Transfer Learning-Based Interpretable Soil Lead Prediction in the Gejiu Mining Area, Yunnan

Ping He , Xianfeng Cheng , Xingping Wen * , Yan Yi , Zailin Chen , Yu Chen *

Posted Date: 29 May 2025

doi: 10.20944/preprints202505.2402.v1

Keywords: Soil Lead (Pb); Transfer Learning; SHAP Analysis; Small Sample Prediction

 doi:10.20944/preprints202505.2402.v1

Article

Transfer Learning-Based Interpretable Soil Lead Prediction in the Gejiu Mining Area, Yunnan

Ping He [1,2,3,†], Xianfeng Cheng [4,5,†], Xingping Wen [1*], Yan Yi [2], Zailin Chen [4,5] and Yu Chen [3,6*]

1 Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China
2 Kunming University, Kunming 650214, China
3 International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China
4 Yunnan Land and Resources Vocational College, Kunming 652501, China
5 Engineering Center of Yunnan Education Department for Health Geological Survey & Evaluation, Kunming 650218, China
6 Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: Xingping Wen. Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China; Email address: wfxyp@qq.com (X.P. Wen); Yu Chen. International Research Center of Big Data for Sustainable Development Goals, No. 9 Dengzhuang South Road, Haidian District, Beijing 100094, China; Email address: chenyu@radi.ac.cn (Y. Chen)
† These authors contributed equally to this work.

Abstract: Accurate prediction of soil lead (Pb) content in small sample scenarios is often limited by data scarcity and variability in soil properties, with traditional spectral modeling methods yielding suboptimal precision. To address this, we propose a transfer learning-based framework integrated with SHAP analysis for predicting soil Pb content in the Gejiu mining area, Yunnan. Using pH data from the European LUCAS soil database as the source domain, spectral features were extracted via a 1D-ResNet model and transferred to the target domain (130 soil samples from Gejiu) for Pb prediction. SHAP analysis was applied to clarify the role of spectral characteristics in cross-component transfer learning, uncovering shared and adaptive features between pH and Pb predictions. The transfer learning model (ResNet-pH-Pb) significantly outperformed direct modeling methods (PLS-Pb, SVM-Pb, and ResNet-Pb), with an R^2 of 0.77, demonstrating superior accuracy. SHAP analysis showed that the model retained key pH-related wavelengths (550-750 nm and 1600-1700 nm) while optimizing Pb-related wavelengths (e.g., 919 nm and 959 nm). This study offers a novel approach for soil heavy metal prediction under small sample constraints and provides a theoretical basis for understanding spectral prediction mechanisms through interpretability analysis.

Keywords: soil lead (Pb); transfer learning; SHAP analysis; small sample prediction

1. Introduction

Lead (Pb), a profoundly toxic, bioaccumulative, and environmentally persistent heavy metal, presents substantial risks to ecosystems and human well-being [1,2]. Due to prolonged mining and smelting activities, Pb content in mining area soils is typically high, and the pollution exhibits significant spatial heterogeneity [3,4]. The Gejiu mining area in Yunnan, one of China's major non-ferrous metal mining regions, suffers from particularly severe Pb pollution due to historical mining activities [5]. Therefore, accurately and rapidly assessing Pb content in the soils of mining areas and identifying the key factors influencing it is of great importance for pollution assessment and environmental management.

Visible-Near Infrared (Vis-NIR) spectroscopy provides non-destructive, swift, and economically efficient benefits for estimating soil Pb levels [6]. Researchers typically enhance Pb spectral signal by selecting specific bands and combine this with machine learning methods such as Partial Least Squares (PLS) and Support Vector Machines (SVM) to build predictive models [7,8]. However, Pb

 doi:10.20944/preprints202505.2402.v1

lacks direct spectral absorption features, making its signal susceptible to soil matrix interference, which limits model accuracy [9]. Deep learning methods, such as 1D-Residual Neural Networks (1D-ResNet), show great potential by extracting complex spectral features through cross-layer connections [10,11]. Yet, their reliance on large sample sizes conflicts with the limited data available in mining areas [12].

Transfer learning presents a novel approach by harnessing knowledge from a related task to boost prediction accuracy despite limited target domain data [13]. Existing studies have shown that models trained on the LUCAS dataset for predicting organic carbon and pH have achieved high accuracy in cross-region transfers [14–16]. Notably, current research has mainly focused on transfer within the same attribute (e.g., pH → pH), while how to use soil physicochemical properties closely related to Pb to assist Pb prediction remains an unexplored issue. Soil pH is a key factor influencing the distribution and mobility of Pb [17,18]. Variations in pH substantially influence the solubility and adsorption capacity of Pb within soil [19]. Compared to Pb, pH is more easily predicted accurately from spectral data [20,21]. This characteristic makes pH an ideal intermediary variable, as large-scale pH data from the LUCAS dataset can be used to train a source domain model and transfer this knowledge to the small sample Pb prediction task, improving the accuracy and stability of Pb predictions.

