Atomic Spectroscopy

Article

Quantitative Determination of Cd Using Energy Dispersion XRF Based on Gaussian Mixture Clustering-Multilevel Model Recalibration

Zhi Zhang,[a,b,#] Yunbing Gao,[a,*,#] Yanan Zhao,[a] Xiaoyang Liu,[d] Xue Li,[e] Xuefei Mao,[e] Yuchun Pan,[a] Wenbin Sun,[b] and Xiande Zhao [c]

[a] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences. Beijing 100097, P. R. China

[b] College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing 100083, P. R. China

[c] Intelligent Equipment Technology Research Center, Beijing Academy of Agricultural and Forestry Sciences. Beijing 100097, P. R. China

[d] Technical Centre for Soil, Agriculture and Rural Ecology and Environment, Ministry of Ecology and Environment, Beijing 100012, P. R. China

[e] Institute of Quality Standard and Testing Technology for Agro-products, Chinese Academy of Agricultural Sciences, Beijing 100081, P. R. China

Received: *February 28, 2024;* ***Revised:*** *April 30, 2024;* ***Accepted:*** *May 03, 2024;* ***Available online:*** *May 03, 2024.*

DOI: 10.46770/AS.2024.038

ABSTRACT: The analysis accuracy of energy dispersion X-ray fluorescence spectrometry (XRF) for detecting heavy metal in agricultural soils is severely depending on complex matrix effect, thereby posing a challenge in fast and precise monitoring soil contamination. To calibrate the XRF detection, a Gaussian mixture clustering-multilevel model (GMC-MLM) was proposed to enhance XRF accuracy for Cd in agricultural soils. Compared with other models such as multiple linear regression (MLR), random forest regression (RF), and support vector machine regression (SVMR), the GMC-MLM effectively disentangled the nested distribution of XRF detection errors. The correlation coefficient between the XRF detection results and ICP-MS test results for the corrected samples can reach 0.9085, with 74% of the corrected samples having a relative error of less than 30%. Notably, according to the GMC-MLM correction method, a knowledge base for localizing corrections in XRF detection has been constructed. When the number of knowledge base sample points is 50, the RMSE (Root Mean Squared Error), and REM (Relative Error of Mean) are 0.7347, 3.7014%, respectively. It can be observed that the model has good extrapolation capability, and with the increase in the number of knowledge base sample points, the correction effect based on the knowledge base gradually stabilizes. This knowledge base-based GMC-MLM calibration method not only can be embedded into XRF detection instruments to correct XRF detection results in different regions of China but also provides theoretical support for the establishment of a nationwide soil sample knowledge base.

INTRODUCTION

Soil heavy metal pollution has emerged as a global environmental and public health concern, and enter the food chain through crop enrichment, posing a threat to human health.[1,2] Soil heavy metal monitoring and surveys are crucial for assessing the environmental quality of agricultural production areas. However, soil heavy metal pollution exhibits spatial heterogeneity, and the high cost associated with traditional soil testing and analysis methods leads

to sparse sampling in monitoring surveys, which is insufficient for precise risk management and efficient remediation. There is an urgent need to develop low-cost methods for soil heavy metal monitoring and surveys. Energy dispersive X-ray fluorescence (EDXRF) is an efficient, nondestructive, and cost-effective analytical method for the detection of multiple heavy metal elements in soil via X-rays. This method has been extensively applied in rapid on-site screening of soil, indicating a promising market outlook. However, these detectors are susceptible to soil matrix effects, leading to low repeatability and accuracy of the results.[3,4,5,6,7] The accuracy of EDXRF detectors limits their widespread adoption in moderate and light soil heavy metal monitoring surveys. Improving the accuracy of XRF detection is of practical significance, particularly for extensive utilization in environmental monitoring surveys within agricultural production areas.

Currently, there are empirical coefficient method, fundamental parameter method, and theoretical alpha coefficient method to achieve quantitative calculation of XRF.[8,9,10,11] The empirical coefficient method, reliant on both the quantity and quality of samples, has the problem of low accuracy because of the complexity and diversity of soil samples.[12] The fundamental parameter method and theoretical alpha coefficient method which considers the principles of X-ray interactions with matter, yields comprehensive mathematical models to correct the XRF results.[13] However, in practice, it is difficult to apply directly to the quantitative analysis of soil because of the heterogeneity of soil and the high content of C, H, O elements in soil samples.[29] With the rapid advancement of spectral deconvolution techniques, machine learning models have been incorporated into quantitative analysis using X-ray fluorescence detection. The Sheibani employed stationary wavelet denoising to address noise issues in XRF detection spectral signals, thereby enhancing the accuracy of XRF quantitative analysis.[14] Additionally, discrete wavelet models have been utilized to correct the relationship between XRF spectral peaks and soil heavy metal concentrations.[15] Furthermore, the Compton compensation method has been adopted to correct the absorption of fluorescence X-rays by soil moisture, enabling quantitative analysis of soil heavy metal elements.[14] The aforementioned quantitative method involves quantitative calculation of XRF detection results through various approaches, including reducing background noise, fitting characteristic peaks of the element under investigation, and correcting these peaks based on influencing factors. However, the content of heavy metals in soil exhibits different gradient distributions influenced by soil type and parent material. Additionally, the complex soil structure and migration are affected by soil texture and organic matter content. Therefore, the quantitative methods mentioned above have demonstrated persisted residuals in practical applications. To enhance the accuracy of XRF detection data, a range of external calibration methods tailored for laboratory-based instrumentation have been introduced. Researchers such as Lei *et al.* explored the comparability of XRF continuous scanning results for rock cores with inductively coupled plasma–optical emission spectrometry (ICP–OES) analysis results. They used the scanning intensity of elemental Compton scattering as a surrogate indicator for the sediment moisture content, facilitating the correction of scanning results for elements such as K, Ca, Ti, and Fe.[17] Poto *et al.* employed regression models to integrate XRF with ICP-MS techniques, resulting in the calibration of XRF results.[18] Gregory utilized ICP-MS test data from samples and employed variables such as water content and element concentration to calibrate XRF test results.[19] The abovementioned studies focused primarily on analyzing the impact of single factors on the XRF measurement error. Few studies have provided a holistic view of the interactions between complex matrices, such as the soil parent material and soil type, and their effects on XRF measurements. Furthermore, the intricate nature of soil precludes the existence of a universal set of parameters and models that can comprehensively address XRF detection errors across different research regions. This limitation restricts the exploitation of the unique advantages and application potential of this advanced XRF technique, hindering its widespread utilization.

