Computers and Electronics in Agriculture 187 (2021) 106257

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

# Measurement of potentially toxic elements in the soil through NIR, MIR, and XRF spectral data fusion

Fang Li [a,b], Li Xu [a], Tianyan You [b], Anxiang Lu [a,*]

[a] Beijing Research Center for Agricultural Standards and Testing, Beijing 100097, China
[b] School of Agricultural Engineering, Institute of Agricultural Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China

ARTICLE INFO

*Keywords:*
X-ray fluorescence
Near-infrared
Mid-infrared
Successive projections algorithm
Partial least square regression

ABSTRACT

This paper aims to investigate the feasibility combination of X-ray fluorescence (XRF), near-infrared (NIR), and mid-infrared (MIR) sensors for the detection of seven key monitoring elements in the soil. Two strategies for data fusion were adopted: (i) the XRF characteristic bands were fused with all spectral data of NIR and MIR separately, and (ii) the XRF characteristic bands were fused with the characteristic bands of NIR and MIR, respectively. Also, different feature extraction methods were compared. The best feature extraction methods for XRF-NIR and XRF-MIR models were principal component analysis (PCA) and successive projections algorithm (SPA). The modeling results showed that strategy (ii) showed better performance, and the XRF-MIR model provided more accurate results than the XRF-NIR model. Both XRF-MIR and XRF-NIR methods improved the accuracy of As, Cr, Cu, Ni, and Zn. The XRF-MIR obtained the best predictions, and the determination coefficients ($R^2$) were 0.93, 0.98, 0.98, 0.95, and 0.98. For Pb and Cd, the measurement could be obtained by XRF alone, and the corresponding $R^2$ were both 0.98. The results confirmed that sensor fusion can effectively improve the accuracy of the spectrometer in detecting metal elements in the soil.

## 1. Introduction

Hazardous elements in agricultural soils have the characteristics of latent, hysteresis, synergism, and non-biodegradability, easily accumulated to toxic levels and causing heavy metal pollution (Huang et al., 2018; Zhang et al., 2018). Heavy metal pollution in agricultural soils becomes severe with the rapid industrialization and urbanization over the last two decades (Wei and Yang, 2010). The contamination of cadmium (Cd), chromium (Cr), lead (Pb), and arsenic (As) poses a considerable threat to human health and the environment (Oliver, 1997; Yang et al., 2018). Although zinc (Zn) and copper (Cu) are essential elements, excessive concentrations in food and feed plants are also of great concern due to toxicity to humans and animals (Kabata-Pendias and Mukherjee, 2007; Adler et al., 2020). Accordingly, a chemical detection method was used to monitor the hazardous soil elements. However, such a chemical detection method is cumbersome. Recently, non-destructive methods have been attracted much attention due to their accuracy and speed (Kumar Sharma et al., 2007; Al Maliki et al., 2014). Eight kinds of metal elements in soil: Cd, Cr, Cu, Zn, Pb, As, Ni and mercury (Hg) have been specified their limited concentration according to "Soil environmental quality risk control standard for soil contamination of agricultural land (GB 15618–2018)". It's difficult to quantify the concentration of Hg with XRF, because Hg can easily become mercury vapor under high-energy X-ray excitation and cannot be accurately measured. So seven metal elements were analyzed in this paper.

As a rapid on-site detection method, X-ray fluorescence (XRF) has been widely used to measure potentially toxic elements. The XRF method has the advantages of fast measurement, low cost, straightforward operation, easy preparation, and non-destructive (Rouillon et al., 2017; Turner et al., 2018; Caporale et al., 2018). The good performance of the XRF spectrometer to measure the soil elements were reported (Rouillon and Taylor, 2016; Paulette et al., 2015). However, when the amount of the elements is very low, the XRF method is difficult to meet the detection requirements (Hu et al., 2017). Thus, improving the accuracy of the XRF method is an essential research direction in XRF-based soil element measurement.

On the other hand, near-infrared (NIR) and mid-infrared (MIR) methods require almost no sample pretreatment and can directly analyze soil/sediment samples combined with chemometric modeling technology. Researchers speculated that the concentration of metal

* Corresponding author.
*E-mail address:* anxxlu@163.com (A. Lu).

https://doi.org/10.1016/j.compag.2021.106257
Received 24 March 2021; Received in revised form 24 May 2021; Accepted 6 June 2021
Available online 12 June 2021

elements is directly related to organic matter, iron oxides, and clay. Thus, metal elements could be detected using the interaction between the metal ions and soil components (Grzegorz et al., 2004; Fan et al., 2010). In recent years, NIR and MIR have been increasingly used for soil metal elements analysis due to fast and non-destructive detection characteristics (O'Rourke et al., 2016; Dx et al., 2019).

Several researchers have studied to combine data from different detection methods, aiming to improve the rapid detection accuracy. The data fusion of some proximal sensors achieved promising results (Ji et al., 2019; Xu et al., 2020). The prediction results of the quantitative models depend on the spectral characteristics of the employed data. Thus, according to XRF, NIR, and MIR spectra characteristics, the influence of inorganic and organic components in the soil can be considered comprehensively by data fusion. The fusion of XRF and Vis-NIR data was previously studied (Ji et al., 2019; Xu et al., 2020; Moros et al., 2009). However, the fusion of XRF and NIR / MIR data was not studied. This paper explores the potential benefits of fusing XRF and NIR data (XRF-NIR) and XRF and MIR data (XRF-MIR). Further, different feature extraction methods based on partial least square regression (PLSR) were compared to obtain the best results, including principal component analysis (PCA), shuffled frog leaping algorithm (SFLA), genetic algorithm (GA), successive projections algorithm (SPA), and competitive adaptive reweighed sampling (CARS).

This study is mainly focused on evaluating the impact of XRF and NIR / MIR data fusion on the prediction of hazardous metal elements in the soil: Cd, Cu, Zn, As, Pb, Cr, and Ni. First, the original spectral data were used to establish PLSR quantitative detection models for seven metal elements in the soil, based on a single sensor (XRF, NIR, MIR) and sensor fusion (XRF-NIR, XRF-MIR). Second, different feature extraction algorithms were compared to determine the most appropriate method for different data fusion models constructed using feature bands. Lastly, the best measurement scheme was selected for estimating metal elements in the soil.

## 2. Materials and methods

### 2.1. Soil samples and apparatus

There are a total of 120 soil samples including 40 standard samples and 80 farmland samples. The standard soil samples were purchased from the National Standards Center. The element contents in the standard samples were collected and verified by the National Standards Center, and the results were accurate and authoritative. Adding standard samples to the experiment can improve the reliability of the analysis results. The rest soil samples were collected from farmland in Shandong from a depth of 0–20 cm from the surface. These samples were collected from 80 different sampling sites, which were at different distances from the mining area, and the range of each sampling site was about 200*200 m. Five sampling points were collected by the quincunx point method. The sampling volume of each point was about 1 kg. The sample quartiles was used to keep the final sample amount of each sampling site at 1 kg. All farmland samples were dried, ground, and sieved indoors after picking out the stones and plant rhizomes. The samples were divided into two parts by the sample quartiles: one part was used for XRF, NIR, and MIR spectral analysis, while the other part was used for chemical detection by inductively coupled plasma mass spectrometry (ICP-MS) (NexION 300, Perkin Elmer Inc., Waltham, MA, USA).