Although transfer learning can enhance prediction performance, the lack of model interpretability limits its practical application. SHAP value helps address the "black box" problem by quantifying feature contributions [22]. While SHAP has shown interpretive potential in fields such as spectral analysis and environmental monitoring [23–25], its application in transfer learning models, particularly in analyzing the reuse of features across components, has not been fully explored.

Therefore, this study innovatively proposes a soil Pb prediction framework that integrates transfer learning with SHAP analysis. Its core contributions include: (1) Constructing a cross-component transfer path from pH to Pb, using the LUCAS dataset (pH prediction) as the source domain, and transferring knowledge from the 1D-ResNet model to address the small sample problem in the Gejiu mining area (Pb prediction). (2) For the first time, using SHAP values to analyze the contribution mechanism of spectral features in cross-component transfer learning models, offering a theoretical foundation for heavy metal spectral prediction.

2. Data and Methods

2.1. Data Sources

This study utilizes two datasets—source domain and target domain data—to predict soil lead (Pb) content via cross-component transfer learning.

The source domain data is derived from the European LUCAS Soil Database, collected by the European Commission, comprising 19,036 surface soil samples (0-20 cm) from multiple European countries [26]. Physicochemical and spectral properties were measured using standardized protocols to ensure data consistency [27]. The initial spectral dataset spans (with a 0.5 nm interval, consisting of 4200 wavelength points) [28]. To align with the target domain data, this study downsampled the data to the range of 400-2499 nm (with a 1 nm interval, consisting of 2100 wavelength points). Given the potential correlation between pH and soil Pb chemical behaviors (such as adsorption and desorption), pH was selected as the source domain task to provide a knowledge foundation for cross-component transfer learning.

The target domain data was collected through field sampling in the Gejiu mining area, Yunnan, China, a region characterized by complex terrain and rich mineral resources, particularly tin mines [29]. Prolonged mining operations have resulted in substantial Pb pollution [30]. From March to April 2024, in the dry season when exposed soil was common, 130 surface soil samples (0-20 cm) were gathered. Sampling was carried out using a grid method: in the northern smelting area, a 1000m×1000m grid was used, with denser sampling (500m×500m) around the smelting plants. In the southern mining region, samples were gathered at 1000-meter intervals along the road. The overall

 doi:10.20944/preprints202505.2402.v1

distribution of sampling points across the study area is illustrated in Figure 1. Soil samples were dried in air, had their impurities removed, and were sieved through a 100-mesh screen. Pb concentrations were determined using inductively coupled plasma mass spectrometry (ICP-MS). Spectral measurements were obtained with an ASD FieldSpec spectrometer, covering 350-2500 nm (1 nm resolution), with five measurements averaged per sample. To reduce noise and align with the source domain, the 350-399 nm range was excluded, retaining only the 400-2499 nm range.

