A Transfer Learning Approach Utilizing Combined Artificial Samples for Improved Robustness of Model to Estimate Heavy Metal Contamination in Soil

YAJIN WANG1, CHAO TAO2 (Member, IEEE), AND BIN ZOU2,3

1School of Geographical Science, Shanxi Normal University, Linfen 041000, China
2Key Laboratory of Metal-organic Prediction of Nonferrous Metals and Geological Environment Monitoring, School of Geosciences and Info-Physic, Central South University, Changsha 410083, China
3Chinese National Engineering Research Center for Control and Treatment of Heavy Metal Pollution, Changsha 410083, China

Corresponding author: Chao Tao (kingtaochao@csu.edu.cn)

This work was supported in part by the National Key Research and Development Projects under Grant 2018YFB0504500, in part by the National Natural Science Foundation of China under Grant 41771458, and in part by the Hunan Natural Science Foundation under Grant 2017JJ3378.

ABSTRACT Benefiting from the nanoscale sampling intervals and subtle spectral information in the visible and near-infrared band, hyperspectral technology is considered as an efficient means for monitoring soil heavy metal contamination whereby the good robustness of prediction model is driven by the increase to spectral dimension in model analysis. Considering the positive correlation between samples size and spectral dimension, we focuses on a novel derivation of enlarging samples size in this study to improve model performance by i) preparing artificial samples taking into account of flexibility and control over the laboratory environment compared with collecting wild samples, and ii) using transfer learning method called transfer component analysis (TCA) for reducing spectral feature differences caused by soil heterogeneity to train model in the same data distribution. The proposed approach was tested on three heavy metals, namely copper (Cu), cadmium (Cd) and lead (Pb), collected in the mining area located in the Xiangjiang Basin, Hunan Province, China. The experiments showed that the initial model constructed by a small number of wild samples performed strong prediction sensitivity as the training samples change. In contrast, a modified model with TCA could showed good robustness with excellent predicted ability, the average prediction accuracy of the determinable coefficient ($R^2$) and the ratio of prediction to deviation (RPD) improved to 0.73 and 1.90, 0.74 and 1.92, 0.72 and 1.73, respectively. The results illustrated there was a more reliable modeling method in potential to predict soil heavy metals based on hyperspectral analysis at low cost.

INDEX TERMS Soil heavy metal contamination, hyperspectral analysis, transfer component analysis, artificially prepared soil samples.

I. INTRODUCTION

In recent years, the frequent incidents related to excessive heavy metal content in grain have indicated the serious situation of soil contamination around the word [1]–[3]. Studies in soil remediation have confirmed the truth that heavy metals are difficult to be effectively degraded due to the complicated physiochemical property [4], [5]. Thus, the heavy metals in agricultural land are prone to accumulate for a long time, which leads to great risk in food security with the decline of soil quality [6], [7]. More important, the public nuisance diseases, in which heavy metals entering the human body through the plant chain, have presented a challenge to health care in continuous accumulation [8]–[10]. Therefore, monitoring the heavy metal concentration in pedosphere has always been the focus of current soil environmental research.

The visible and near-infrared reflectivity (VNIR) spectroscopy have been applied to the quantitative estimation of heavy metal in soil for nearly two decades [11]. In tradition, most researchers prepared experimental data through
collecting a large number of soil samples in the wild environment as the basis, and thereafter, they tended to analyze quantitative relationship between heavy metal concentration and spectral reflectance relied on statistical model in the study; ultimately, the constructed model was expected to make valid prediction in real environment for practical application [12], [13]. There were plenty of research scholars having done detailed investigation on this. For example, Choe [14] applied the stepwise multiple linear regression (SMLR) model to predict respectively the content of heavy metals (i.e., lead (Pb), zinc (Zn), arsenic (As)) in Rodalquilar mining area located in Spain, and this work also showed the feasibility of mapping the heavy metal distribution combining with hyperspectral remote sensing. Likewise, Sun [15] presented a combination of partial least squares regression and genetic algorithm (GA-PLSR) to predict soil nickel (Ni) concentration in two different regions located in Hunan Province, China, and the spectral bands in connection with organic matter and clay minerals were deemed to be more effective in prediction model compared with entire VNIR spectroscopy. Moreover, Malmir et al. [16] extracted the data from hyperspectral imagine to develop partial least squares regression (PLSR) models to predict micro-elements in sieved and ground air-dried soils, and this study revealed that copper (Cu), magnesium (Mg), Zn associated with soil carbon and nitrogen could be successfully predicted with excellent accuracy.

Although aforesaid researches have shown satisfactory predictability to different heavy meals nowadays, most of them adopted statistical models such as PLSR, SMLR and SVR as the main approach to make quantitative predictions [17]. In this case, one fact that needs to be noted is that statistical models tend to perform good prediction only when the training sample data is a good representative of test samples in the process to model [18]–[20]. As a result, numerous soil samples must be prepared to ensure the reliability and robustness of the experimental results at the beginning of model construction. However, this will not embody the superiority of visible and near-infrared spectroscopy in the matter of minoring heavy metal contamination in soil. On the one hand, collecting substantial soil samples in the field will increase capital and time costs, and this may well occur similar predicament compared with inchoate geospatial interpolation technology in early research. On the other hand, the reflectance spectrum influenced by field environment will be interfered sharply by multifarious components in soil, which may lead to difficult feature extraction in reflectivity signal of heavy metals [21]. Furthermore, some studies attempted to demonstrate the model robustness on the condition of small sample data; nevertheless, these statistical models usually showed neither favorable robustness ability nor high prediction accuracy in this context [22], [23]. In addition, models trained with small sample data were likely to manifest sensitivity to the test samples, because the insufficient sample data was apt to a partial distribution in training set. To break through the limitation of existing method, this paper would put forward to artificially prepare soil samples contaminated by heavy metals to spike train samples with a view for expanding experimental data set at low cost.