For XRF analysis, 5 g of soil was compacted in a vinyl sample cup (diameter × height: 2 cm × 3 cm) (NCS Testing Technology Co., Ltd., Beijing, China). The cup was then placed into the sample chamber of the XRF spectrometer (ATFM-100, NCS Testing Technology Co., Ltd., Beijing, China). Each sample was scanned three times, and the average spectrum was calculated for later analysis. A portable instrument (XL410, Axsun Technologies, Billerica, MA, USA) was used to collect the NIR spectra of soil samples. The wavelength range was 1350–1800 nm, and the spectral resolution was 0.8 nm. A glass beaker containing 1 g of soil sample was placed on the detection window to obtain NIR spectra. The average spectrum of 32 scans of each sample was obtained as the analysis spectrum. The MIR spectra of soil samples were collected using Spectrum 400 (Perkin Elmer Inc., Waltham, MA, USA). The wavenumber range was 4000–650 $cm^{-1}$ with a resolution of 4 $cm^{-1}$. In order to achieve the MIR spectra, weigh 0.5 g sample and presse it on the sample stage for testing. Each sample was scanned 16 times to get the average spectrum for analysis.

### 2.2. Data analysis

After spectra were collected from soil samples, two spectral preprocessing were applied before further analysis. The standard normal variate transform (SNV) (Barnes et al., 1989) was adopted to eliminate the influence of solid particle size, surface scattering, and optical path changes on the spectra. The wavelet transform (WT) was used for spectral smoothing with Coif3 wavelet basis for three-layer decomposition.

The PLSR models were built using the preprocessed data. PLSR is a classic and commonly used linear multivariate statistical method, which has been widely used to establish spectral detection models. PLSR combines factor analysis and regression analysis, performs principal component decomposition on the spectra matrix X and the concentration matrix Y, so that the principal component is the largest correlation with the concentration. The leave-one-out cross-validation is used to calculate the predicted residual sum of squares (PRESS) and to avoid overfitting. The optimal number of latent variables is determined by PRESS and the cumulative contribution rate of the latent variables. Finally, a linear regression model between X and Y is built (Wold et al., 2001; Cheng and Sun, 2016). Elements have strong characteristic intensity in the XRF spectra, so the XRF models and the fusion models of XRF with NIR and MIR were established to evaluate the influence of data fusion on the modeling results. Two different strategies were adopted for data fusion. First, XRF feature data were fused with all MIR and NIR spectral data. Second, the XRF feature data were fused with extracted feature bands from MIR and NIR data. The data matrix [**x, y**] of the XRF-MIR fusion model is used as independent variables, where **x** is the variables of XRF spectra and **y** is the variables of MIR spectra. Similarly, the data matrix [**x, z**] represents the input variables of the XRF-NIR fusion model, where **z** indicates the variables of NIR spectra.

Five methods were used for feature extraction for the NIR and MIR spectral characteristic bands acquisition, including PCA, SFLA, GA, SPA, and CARS. Feature extraction removes redundant variables, eliminates excessive noise spectral bands, and reduces the computational complexity of the model while improving the model in terms of anti-interference ability and accuracy (Ding et al., 2011; Guyon et al., 2008; Nixon and Aguado, 2019). PCA transforms a group of potentially correlated variables into a group of linearly uncorrelated variables through orthogonal transformation, so that a few new variables can express the data characteristics of the original variables as much as possible without losing information (Abdi and Williams, 2010). SFLA has few adjustment parameters, fast calculation speed, strong global search and optimization ability, and is easy to implement. The strategy of not updating individual components that do not meet the threshold condition reduces the individual spatial differences, thereby improving the performance of the algorithm (Eusuff and Lansey, 2003). The main feature of GA is to directly operate on structural objects, without the limitation of derivation and function continuity; it has inherent implicit parallelism and good global optimization ability. Using the probabilistic optimization method, the optimized search space can be automatically obtained and guided without definite rules, and the search direction can be adjusted adaptively (Whitley, 1994). SPA can find the variable group containing the minimum redundant information from the spectra, so that the collinearity between the variables is minimized, thereby greatly reducing the number of variables used in modeling, and improving the

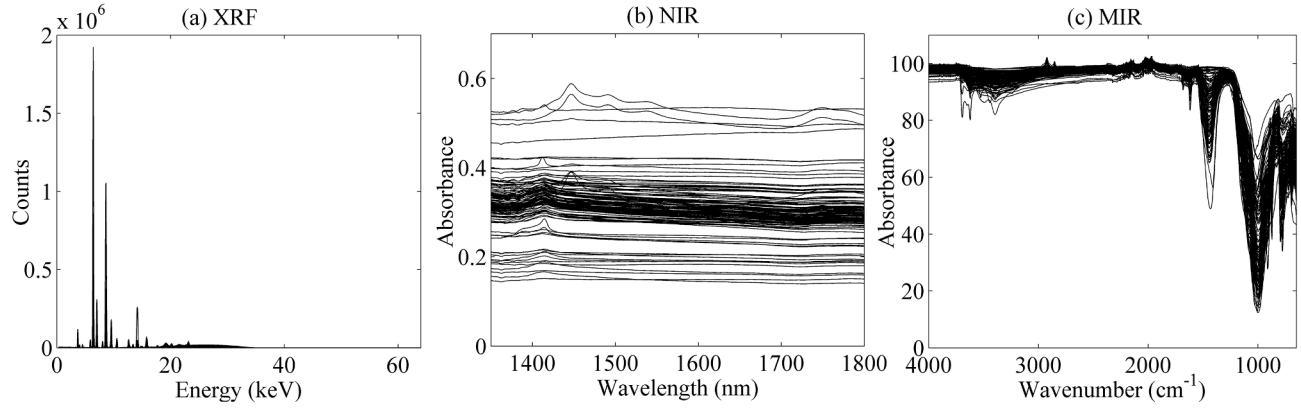


**Fig. 1.** Raw (a) XRF spectra, (b) NIR spectra, and (c) MIR spectra for the soil samples.

**Table 1**
Statistics of the soil elements detected by ICP-MS (mg/kg).

| Elements | Total number | Min. | Max. | Mean | SD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| As | 116 | 1.70 | 234 | 36.25 | 53.901 | 2.86 | 7.22 |
| Cd | 121 | 0.03 | 870 | 39.19 | 114.86 | 4.61 | 25.53 |
| Cr | 120 | 2.93 | 410 | 54.44 | 41.39 | 5.97 | 47.70 |
| Cu | 118 | 2.80 | 1740 | 127.38 | 291.11 | 3.77 | 14.93 |
| Ni | 111 | 1.54 | 349 | 33.02 | 40.52 | 6.41 | 44.95 |
| Pb | 116 | 7.72 | 3490 | 184.88 | 552.87 | 4.44 | 20.38 |
| Zn | 108 | 0.26 | 4930 | 555.16 | 970.51 | 2.65 | 6.75 |

**Table 2**
XRF modeling information for soil elements.

| Elements | Peak position / keV | Energy range / keV |
|---|---|---|
| Cr | 5.414(Kα1) | 5.34–5.53 |
| Ni | 7.477(Kα1) | 7.38–7.60 |
| As | 10.543(Kα1) | 10.39–10.73 |
| Pb | 12.611(Lβ1) | 12.48–12.77 |
| Cu | 8.047(Kα1) | 7.88–8.19 |
| | 8.904(Kβ1) | 8.88–9.01 |
| Zn | 8.638(Kα1) | 8.51–8.79 |
| | 9.571(Kβ1) | 9.48–9.70 |
| Cd | 23.172(Kα1) | 23.00–23.31 |
| | 26.093(Kβ1) | 26.07–26.13 |

speed and efficiency of modeling (Araújo et al., 2001). CARS uses adaptive re-weighting sampling technology to select the wavelength points with large absolute values of regression coefficients in the PLS model, remove the wavelength points with small weights, and use cross-validation to select the subset with the lowest root mean square error value, which can effectively find the optimal variable combination (Li et al., 2009).