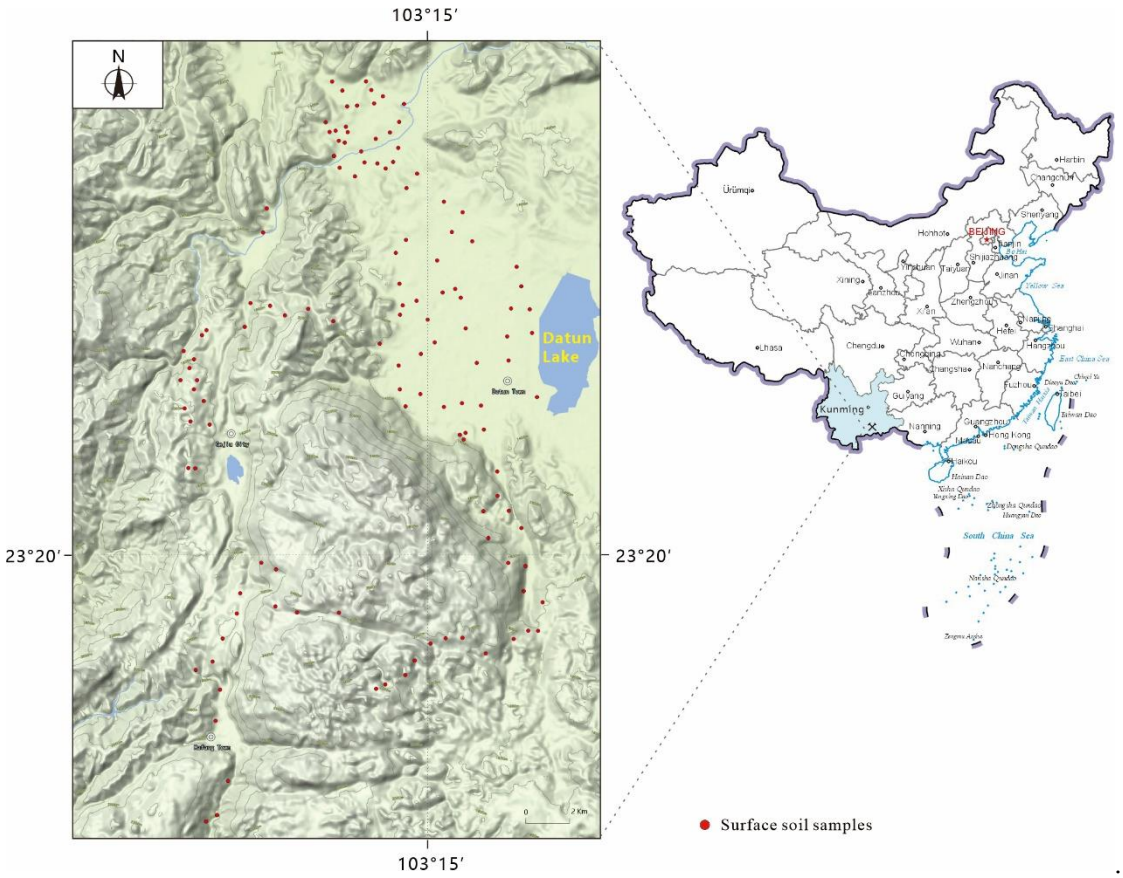


Figure 1. Target Domain Soil Sample Distribution.

2.2. Methods

2.2.1. ResNet Model Architecture

ResNet optimizes deep network performance through a unique residual learning mechanism, effectively addressing the degradation problem that arises with increasing network depth in traditional architectures [10]. The core design relies on residual blocks, where skip connections merge input and convolutional features, enhancing both training efficiency and model expressiveness. This research developed a 1D-ResNet model utilizing spectral data characteristics. The model's input consists of log-transformed (logR) spectral reflectance data, with a dimensionality of 2100. To improve computational efficiency and extract key features, the first layer applies average pooling with a window size of 10, reducing the dimension to 210. The data then passes through a 1D convolutional layer (48 filters, kernel size 3, stride 1, Leaky ReLU activation function with $\alpha = 0.01$) for feature extraction, followed by batch normalization to optimize data distribution. The data subsequently undergoes max pooling with a window size of 2, followed by two residual blocks (with ReLU activation and skip connections). Afterward, it advances through a 1D convolutional layer (32 filters), a flattening layer, and two dense layers (16 and 10 nodes, using Leaky ReLU activation). The output layer consists of a single-node dense layer with ReLU activation. All convolutional and fully connected layers use L2 regularization ($\lambda = 0.0004$). The specific network architecture is shown in Figure 2.

 doi:10.20944/preprints202505.2402.v1

The model was constructed using Python 3.9 and TensorFlow 2.0, employing the Adam optimizer (learning rate = 0.001) for training and utilizing mean squared error (MSE) as the loss metric. Training runs for 2000 iterations, with early stopping enabled (patience = 120), meaning training halts if no improvement is observed after 120 epochs.

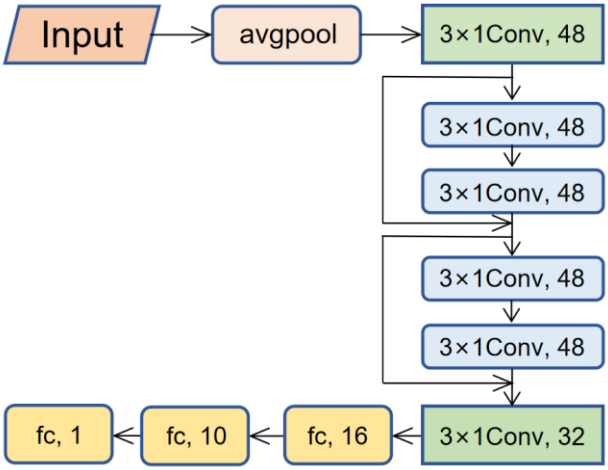


Figure 2. 1D-ResNet Model Architecture.

2.2.2. Transfer Learning Process

Transfer learning enhances target domain prediction by using rich source domain data, making it ideal for cross-component predictions such as from pH to Pb. This study employs a 1D-ResNet model with a fine-tuning strategy to transfer spectral features from the LUCAS dataset (source domain, 19,036 samples) to the Gejiu mining area (target domain, 130 samples) for improved soil Pb content prediction.

First, the source domain model, ResNet-pH, is trained on the LUCAS dataset to predict pH values, using 14,277 samples (75%) for training and 4,759 samples (25%) for testing. Next, a baseline model, ResNet-Pb, is independently trained on the Gejiu dataset to directly predict Pb content, using 98 samples (75%) for training and 32 samples (25%) for testing, without relying on source domain information. For transfer learning, pre-trained weights from ResNet-pH are loaded into the target domain model. Convolutional layers and residual blocks are frozen to preserve general spectral features, while the max-pooling layer and fully connected layers are trained to adapt to Pb prediction, yielding the transfer learning model, ResNet-pH-Pb.

Both datasets are randomly split into 75% training and 25% testing sets, and this process is repeated over 10 rounds to ensure robust evaluation. The final evaluation of the model is based on the average of the evaluation metrics from the 10 test rounds. Model performance is assessed using metrics detailed in Section 2.2.4.

2.2.3. Interpretability Analysis

SHAP values, based on game theory, quantify feature impacts on predictions, enhancing model interpretability [31]. This research utilizes the GradientExplainer module from the SHAP library to calculate SHAP values and assess feature importance for ResNet-based models.