Spiked samples were prepared artificially to increase the amount of data in experiment, which was proved to be feasible in different soil element analyses [24], [25]. A good case in point was soil organic carbon investigation in which large national soil spectral libraries were used to increase sample density in regional scale calibrations [26], [27]. In the process of model training, spiking calibration samples from libraries caused a decrease in bias, which improved effectively prediction accuracy compared to those constructed with small sample data. However, this mode is rarely adopted in the field of soil heavy metal assessment. The prime reason is that there is a significant spectral reflectance difference between the artificial samples in the laboratory and natural samples in the wild. To this issue, we proposed a transfer learning method called transfer component analysis (TCA) to reduce spectral differences. For one thing, transfer learning is an important branch in the realm of machine learning, which shows good applicability in recognize knowledge form previous domain to a new one. This is similar to how to learn the spectral characteristics commonality between the wild and laboratory environment. For another, transfer component analysis, a classic method of transfer learning, mainly lies in projecting data from diverse domains to a same feature space, which will facilitate the original data sets with different characteristics submit the identical distribution as much as possible in the same dimensional space. So long as both of soil spectrum from wild and laboratory show similarities, artificially prepared samples could be spiked in the training set consisted of wild samples to enhance model robustness. Therefore, this article takes the mining area located in Xiangjiang River Basin, Hunan Province, China as a case analysis, and regards three different heavy metals included cadmium (Cd), copper (Cu), and lead (Pb) as research examples. It aims to: (i) analyze the sensitivity of statistical models to test sets trained by small sample data from wild collection; (ii) compare and contrast the data on soil spectral reflectance in the laboratory and field environment, and reduce the characteristic differences through TCA transformation; (iii) use artificially prepared sample data to spike wild samples to train model after spectral transformation with TCA method; (iv) compare and contrast the prediction accuracy and robustness of models constructed with different training samples to provide a more reliable modeling approach in practical environment.

II. DATA AND METHOD
A. ARTIFICIAL PREPARATION OF SOIL SAMPLES CONTAMINATED WITH HEAVY METAL
In this article, we provided three different types of heavy metal-spiked soil samples, namely cadmium, copper and lead. They were divided into two groups of experiments according to the differences in background soil sources. One
group included cadmium, and the other group covered copper and lead.

To prepare Cd-spiked soil samples, there was around 30kg clear background soil collected in Hengyang City, Hunan Province, China. After the process of stones eliminating, air drying, fine girding and using 100-mesh polyethylene to sieve, the processed background soil was artificially contaminated by heavy metal standard solution. More specifically, the collected soil was equally divided into 66 parts in which each of them was 100g. One was used to examine the initial Cd concentration in the background soil, and the remaining sixty-five parts were sequentially added with the Cd standard solution under the guidance of designed concentration value (GSBG 62040-90(4801), Concentration: 1000 ug/L, Medium: 10% HCl). The operation of experiment was on the basic of work directive of reference materials (3) (GB/T 15000.3-2008/ISO Guide 35:2006) promulgated by the People’s Republic of China. Only the concentration of Cd was artificially adjusted in the samples, and the other components (e.g. pH, total iron and organic matter) in the soil were controlled equally. An overview of design content values of Cd-spiked samples was shown in Table 1.

In the design of artificially prepared soil samples, the spiked cadmium concentration referred to (1)

\[ C = 10^{-3} \times (A - B)/100 \times \rho \]  

where \( C \) represents the standard solution volume (mL) added to artificially prepared soil samples, \( A \) represents the design content value in Cd-spiked soil samples (mg/kg), \( B \) represents the initial Cd concentration in the collected background soil, and \( \rho \) represents Cd standard solution concentration (In our experimental preparation, \( \rho = 100 \text{ug/L} \)).

Different with Cd-spiked samples, the background soil spiked by heavy metal Cu and Pb derived from Changsha City, Hunan Province, China. In contrast, background soil types in both cities were dominated by latosol and red soil according to Chinese Soil Taxonomy (GBT 17296-2009). In the artificial preparation as for spiked samples of Cu and Pb, the concentration design principle was referred the work directive of reference materials (3) (GB/T 15000.3-2008/ISO Guide 35:2006). Similarly, the preparation process was submitted to the pattern mentioned above.

A closer look at the table 1 highlighted the fact that the work collected soil samples in the field was independent of the preparation of artificial heavy metal-spiked samples, which meant sample interpolation could be implemented in theory whereas the concentration distribution of heavy metal in samples was indeterminate in the wild. In addition, the background soil varied with the type of heavy metal in artificial preparation, which provided the experimental data on exploration whether it could break through regional limitations in the artificial sample preparation.

**B. OVERVIEW OF THE STUDY AREA**

The research area in this paper is located in Xiangjiang River Basin, Hunan Province, China. Hunan province is in the front rank of non-ferrous metal production with a long mining history, and there are a mass of mining areas in the Xiangjiang River Basin to support economic development [28]–[30]. Due to the significant role of Xiangjiang River in agricultural production and economic construction, the heavy metal contamination in this basin needs to be paid more attention for the sake of river ecosystem health.
A total of 40 samples were collected in a lead-zinc mining area as data set in the wild. The sampling range was from 112°35'24" E to 113°36'37" E and 26°32'37" N to 25°34'12" N. According to Chinese soil taxonomic criteria (GBT 17296-2009), red soil and latosol were the main soil types in the region. It should be noted that the sample soil collected in the field derived from the Hengyang city, which was identical region to Cd-spiked soil, but was different from the background soil of Cu-spiked and Pb-spiked artificial samples.

