, J., Wu, G., 2016. Estimation of arsenic in agricultural soils using hyperspectral vegetation indices of rice. J. Hazard. Mater. 308, 243–252.

Shi, T., Chen, Y., Liu, Y., Wu, G., 2014. Visible and near-infrared reflectance spectroscopy—an alternative for monitoring soil contamination by heavy metals. J. Hazard. Mater. 265, 166–176.

Shi, T., Liu, H., Chen, Y., Fei, T., Wang, J., Wu, G., 2017. Spectroscopic diagnosis of arsenic contamination in agricultural soils. Sensors 17, 1036.

Smith, A.M.S., Lentile, L.B., Hudak, A.T., Morgan, P., 2007. Evaluation of linear spectral unmixing and ΔNBR for predicting post-fire recovery in a North American ponderosa pine forest. Int. J. Remote Sens. 28, 5159–5166.

Sun, W., Xia, Z., 2017. Estimating soil zinc concentrations using reflectance spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 58, 126–133.

Svetnik, V., Liaw, A., Tong, C., Wang, T., 2004. Application of breiman's random Forest to modeling structure-activity relationships of pharmaceutical molecules. Multiple Classifier Systems, International Workshop, Mcs 2004. Springer Nature, Cagliari, Italy, pp. 334–343.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. Series B Stat. Methodol. 73, 273–282.

Tuia, D., Camps-Valls, G., Matasci, G., Kanevski, M., 2010. Learning relevant image features with multiple-kernel classification. IEEE Trans. Geosci. Remote. Sens. 48, 3780–3791.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., Leo, G.A.D., Torricelli, P., 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecol. Modell. 222, 1471–1478.

Wang, J., Li, X., Bai, Z., Huang, L., 2017. The effects of coal gangue and fly ash on the hydraulic properties and water content distribution in reconstructed soil profiles of coal-mined land with a high groundwater table. Hydrol. Process. 31, 687–697.

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. Environ. Pollut. 206, 227–235.

Ye, S., Zeng, G., Wu, H., Liang, J., Zhang, C., Dai, J., Xiong, W., Song, B., Wu, S., Yu, J., 2019. The effects of activated biochar addition on remediation efficiency of co-composting with contaminated wetland soil, Resources. Conserv. Recycl. 140, 278–285.

Ye, S., Zeng, G., Wu, H., Zhang, C., Dai, J., Liang, J., Yu, J., Ren, X., Yi, H., Cheng, M., 2017. Biological technologies for the remediation of co-contaminated soil. Crit. Rev. Biotechnol. 37, 1–15.

Zou, H., Hastie, T., 2005. Addendum: regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. 67, 768.