For the source domain model, ResNet-pH, given the large LUCAS dataset, 1,000 samples are randomly selected as background data using a fixed seed for reproducibility, balancing computational efficiency and representativeness. From the test set, 1000 samples are selected to compute SHAP values, with their mean absolute values used to assess each wavelength's contribution to pH prediction.

For the transfer learning model, ResNet-pH-Pb, all 130 samples from the Gejiu mining area dataset are used as background data to compute SHAP values for the test set, revealing the impact of transfer learning on Pb content prediction.

 doi:10.20944/preprints202505.2402.v1

By comparing the SHAP values of ResNet-pH and ResNet-pH-Pb, we analyze differences in wavelength contributions between pH and Pb prediction tasks, elucidating how transfer learning adjusts shared and task-specific wavelengths to enhance model predictions.

2.2.4. Comparison Experiments and Evaluation Metrics

To validate the effectiveness of cross-component transfer learning for soil Pb prediction in the target domain, this study trains PLS and SVM models on the target domain dataset, labeled as PLS-Pb and SVM-Pb, and compares their performance with the transfer learning model ResNet-pH-Pb.

PLS conducts modeling by constructing a linear correlation between spectral measurements and Pb levels, which is computationally simple and effective in addressing multicollinearity issues in multivariate data [32]. SVM utilizes a kernel function to model non-linear patterns, exhibiting robust performance and generalization ability, and is widely applied in soil property prediction [33]. PLS employs cross-validation and grid search to identify the best number of principal components (11), while SVM applies Bayesian optimization to determine optimal parameters (C = 0.26, gamma = 'scale', kernel = 'rbf').

Model performance is assessed using the coefficient of determination (R^2), root mean square error (RMSE), and residual prediction deviation (RPD). R^2 indicates the extent of variance in the data accounted for by the model, with values nearer to 1 signifying improved model fit. RMSE quantifies the average discrepancy between predicted and actual values, with lower values reflecting greater prediction precision. RPD evaluates the ratio of standard deviation to residuals to gauge prediction reliability, with RPD < 1.4 indicating inadequate prediction, 1.4 ≤ RPD < 2.0 suggesting acceptable performance, and RPD ≥ 2.0 representing superior prediction [34].

3. Results

3.1. Statistical Characterization of Soil Pb Concentrations

This research conducted statistical analysis on soil Pb levels in the Gejiu mining region (Table 1). Pb levels varied from 34.6 mg/kg to 9720 mg/kg, highlighting considerable differences in Pb concentrations across sampling locations. The average Pb level was 974.06 mg/kg, with a median of 232 mg/kg. The skewness of 3.12 and kurtosis of 9.1 suggest a markedly positively skewed distribution, indicating that most locations exhibited relatively low Pb levels, while a few showed exceptionally high concentrations. Furthermore, the standard deviation (SD) was 1969.17 mg/kg, and the coefficient of variation (CV) of 2.02 indicates substantial variability, pointing to significant spatial heterogeneity. This may be attributed to localized contamination hotspots resulting from mining activities. To mitigate the effects of skewness and variability, a logarithmic transformation was applied to the Pb levels.

Table 1. Summary Statistics of Pb Levels in Soil Samples from the Gejiu Mining Region.

Sample size	Min	Max	Mean	SD	CV	Skew	Kurt	Median
130	34.6	9720	974.06	1969.17	2.02	3.12	9.1	232

3.2. Source Domain Modeling

The scatter plot depicting the results of 10 rounds of random sampling tests for the source domain model (ResNet-pH) is presented in Figure 3. The model's average R^2 is 0.91, RMSE is 0.42, and RPD is 3.24, indicating strong predictive capability. This high-precision source domain model effectively captured the spectral features of pH values, providing a reliable foundation for subsequent target domain Pb prediction and wavelength contribution analysis.

 doi:10.20944/preprints202505.2402.v1

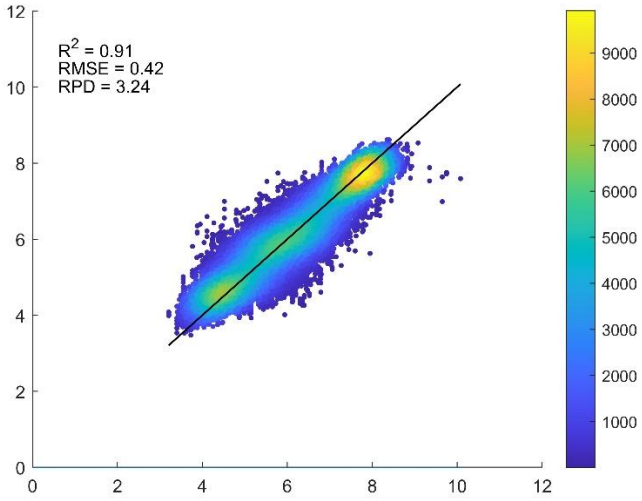
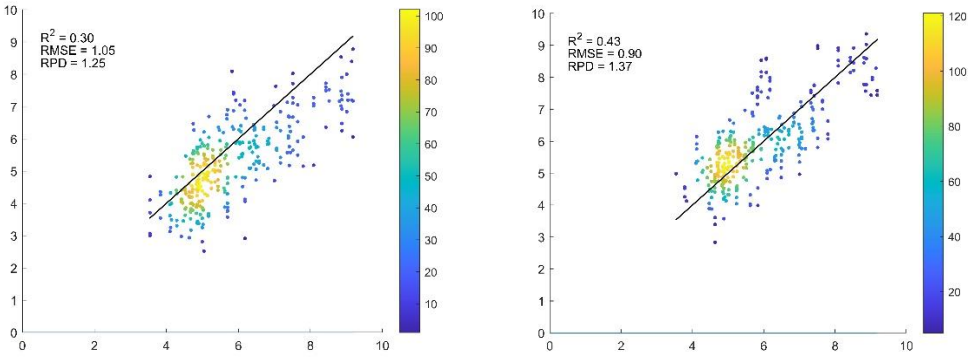