Based on the analysis presented above, this study considers the interactive effects of organic matter, soil type, parent material, and texture during XRF detection. Using a Gaussian Mixture Clustering Model (GMC) to pre-classify the sample detection results, we adopt laboratory measurements as the true values and employ a multi-level modeling (MLM) approach for integrated modeling. This approach aims to further eliminate residual errors in the quantitative analysis of instrumental detection. Additionally, to address the complexity of soil matrices in different regions, we propose a technical method for constructing localized knowledge bases to address the inconsistency of modeling parameters across regions. Our objective is to provide technical references for the widespread application of EDXRF.

EXPERIMENTAL

Sample preparation. The research area is located at the central part of China, encompassing approximately 3,855km^2 of cultivated land, which accounts for approximately 20% of the total land area. The area is divided into paddy fields, irrigated land, and dryland, covering 1349 km^2, 194 km^2, and 2312 km^2 respectively. The primary crops cultivated in this region include rice, corn, oilseed rape, and fruits. The predominant topography is hilly terrain, encompassing four types of zonal soils, namely, yellow soil, yellow–brown soil, brown soil, and red soil, and four types of nonzonal soils, namely, purple soil, lime (rocky) soil, hydric soil, and paddy soil. Yellow soil, yellow–brown soil, and lime (rocky) soil are the most common in terms of area, accounting for approximately 60% of the total area, and the parent material of the

Fig. 1 Sampling sites.

soil-forming materials mainly includes sedimentary rocks, such as limestone, sandy conglomerate, and dacite as well as medium-acidity intrusive rocks. The eastern-central region mostly comprises flood-alluvial deposits, and the northwestern region is dominated by the coexisting parent materials of carbonate and intrusive rocks. The complex parent material characteristics result in varying degrees of metal element enrichment in the soil, leading to elemental content differences. The soil texture is predominantly loamy clay, with an organic matter content ranging from 2% to 10%. The sampling points were uniformly distributed across paddy fields, irrigated land, and rainfed soil.

In January 2019, a grid-based sampling method was adopted to collect soil samples from paddy fields and dryland areas within the study area (Fig. 1). A total of 350 soil samples were collected, with approximately 1 kg of soil collected at each sampling point. These samples were obtained from the surface soil layer (0-20 cm), and their locations were determined via GPS technology. Five subsamples uniformly distributed across a 10 m × 10 m representative open space were mixed to form a composite soil sample, and the collected samples were packed in polyethylene sample bags and sent to the laboratory, where they were naturally dried at room temperature under dry and ventilated conditions. Stones and other debris in the samples were removed, and after passing through a 100-mesh sieve (<0.15 mm particle size), the samples were thoroughly mixed and divided into two equal portions for separate ICP-MS testing and XRF detection.

Analytical Methods. The ICP-MS (Agilent, 7700X) method is as follows (Soil Waste – Determination of metals – Inductively coupled plasma mass spectrometry (ICP-MS) HJ 766-2015): Weighing 100mg of the sample with one over Ten - thousand analytical balance, followed by the addition of a small amount of purified water (Millipore, Billerica, MA, USA) for moistening. The sample is then placed into a fully automated graphite digestion instrument (DEENA II), and undergoes a digestion process through the application of heat, utilizing a mixture of hydrochloric acid (HCl, ρ =1.19g/ml), nitric acid (HNO3, ρ=1.42g/ml), and hydrofluoric acid (HF, ρ=1.49g/ml) in a ratio of 3:1:1. The concentration of Cd in the digested solution was determined using inductively coupled plasma mass spectrometry (ICP-MS). The laboratory test result was calculated as the average of three measurements. To guarantee the quality of the measurements, standard materials GSS-1 and GSS-4 were incorporated into the laboratory's analytical quality assurance and quality control procedures.

The other portion of the soil samples (300 mg) was analyzed by an NX-200S spectrometer (EDXRF, with a detector resolution up to 125 eV, the detection limit for Cd was determined to be 0.16 mg kg^{-1}.). For this analysis, 300 milligrams of soil samples were placed in plastic cups, spread evenly over a degreased cotton substrate, compacted, covered with a bottom lid to ensure a flat measurement surface, and prepared as samples, the thickness of the sample should be impenetrable to X-rays. The probe window was aligned vertically with the surface of the thin film test, select soil as the measurement type, and the determination time was set to 350 seconds. To guarantee the precision and stability of the test outcomes, standard materials GSS-1 and GSS-4 were incorporated into the XRF analytical quality assurance and quality control procedures. Concurrently, each test sample underwent testing three separate times, and the result was determined by calculating the mean of these measurements.

Gaussian mixture clustering-multilevel model construction. The GMC-MLM is an XRF detection error correction methods,

which are constructed by hierarchical segmentation using the Gaussian mixture algorithm integrated with a multilevel model. The GMC model, a composite model of multiple Gaussian distribution probability density functions. It enables pre-classification of samples based on the detected content at monitoring points and environmental variable indicators, achieving multi-level categorization. The GMC model effectively addresses detection errors arising from the mixing of soil environmental factors within pre-classified samples. Specifically, the MLM employs random slopes to account for the intricate interactions among various soil environmental factors, enabling modeling under the cumulative influence of multiple variables. Therefore, introducing Gaussian clustering based on a multilevel model not only addresses the limitations of traditional statistical methods in analyzing multilevel data but also captures the complexity of mixed soil environmental variables.

When the considered elements exhibit high concentrations during XRF detection, low detection errors occur due to the substantial spectral peak intensity and relatively minor impact of soil matrix effects.[20] However, for elements with low concentrations, the influences of both soil matrix phenomena and soil background values become significant, leading to issues such as low concentration-high detection error and high concentration-low detection error phenomena. Moreover, under similar error conditions, there is a certain degree of similarity among soil environments. To address this phenomenon, the GMC-MLM was employed to categorize samples based on the concentration–error-soil environment relationships detected via XRF. The XRF detection error can be calculated by Equation (1). The soil environmental similarity can be expressed as Equation (2). The algorithm calls the 'GaussianMixture' function in Python, and the parameters of GMC are estimated by expectation maximization EM.

$$I_n(x)=(F_x - C_x)/C_x \quad (1)$$

$$\mathrm{T_S} = 3 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \left[\sum_{k'}^{t} I_{\{x_{i,k'} \neq y_{j,k'}\}}\right]^{1/2}}{n} \quad (2)$$

Where $I_n(x)$ is the relative error of XRF detection, C_x is the ICP-MS test result of sample point x, and F_x is the XRF test result of sample point x. A positive value of $I_n(x)$ indicates that the XRF detection result is higher than the ICP-MS test result (truth value), while a negative value of $I_n(x)$ indicates that the XRF detection result is lower than the ICP-MS test result (truth value). $\mathrm{T_S}$ represents the quantitative attribute (soil type, parent material, texture) in environmental factors in similarity calculation, $x_{i,k}$ represents the k-th attribute value of the i-th point, $y_{j,k}$ represents the k-th attribute value of the j-th point, except for sample i.