C. SPECTRAL DATA PREPROCESSING

First of all, the collected samples were removed impurities such as debris and plant residues after air-drying. Second, they were ground in a ceramic bowl and passed through a 100-mesh sieve. At last, each processed sample was divided into two parts on average, which were used for spectral measurement and heavy metal concentration determination in the laboratory, respectively. In particular, the spectrometer PSR-3500 (spectrum range: 350 to 2500 nm) was used to measure the soil spectral reflectance for hyperspectral data. Due to the darkroom environment in the laboratory, a 1000W halogen lamp was provided as a test source to secure bare fiber. The illumination direction of the light source was at an angle of 15° to the vertical direction, the distance of the light source was set to 30 cm, and the probe was kept at 45° to the plane of the soil sample. Before the measurement, the standard whiteboard was calibrated and optimized in the first, and each soil sample was repeatedly measured 5 times, then the arithmetic mean value was taken as the actual reflection spectrum data of the soil sample. The spectral resolution was 1.5 nm at 350-1000 nm, 3.8 nm at 1000-1900 nm, and 2.5 nm at 1900-2500 nm. Ultimately, there were a total of 1024 band values in soil spectrum measurement. Meanwhile, the graphite furnace atomic absorption spectrophotometry was adopted to determine the concentrations of cadmium and lead in soil (GBT 17141-1997), and the detection limits in which it was calculated by digesting 0.5g sample to a fixed volume to 50 ml were 0.01 mg/kg and 0.1 mg/kg in the standard specification, respectively. Furthermore, the concentration determination of copper was based on flame atomic adsorption spectrophotometry (GB/T 17138-1997), and the detection limit in which the calculation method was 1 mg/kg as well as before.

The soil samples spiked with three different heavy metals in the laboratory were compared with those in the wild concerning spectral reflectance. It was anticipated that spectral curves tended to be similar in the artificially prepared soil...
samples whereas there were significant differences of spectrum in the wild samples measured in the laboratory. In order to improve the signal-to-noise ratio, first-order difference was put into use to heighten potential response characteristics of heavy metals in soil spectrum. It could be clearly seen that the spectrum was enhanced in the wavelength of 600 nm, 700 nm, 900 nm and 2200 nm approximately. These bands coincided with the spectral response intervals of total iron, clay minerals and organic matter, and might contain potential heavy metal signatures due to significant correlation in the statistical.

Although artificially prepared samples behaved similarly in spectral reflectance, there were obvious discrepancy as compared to the spectra from wild samples. For instance, the number of band peaks was not the same in the wavelength range of 500nm to 900nm. In general, soil spectrum from different sources might vary in characteristic distribution, which would lead to application difficulty for artificially prepared samples in model training. To this end, this paper proposed the method of transfer component analysis for another spectral prepossessing.

D. TRANSFER COMPONENT ANALYSIS

The target of transfer learning is to recognize and apply the knowledge or patterns learned in a previous domain to different but related domains [31]. In particular, transfer component analysis (TCA) is a classic method of transfer learning proposed in recent years [32], [33]. The main purpose of using TCA in this paper is to reduce the difference in spectrum characteristics from two domains, and wild samples and artificially prepared samples are defined as source and target domains, respectively. The principle of TCA is to project the spectral reflectance data of source domain and target domain to the Reproducing Kernel Hilbert Space (RKHS), and minimize the Maximum Mean Discrepancy (MMD) as relevant criterion for estimating the distance between distributions. The mathematical process is explained as follows:

Assuming that the data set from source domain is defined as \( X_S = \{x_{S1}, \ldots, x_{Sn}\} \), where \( x_{Si} \in \chi \) corresponds to the spectral reflectance in wild samples, and \( n_1 \) represents the sample number in the data set; similarly, the data set related to target domain is defined as \( X_T = \{x_{T1}, \ldots, x_{Tn}\} \), where \( x_{Ti} \in \chi \) corresponds to the spectral reflectance in artificially prepared samples, and \( n_2 \) represents a certain heavy metal-spiked sample number in artificially preparation. Meanwhile, \( P(X_S) \) and \( Q(X_T) \) (or \( P \) and \( Q \) for short) are defined as the marginal probability distribution of \( X_S \) and \( X_T \), respectively. In the case where the data characteristics of the two domains are inconsistent, there is usually \( P \neq Q \). The nearest point in our experiment is to make \( P \) as similar as possible to \( Q \). Borgwardt et al. [34] proposed the distance between two distributions \( P \) and \( Q \) could be measured by the squared distance between the empirical means of the two domains, and it was defined as MMD to estimate the distance of domains mapped into a RKHS. The calculation formula is as follow:

\[
\text{Dist}(X_S, X_T) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \Phi(x_{Si}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \Phi(x_{Ti}) \right\|^2_H
\]  

(2)

where \( H \) represents the feature space of RKHS, and \( \Phi : \chi \rightarrow H \).

Due to the difficulty in finding nonlinear mapping \( \Phi \), the TCA proposed by Pan et al. [35], Pan and Yang [31] transformed this problem into a nuclear learning problem by defining a kernel function \( k(x_i, x_j) = \Phi(x_i)\Phi(x_j) \). Hence, Equation (2) can be reformulated as:

\[
\text{Dist}(X_S, X_T) = tr(KL) \tag{3}
\]

where \( K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix} \) is a \((n_1 + n_2) \times (n_1 + n_2)\) kernel matrix, and \( K_{S,S}, K_{S,T}, K_{T,S}, K_{T,T} \) are defined by virtue of the kernel function \( k(\cdot, \cdot) \), respectively. Furthermore, \( L = [L_{ij}] \geq 0 \) with

\[
L_{ij} = \begin{cases} 
\frac{1}{n_1}, & x_i, x_j \in X_S \\
\frac{1}{n_2}, & x_i, x_j \in X_T \\
-\frac{1}{n_1n_2}, & \text{otherwise}
\end{cases}
\]

Instead of learning \( k(\cdot, \cdot) \), this problem can be well addressed through learning kernel matrix \( K \), which is considered a semi – definite program (SDP) in the end [31]; then, Principal Component Analysis (PCA) is used on the learned kernel matrix \( K \) to acquire a low-dimensional latent space across domains. To sum up, TCA method could reduce the spectral differences from two domains in low dimension space and retain the original spectra characteristic as much as possible. In this paper, spectral dimension form source and target domain dropped from 1024 to around 100 dimensions in three different heavy metal spectra transform. One further point to note was that the decrease of the spectral dimension avoided the redundancy of high-dimensional data in the model calibration.