Figure 3. Predicted vs. Observed pH for ResNet-pH Model Based on 10 Random Sampling Tests. The average evaluation results of 10 test rounds are displayed in the upper-left section of the figure.

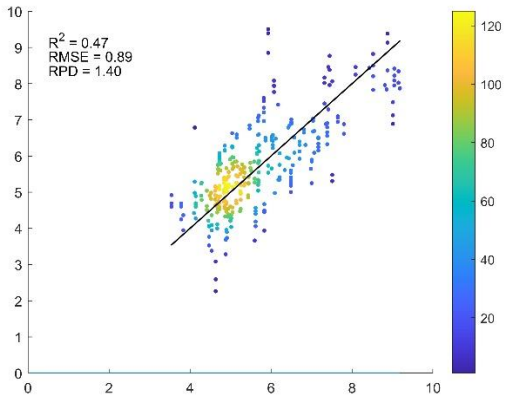
3.3. Direct Modeling in the Target Domain

For the target domain Gejiu dataset, ResNet, PLS, and SVM models were directly trained to predict Pb levels, with 10 rounds of random sampling evaluations performed. The findings are displayed in Figure 4. From the model performance, SVM-Pb achieved the best prediction results (R^2 = 0.47, RMSE = 0.89, RPD = 1.40), followed by PLS-Pb (R^2 = 0.43, RMSE = 0.90, RPD = 1.37), and ResNet-Pb had the worst performance (R^2 = 0.30, RMSE = 1.05, RPD = 1.25). Compared to SVM-Pb and PLS-Pb, the R^2 of ResNet-Pb decreased by 0.17 and 0.13, respectively. This indicates that for small sample target domains, deep learning models do not perform as well as traditional methods. Moreover, the R^2 of all models is below 0.5, suggesting that direct modeling has limited applicability in small sample target domains.



(a) ResNet-Pb (b) PLS-Pb

 doi:10.20944/preprints202505.2402.v1



(c) SVM-Pb

Figure 4. Predicted vs. Observed Pb Values for ResNet, PLS, and SVM Models Based on 10 Random Sampling Tests in the Target Domain.

3.4. Performance of the Transfer Learning Model

The results of 10 rounds of random sampling tests for the transfer learning model ResNet-pH-Pb are shown in Figure 5, yielding R^2=0.77, RMSE=0.59, and RPD=2.12, significantly outperforming the direct modeling methods. The R^2 box plot in Figure 6 shows that the average R^2 of ResNet-pH-Pb is 0.47, 0.34, and 0.30 higher than that of ResNet-Pb, PLS-Pb, and SVM-Pb, respectively. This indicates that transfer learning not only improves Pb content prediction accuracy but also significantly enhances model stability and generalization.

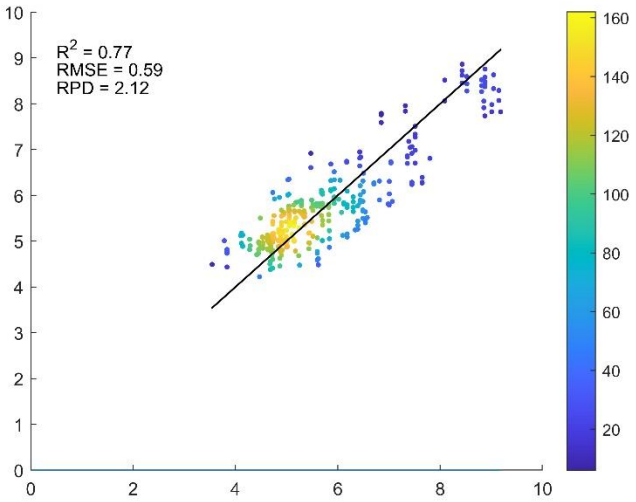


Figure 5. Performance of ResNet-pH-Pb Model in 10 Random Sampling Tests.