When the dependent variable (Y) of the model is the XRF detection error (as shown in the equation 5) and the independent variables are the XRF detection results and soil environmental variables (X), the model can be represented as a linear model, as expressed in Equation (3).

$$Y = b_0 + \beta_1 X1 + \beta_2 X2 + \cdots + \beta_n Xn + \varepsilon \quad (3)$$

Where b_0 represents the intercept, $\beta_1, \beta_2, \ldots \beta_n$ stand for the regression coefficients, Xn denotes dependent variable, and ε represents the random error.

Due to the interdependence and interactions among the factors influencing XRF detection, the construction of a linear model that ignores the correlations among the data is inaccurate. Therefore, in order to correct the errors of XRF, considering the soil sample as level 1 (the differences among individual soil samples) and category division by the Gaussian clustering model as level 2 (the differences among different categories of samples affect the results of individual soil samples), with varying intercepts and slopes across levels, the introduction of random variables on top of the linear model is needed, as expressed in Equation (4), the algorithm is implemented by lme4, R language.

$$\begin{cases} Y = b_0 + (\beta_1 + u_{1j})X1_{ij} + (\beta_2 + u_{2j})X2_{ij} + \cdots + (\beta_n + u_{nj})Xn_{ij} + u_{0j} + \varepsilon_{0ij} \\ u_{0j} \sim N(0, \sigma_{u0}^2) \\ \varepsilon_{0ij} \sim N(0, \sigma_{\varepsilon}^2) \\ i = 1, 2, 3, \ldots, n; j = 1, 2, 3, \ldots, m \end{cases} \quad (4)$$

$$Y = \triangle_{xrf-icpms} = ln(f_{xrf}) - ln(f_{icpms}) \quad (5)$$

Where i represents a unit at Level 1, j denotes a unit at Level 2, b_0 signifies the overall mean intercept for Level 2 units, representing the average intercept across all units at this level, u_{0j} represents the random intercept, which captures the variation in the intercept for each specific sample or category from the overall mean. β_n represents the overall mean slope for Level 2 units, u_{nj} represents the random slope, signifying the difference in slope for each category or sample point in comparison to the overall mean slope.

RESULTS AND DISCUSSION

Descriptive statistics of the soil sampling points. In this study, we selected soil cadmium (Cd) with large detection errors from X-ray fluorescence (XRF) as the research object. The original sampling points were subjected to outlier removal using three times the standard deviation as a criterion, resulting in 249 soil sampling points with Cd concentration data. Descriptive statistical analysis was performed of these data to elucidate the characteristics of the heavy metal content in the soil samples. ICP-MS analysis and XRF detection are provided in Table 1. The average Cd concentration detected using ICP-MS was 0.776 mg kg^{-1}, with a skewness coefficient of 16.65, a kurtosis coefficient of 3.86, and a coefficient of variation of 129.01%. The average concentration of Cd using XRF detection (NX-200S spectrometer)

Table 1. Descriptive statistics of the Cd content at the soil sampling sites (mg kg^{-1})

Properties	Number	Minimum	Maximum	Mean	Standard deviation (SD)	Kurtosis	Skewness	CV (%)
ED-XRF	249	0.176	6.871	1.051	1.660	13.6155	3.4578	157.85
ICP-MS	249	0.160	9.020	0.776	1.001	16.6494	3.8569	129.01

Fig. 2 Distribution of the XRF detection errors in different soil environments. (a), (b), (c) and (d) represent XRF detection errors for different soil parent materials, soil types, soil textures, and organic matter content, respectively.

was 1.051 mg kg^{-1}, with a skewness coefficient of 13.61, a kurtosis coefficient of 3.46, and a coefficient of variation of 157.85%. The data in indicated that, compared to the ICP-MS test results (truth value), the XRF detection method exhibited a degree of error. Additionally, the coefficient of variation by the XRF detection was significantly higher than that under the ICP-MS test. The results of portable XRF detection are influenced by the soil environment (including soil texture, soil organic matter, soil parent material, and soil type). Therefore, the XRF detection results should be calibrated based on the soil environment.

Influence analysis of environment variables on XRF test results. When XRF detection is used to assess the content of target pollutants, XRF detection error is influenced by various environmental factors, including the soil type, parent material type, soil organic matter, and soil texture.[21,22,23] Studies have suggested that the soil parent material type can effectively reflect the elemental characteristics of soil, particularly the elemental content in deep soil layers. Among the different parent material types, there are significant differences in the background concentrations of soil heavy metals.[24] For instance, in areas with black shale, Cd and As exhibit particularly high background concentrations, while areas with basalt exhibit high background values of Cu and Ni. The soil type is developed on the foundation of the parent material and is influenced by various factors, such as clarification and iron-aluminum enrichment. Studies have demonstrated that different soil types exhibit distinct elemental compositions.[25,26] For instance, red soils are characterized by a substantial presence of iron (Fe), while brown soils contain silicon and aluminum. The enhanced absorption effects between these elements and the target elements could introduce variability in the XRF detection results. In addition to the aforementioned influencing factors, variations in the sample texture during XRF irradiation could increase the scattering of X-rays, subsequently affecting the intensity of the obtained elemental signals.[27,28] Soils with higher organic matter contents are enriched in light elements such as C, H, and O, altering the sample density and effective volume and leading to X-ray scattering and attenuation.[29,30,31,32]

The impacts of the aforementioned environmental variables on the XRF detection errors were analyzed separately in this study, and the results are shown in Fig. 2. Research has revealed that detection

Fig. 3 Gaussian mixture model clustering chart (a), multilevel distribution of the XRF detection results and ICP-MS test results for the soil sample points (b). Grey, red, blue, and green represent sample points of LP (A), HP (B), LN (C), and HN (D) categories, respectively.