E. MODEL CALIBRATION AND EVALUATION

Benefiting from the advantages of principal component analysis and canonical correlation analysis [36], partial least squares regression (PLSR) coupled with the latent variables (LVs) number determined by cross-validation was applied for model calibration. We developed three types of models for different purpose mentioned in introduction section.

(I) Initial model: To verify the robustness of model contrasted with small amounts of samples, we set up three kinds of initial models based on samples collected in the wild: Cd model, Cu model and Pb model. For each heavy metal initial model, we carried out 50 experiments through changing randomly sample order of whole datasets in order to avoid the contigency in the sample selection. In every experiment, one third of samples as calibration set were used to build model, one-third samples were as validation set to determine
the optimal number of LVs by leave-one-out cross-validation for avoiding over-fitting, and residual one-third samples were as test set to evaluate the prediction accuracy of initial models.

(II) Modified model without TCA: To explore the effect of spiked samples without TCA spectral transformation, we set this model as a control to allow calibration set added with artificially prepared samples. There are also three kinds of heavy metal-spiked experimental groups, namely Cd modified model without TCA, Cu modified model without TCA and Pb modified model without TCA. In each group, a total of 50 experiments in which validation set and test set were consistent with the former were carried out in comparison with initial models.

(III) Modified model with TCA: To study the feasibility of TCA spectral transformation between field and artificially prepared samples, we adopted the spectral dataset of reducing dimensionality via TCA method to train model. To compare and contrast, the calibration set, validation set and test set remained the same with Modified model without TCA in three heavy metal experiments, except TCA spectral preprocessing.

The evaluation indexes of model accuracy are the coefficient of determination (R²) (4), the root mean square error (RMSE) (5), and the ratio of prediction to deviation (RPD) (6), respectively. To be specific, R² illuminates the correlative degree of predicted values and measured values, if there is a high level of correlation between them, the calculated value of R² is closer to 1. RMSE represents the deviation between predicted and measured values, and a small value of RMSE indicates the small prediction error variability due to sensitivity to outliers. The RPD corresponds to the ratio between standard deviation (S.D) and RMSR, which was used to evaluate the reliability of prediction in the external test.

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

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

\[ RPD = \frac{S.D}{RMSE} \]

where \( y_i \) = measured value, \( \hat{y}_i \) = predicted value, \( \bar{y} \) = mean measured value, and \( n \) = the number of sample in the test set with \( i = 1, 2, \ldots, n \).

According to the model accuracy assessment theory proposed by Williams [37] and Saeys [38], the criteria to evaluated the model prediction ability were as follows: the values of R² and RPD are great than 0.9 and 3.0 represents the model was excellent in prediction; good prediction is defined by R² value at 0.82 to 0.9 and RPD value at 2.5 to 3.0; approximate prediction is with R² value at 0.66 to 0.81 and RPD values at 2.0 to 2.5; the possibility to distinguish high and low values is identified with R² value at 0.50 to 0.65 and RPD values at 1.5 to 2.0; meanwhile, not only is the value of R² greater than 0.5, but also the RPD value is more than 1.5 is recognized as a successful model. A model with good robustness is supposed to meet high R² and RPD values, and low RMSE values in prediction accuracy. All model construction and accuracy evaluation were performed in MATLAB 2014a.

### III. RESULTS AND DISCUSSION

#### A. CONTAMINATION ANALYSIS OF SOIL SAMPLE

To investigate the model performance under the condition of small data, a total of 40 samples were collected in the wild for experimental data preparation. Two of them were eliminated due to the spectral noise interference in the measurement, and there were 38 remained to model analysis in the following experimental research. According to the soil environmental quality risk control standard for soil contamination of agricultural land promulgated by the Chinese government (GB15618-2018), the risk screening value was identified as maximum concentration of heavy metal that threaten crop growth and soil ecological environment. For heavy metal Cd, Cu and Pb, they were 0.3 mg/kg, 50 mg/kg and 70 mg/kg in acidic soil, respectively. The given data in Table 2 indicated that the heavy metal contamination in case area was a serious challenge, because the mean concentrations of three heavy metals exceeded the defined risk screening values to a large extent.

Table 2 revealed data on three kinds of sample preparation, namely Wild Sample (WS), Artificially Prepared Samples (APS), and Wild Samples plus Artificially Prepared Samples (WS+APS). As the heavy metal-spiked samples (i.e. Artificially Prepared Samples) was not prepared in the premise of knowing field sample concentration distribution, maximum values of heavy metal in WS and APS performed inaccuracy as well as minimum values. One further point to note

| Sample source                  | Heavy metal | N  | Min  | Max  | Mean | Std |
|-------------------------------|-------------|----|------|------|------|-----|
| Wild Samples (WS)             | Cd          | 38 | 0.7  | 56.9 | 13.3 | 14.4|
|                              | Ca          | 38 | 14.4 | 725.0| 156.2| 179.5|
|                              | Pb          | 38 | 41.3 | 6174.8| 1594.4| 1816.1|
| Artificially Prepared Samples (APS) | Cd      | 65 | 0.63 | 58.95| 25.50| 19.11|
|                              | Cu          | 39 | 43.2 | 573.1| 302.64| 130.08|
|                              | Pb          | 40 | 559.8| 8081.5| 4040.19| 1886.07|
| Wild Samples plus Artificially Prepared Samples (WS+APS) | Cd       | 103| 0.63 | 58.95| 21.00| 18.41|
|                              | Cu          | 77 | 14.4 | 725.0| 230.42| 171.99|
|                              | Pb          | 78 | 41.3 | 8081.5| 2848.66| 2213.70|

Note: N refers to the number of samples, Min refers to the minimum concentration in samples, Max refers to the maximum concentration in samples, Mean refers to the mean concentration of samples, and Std refers to standard deviation in samples.
was that the values of standard deviation in samples were high whereas they differed in three types of sample data. This might be related to two reasons: one was that the heavy metal concentration from field samples did not behave normal distribution usually, which varied a lot due to complicated contamination cases; the other was that the chief purpose of AFS was to spike sample data from field environment for increases in data size. In theory, there was no requirement that the spiked samples (i.e. Artificial Preparation Samples) themselves followed a normal distribution.