Different methods were analyzed to select the optimal feature extraction algorithm for each fused model. In order to evaluate the performance of the prediction model, the root mean square error estimated by calibration (RMSEC), cross-validation (RMSECV), and prediction (RMSEP) and determination coefficient of calibration ($R_C^2$), cross-validation ($R_{CV}^2$) and prediction ($R_P^2$) were used as evaluation indicators. Higher $R^2$ and low RMSE indicate a better prediction model.

## 3. Results and discussion

### 3.1. Statistics of samples

The spectral information was obtained by scanning all soil samples by spectroscopy (Fig. 1), and the element values were obtained by ICP-MS. Table 1 summarizes the statistics of the samples used in modeling each element. Note that outliers were eliminated during modeling. As shown in Table 1, the seven elements were positive skewness, and there was a big range between the minimum and maximum element content of all soil elements. A large standard deviation (SD) value indicates that the data had a high degree of dispersion. The amounts of the element in some sites were very high due to their close location to the mining area. The high-content points can be used to study the possibility of applying the model to soil environments with high element concentrations.

**Table 3**
Results of the PLSR models by single sensors.

| Models | Elements | Calibration set | | Cross-validation set | | Prediction set | |
|---|---|---|---|---|---|---|---|
| | | $R_C^2$ | RMSEC | $R_{CV}^2$ | RMSECV | $R_P^2$ | RMSEP |
| XRF | As | 0.91 | 16.12 | 0.90 | 19.35 | 0.89 | 22.54 |
| | Cd | 0.98 | 15.72 | 0.98 | 16.01 | 0.97 | 16.92 |
| | Cr | 0.95 | 9.54 | 0.94 | 10.12 | 0.94 | 10.87 |
| | Cu | 0.96 | 31.65 | 0.95 | 37.58 | 0.94 | 44.37 |
| | Ni | 0.82 | 16.96 | 0.83 | 16.34 | 0.80 | 17.64 |
| | Pb | 0.98 | 78.63 | 0.97 | 81.02 | 0.97 | 83.15 |
| | Zn | 0.94 | 256.53 | 0.94 | 253.87 | 0.92 | 287.25 |
| NIR | As | 0.30 | 44.96 | 0.27 | 61.42 | 0.22 | 85.38 |
| | Cd | 0.01 | 113.88 | 0.01 | 110.54 | 0.01 | 117.24 |
| | Cr | 0.01 | 41.00 | 0.01 | 42.31 | 0.01 | 51.02 |
| | Cu | 0.34 | 131.64 | 0.31 | 154.27 | 0.29 | 170.35 |
| | Ni | 0.20 | 36.10 | 0.21 | 32.89 | 0.12 | 61.61 |
| | Pb | 0.50 | 391.16 | 0.46 | 445.26 | 0.41 | 495.82 |
| | Zn | 0.21 | 942.52 | 0.19 | 1000.24 | 0.15 | 1205.86 |
| MIR | As | 0.38 | 42.37 | 0.34 | 56.14 | 0.29 | 79.18 |
| | Cd | 0.75 | 56.99 | 0.73 | 64.22 | 0.69 | 73.48 |
| | Cr | 0.89 | 13.88 | 0.88 | 18.26 | 0.86 | 22.34 |
| | Cu | 0.42 | 122.50 | 0.40 | 137.92 | 0.36 | 160.27 |
| | Ni | 0.97 | 7.38 | 0.97 | 7.31 | 0.96 | 9.28 |
| | Pb | 0.73 | 285.42 | 0.72 | 291.54 | 0.68 | 335.85 |
| | Zn | 0.37 | 839.37 | 0.34 | 963.81 | 0.30 | 1014.29 |