errors are larger when the parent material types are impure carbonates or Quaternary deposits. This difference is attributed mainly to the secondary enrichment process during mineral formation in certain rocks with inherently low metal contents.[33] This process leads to anomalous enrichment of elements such as Cd, Pb, and As, resulting in the absorption-enhancing effects of high concentrations of other heavy metal elements on the target elements, thereby affecting the detection of the target elements by XRF. When the soil types are burozem or brown soil, the XRF detection errors are significantly increased. This is primarily due to the higher contents of organic matter, nitrogen, phosphorus, and potassium in the surface layers of these soil types. The detection of these light elements using X-rays could cause interference, thereby affecting XRF detection of Cd. Through the study of the soil texture, it has been observed that when the soil texture is intrusive rock, the error is larger than that for other soil textures. Additionally, analysis of the sand content in intrusive rocks has revealed that a higher sand content corresponds to larger detection errors. The analysis of the impact of the organic matter content on XRF detection indicated that when the organic matter content is low, the impact on XRF detection is relatively minimal. However, with increasing organic matter content, the impact gradually intensifies. Once the organic matter content reaches a certain level, the impact on XRF detection decreases. These findings conform with the principles of XRF detection. However, spatial heterogeneity in the distributions of soil types, parent materials, textures, and organic matter introduces complexity. The interactions between these variables indicate that a single environmental factor cannot fully explain the causes of XRF detection errors.[34] For example, in the central and southern regions of China, characterized by abundant precipitation and natural coniferous and shrub vegetation, the organic matter content is significantly higher than that in the eastern regions. This ecological setting contributes to the development of soil types such as yellow soil and lime stone. The complex matrix characteristics arise from the interplay between soil types and organic matter. Therefore, considering the impact of soil–environmental factor interactions on XRF detection is important, given the complex and variable nature of these factors.

Construction and validation of Gaussian mixture clustering-multilevel models. The distribution of the XRF detection errors of the soil sample points exhibits nesting construction, and the generation of detection errors is influenced by the confounding effects of soil environment al factors. Therefore, the Gaussian Mixture Clustering Model is adopted to cluster the soil sample point data reasonably, and a Multilevel modeling approach is employed to comprehensively consider the correction modeling of XRF results under the influence of different sample categories and environmental factors. When detecting low concentrations of analytes, the excited X-ray fluorescence is relatively weak, resulting in a distribution pattern of low concentration with high errors and high concentration with low errors. Additionally, soils with high organic matter content absorb and scatter X-rays, reducing the penetration depth of the XRF signal and leading to negative errors.[29] On the other hand, an increase in soil sample particle size results in an increase in fluorescence intensity, presenting positive errors.[35] Consequently, the concentration of the target substance exhibited a distribution pattern of high concentration-low detection error, low concentration-high detection error (as shown in Fig. 3) and homogeneous errors - similar soil environments were observed.

Following this principle, we established the GMC-MLM by dividing the data from 249 soil samples into training and testing sets at 150 and 99. Firstly, the XRF detection results and XRF detection errors (as defined in Equation 1) were used as input data. The EM algorithm was employed to iteratively solve for the unknown

Table 2. Comparison of the GMC-MLM correction results before and after correction for the different sample categories

Zone	Mean (ICP-MS Cd)	SD (ICP-MS Cd)	Mean (XRF Cd (before))	SD (XRF Cd (before))	Mean (XRF Cd (after))	SD (XRF Cd (after))	Mean ($I_n(x)$)
LP	0.2681	0.0872	0.4979	0.2269	0.2718	0.0884	0.8278
HP	0.7573	0.7556	1.0374	1.1146	0.7857	0.7189	0.3245
LN	1.2430	1.5407	0.3173	0.1253	0.9973	0.6695	-0.5146
HN	2.3991	2.6847	1.5165	1.6061	2.4853	3.0452	-0.3006

Note: $I_n(x)$ is shown in Equation 1.

parameters of the Gaussian Mixture Model until the sub-datasets satisfied the maximum probability of normal distribution. Subsequently, environmental factors such as soil parent material, soil type, soil texture, and organic matter content were considered. The positive and negative errors in the XRF detection results of soil samples were classified based on environmental similarity, as shown in Equation 2. Ultimately, the soil samples were categorized into four types: low concentration with positive errors (LP), low concentration with negative errors (LN), high concentration with positive errors (HP), and high concentration with negative errors (HN). Secondly, based on the exploration of environmental variables, four environmental variables, namely, the soil type, parent material type, soil organic matter content, and soil texture, were selected and their interaction were analyzed. A random slope effect model was employed to calibrate the XRF results.

The number of soil samplings in the LP, HP, LN, and HN is 83, 53, 62, and 51, respectively. Types LP and LN exhibited relatively low mean concentrations of the target substance under both detection methods, at 0.4979 and 0.3173 mg/kg, respectively, with relatively large mean relative errors of 0.8278 and -0.5146, respectively. Conversely, types HP and HP demonstrated relatively high mean concentrations of the target substance under both detection methods, at 1.0374 and 1.5165 mg/kg, respectively, with smaller mean relative errors of 0.3245 and -0.3006, respectively. The intercepts and slopes of the different error categories exhibited varying linear relationships between the ICP-MS test results and XRF detection results, as shown in Fig. 3.

The interaction effects of the soil type, parent material type, soil organic matter content, and soil texture, were assessed using the geographic detector model. The findings indicated variations in the interaction results among the different clustering points. In the low concentration-high error category (LP and LN clusters), there was an enhanced interaction effect between organic matter and the parent material type, between the soil type and parent material, and between the soil texture and organic matter. This suggests that, during XRF testing of soil samples with low concentrations and high errors, the factors influencing the detection errors are more notably associated with the interaction between the soil type, parent material type, and organic matter. In the high concentration-low error category (HP and HN clusters), the interaction effect of organic matter with the other soil environmental variables mainly exhibited a diminishing interactive effect. Notably, at high concentrations, the influence of the organic matter content on XRF detection was not consistent with that of the other environmental variables. The single variables exhibited relatively low explanatory power for the XRF detection errors, indicating a lower dependency of the XRF detection errors on the individual variables. However, when the soil environmental variables interact, the explanatory power significantly increases. This finding suggested that within the context of complex matrices, the soil particle size, soil element type, and organic matter content interact synergistically, collectively influencing the XRF detection errors. Therefore, removing matrix effect errors in XRF rapid detection instruments without considering the localized soil environment impact can be difficult and challenging. Therefore, the GMC was utilized to classify the soil samples, while the MLM was employed to integrate the GMM-MLM framework with the interaction of environmental variables. The R^2 value of 0.758 and the AIC value of 288.42 demonstrated that the GMC-MLM model exhibited a satisfactory fitting effect, indicating its potential for calibrating XRF test results.