 doi:10.20944/preprints202505.2402.v1

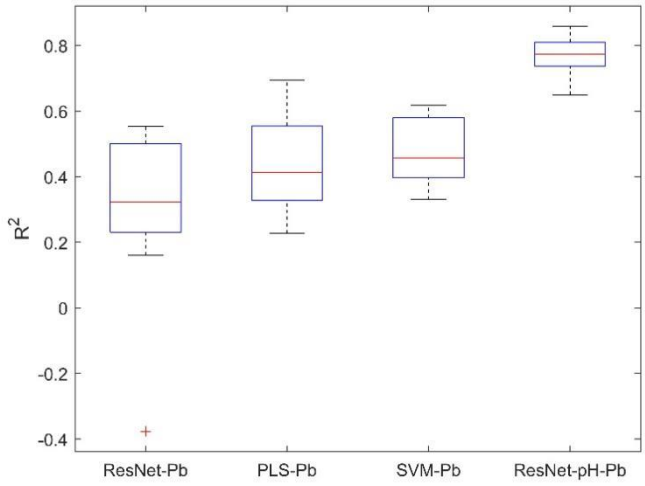


Figure 6. R^2 Boxplot Comparison of ResNet-pH-Pb with PLS-Pb, SVM-Pb, and ResNet-Pb.

3.5. Wavelength Contribution in ResNet Modeling

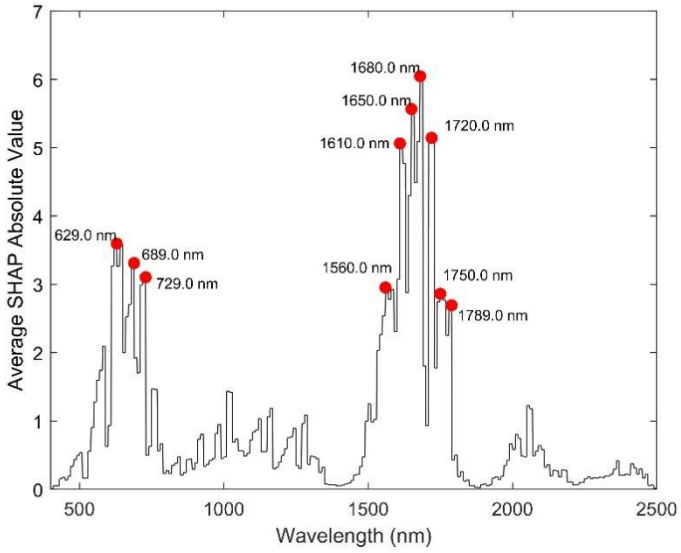
To explore the contribution of different wavelengths in ResNet modeling, SHAP values were used to interpret the source domain model ResNet-pH and the transfer learning model ResNet-pH-Pb, as shown in Figure 7.

In the ResNet-pH model (Fig. 7a), the wavelengths with significant contributions to pH prediction are mainly concentrated at 629 nm, 689 nm, 729 nm, 1560 nm, 1650 nm, 1680 nm, 1750 nm, 1789 nm, and other positions, suggesting that these wavelengths may be closely related to the spectral features of soil pH.

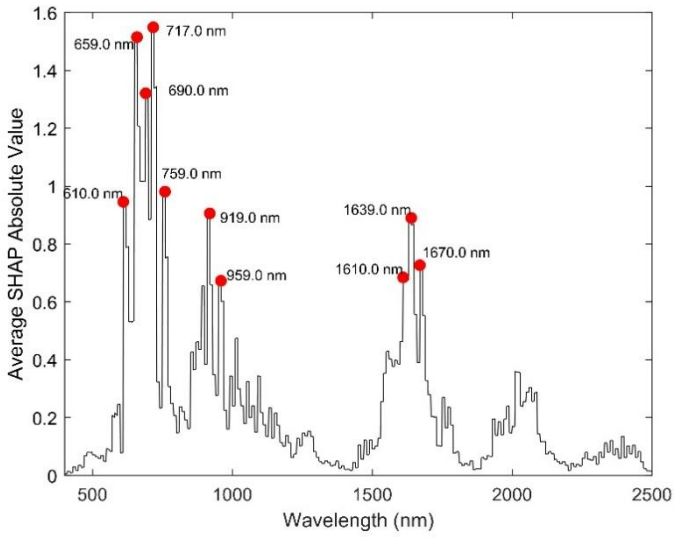
In contrast, in the ResNet-pH-Pb model (Fig. 7b), the wavelength contributions have shifted. New key wavelengths have emerged, including 610 nm, 659 nm, 717 nm, and 759 nm in the visible range, and 919 nm, 959 nm, 1639 nm, and 1670 nm in the near-infrared range. Meanwhile, some key wavelengths from the source domain model, like 689 nm and 1610 nm, are still retained. This indicates that in the Pb prediction task, the model not only inherited part of the spectral information from the pH modeling but also focused on new wavelength regions related to Pb, especially in the near-infrared range (e.g., 919 nm and 959 nm).