Figure 3 detailed the differences on three heavy metal concentration distribution, there were two conclusions drawn from it: (1) In the wild samples, most of concentration values gathered the range of low pollution. Although a few samples distributed in the scope of high concentration, the number of samples was insufficient to model in reality; (2) The role of artificial preparation samples was to increase samples in training set at high concentration to improve model learning capability. Accordingly, three types of model would be explored whether they behaved robustness or not, namely initial model, modified model without TCA, and modified model with TCA.

B. PREDICTIVE ANALYSIS OF THREE TYPES OF MODEL FOR HEAVY METAL Cd

As the background soil to prepare Cd-spiked samples was from the same region with samples collected in the wild, heavy metal Cd was first studied as experimental object, and PLSR as statistical model was trained to predict the content of heavy metal in the test set. To compare with modified models with TCA in later, the spectral reflectance after first-order transformation within all bands was used as independent variable (X) in initial model and modified model without TCA, and the content value of heavy metal after logarithmic transformation was regarded as dependent variable (Y) in order that the standard deviation of three heavy metals in different soil samples was large in Table 2. Meanwhile, we set validation set resembled machine learning to determine the optimal parameters (i.e. the optimal number of LVs) of PLSR to prevent out-fitting. For every type of model, we carried out 50 experiments in case of occasional result, with each experiment, different samples were randomly selected as training set, validation set and test set to build the model and measure predictive performance. In particular, the standard deviations in 50 accuracy results were calculated to evaluate the robustness for reference in three models. Table 3 revealed the data on average prediction results in 50 random experiments.

| Model category | Training set | Validation set | Test set | $R^2$ (mean±std) | RPD (mean±std) | RMSE (mean±std) |
|----------------|--------------|----------------|----------|------------------|----------------|------------------|
| Initial model | 13           | 13             | 12       | 0.45±0.04        | 1.16±0.10      | 0.99±0.14        |
| Modified model without TCA | 78           | 13             | 12       | 0.43±0.04        | 1.12±0.08      | 1.00±0.09        |
| Modified model with TCA | 78           | 13             | 12       | 0.73±0.01        | 1.90±0.09      | 0.51±0.01        |

A closer look at the table highlighted the fact that the accuracy performance of modified model with TCA was superior to the other two, while the average values of $R^2$ and RPD were less than 0.5 and 1.5, respectively. This indicated the first two models were unsuccessful in predict performance of test set, and in addition, the value of RMSE in modified model with TCA was nearly half of those in initial model and modified model without TCA, and it admitted the superiority in evade predicted deviation in tertiary model. As far as model robustness was concerned, an overview of the table showed that the standard deviations tended to decrease gradually in terms of $R^2$, RPD and RMSE, which, on the other hand, meant that the smaller the value was in average prediction results, the more stable the predict model was in robustness. More specifically, in 50 random experiments, the predict performance of model should be consistent in accordance with different data set.
According to the accuracy factors to assess whether the model had a successful prediction ability, it was clear from the pie charts in Figure 4 that the greatest predictive success rate was in modified model with TCA, with exactly 90%, whereas less than 20% of successful prediction rate was in initial model and modified model without TCA, which showed a majority of Cd predicted results came from low accuracy defined as unsuccessful prediction. There were three premier factors that could account for such phenomenon. First and foremost, when only thirteen samples were trained to model in initial model, our experimental results showed that it was hard to guarantee the training set could learn effectively spectral features with samples variation in test set. Moreover, although the number of training samples increased to 78 in the modified model without TCA, the spectrum characteristics from training set could not represent those in test set on the condition of direct supplement by artificial preparation samples. Last but not least, since both of spectrum between training and test set had been transformed via TCA, the model performed a significant improvement of predictive results, which implied the feasibility of modified model with TCA and deserved to study further via different heavy metals in various ways.

C. PREDICTIVE ANALYSIS OF THREE TYPES OF MODEL FOR HEAVY METAL Cu AND Pb
Considering the practical application, we hope that the background soil collection of artificial samples is more convenient, which does not necessarily come from a specific research area. To enhance the feasibility of research, we prepared heavy meal Cu-spiked and Pb-spiked samples on the basis of background soil from Changsha city. The twofold purposes of this section could attribute to: (1) verify the three types of model discussed above whether performed consistently in the case of different heavy metals; (2) explore that the background soil of artificial spiked samples whether could derive from different area compared with wild samples. To this end, according to the random experiment instructions, we estimated soil Cu and Pb concentrations in three types of model followed a very similar pattern.

Table 4 gave the average prediction results of Cu and Pb in 50 random experiments, compared to statistical results of heavy metal Cd listed in Table 3, there were three similar experimental phenomenon that could be summarized: (1) Unlike the low-precision performance of the first two models, the values of $R^2$ and RPD increased dramatically in the modified model with TCA; (2) in contrast, there was a rapid decrease in the value of RMSE in the third model; for instance, the value of RMSE had dropped from 1.09 to 0.54 in the prediction of heavy metal Cu; (3) in 50 random experiments, standard deviation took a gradual downward trend in terms of $R^2$ and RMSE in three types of model. From what have been discussed above, we might reasonably come to the conclusion that only the modified model with TCA could make the valid prediction with higher accuracy. Otherwise, even if quite a few artificial prepared samples we spiked in the training set, it could still by no means show feasibility and robustness in prediction for test set.

The pie charts shown in Figure 5 gave the information about the prediction success rate in 50 random experiments in three models. According to the sample data we collected in the wild real environment, which had been explained in detail in Table 2, it was considerable large in standard deviation as regards the concentration values of heavy meal Cu and Pb, which, indeed, exacerbated the demerit of modeling with small samples. More specifically, observing the prediction statistics about initial model in the Fig. 5(a) and Fig. 5(d), there was a sheer impossibility to model successfully based on thirteen training samples in the prediction of Cu and Pb, while a slight rise of prediction success rate took place in the modified model without TCA. However, the modified model with TCA witnessed a dramatic increase resembled statistical result of Cd experiment. It was clear from the pie chart in the Fig. 5(c) that a majority of statistical results came from

![Pie chart statistics for Cd predicted results in 50 random experiments.](image-url)
successful prediction, with exactly 94%; similarly, for heavy metal Pb, 84% of successful rate contributed to the prediction results.