To validate the correction effect of the GMC-MLM across different sample categories, the ICP-MS test results, and mean values of the XRF detection results before and after model correction were separately analyzed for each category, as listed in Table 2. The investigation revealed that, at varying concentrations of the target substance, the XRF results exhibited a marked proximity to the ICP-MS test results following GMC-MLM correction. The soil sample detection results, under the different error conditions, closely agreed with the ICP-MS test results, with greater improvement in the correction results at lower concentrations of the target substance than at higher concentrations.

The correction effect of the model was verified by utilizing kriging interpolation in simulating the spatial distribution, which involves comparing the spatial trends of the postcorrection heavy metal concentrations with those observed in the ICP-MS tests.[36,37] Spatial simulations were conducted using Cd ICP-MS results, XRF detection results, and three-dimensional stereograms of the spatial distribution after GMC-MLM correction (Fig. 4). Notably, the spatial distribution trend of the results after GMC-MLM correction was the closest to that of the observed values, but the correction effect is poor for samples with low XRF detection values (Area b in Fig. 4), and good for samples significantly above

Fig. 4 Simulated spatial distribution of soil heavy metal Cd via kriging interpolation for different results.

Table 3. Comparison of the calibration results for the different models

Model	R^2	MAE	RMSE	*REM*	RED	RES
MLR	0.698	0.541	1.157	0.2261	0.2843	0.0971
GMC-MLM	0.766	0.283	0.753	0.1051	0.0986	0.0672
RF	0.472	0.508	1.198	02704	0.3570	0.0255
SVMR	0.431	0.549	1.216	0.3589	0.2104	0.1229

the quantification limit (Area b in Fig. 4). The XRF detection results showed overall underestimation in the region, with inaccuracy for both lower and higher values. The spatial distribution of the results after model correction exhibited a high degree of similarity with the ICP-MS test results, indicating elevated Cd concentrations in the central and western regions and lower Cd concentrations in the eastern region. This model effectively corrected the XRF detection results across different concentration gradients by considering the soil environmental conditions and yielding global results that were closer to the ICP-MS test results.[7,38,38]

In addition, according to GB/T 15618-2018 (Soil Environment Quality Risk Control Standard for Soil Contamination of Agricultural Land), which specifies screening and control values for the soil pollution risk in agricultural land, soil samples with heavy metal content ranging from 0.3 to 2.0 mg/kg are categorized as the risk screening zone, while those exceeding 2.0 mg/kg are designated as the risk control zone. Comparison between the XRF and ICP-MS Cd test results revealed that XRF detection exhibited misclassification of 94 points in the soil pollution risk screening, accounting for 37.75% of the total sample points. Additionally, 24 points in the risk control zone were incorrectly identified as noncontrol zone points, accounting for 9.64% of the total sample points. After correction using the GMC-MLM, the XRF detection results revealed 32 misclassified points in the soil pollution risk screening zones, 4 points in the risk control zone were incorrectly identified as noncontrol zone points. After the correction of XRF results, the misclassification rate of sample points has significantly decreased from 47.39% to 14.46%. Nonetheless, given the constraints imposed by the accuracy of auxiliary variables in the study area, there remains potential for further enhancing the correction outcomes. Therefore, the use of corrected XRF detection results could reduce the possibility of misjudging the environmental quality, providing a lower-cost and more accurate detection method for categorizing agricultural land soil environments and soil pollution remediation.

Analysis of the performance of the correction by models. The XRF detection results of the 249 soil sampling points were corrected using four models, namely, Gaussian mixture clustering-multilevel models (GMC-MLM), Multiple Linear Regression (MLR), Random Forest (RF), and Support Vector Machine Regression (SVMR). The training and testing sets were divided in a ratio of 6:4 (150 points and 99 points), with soil type, parent material type, soil texture, and organic matter content serving as input variables. To compare the performance levels of the four models, we used MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), REM (Relative Error of Mean), RED (Relative Error of Dispersion Coefficient), and RES (Residual Error of Skewness) as evaluation metrics. RMSE and MAE assess the fitting accuracy of the model by quantifying the mean error and the square root of the error, respectively, between the model's predicted values and the dependent variable. RME, RED, and RES evaluate the stability of the model's predictions from distinct perspectives: the overall predictive trend, the dispersion level of the data distribution, and the shape characteristics of the data distribution, respectively. The results are summarized in Table 3. The RF model is constructed by Python's 'Randomforestregressor' function, in which two important parameters, 'mtry' and 'ntree', are matched by 'RandomizedSearchCV' to get the best parameters. The SVMR model is implemented as a "SVM" in Python's "Scikit-learn" library. In this model, the radial kernel function is selected, and the gamma and cost values are adjusted to optimize

Fig. 5 Distribution of the XRF detection results after correction by the different models compared with the ICP-MS test results, with the dashed line representing the equality line between the actual and corrected XRF values. Panels (a), (b), (c), (d), and (e) respectively depict the relationships between the uncorrected XRF detection results and ICP-MS results, as well as the distributions after correction using MLR, GMC-MLM, RF, and SVMR models.

the result. The MAE values of the linear models MLR and GMC-MLM were 0.541 and 0.255 mg/kg, respectively. The RMSE values were 1.157 mg/kg for the MLR model and 0.751 mg/kg for the GMC-MLM. The MAE values of the nonlinear RF and SVMR models were 0.549 and 0.508 mg/kg, respectively, with RMSE values of 1.216 and 1.198 mg/kg, respectively. The adoption of GMC-MLM yielded a correlation coefficient of 0.9085 between XRF detections and ICP-MS results. This represents a significant improvement in detection accuracy compared to the original correlation of 0.7390 between raw XRF data and ICP-MS measurements. Furthermore, the REM, RED, and RES of GMC-MLM were notably lower than the MLR, reflecting enhanced model stability. This superiority was particularly evident when contrasted with the MLR model, which does not consider hierarchical structure or environmental confounding effects. Regarding the nonlinear models, the RF model, which leverages multiple decision trees for regression analysis, exhibited an MAE of 0.508 and an RMSE of 1.198. The calibration effect surpassed that of the SVM, with smaller REM and RES values indicating enhanced stability. Thus, the GMC-MLM exhibits not only high fitting accuracy but also effective calibration, coupled with low REM and RES values, which collectively indicate its excellent stability. These distinctive features clearly demonstrate the superior fitting performance of the GMC-MLM.