### 3.2. Predictability of single sensor model

Different characteristic peaks and energy ranges of soil elements

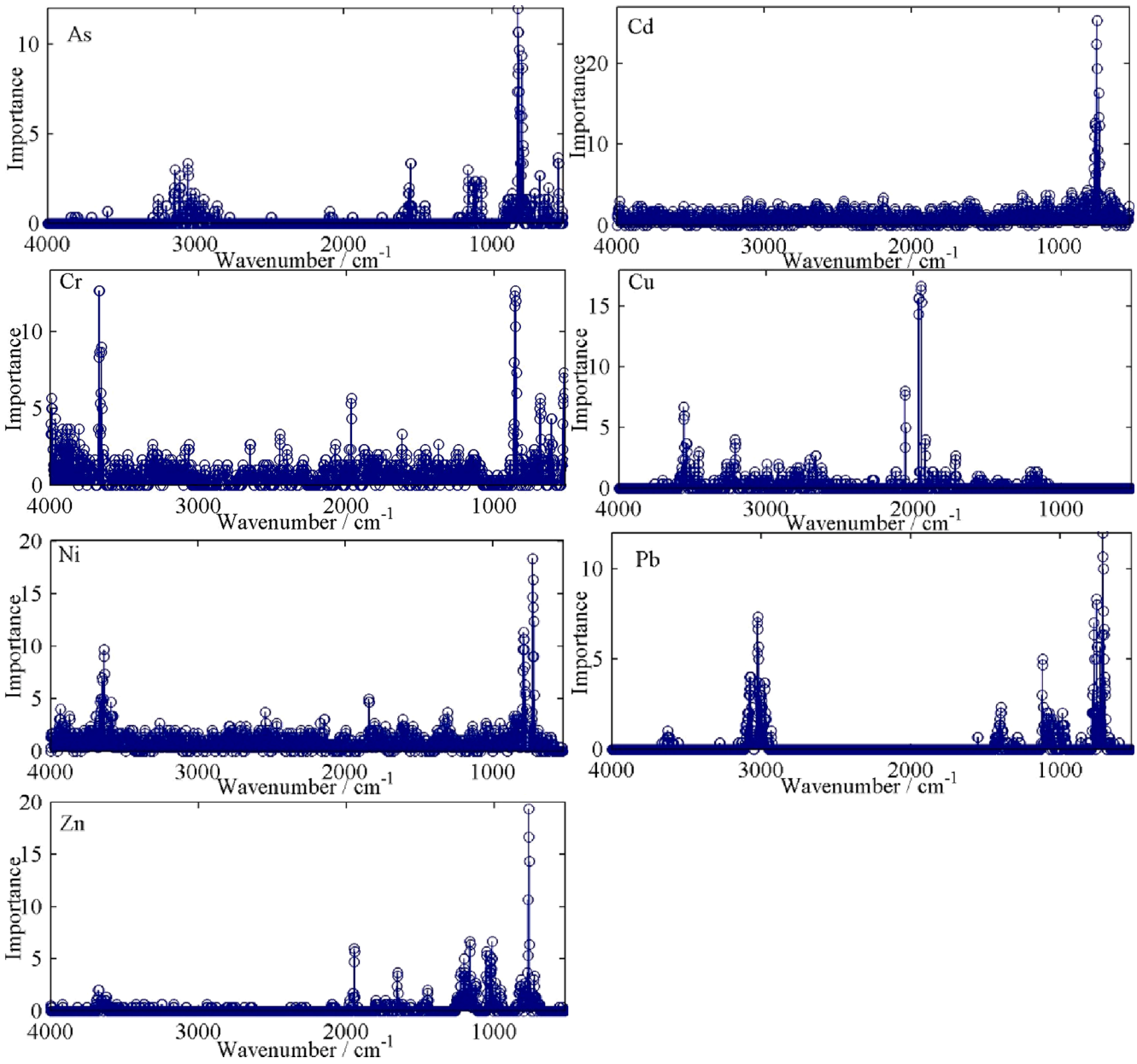


**Fig. 2.** Bands importance of MIR spectra for soil elements.

were selected to improve the accuracy of the XRF models, as shown in Table 2. Further, models based on all MIR spectral bands and NIR spectral bands were constructed. Table 3 compares the prediction results of XRF, MIR, and NIR models. The MIR model achieved a good prediction on Cr and Ni, while the NIR model provided poor prediction accuracy for the seven elements. These results are comparable to the previous study (Grzegorz et al., 2004; Soriano-Disla et al., 2013). As shown in Table 3, the MIR model obtained the best estimation for Ni, while the XRF model outperformed the MIR and NIR models for the other six elements since the XRF spectra contain clear and strong characteristic peaks for each element (Turner and Taylor, 2018). Modeling with characteristic bands achieved an excellent correlation between XRF signals and element concentrations.

The intensity of infrared spectra is relatively weak, and inorganic ions generally have no response. Also, metal elements are quantified through the combination or correlation with organic matter, carbonate, Fe minerals, and clay minerals (Moros et al., 2009; Malley and Williams, 1997). The NIR model provided poor estimation results, while the MIR model shows potential. Thus, the importance of MIR spectral bands was analyzed to study the relationship further. As shown in Fig. 2, the spectral absorption characteristics of organic matter in the MIR bands are mainly derived from the strong fundamental molecular vibrations of C-H, C-O, C = O, and N-H (Śliwińska et al., 2019; Song et al., 2012; Wu et al., 2007; Gredilla et al., 2016). The most important variables in the MIR spectra were mainly around 3600–3620, 2900–3100, 1900–2000, 1200, and 700–1100 $cm^{-1}$. These bands were closely related to the absorption of N-H, C-H, C-O, and C = O bonds of the organic matter. The bands around 3620, 3450, and 1200 $cm^{-1}$ were mainly affected by clay minerals. The important variables were consistent with the previously reported ones (Wu et al., 2007; Bertrand et al., 2002).

### 3.3. Predictability of sensor fusion model

#### 3.3.1. Prediction with full spectral data

After SNV and WT preprocessing were applied, the sensor fusion models were constructed by combining the XRF characteristic peak

**Table 4**
PLSR results for the fused XRF-NIR and XRF-MIR.

| Models | Elements | Calibration set | | Cross-validation set | | Prediction set | |
|---|---|---|---|---|---|---|---|
| | | $R_C^2$ | RMSEC | $R_{CV}^2$ | RMSECV | $R_P^2$ | RMSEP |
| XRF-NIR | As | 0.91 | 16.12 | 0.90 | 19.35 | 0.89 | 22.54 |
| | Cd | 0.98 | 15.72 | 0.98 | 16.01 | 0.97 | 16.92 |
| | Cr | 0.95 | 9.54 | 0.94 | 10.12 | 0.94 | 10.87 |
| | Cu | 0.96 | 31.34 | 0.95 | 37.21 | 0.94 | 44.07 |
| | Ni | 0.82 | 16.96 | 0.83 | 15.84 | 0.80 | 21.64 |
| | Pb | 0.98 | 78.63 | 0.97 | 81.02 | 0.97 | 83.15 |
| | Zn | 0.94 | 256.53 | 0.94 | 253.87 | 0.92 | 287.25 |
| XRF-MIR | As | 0.91 | 16.02 | 0.90 | 18.85 | 0.89 | 21.67 |
| | Cd | 0.98 | 15.72 | 0.98 | 16.01 | 0.97 | 16.92 |
| | Cr | 0.97 | 7.57 | 0.96 | 8.34 | 0.96 | 8.67 |
| | Cu | 0.96 | 30.35 | 0.95 | 35.87 | 0.94 | 40.25 |
| | Ni | 0.85 | 15.48 | 0.85 | 15.84 | 0.83 | 16.98 |
| | Pb | 0.98 | 78.52 | 0.97 | 81.01 | 0.97 | 82.78 |
| | Zn | 0.99 | 111.94 | 0.98 | 132.47 | 0.98 | 137.25 |

variables and all NIR and MIR spectral variables. The XRF model was used as the control group since its prediction accuracy is the best. Then, the fusion models were compared to the XRF model in terms of prediction accuracy. Table 4 presents the prediction results for the fused XRF-NIR and XRF-MIR models. The XRF-NIR model achieved no accuracy increase for all soil elements over the XRF model (Table 3). It may be because there were few NIR spectral peaks, and the difference between spectral data was minimal. When all spectral data were involved in the modeling, it may cause data redundancy and fail to improve the accuracy (Soriano-Disla et al., 2013).

In contrast, the XRF-MIR fusion model improved prediction accuracy for five elements of As, Cr, Cu, Ni, and Zn in the soil. The improvement of Cr, Ni and Zn was obvious, while As and Cu were relatively small. There was no improvement for Cd and Pb. This may be because the Cd and Pb in the analyzed soil samples mainly existed in inorganic form; thus, there were no obvious intensity changes in the MIR spectra. For As, Cr, Cu, Ni, and Zn, they were easily combined with the organic components in the soil, which can be reflected in the spectral changes to detect. These results were consistent with the previous research (Śliwińska et al., 2019).

### 3.3.2. *Prediction with feature bands*

The characteristic bands of the spectral were extracted by five feature extraction algorithms, including PCA, GA, SFLA, SPA, and CARS. As shown in Table 5 and Table 6, the extracted feature bands by all the five methods provided improved accuracy for As, Cr, Cu, Ni, and Zn. However, it could not improve the accuracy for Pb and Cd, which is consistent with the results obtained with the full spectral data fusion models. Among the five feature extraction methods, the PCA and SPA showed the best performance for XRF-NIR and XRF-MIR, respectively. For the combination of the XRF-NIR model and the PCA feature extraction, the accuracies of Cr and Ni were improved the most, followed by As, Cu, and Zn. The prediction accuracy of the seven elements in descending order was: Cd > Pb > Cr > Cu > Zn > As > Ni. It is worth mentioning that, considering a single element, the SFLA showed the best prediction performance for Zn in the XRF-NIR models. Therefore, SFLA can be used to predict Zn, and PCA can be used to predict other elements. However, note that such an adaptive scheme could increase the computational complexity.