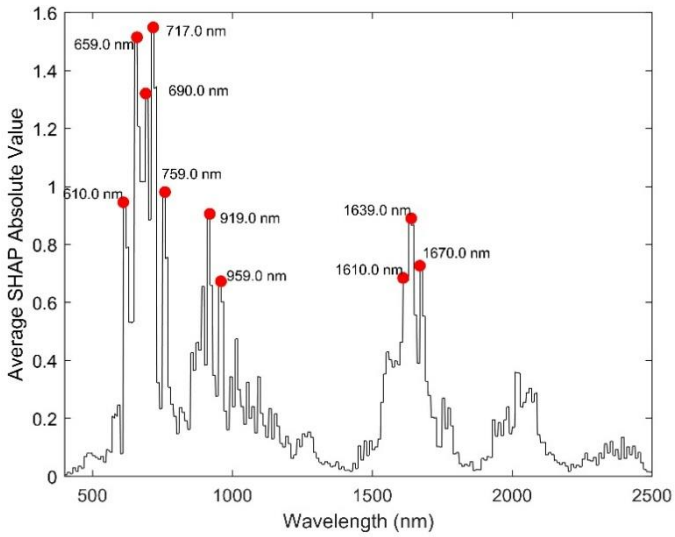
Overall, the transfer learning model ResNet-pH-Pb has undergone significant adjustments in its wavelength contributions. This indicates that while the spectral behavior of soil Pb is partially related to pH, they are not identical. With the introduction of transfer learning, the model is able to more effectively focus on the important wavelengths needed for Pb prediction, enhancing the reliability of Pb content prediction.

 doi:10.20944/preprints202505.2402.v1



(a) ResNet-pH





(b) ResNet-pH-Pb

 doi:10.20944/preprints202505.2402.v1

Figure 7. Wavelength Contribution in ResNet-pH and ResNet-pH-Pb Models.

4. Discussion

4.1. Improvement in Prediction Performance for Small Sample Target Domains Using Transfer Learning

In soil spectroscopy, the accuracy of Pb prediction is often limited by small sample sizes and high heterogeneity. In this study, direct modeling was conducted on a target domain with 130 samples from the Gejiu region. The findings indicate that the R^2 values for SVM-Pb, PLS-Pb, and ResNet-Pb are 0.47, 0.43, and 0.30, respectively. This indicates that both traditional modeling methods and deep learning approaches struggle to achieve satisfactory Pb prediction results under small sample conditions. These findings are consistent with existing research. Specifically, Tan et al. (2022) applied CARS in feature selection and PLS in modeling to predict soil Pb content, achieving a maximum R^2 of 0.60 in the validation set [35]. Arif et al. (2022) selected feature wavelengths and applied PLS for modeling, achieving an optimal R^2 of 0.66 [36]. Chen et al. (2022) used fractional-order derivatives for feature selection and combined PLS and SVM regression to predict Pb content, with R^2 ranging between 0.54 and 0.59, which is insufficient for high-accuracy predictions [7].

In contrast, the transfer learning model ResNet-pH-Pb in this study significantly improved Pb prediction performance in the target domain (R^2 = 0.77). The R^2 box plot (Figure 6) shows that the R^2 distribution of ResNet-pH-Pb is more stable, with improvements of 0.30, 0.34, and 0.47 over SVM-Pb, PLS-Pb, and ResNet-Pb, respectively. Transfer learning has been explored in soil spectroscopy to some extent. For instance, Kok et al. (2024) applied transfer learning to improve pH prediction, achieving an R^2 of 0.66 [15]. However, most of these studies focus on predicting single soil properties, with fewer studies addressing cross-property transfer learning applications. In this study, we innovatively implemented cross-component transfer from pH to Pb, leveraging the high accuracy of the source domain model ResNet-pH (R^2 = 0.91). By sharing spectral features (e.g., both pH and Pb are governed by soil organic matter and iron oxides), the model reduces overfitting in the small sample target domain and improves its generalization ability.

4.2. Feature Analysis of Wavelength Contribution

Wavelength contribution analysis further reveals the predictive mechanism of transfer learning. Here, SHAP values were employed to evaluate wavelength contributions in the source domain model ResNet-pH (Fig. 7a) and the transfer learning model ResNet-pH-Pb (Fig. 7b). It was found that both models showed high contributions in the 550-750 nm and 1600-1700 nm bands. These bands are associated with the absorption characteristics of soil organic matter, iron oxides (550-750 nm), and water and organic matter (1600-1700 nm) [37], while organic matter and iron oxides are the main adsorbents of Pb [38]. This finding suggests that, although pH and Pb are different components, the spectral features learned by the source domain model can still be effectively transferred to the target domain. This feature sharing originates from the potential correlation between pH and Pb in soil. In the tin mining area, the soil pH is typically acidic, primarily driven by sulfide mineral oxidation generating sulfuric acid [3,39,40]. Low pH values lead to the dissolution of iron oxides, releasing adsorbed Pb, which increases its mobility [41,42]. This acidic environment provides the chemical basis for cross-component transfer, and the spectral features learned by the source domain model ResNet-pH can effectively transfer to the target domain, alleviating overfitting in small sample scenarios and significantly improving Pb prediction accuracy.