From the factors mentioned above, we might safely draw conclusions on three aspects: first, the small sample data collected under natural environment, in which the training samples could by no means represent the data feature in test set, was difficult to construct the prediction model in an efficient way. With the change in training samples, our experiments showed the initial model had a significant variation as for predictive capability, and was virtually impossible to predict extremum such as high concentration of Pb in wild samples. This work demonstrated that the estimation of heavy metal in soil was not appropriate in the condition of small sample data, it was necessary to supplement a considerable number of samples in training set. Second, if the artificially prepared heavy metal samples were directly spiked to the training set, there was less possibility for improvement in predicted accuracy. Because of the obvious spectral reflectance differences of both, the data features learned from artificial samples was not compatible with test samples in the wild. Therefore, in order to allow artificial samples to effectively participate in model training, spectral transformation must be performed to make the spectral reflectance from artificial and field samples as similar as possible. Third, TCA had a good performance in soil spectrum conversion. Our experiments showed that after the TCA transformation, artificially prepared samples could be spiked into the training set, and improve dramatically the predicted accuracy of model. Compared to the initial model and modified model without TCA, the modified model with TCA achieved high-precision prediction with good robustness, in spite of random variation of test samples. Since the above conclusions were based on the average prediction results in 50 random experiments, we preferred to provide more convincing experimental analysis to support the viewpoints in detail.

**D. SENSITIVITY ANALYSIS OF INITIAL MODEL TO TRAINING SAMPLES**

In the case of a small sample of data, how much did the training sample affect the predictive capacity of the model? We discussed it with two groups of experiments by virtue of samples collected in the wild. The first set of experiments was similar to what mentioned in section B, samples were randomly selected by means of the randperm function on the MATLAB. In the second experiment, the collected samples were sorted according to the laboratory concentration analysis, and the selected training samples were as consistent as possible with the test set in the concentration range. Moreover, the above two experiments were also conducted 50 random tests to count the best and worst predictions, accompanied by the changes of training and test samples.

Table 5 compared and contrasted the data on the worst and best prediction results in 50 experiments based on different training and test samples for estimating the content of three heavy metals. An interesting phenomenon was that whatever way the training samples was selected to model, the constructed model had similar prediction performance for the test set. For example, the values of accuracy evaluation factors, namely $R^2$, RPD and RMSR, had fluctuated slightly around the level of 0.7, 1.7 and 0.7 in the best prediction as for heavy metal Cd. In addition, one further point deserved to be noted was that there were significant differences between the worst and best prediction result. In the case for predicting the concentration of Cd, the difference values of $R^2$, RPD and RMSR...
and RMSR shown in the two sample selection approaches were around 0.7, 0.8 and 0.5, respectively. Although this situation can be changed to a certain extent in the improvement of statistical models, the negative effects caused by a small number of training samples could not be ignored in a real environment. In another word, due to the complexity of pollution causes in the natural environment, there were large concentration differences inescapably in the collected wiled samples, which, on the other hand, implied the a satisfied accuracy could not appear in the prediction for test set, unless a huge number of training samples were applied to model in the premise of same feature distribution.

E. ROBUSTNESS ANALYSIS OF THE MODIFIED MODEL WITH TCA

Huber had systematically given the robustness concept in three levels from the perspective of robust statistics [39]. In the first place, the model was supposed to show a higher accuracy or a better effectiveness than that in general, and in our research, a robustness model was recognized with a good predictive ability according to the appraised criterion. In the second place, assuming that a small disturbance such as the spectral noise appeared ineluctably in measurement, and theoretically, there were small influence in algorithm performance of model. As far as predictive model of heavy metal was concerned, the spectral differences between wild samples and artificially prepared samples could by no means effect the accuracy result in test set. In the last, considering large deviation in model assumption, for instance, outliers, it would not have a catastrophic impact in model prediction; to be specific, in each random experiment, high concentration values of heavy metal in the training set could not change the prediction accuracy of the model to a large extent.

The charts below gave information about accuracy assessment factors distribution in 50 random experiments. According to the evaluation criteria, if the values of $R^2$ and RPD were respectively less than 0.5 and 1.5, the model prediction was deemed to be unsuccessful in our experiments. In the bar charts shown in Figure 5, the values of $R^2$, as well as RPD, had similar distribution in initial model and modified model without TCA. Specifically, among the predicted results of the three heavy metals, the values of $R^2$ were mostly concentrated in the range of less than 0.5, and the values of RPD almost distributed in the range of less than 1.5. Such low accuracy prediction performance indicated the two models might lacked good robustness in term of the first level mentioned above.

In contrast, the modified model with TCA showed a higher accuracy or a better effectiveness in the case of result statistic. In particular, the values of $R^2$ were chiefly distributed in the effective rang of 0.66 to 0.81; likewise, most RPD values were concentrated in the range of 1.5 to 2.0 and 2.0 to 2.5. When the modified model with TCA compared with the modified model without TCA, an important distinction was that the robustness of model could not be affect sharply due to the spectral difference in training set, which, from the second level mentioned above in the first paragraph, confirmed the robustness of the modified model with TCA.
However, whether it had better capability in the prediction of outliers deserved further discussion. In this section, the scatter plots below showed the best accuracy result in 50 random experiments for three models.