For the combination of the XRF-MIR models and SPA feature extraction, Ni had the largest improvement in accuracy, followed by Zn, Cr, and Cu, while the smallest improvement was achieved for As. The

**Table 5**
PLSR results for the XRF-NIR models with different feature extraction methods.

| Feature extraction method | Elements | Calibration set | | Cross-validation set | | Prediction set | |
|---|---|---|---|---|---|---|---|
| | | $R_C^2$ | RMSEC | $R_{CV}^2$ | RMSECV | $R_P^2$ | RMSEP |
| PCA | As | 0.92 | 14.74 | 0.91 | 16.59 | 0.91 | 16.94 |
| | Cd | 0.98 | 14.85 | 0.98 | 15.16 | 0.97 | 15.98 |
| | Cr | 0.97 | 7.01 | 0.96 | 8.13 | 0.95 | 9.23 |
| | Cu | 0.95 | 34.36 | 0.96 | 31.75 | 0.94 | 39.57 |
| | Ni | 0.87 | 14.55 | 0.86 | 15.04 | 0.85 | 15.52 |
| | Pb | 0.98 | 78.49 | 0.97 | 80.82 | 0.97 | 83.07 |
| | Zn | 0.93 | 277.13 | 0.94 | 258.45 | 0.92 | 281.19 |
| SFLA | As | 0.91 | 15.92 | 0.90 | 19.14 | 0.90 | 20.29 |
| | Cd | 0.98 | 15.25 | 0.98 | 15.71 | 0.97 | 16.47 |
| | Cr | 0.95 | 8.81 | 0.94 | 9.68 | 0.93 | 10.16 |
| | Cu | 0.94 | 38.88 | 0.95 | 34.08 | 0.93 | 43.30 |
| | Ni | 0.83 | 16.44 | 0.82 | 17.06 | 0.81 | 18.75 |
| | Pb | 0.98 | 79.57 | 0.97 | 82.15 | 0.97 | 83.82 |
| | Zn | 0.96 | 224.20 | 0.95 | 241.78 | 0.93 | 276.83 |
| GA | As | 0.91 | 15.86 | 0.90 | 18.99 | 0.90 | 20.06 |
| | Cd | 0.98 | 15.27 | 0.98 | 15.86 | 0.97 | 16.71 |
| | Cr | 0.96 | 7.73 | 0.95 | 8.34 | 0.95 | 9.52 |
| | Cu | 0.94 | 39.29 | 0.94 | 39.93 | 0.93 | 42.15 |
| | Ni | 0.83 | 16.53 | 0.82 | 17.30 | 0.80 | 22.78 |
| | Pb | 0.98 | 79.53 | 0.97 | 82.01 | 0.97 | 83.62 |
| | Zn | 0.95 | 234.12 | 0.94 | 250.32 | 0.93 | 279.15 |
| SPA | As | 0.91 | 15.88 | 0.90 | 18.84 | 0.89 | 21.01 |
| | Cd | 0.98 | 15.20 | 0.98 | 15.31 | 0.97 | 16.35 |
| | Cr | 0.96 | 8.46 | 0.95 | 9.67 | 0.95 | 10.01 |
| | Cu | 0.94 | 38.75 | 0.95 | 38.24 | 0.93 | 41.90 |
| | Ni | 0.83 | 16.42 | 0.83 | 16.35 | 0.81 | 18.77 |
| | Pb | 0.98 | 79.63 | 0.97 | 80.81 | 0.97 | 83.74 |
| | Zn | 0.95 | 227.43 | 0.94 | 248.44 | 0.94 | 251.71 |
| CARS | As | 0.91 | 15.89 | 0.90 | 18.80 | 0.90 | 20.31 |
| | Cd | 0.98 | 15.23 | 0.98 | 15.63 | 0.97 | 16.86 |
| | Cr | 0.96 | 8.73 | 0.96 | 9.45 | 0.95 | 10.24 |
| | Cu | 0.94 | 38.42 | 0.94 | 38.68 | 0.92 | 45.78 |
| | Ni | 0.83 | 16.53 | 0.82 | 17.39 | 0.80 | 22.84 |
| | Pb | 0.98 | 79.64 | 0.97 | 81.00 | 0.97 | 83.92 |
| | Zn | 0.93 | 289.02 | 0.93 | 293.14 | 0.92 | 299.56 |

**Table 6**
PLSR results for the XRF-MIR models with different feature extraction methods.

| Feature extraction method | Elements | Calibration set | | Cross-validation set | | Prediction set | |
|---|---|---|---|---|---|---|---|
| | | $R_C^2$ | RMSEC | $R_{CV}^2$ | RMSECV | $R_P^2$ | RMSEP |
| PCA | As | 0.93 | 13.96 | 0.92 | 14.79 | 0.92 | 15.33 |
| | Cd | 0.98 | 14.47 | 0.98 | 14.89 | 0.97 | 15.12 |
| | Cr | 0.98 | 6.10 | 0.97 | 6.99 | 0.97 | 7.87 |
| | Cu | 0.98 | 24.14 | 0.98 | 24.01 | 0.97 | 26.42 |
| | Ni | 0.90 | 12.71 | 0.89 | 13.15 | 0.88 | 14.29 |
| | Pb | 0.97 | 94.85 | 0.96 | 98.27 | 0.96 | 100.35 |
| | Zn | 0.99 | 127.26 | 0.98 | 141.54 | 0.98 | 141.86 |
| SFLA | As | 0.93 | 14.67 | 0.92 | 15.48 | 0.91 | 16.58 |
| | Cd | 0.98 | 14.35 | 0.98 | 14.72 | 0.97 | 15.03 |
| | Cr | 0.96 | 8.05 | 0.95 | 8.97 | 0.95 | 9.16 |
| | Cu | 0.97 | 27.95 | 0.96 | 31.02 | 0.95 | 34.79 |
| | Ni | 0.92 | 11.10 | 0.91 | 12.05 | 0.90 | 12.84 |
| | Pb | 0.97 | 97.49 | 0.97 | 95.27 | 0.96 | 99.64 |
| | Zn | 0.98 | 144.62 | 0.97 | 154.81 | 0.97 | 160.57 |
| GA | As | 0.92 | 15.31 | 0.91 | 15.93 | 0.90 | 18.76 |
| | Cd | 0.98 | 14.48 | 0.98 | 15.17 | 0.97 | 16.85 |
| | Cr | 0.97 | 6.82 | 0.96 | 7.26 | 0.95 | 8.78 |
| | Cu | 0.97 | 27.63 | 0.97 | 27.41 | 0.96 | 30.05 |
| | Ni | 0.90 | 12.85 | 0.89 | 13.20 | 0.88 | 14.37 |
| | Pb | 0.96 | 105.23 | 0.96 | 101.58 | 0.95 | 109.41 |
| | Zn | 0.98 | 146.82 | 0.97 | 158.14 | 0.97 | 161.20 |
| SPA | As | 0.93 | 14.53 | 0.92 | 14.88 | 0.92 | 15.37 |
| | Cd | 0.98 | 14.65 | 0.98 | 15.02 | 0.97 | 15.51 |
| | Cr | 0.98 | 5.10 | 0.98 | 5.34 | 0.97 | 6.28 |
| | Cu | 0.98 | 20.96 | 0.98 | 18.59 | 0.97 | 20.80 |
| | Ni | 0.95 | 9.22 | 0.94 | 9.94 | 0.93 | 10.46 |
| | Pb | 0.97 | 100.29 | 0.96 | 105.87 | 0.96 | 108.52 |
| | Zn | 0.98 | 133.5 | 0.97 | 139.45 | 0.97 | 141.02 |
| CARS | As | 0.94 | 12.93 | 0.93 | 13.94 | 0.92 | 15.08 |
| | Cd | 0.98 | 14.27 | 0.98 | 14.79 | 0.97 | 15.04 |
| | Cr | 0.98 | 6.16 | 0.97 | 6.99 | 0.97 | 7.45 |
| | Cu | 0.98 | 21.78 | 0.98 | 22.30 | 0.97 | 25.19 |
| | Ni | 0.93 | 10.89 | 0.92 | 11.24 | 0.91 | 12.36 |
| | Pb | 0.98 | 83.62 | 0.97 | 90.67 | 0.97 | 93.06 |
| | Zn | 0.97 | 168.67 | 0.96 | 171.29 | 0.96 | 175.11 |