In summary, compared with the other three approaches, the GMC-MLM can address the aggregation phenomenon of the XRF detection error distribution of soil sample data through the GMC algorithm to achieve a reasonable hierarchical structure. Moreover, it can capture the interactions between environmental variables that cannot be represented by the linear model and machine learning models, effectively improve the accuracy of XRF detection, and the corrected XRF detection results are closest to the actual measurement data (as shown in Fig. 5). This method can be used for correcting soil heavy metal contents with hierarchical structures and for improving the accuracy of XRF detection under the interactive and complex effects of the soil environment.

Extrapolation capability analysis of Localized Knowledge Repository. Given the significant variations in soil influencing factors, including parent material type, soil type, and texture, across different regions, it is insufficient for XRF instrument manufacturers to rely solely on standard substances to improve accuracy. Soil environmental information exhibits regional specificity and relative stability, remaining generally unchanged unless subject to abrupt factors such as soil replacement. Therefore, this study proposes the establishment of a localized knowledge base that stores accurate laboratory values, environmental factors, and XRF detections from existing sample points. This knowledge

Table 4. Descriptive Statistics of Soil Sampling Points Based on Knowledge Base and Correction Results of GMC-MLM Model

Number	Descriptive Statistics of Knowledge Base Sample Points					Validation Results of Model Extrapolation Capability				
	(Mean$_{XRF}$, Mean$_{ICM-MS}$)	(SD$_{XRF}$, SD$_{ICM-MS}$)	(Kurtosis$_{XRF}$, Kurtosis$_{ICM-MS}$)	(Skewness$_{XRF}$, Skewness$_{ICM-MS}$)	(CV $_{XRF}$, CV $_{ICM-MS}$) (%)	MAE	RMSE	*REM*	RED	RES
50	(0.8539,1.4296)	(1.0356,2.2934)	(13.0754,7.7003)	(3.3018,2.7326)	(85.39,142.96)	0.7347	1.352	3.7014%	5.2514%	37.43%
75	(0.8097,1.2583)	(1.0009,2.0254)	(13.5306,10.3378)	(3.4246,3.0622)	(123.62,160.96)	0.3518	0.8312	0.9418%	4.7678%	31.26%
100	(0.8075,1.0439)	(0.9391,1.3641)	(10.3806,10.2850)	(3.0235,2.9048)	(116.30,130.67)	0.3308	0.8578	2.3982%	5.1085%	31.35%
125	(0.8482,1.1418)	(1.1003,1.7852)	(15.9740,13.6503)	(3.7572,3.4407)	(129.72,156.34)	0.3126	0.7933	1.3034%	0.0654%	34.05%
150	(0.7415,1.0433)	(0.9170,1.5803)	(15.5785,10.9209)	(3.6637,3.1460)	(123.67,151.47)	0.3223	0.8138	1.6292%	9.96%	28.27%

base can serve as a source of prior knowledge for subsequent detections, enhancing detection accuracy and reducing errors. According to the aforementioned GMC-MLM, a knowledge base for XRF detection localization correction can be constructed. The knowledge entities include sample attributes ($V_{XRF}, V_{Lab}, V_{Env}$), instrument parameters (P_{xrf}, P_{Lab}), and study regions (region). The entity relationship is represented by the model-based XRF detection result correction method ($f()$), as shown in equation (6). Generally, it is required that the localization knowledge repository should have no fewer than 50 records. By dynamically increasing a certain number of sample pairs and soil environmental variables, the re-calculation of model parameters and knowledge base updates can be achieved. Through validating the extrapolation ability of the GMC-MLM model based on this localization knowledge repository, the model can be reused and promoted in different study areas. It can also provide theoretical support for the more accurate calibration of XRF rapid detection results based on different XRF instruments (ED-XRF) and laboratory detection methods (ICP-MS, AAS, *etc.*).

$$\mathrm{R}_{region}:\{\mathrm{V}=(V_{XRF},V_{Lab},V_{Env})^{T},P=(P_{xrf},P_{Lab}),region,f(V)\} \quad (6)$$

Where R_{region} denotes the localized knowledge repository of the research area, V represents the attribute values of the sample points, including XRF detection results (V_{XRF}), ICP-MS test results (V_{Lab}), and soil environmental variables of the sample points (V_{Env}). P_{xrf} and P_{Lab} are the parameters of the XRF detection instrument and ICP-MS test instrument, respectively. The region represents the applicable area of the knowledge repository, $f()$represents the calibration model, and this article is the GMC-MLM model.

In the study area of 249 XRF and ICP-MS tested soil sampling points, randomly selecting 50, 75, 100, 125, and 150, a total of 5 groups of soil sampling points as the localization knowledge base detection value sample pairs, using GMC-MLM + corresponding parameters as the XRF soil sample detection value correction framework, its extrapolation ability and accuracy improvement are verified, with verification set quantities of 199, 174, 149, 124, and 99, respectively. The selected knowledge base sample points need to fully consider the soil environment and spatial distribution, ensuring coverage of the concentration and error category range of the target substance in the study area. Descriptive statistics of different localization knowledge base sample points and model extrapolation ability results are shown in Table 4, using MAE, RMSE, REM, RED, RES, *etc.*, as accuracy verification indicators. When the knowledge base sample points are 50, RMSE, MAE, REM are 0.7347, 1.352, and 3.7014%, respectively, indicating that the corrected XRF detection results have a small error compared to the ICP-MS test results. As the number of knowledge base sample points dynamically increases, the post-correction RMSE and MAE gradually decrease, and the XRF post-correction results gradually approach the ICP-MS test results. When the number of sample points in the knowledge base reaches 75, the model fitting effect is relatively good, and the correction effect tends to be stable. It can be seen that the proposed method based on Gaussian Mixture-Multi-Level model correction and localization knowledge base has correction extrapolation ability, which can be used for other batches of XRF on-site sampling and detection in the local area. With the increase of knowledge base sample points, the extrapolation ability of the GMC-MLM model continues to strengthen. Therefore, the establishment of a soil environmental localization knowledge base is crucial. Embedding this knowledge base into XRF rapid detection instruments and ensuring the dynamic update of soil detection sample points provide a guarantee for the construction of XRF detection result correction models, greatly improving the accuracy of XRF detection.

Analysis of the model applicability based on environmental similarity. The analysis of the similarity of soil environmental variables across different sample categories not only validates the feasibility of the GMC model for sample classification, which demonstrates that XRF detection errors among similar samples share similar soil environments, but also proves the usability of the soil sample knowledge base constructed in this study. The similarity function T_S was employed to quantify the environmental similarity between the LP, LN, HP, and HN points. Considering the conditions in the study area, the organic matter content was graded based on the ranges of 0-6 g/kg and 6-10 g/kg.