Figure 7 demonstrated the best prediction in three types of model as for three different heavy metals, respectively. Refer to the third level definition mentioned earlier, model with good robustness ought to be insensitivity along with parameter perturbation caused by training sample change. In reality, the intervention of outliers usually gave rise to large fluctuations in prediction result. Since the experimental sample data used in this article, heavy metals content varied greatly in the high concentration range, the model especially needed to pay attention to the prediction effect of high concentration values. It was clear from the scatter plots above that the modified model with TCA was more superior than the initial model in the prediction of high concentration values. Specifically speaking, the predicted points were closer to the regression line in high concentration interval, so it was anticipated that the values of RMSE was minimum compared to the other two models, which indicated the prediction capability of the modified model with TCA would not be affected by abnormal values such as high concentration to a certain extent. In particular, refer to the evaluation criteria of model prediction capability, the modified model with TCA had excellent prediction in the experiment of heavy metal Cu, with the values of $R^2$ and RPD exactly 0.93 and 3.23, whereas $R^2$ and RPD values were respectively around 0.5 and 1.5 in the best accuracy results of another two models. In practical application, the model with good robustness would enhance the reliability in model predicted result. The modified model based on TCA spectral transformation was proposed in this paper through experimental exploration, which made up the insufficient of sample collected in the natural environment by means of the supplement of artificially prepared samples in the laboratory. The robustness of model was demonstrated by the improved prediction accuracy in the case of low cost. In the further study, it will be more fully verified whether the artificially prepared samples can be applied in a wider area.

IV. CONCLUSION
Considering the high sampling cost in the natural environment, studies on soil heavy metal contamination were usually based on small sample data, which made questionable in the...
matter of model predictive accuracy and robustness. In this paper, three kinds of heavy metals, namely cadmium, copper and lead, were explored experimentally to draw a conclusion: the model trained by a small number of samples did not have good prediction accuracy and robustness. In this case, artificially prepared samples in the laboratory were proposed to enhance robustness in the condition that the training set was supplemented with artificially prepared samples. The research in this article showed in the case of a small number of natural samples, artificially prepared samples was able to be adopted for model training by virtue of spectral transformation. A model with good robustness and accuracy assessment capability could be anticipated on the basis of low cost.

REFERENCES

[1] X. Zhang, W. Sun, Y. Cen, L. Zhang, and N. Wang, “Predicting cadmium concentration in soils using laboratory and field reflectance spectroscopy,” Sci. Total Environ., vol. 650, pp. 321–334, Feb. 2019.

[2] X. Jiang, B. Zou, H. Feng, J. Tang, Y. Tu, and X. Zhao, “Spatial distribution mapping of Hg contamination in subclass agricultural soils using GIS enhanced multiple linear regression,” J. Geochem. Explor., vol. 196, pp. 1–7, Jan. 2019.

[3] B. Mugoša, D. Dušović, M. Nedović-Vuković, S. Barjaktarović-Labović, and M. Vrvić, “Assessment of ecological risk of heavy metal contamination in coastal municipalities of montenegro,” Int. J. Environ. Res. Public Health, vol. 13, no. 4, p. 393, Mar. 2016.

[4] X. Hou, Y. Li, Y. Pan, Y. Jin, and H. Xiao, “Controlled release of agrochemicals and heavy metal ion capture dual-functional reductase-responsive hydrogel for soil remediation,” Chem. Commun., vol. 54, no. 97, pp. 13714–13717, 2018.

[5] J. Zhou, S. Chen, J. Liu, and R. L. Frost, “Adsorption kinetic and species variation of arsenic for As(V) and As(III) by biologically mackinawite (Fe5S7),” Chem. Eng. J., vol. 354, pp. 237–244, Dec. 2018.

[6] J. Kowalska, R. Mazurek, M. Gąsiorek, M. Setlak, T. Zaleski, and J. Waroszewski, “Soil pollution indices conditioned by medieval metal-lurgical activity—A case study from Krakow (Poland),” Environ. Pollut., vol. 218, pp. 1023–1036, Nov. 2016.

[7] M. A. Rahman, A. Rahman, M. Z. K. Khan, and A. M. N. Renzaho, “Human health risks and socio-economic perspectives of arsenic exposure in bangladesh: A scoping review,” Ecotoxicology Environ. Saf., vol. 150, pp. 335–343, Apr. 2018.

[8] T. Shi, Y. Chen, Y. Liu, and G. Wu, “Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals,” J. Hazard. Mater., vol. 265, pp. 166–176, Jan. 2014.

[9] W. Sun, A. K. Skidmore, T. Wang, and X. Zhang, “Heavy metal pollution at mine sites estimated from reflectance spectroscopy following correction for skewed data,” Environ. Pollut., vol. 252, pp. 1117–1124, Sep. 2019.

[10] Q. Yang, Z. Li, X. Lu, Q. Duan, L. Huang, and J. Bi, “A review of soil heavy metal pollution from industrial and agricultural regions in China: Pollution and risk assessment,” Sci. Total Environ., vol. 620, pp. 690–700, Nov. 2018.

[11] D. F. Malley and P. C. Williams, “Use of near-infrared reflectance spectroscopy in prediction of heavy metals in freshwater sediment by their association with organic matter,” Environ. Sci. Technol., vol. 31, no. 12, pp. 3461–3467, Dec. 1997.

[12] F. Wang, J. Gao, and Y. Zha, “Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges,” ISPRS J. Photogramm. Remote Sens., vol. 136, pp. 73–84, Feb. 2018.

[13] K. Tan, Y. Y. Ye, P. Du, and Q. Q. Zhang, “Estimation of heavy metal concentrations in reclaimed mining soils using reflectance spectroscopy,” Spectrosc. Spect. Anal., vol. 34, no. 12, p. 3317, 2014.

[14] E. Choe, F. van der Meer, F. van Ruitenberg, H. van der Werff, B. de Smeth, and K.-W. Kim, “Mapping of heavy metal pollution in stream sediments using combining field spectrometry, hyperspectral remote sensing: A case study of the roudalquiver mining area, SE spain,” Remote Sens. Environ., vol. 112, no. 7, pp. 3222–3233, Jul. 2008.

[15] W. Sun, X. Zhang, X. Sun, Y. Sun, and Y. Cen, “Predicting nickel concentration in soil using reflectance spectroscopy associated with organic matter and clay minerals,” Geoderma, vol. 327, pp. 25–35, Oct. 2018.