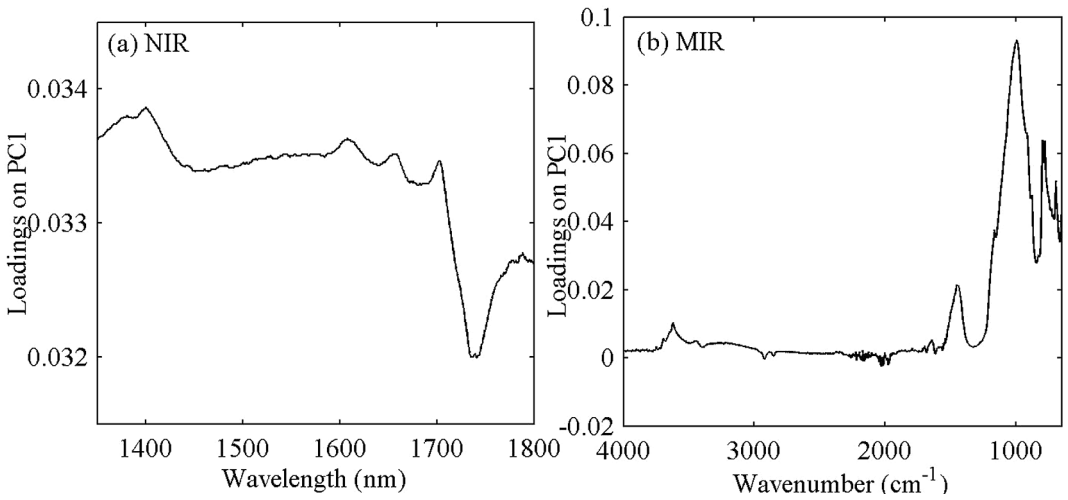


**Fig. 3.** Loading values on PC1 for (a) the NIR spectra and (b) MIR spectra. The preprocessing SNV correction was applied.

prediction accuracy of the seven elements from high to low was: Cr > Zn > Cd > Cu > Pb > Ni > As. The XRF-MIR models significantly outperform the XRF-NIR models in terms of the prediction accuracy because of more relevant information in the MIR spectra elements (Dong et al., 2011; Chakraborty et al., 2015). In the XRF-MIR models of As, the PCA and CARS methods had better accuracy than the SPA method. However, since the difference is slight, considering the additional computational complexity for an adaptive algorithm, it is suggested to use SPA to build As model.

### 3.3.3. *Spectral analysis using PCA*

For the XRF-NIR fusion models, PCA showed the best performance, and the first two PCs were selected to establish the models. Fig. 3 shows the loading values on PC1 for the NIR and MIR spectra after SNV preprocessing. The first two PCs accounted for 99.0% and 85.7% of NIR and MIR spectral data variance. The importance of a variable in each PC is measured based on the loadings, where the maximum and minimum loading values indicate the most important NIR and MIR bands (Alamprese et al., 2013; Song et al., 2010). The significant variables in

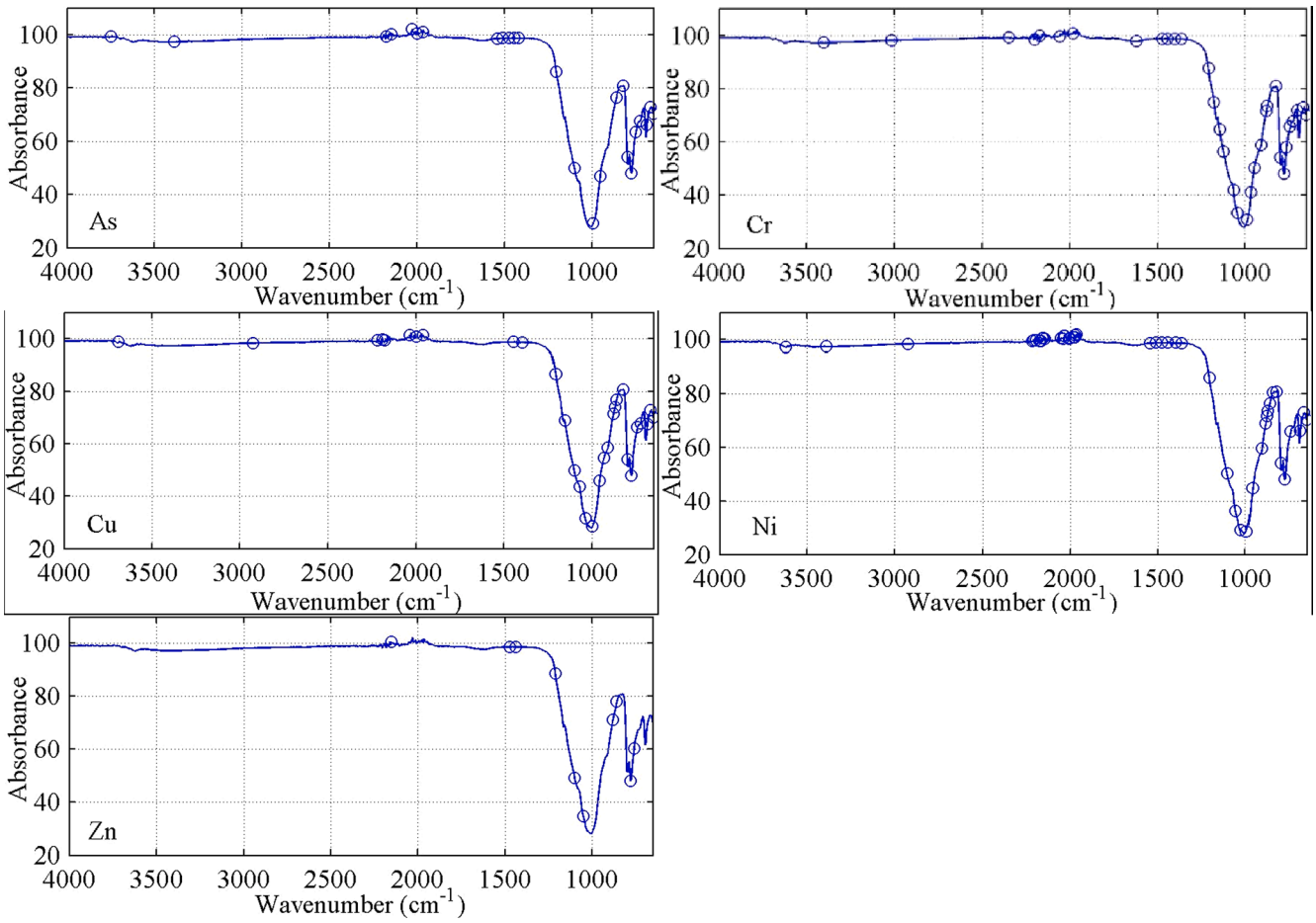


**Fig. 4.** Selected characteristic bands of MIR by SPA.

NIR spectra around 1400 were associated with the O-H bonds of the free water, and the bands around 1741 nm were affected by C-H bonds of organic matter (Xu et al., 2020; Soriano-Disla et al., 2013). 3614, 991, 796, and 694 $cm^{-1}$ in MIR spectra corresponded to the absorption characteristic of C-H, C-C, C-O, C-N (Ji et al., 2019). The results verified that the metal elements were combined with the soil components to be expressed on the spectra.