Fig. 6 Calculated environmental similarity among soil samples with different XRF detection errors. LP (A), HP (B), LN (C), and HN (D) represent four error categories. The black borders delineate the extent of similarity in soil environmental variables within the same error category.

Additionally, similarity calculations were performed based on the soil texture, parent material type, and soil type classification standards, as expressed in Equation 2.

By computing the similarity of soil environmental variables across different categories of errors, it is evident that a higher level of environmental similarity corresponds to potentially similar causes and magnitudes of detection errors. The results indicate that soil samples belonging to the same error category exhibit greater soil environmental similarity compared to those belonging to different error categories. Specifically, a value closer to 3 signifies a higher degree of similarity. Using the similarity calculation equation, the soil environmental similarity among LP-category samples is found to be 1.7158, while the environmental similarity between HN-category soil samples is 2.1014. The environmental similarity between HN-category and LP-category samples is 1.5310. These results indicate that under the same error category, the environmental conditions of the samples exhibit consistency, whereas under different error categories, the environmental conditions of the samples demonstrate variability. Specifically, at the Class LP sampling points, yellow–brown soil dominated, with the parent material primarily comprising carbonate rock and the organic matter concentration varying between 0 and 6 mg/kg. In contrast, at the Class HN sampling points, yellow soil dominated, the parent material mainly comprised impure carbonate, and the organic matter concentration ranged from 6-10 mg/kg. Five samples were randomly selected from the different categories, the similarity was calculated, and the environmental conditions of the samples were analyzed. The results are shown in Fig. 6. The analysis results are consistent with the overall trend, providing ample evidence that the model provides high adaptability in areas with complex soil environments. Clearly, based on the above analysis, using historical XRF ICP-MS sample detection values and environmental variables as localized parameters/knowledge references for XRF detection, and simulating the similarity of environmental variables to select specific calibration models, the algorithm is applied to the instrument, forming a separate secondary correction module. It could enhance the monitoring efficiency and accuracy of XRF instruments in previously untested areas. This provides a new perspective for the widespread application of rapid detection instruments.

CONCLUSION

Addressing the significant impact of complex soil matrix effects on XRF detection results, this study introduces the GMC-MLM model, effectively mitigating the confounding influence of multiple factors on XRF outcomes. By leveraging a subset of laboratory results as truth values, the pre-classification of XRF detection data facilitates the development of complex factor modeling. These models can be utilized for farmland environmental monitoring and surveys. Through rigorous environmental similarity analysis and validation of the model's extrapolation capabilities, the knowledge base proposed in this study can be effectively employed for XRF detection correction. This innovative methodology has the potential to be integrated into XRF instruments, significantly enhancing the widespread application and reliability of XRF detection technology in farmland monitoring and surveys.

AUTHOR INFORMATION

Yunbing Gao received the Ph.D. in 2016 from the College of Information and Electrical Engineering, China Agricultural University. He is a professor of engineering at Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences. His major research interests are soil environmental quality monitoring and evaluation, spectroscopy instrument application, *etc.* Yunbing Gao is author or co-author of over 50 articles published in peer-reviewed scientific journals. In 2021 and 2022, he received two science and technology awards from Henan provincial government and Chinese Society for Geodesy Photogrammetry and Cartography.

Corresponding Author

* Y. B. Gao

Email address: gybgis@163.com

# These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank China Three Gorges University Renshun associate professor for providing excellent data base. This work was supported financially by the National Key Research and Development Program of China (2022YFC3700805) and the National Key R&D Program for Young Scientists (2023YFD1700016).

REFERENCES

1. X. Gao, S. Kang, Q. Liu, P. Chen, and Z. Duan, *J. Geogr. Sci.*, 2020, **30**, 1481-1494. https://doi.org/10.1007/s11442-020-1794-8
2. G. Tepanosyan, N. Harutyunyan, and L. Sahakyan, *Environ. Geochem. Hlth.*, 2022, **44**, 1739-1750. https://doi.org/10.1007/s10653-021-01079-7
3. W. Zhang, Y. J. Zhang, D. Chen, R. Zhang, X. Y. Yu, Y. W. Gao, C. L. Wang, J. Liu, N. J. Zhao, and W. Q. Liu, *Adv. Mat. Res.*, 2013, **705**, 70-74. https://doi.org/10.4028/www.scientific.net/AMR.705.70
4. S. -J. Baek, A. Park, Y. -J. Ahn, and J. Choo, *Analyst.*, 2015, **140**, 250-257. https://doi.org/10.1039/c4an01061b
5. E. N. Han and E. Boydaş, *Instrum. Sci. Technol.*, 2021, **49**, 616-628. https://doi.org/10.1080/10739149.2021.1923029
6. S. P. Verma, M. Rosales-Rivera, M. A. Rivera-Gómez, and S. K. Verma, *Spectrochim. Acta. B.*, 2019, **162**, 9-14. https://doi.org/10.1016/j.sab.2019.105714
7. Y. Declercq, N. Delbecque, J. De Grave, P. De Smedt, P. Finke, A.M. Mouazen, S. Nawar, D. Vandenberghe, M. Van Meirvenne, and A. Verdoodt, *Remote Sens.*, 2019, **11**, 2490. https://doi.org/10.3390/rs11212490
8. X. H. Ying, Q. Wang, G. Z. Cao, and M. F. Yu, *Spectrosc. Spect. Anal.*, 2004, **24**, 1681-1683. https://doi.org/10.1016/j.jco.2003.08.015
9. H. M. Abdelmigid, M. A. Baz, M. A. AlZain, J. F. Al-Amri, H. G. Zaini, M. M. Morsi, M. Abualnaja, and E. A. Althagafi, *Agronomy*, 2022, **12**, 895. https://doi.org/10.3390/agronomy12040895
10. X. G. Tuo, B. Cheng, K. L. Mu, and Z. Li, *Nucl. Sci. Tech.*, 2008, **19**, 278-281. https://doi.org/10.1016/S1001-8042(09)60004-X
11. M. B. Liu, X. L. Liao, D. W. Cheng, Z. Y. Ni, and H. Z. Wang, *Spectrosc. Spect. Anal.,* 2021, **41**, 2807-2811. https://doi.org/10.3964/j.issn.1000-0593(2021)09-2807-05
12. A. Markowicz., *Pramana-J Phys*, 2011, **76**, 321-329. https://doi.org/10.1007/s12043-011-0045-z
13. S. K. Le and Y. M. Duan, *J. Inorg. Anal. Chem.*, 2015, **5**, 16-19. https://doi.org/10.3969/j.issn.2095-1035.2015.03.005
14. A. Sheibani, H. Saeedi-Sourck, and M. Tabrizchi, *J. Chemometr.*, 2017, **31**, e2911. https://doi.org/10.1002/cem.2911
15. F. Li, A. X. Lu, and J. H. Wang, *Int. Environ. Res. Public Health*, 2017, **14**, 1163. https://doi.org/10.3390/ijerph14101163
16. R. Y. Gu, M. Lei, T. B. Chen, X. M. Wan, Z. P. Dong, Y. T. Wang, and P. W. Qiao, *Anal. Chem.*, 2019, **91**, 5858-5865. https://doi.org/10.1021/acs.analchem.9b00201
17. G. L. Lei, H. C. Zhang, F. Q. Chang, Y. Zhu, C. H. Li, X. Xie, Y. B. Lei, W. X. Zhang, and P. Yang, *Journal of Lake Sciences*, 2011, **23**, 287-294. https://doi.org/10.18307/2011.0220
18. L. Poto, J. Gabrieli, S. Crowhurst, C. Agostinelli, A. Spolaor, W. R. L. Cairns, G. Cozzi, and C. Barbante, *Anal. Bioanal. Chem.*, 2014, **407**, 379-385. https://doi.org/10.1007/s00216-014-8289-3
19. R. B. Gregory, R. Timothy Patterson, Eduard G. Reinhardt, Jennifer M. Galloway, and H. M. Roe, *Chem. Geol.*, 2019, **521**, 12-27. https://doi.org/10.1016/j.chemgeo.2019.05.008