[16] M. Malmir, I. Tahmasbian, Z. Xu, M. B. Farrar, and S. H. Bai, “Prediction of soil macro- and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique,” Geoderma, vol. 340, pp. 70–80, Apr. 2019.

[17] T. Shi, H. Liu, Y. Chen, T. Fei, J. Wang, and G. Wu, “Spectroscopic diagnosis of arsenic contamination in agricultural soils,” Sensors, vol. 17, no. 5, p. 1036, May 2017.

[18] D. McNieish, “Missing data methods for arbitrary missingness with small samples,” J. Appl. Statist., vol. 44, no. 1, pp. 24–39, Jan. 2017.

[19] D. McNieish and J. R. Harring, “Correcting model fit criteria for small sample latent growth models with incomplete data,” Educ. Psychol. Meas., vol. 77, no. 6, pp. 990–1018, Dec. 2017.

[20] P. T. von Hippel, “New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples,” Struct. Equat. Model. A Multidisciplinary J., vol. 23, no. 3, pp. 422–437, May 2016.

[21] B. Zou, X. Jiang, H. Feng, Y. Tu, and C. Tao, “Multisource spectral-integrated estimation of cadmium concentrations in soil using a direct standardization and spiking algorithm,” Sci. Total Environ., vol. 701, Jan. 2020, Art. no. 134890.

[22] C. Tao, Y. Wang, W. Cui, B. Zou, Z. Zou, and Y. Tu, “A transferable spectroscopic diagnosis model for predicting arsenic contamination in soil,” Sci. Total Environ., vol. 669, pp. 964–972, Jun. 2019.

[23] T. Chao, W. Ya-Jin, Z. Bin, T. Yu-Long, and J. Xiao-Lu, “Assessment and analysis of migrations of heavy metal lead and zinc in soil with hyperspectral inversion model,” Spectrosc. Spect. Anal., vol. 38, no. 6, pp. 1850–1855, 2018.

[24] S. Najafi and M. Jalali, “Effect of heavy metals on pH buffering capacity and solubility of Ca, Mg, K, and P in non-spiked and heavy metal-spiked soils,” Environ. Monit. Assess., vol. 188, no. 6, p. 342, Jun. 2016.

[25] S. Nawar and A. Mouaazen, “Predictive performance of mobile visible-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques,” Catena, vol. 151, pp. 118–129, Apr. 2017.

[26] J. Wetterlind and B. Stenberg, “Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples,” Eur. J. Soil Sci., vol. 61, no. 6, pp. 823–843, 2010.

[27] Z. Shi, W. Ji, R. A. V. Rossel, S. Chen, and Y. Zhou, “Prediction of soil organic matter using a spatially constrained partial least squares regression and the chinese vis-NIR spectral library,” Eur. J. Soil Sci., vol. 66, no. 4, pp. 679–687, Jul. 2015.

[28] H. Liu, A. Probst, and B. Liao, “Metal contamination of soils and crops affected by the Chenzhou lead/zinc mine spill (Hunan, China),” Sci. Total Environ., vol. 339, no. 1, pp. 153–166, 2005.

[29] L. Wang, Z. Guo, X. Xiao, T. Chen, X. Liao, J. Song, and B. Wu, “Heavy metal pollution of soils and vegetables in the midstream and downstream of the Xiangjiang river, Hunan province,” J. Geographical Sci., vol. 18, no. 3, pp. 353–362, Aug. 2008.

[30] X. Li, Z. Zhao, Y. Yuan, X. Wang, and X. Li, “Heavy metal accumulation and its spatial distribution in agricultural soils: Evidence from Hunan province, China,” J. Appl. Stress., vol. 44, no. 1, pp. 24–39, Jan. 2017.

[31] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in Proc. Int. Conf. Comput. Vis., 2013, pp. 2960–2967.

[32] M. Long, Y. Cao, J. Wang, and M. I. Jordan, “Learning transferable features with deep adaptation networks,” 2015, arXiv:1502.02791 [Online]. Available: http://arxiv.org/abs/1502.02791.

[33] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, B. Schölkopf, and A. Smola, “Integrating structured biological data by kernel maximum mean discrepancy,” Bioinformatics, vol. 22, pp. 49–57, Jul. 2006.
Yajin Wang received the B.S. degree from the School of Prospecting and Surveying Engineering, Changchun Institute of Technology, and the M.S. degree from the School of Geosciences and Info-Physics, Central South University. She is currently an Associate Teacher with the School of Geographical Science, Shanxi Normal University. She has published articles in many international journals, such as Science of the Total Environment. Her research interests include hyperspectral information analysis and remote sensing detection of soil heavy metals.

Chao Tao (Member, IEEE) received the B.S. degree from the School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, in 2012. He is currently an Associate Processor with the School of Geosciences and Info-Physics, Central South University, Changsha, China. His research interests include computer vision, machine learning, high-resolution remote sensing image information extraction, and automatic target recognition. He has been frequently serving as a Reviewer for international journals, such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOENGINEERING AND REMOTE SENSING SOCIETY LETTERS, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, PERS, and the ISPRS Journal of Photogrammetry and Remote Sensing.

Bin Zou received the B.S. degree in geographic information engineering from Hunan Normal University, the master’s degree specializing in cartography and geographic information system from Northeast Normal University, and the Ph.D. degree in cartography and geographic information engineering from Central South University. His research interests include remote sensing monitoring and analysis of air pollution, remote sensing detection of soil heavy metals and environmental exposure risk assessment, and early warning. He is a member of the Health-GIS Committee of the International Chinese Geographic Information Science Association (CPGIS) and the Atmospheric Remote Sensing Working Group, International Society of Photogrammetry and Remote Sensing (ISPRS). He has been invited as a Reviewer for international journals, such as Remote Sensing of Environment, Environment Science and Technology, the International Journal of Geographical Information Science, and the International Journal of Geographical Information Science.