### 3.3.4. *The selected feature bands by SPA*

The SPA was proved as the optimum feature extraction method for the XRF-MIR fusion models in the previous section. The MIR spectra of the elements of which the XRF-MIR fusion model improved predictability were analyzed: As, Cr, Cu, Ni, and Zn. The numbers of characteristic bands corresponding to As, Cr, Cu, Ni, and Zn elements were 25, 33, 31, 46, and 10, respectively. The ratio of the characteristic bands to the total variables of the five elements were 1.44%, 1.90%, 1.78%, 2.65%, and 0.58%. The characteristic bands were located around 3600, 3000, 2100–1900, 1600–650 $cm^{-1}$, as shown in Fig. 4, similar to the importance of variables. These bands were associated with organic matter, clay minerals, and water (Wu et al., 2007; Bertrand et al., 2002). The characteristic bands of Zn were less and had good predictability. Since Zn was easily combined with soil components, and the combined product had strong signal characteristics in the spectrum, it can be inferred that several feature bands can effectively improve the accuracy of the fusion model.

### 3.3.5. *Prediction results*

In order to obtain the best spectral estimation results for seven elements, the XRF-MIR model and the XRF model were jointly used. The PLSR predicted contents for As, Cr, Cu, Ni, and Zn were generated by the XRF-MIR model, while those for Pb and Cd were generated by the XRF. Fig. 5 compares the actually measured chemical contents and the predicted PLSR contents with 95% prediction intervals. For low contents, the prediction interval can completely cover the chemical detection contents. Even though the measured chemical contents were located at the boundary of the prediction interval for some high content points, the high prediction accuracy was achieved overall.

## 4. Conclusion

In this paper, the spectral data fusion models, XRF-NIR and XRF-MIR, were proposed for soil elements monitoring, investigating their potential benefits. Various combinations of XRF spectral characteristic bands and MIR and NIR spectra and feature extraction methods were investigated to determine the optimal models for each element. For the feature extraction, the SPA showed the best performance in combination with the XRF-MIR model. Finally, the XRF-MIR model was selected as the prediction models for As, Cr, Cu, Ni, and Zn, while the XRF model was selected for Pb and Cd. The experimental results showed that the data fusion with different sensors effectively improved the accuracy of five elements. When this method is used for the measurement of these 5 elements in the actual soil, the samples can be sent to the laboratory, measured with different instruments, and then used the established model to predict the concentration. By using portable apparatuses, this method can also be used for direct on-site measurements. Future works will include exploring other spectral fusion models for improving the detection accuracy of Pb and Cd.

**Funding**

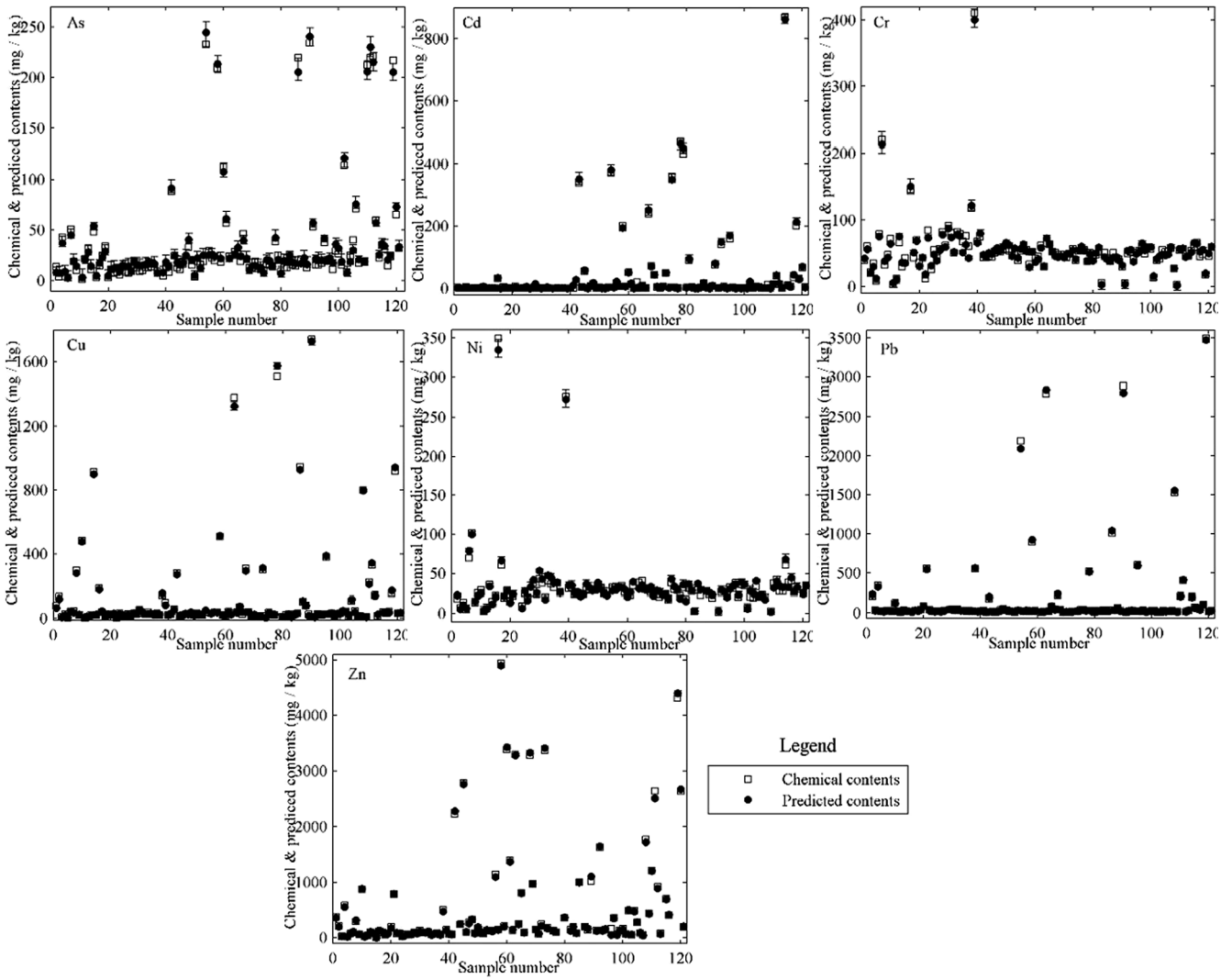


**Fig. 5.** Comparison of the chemical contents measured by ICP-MS and the predicted contents of the PLSR models. Line bars indicate 95% prediction intervals.

This work was financially supported by the Special Projects of Construction of Science and Technology Innovation Ability of Beijing Academy of Agriculture and Forestry Sciences (KJCX20180406).

## CRediT authorship contribution statement

**Fang Li:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Li Xu:** Software, Validation, Investigation, Writing - review & editing, Funding acquisition. **Tianyan You:** Conceptualization, Formal analysis, Writing - review & editing, Project administration. **Anxiang Lu:** Conceptualization, Methodology, Validation, Resources, Writing - review & editing.

## Declaration of Competing Interest

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

## References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2, 433.

Adler, K., Piikki, K., Söderström, M., Eriksson, J., Alshihabi, O., 2020. Predictions of Cu, Zn, and Cd concentrations in soil using portable X-ray fluorescence measurements. Sensors 20, 474.

Al Maliki, A., Bruce, D., Owens, G., 2014. Prediction of lead concentration in soil using reflectance spectroscopy. Environ. Technol. Innovation 1–2, 8.