20. S. L. Walser, E. C. Sirkovich, J. B. Richardson, A. E. McStay, and N. Perdrial, *X-ray Spectrom.*, 2022, **52**, 72-82. https://doi.org/10.1002/xrs.3321
21. Z. Mao, *Spectrosc. Spect. Anal*, 1999, **19**, 738-741. https://doi.org/10.1016/S1386-1425(99)00115-8
22. L. Lupi, L. Bertrand, M. V. Monferrán, M. V. Amé and M. D. P. Diaz, *J. Hydrol.*, 2019, **572**, 403-413. https://doi.org/10.1016/j.jhydrol.2019.03.019
23. A. -X. Lu, J. -X. Wang, L. -G. Pan, P. Han, and Y. Han, *Spectrosc. Spect. Anal.*, 2010, **30**, 2848-2852. https://doi.org/10.3964/j.issn.1000-0593(2010)10-2848-05
24. P. J. Hu, J. Zhang, J. Liu, X. Y. Li, Y. P. Du, L. H. Wu, and Y. M Luo, *Acta Pedologica Sinica*, 2023, **60**, 1363-1377. https://doi.org/10.11766/trxb202307090262
25. Y. -B. da Silva, C. -A. do Nascimento, C. M. Biondi, P. van Straaten, Y. J. Agra B. da Silva, V. S. de Souza, J. -T. de Araújo, V. C. Alcantara, F. L. da Silva, and R. -B. da Silva, *Catena*, 2020, **193**, 104641. https://doi.org/10.1016/j.catena.2020.104641
26. A. Takeda, K. Kimura, and S. Yamasaki, *Geoderma*, 2004, **199**, 291-307. https://doi.org/10.1016/j.geoderma.2003.08.006
27. D. Gallhofer and B. G. Lottermoser, *Miner. Eng.*, 2018, **8**, 320. https://doi.org/10.3390/min8080320
28. E. Hangen and F. Vieten, *Water Air Soil Pollut.*, 2016, **227**, 143. https://doi.org/10.1007/s11270-016-2844-9
29. R. Ravansari, S. C. Wilson, and M. Tighe, *Environ. Int.*, 2020, **134**, 105250. https://doi.org/10.1016/j.envint.2019.105250
30. Y. -T. Shen, *Spectrosc. Spect. Anal.*, 2012, **32**, 3117-3122. https://doi.org/10.3964/j.issn.1000-0593(2012)11-3117-06
31. T. R. Tavares, J. P. Molin, L. C. Nunes, E. E. Novais Alves, F. L. Melquiades, H. W. Pereira de Carvalho, and A. M. Mouazen, *Remote Sens.*, 2020, **12**, 963. https://doi.org/10.3390/rs12060963
32. T. R. Tavares, B. Minasny, A. McBratney, M. R. Cherubin, G. T. Marques, M. M. Ragagnin, E. E. N. Alves, J. Padarian, J. Lavres, and H. W. P. de Carvalho, *Geoderma*, 2023, **439**, 116701. https://doi.org/10.1016/j.geoderma.2023.116701
33. F. Xu, B. Hu, C. Wang, J. Zhao, F. Wang, X. Ding, Q. Li, and J. Guo, *J. Ocean. U. China*, 2021, **20**, 848-856. https://doi.org/10.1007/s11802-021-4554-1
34. F. Guo, S. C. Clemens, T. Wang, Y. Wang, Y.M. Liu, F. Wu, X. X. Liu, Z.D. Jin, and Y. B. Sun, *Catena*, 2021, **198**, 105019. https://doi.org/10.1016/j.catena.2020.105019
35. A. X Lu, J.H. Wang, L. G. Pan, P. Han, and Y. Han, 2010, **30**, 2848-2852. https://doi.org/10.3964/j.issn.1000-0593(2010)10-2848-05
36. Y. Chi, J. Sun, T. Li, and X. Ma, *Ecol. Indic.*, 2023, **146**, 109774. https://doi.org/10.1016/j.ecolind.2022.109774
37. S. Nawar, N. Delbecque, Y. Declercq, P. De Smedt, P. Finke, A. Verdoodt, M. Van Meirvenne, and A. M. Mouazen, *Geoderma*, 2019, **350**, 29-39. https://doi.org/10.1016/j.geoderma.2019.05.002
38. S. M. O'Rourke, U. Stockmann, N. M. Holden, A. B. McBratney, and B. Minasny, *Geoderm*a, 2016, **279**, 31-44. https://doi.org/10.1016/j.geoderma.2016.05.005
39. X. C. Zhang, *X-ray Spectrom.*, 2005, **34**, 207-212. https://doi.org/10.1002/xrs.794