Alamprese, C., Casale, M., Sinelli, N., Lanteri, S., Casiraghi, E., 2013. Detection of minced beef adulteration with turkey meat by UV–vis, NIR and MIR spectroscopy. LWT - Food Sci. Technol. 53, 225.

Araújo, M.C.U., Saldanha, T.C.B., Galvao, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics Intelligent Laboratory Syst. 57, 65.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc.

Bertrand, I., Janik, L.J., Holloway, R.E., Armstrong, R.D., Mclaughlin, M.J., 2002. The rapid assessment of concentrations and solid phase associations of macro- and micronutrients in alkaline soils by mid-infrared diffuse reflectance spectroscopy. Soil Res. 40.

Caporale, A.G., Adamo, P., Capozzi, F., Langella, G., Terribile, F., Vingiani, S., 2018. Monitoring metal pollution in soils using portable-XRF and conventional laboratory-based techniques: Evaluation of the performance and limitations according to metal properties and sources. Sci. Total Environ. 643, 516.

Chakraborty, S., Weindorf, D.C., Li, B., Ali Aldabaa, A.A., Ghosh, R.K., Paul, S., et al., 2015. Development of a hybrid proximal sensing method for rapid identification of petroleum contaminated soils. Sci. Total Environ. 514, 399.

Cheng, J.-H., Sun, D.-W., 2016. Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. Food Eng. Rev. 9, 36.

Ding, S., Zhu, H., Jia, W., Su, C., 2011. A survey on feature extraction for pattern recognition. Artif. Intell. Rev. 37, 169.

Dong, Y.-W., Yang, S.-Q., Xu, C.-Y., Li, Y.-Z., Bai, W., Fan, Z.-N., et al., 2011. Determination of soil parameters in apple-growing regions by near- and mid-infrared spectroscopy. Pedosphere 21, 591.

Dx, A., Sc, B., Ravr, C., D AB, Sl, E., Yin, Z.A., et al. X-ray fluorescence and visible near infrared sensor fusion for predicting soil chromium content. Geoderma 2019;352:61.
Eusuff, M.M., Lansey, K.E., 2003. Optimization of water distribution network design using the shuffled frog leaping algorithm. J. Water Resour. Plann. Manage. 129, 210.
Fan, Q., Wang, Y., Sun, P., Liu, S., Li, Y., 2010. Discrimination of Ephedra plants with diffuse reflectance FT-NIRS and multivariate analysis. Talanta 80, 1245.
Gredilla, A., Fdez-Ortiz de Vallejuelo, S., Elejoste, N., de Diego, A., Madariaga, J.M., 2016. Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: A review. TrAC Trends Anal. Chem., 76:30.
Grzegorz, S., Mccarty, G.W., Stuczynski, T.I., Reeves, J.B., 2004. Near- and mid-infrared diffuse reflectance spectroscopy for measuring soil metal content. J. Environ. Qual. 33, 2056.
Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A., 2008. Feature extraction: foundations and applications. Springer.
Hu, B., Chen, S., Hu, J., Xia, F., Xu, J., Li, Y., et al., 2017. Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. PLoS ONE 12, e0172438.
Huang, Y., Deng, M., Wu, S., Japenga, J., Li, T., Yang, X., et al., 2018. A modified receptor model for source apportionment of heavy metal pollution in soil. J. Hazard. Mater. 354, 161.
Ji, W., Adamchuk, V.I., Chen, S., Mat Su, A.S., Ismail, A., Gan, Q., et al., 2019. Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. Geoderma 341, 111.
Kabata-Pendias, H.A., Mukherjee, A.B., 2007. Trace Elements from Soil to Human. Springer, Berlin Heidelberg.
Kumar Sharma, R., Agrawal, M., Marshall, F., 2007. Heavy metal contamination of soil and vegetables in suburban areas of Varanasi, India. Ecotoxicol. Environ. Saf. 66, 258.
Li, H., Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta 648, 77.
Malley, D.F., Williams, P.C., 1997. Use of Near-Infrared Reflectance Spectroscopy in Prediction of Heavy Metals in Freshwater Sediment by Their Association with Organic Matter. Environ. Sci. Technol., 31:3461.
Moros, J., Vallejuelo, S.F.-Od, Gredilla, A., Diego, A.d., Madariaga, J.M., Garrigues, S., et al., 2009. Use of reflectance infrared spectroscopy for monitoring the metal content of the estuarine sediments of the Nerbioi-Ibaizabal River (Metropolitan Bilbao, Bay of Biscay, Basque Country). Environ. Sci. Technol., 43:9314.
Nixon, M., Aguado, A., 2019. Feature extraction and image processing for computer vision. Academic Press.
Oliver, M.A., 1997. Soil and human health: a review. Eur. J. Soil Sci. 48, 573.
O'Rourke, S.M., Minasny, B., Holden, N.M., McBratney, A.B., 2016. Synergistic use of Vis-NIR, MIR, and XRF spectroscopy for the determination of soil geochemistry. Soil Sci. Soc. Am. J. 80, 888.
Paulette, L., Man, T., Weindorf, D.C., Person, T., 2015. Rapid assessment of soil and contaminant variability via portable x-ray fluorescence spectroscopy: Copşa Mică. Romania. Geoderma 243–244, 130.
Rouillon, M., Taylor, M.P., 2016. Can field portable X-ray fluorescence (pXRF) produce high quality data for application in environmental contamination research? Environ. Pollut. 214, 255.
Rouillon, M., Taylor, M.P., Dong, C., 2017. Reducing risk and increasing confidence of decision making at a lower cost: In-situ pXRF assessment of metal-contaminated sites. Environ. Pollut. 229, 780.
Śliwińska, A., Smolinski, A., Kucharski, P., 2019. Simultaneous analysis of heavy metal concentration in soil samples. Appl. Sci. 9, 4705.
Song, F., Guo, Z., Mei, D., 2010. Feature selection using principal component analysis. In: 2010 international conference on system science, engineering design and manufacturing informatization: IEEE, p. 27.
Song, Y., Li, F., Yang, Z., Ayoko, G.A., Frost, R.L., Ji, J., 2012. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. Appl. Clay Sci. 64, 75.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M. J., 2013. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49, 139.
Turner, A., Chan, C.C., Brown, M.T., 2018. Application of field-portable-XRF for the determination of trace elements in deciduous leaves from a mine-impacted region. Chemosphere 209, 928.
Turner, A., Taylor, A., 2018. On site determination of trace metals in estuarine sediments by field-portable-XRF. Talanta 190, 498.
Wei, B., Yang, L., 2010. A review of heavy metal contaminations in urban soils, urban road dusts and agricultural soils from China. Microchem. J. 94, 99.
Whitley, D., 1994. A genetic algorithm tutorial. Statistics Comput. 4, 65.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics Intelligent Laboratory Syst. 58, 109.
Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., et al., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71, 918.
Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., Shi, Z., 2020. Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. Environ. Pollut. 263, 114649.
Yang, Q., Li, Z., Lu, X., Duan, Q., Huang, L., Bi, J., 2018. A review of soil heavy metal pollution from industrial and agricultural regions in China: Pollution and risk assessment. Sci. Total Environ. 642, 690.
Zhang, Pengyan, Qin, Chengzhe, Hong, Xin, et al., 2018. Risk assessment and source analysis of soil heavy metal pollution from lower reaches of Yellow River irrigation in China. Sci. Total Environ.