Although pH and Pb predictions share some key wavelengths, transfer learning also optimizes the model's adaptability to the target domain. The contribution of ResNet-pH-Pb significantly increased in the 919 nm, 959 nm, 1639 nm, and 1670 nm bands. The 919 nm and 959 nm bands may be related to the absorption characteristics of carbonates and clay minerals in the soil [43], components that play an important role in Pb prediction, as Pb often exists in the form of lead carbonates or adsorbed on clay minerals [44]. The 1639 nm and 1670 nm bands further strengthened the contributions of water and organic matter, possibly associated with Pb hydrolysis and organic matter complexation [43]. Furthermore, the SHAP contribution to Pb prediction slightly increased in

 doi:10.20944/preprints202505.2402.v1

the visible light region (600-700 nm), potentially linked to enhanced spectral interactions of Pb with soil iron oxides and organic matter [38]. Existing studies have also shown that key wavelengths for Pb prediction typically include 600-800 nm, 1390-1460 nm and 1870-1960 nm [45,46]. The 550-750 nm and 1600-1700 nm bands identified here align closely with prior research, confirming their association with Pb adsorption mechanisms. Compared with the SHAP values analyzed for single-component predictions by Zhong et al. (2024), this study innovatively reveals the feature sharing and adaptation mechanisms of cross-component transfer through SHAP value analysis [22]. However, this study still has limitations, as the small sample size in the target domain may affect the model's ability to predict extreme Pb contents. Nevertheless, the present work has limitations; due to the limited sample size in the target domain, the model's capacity to predict extreme Pb content may be constrained. Future research could explore other heavy metals and integrate soil physicochemical properties, such as organic matter, to enhance the understanding of spectral contributions, thereby improving the model's generalizability and stability.

5. Conclusion

This study successfully achieved cross-component transfer prediction from pH to Pb and, for the first time, utilized SHAP values to analyze wavelength contributions in transfer learning models, innovatively broadening the use of soil spectroscopy for heavy metal assessment. The source domain model ResNet-pH demonstrated high accuracy on the LUCAS dataset ($R^2 = 0.91$). In the target domain of the Gejiu mining area, traditional direct modeling methods (SVM-Pb, PLS-Pb, ResNet-Pb) showed low prediction performance ($R^2 < 0.5$ for all). The transfer learning model ResNet-pH-Pb significantly improved prediction accuracy ($R^2 = 0.77$), with R^2 values 0.30, 0.34, and 0.47 higher than SVM-Pb, PLS-Pb, and ResNet-Pb, respectively, confirming the advantages of transfer learning in small sample target domains. Wavelength contribution analysis revealed high contributions in the 550-750 nm and 1600-1700 nm bands, further demonstrating the feature sharing mechanism in cross-component transfer. This study demonstrates that transfer learning methods not only markedly enhance the prediction accuracy of small sample target domains but also provide a robust and efficient approach for evaluating soil heavy metal contamination, offering significant applied potential.

Acknowledgments: The research was supported by the International Research Centre of Big Data for Sustainable Development Goals (CBAS) [Grant No. CBASYX0906], the National Natural Science Foundation of China (42271422) and the key project of sustainable development international cooperation program by NSFC (Grant No.42361144883), the Engineering Center of Yunnan Education Department for Health Geological Survey & Evaluation (9135009009), Science and Technology Innovation Team for Highland Ecological Agriculture Geological Survey and Evaluation of Yunnan Education Department.

References

1. Gholizadeh, A.; Borůvka, L.; Saberioon, M. M.; Kozák, J.; Vašát, R.; Němeček, K., Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil and Water Research* **2015,** 10, (4), 218-227.
2. Bradl, H. B., Adsorption of heavy metal ions on soils and soils constituents. *Journal of Colloid and Interface Science* **2004,** 277, (1), 1-18.
3. Li, Z.; Ma, Z.; van der Kuijp, T. J.; Yuan, Z.; Huang, L., A review of soil heavy metal pollution from mines in China: Pollution and health risk assessment. *Science of The Total Environment* **2014,** 468-469, 843-853.
4. Luo, X.; Wu, C.; Lin, Y.; Li, W.; Deng, M.; Tan, J.; Xue, S., Soil heavy metal pollution from Pb/Zn smelting regions in China and the remediation potential of biomineralization. *Journal of Environmental Sciences* **2023,** 125, 662-677.
5. Guo, G.; Zhang, D.; Wang, Y., Probabilistic Human Health Risk Assessment of Heavy Metal Intake via Vegetable Consumption around Pb/Zn Smelters in Southwest China. In *International Journal of Environmental Research and Public Health*, 2019; Vol. 16.

 doi:10.20944/preprints202505.2402.v1

6. Shi, T.; Chen, Y.; Liu, Y.; Wu, G., Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. *Journal of Hazardous Materials* **2014,** 265, 166-176.
7. Chen, L.; Lai, J.; Tan, K.; Wang, X.; Chen, Y.; Ding, J., Development of a soil heavy metal estimation method based on a spectral index: Combining fractional-order derivative pretreatment and the absorption mechanism. *Science of The Total Environment* **2022,** 813, 151882.
8. He, P.; Cheng, X.; Wen, X.; Cao, Y.; Chen, Y., Improving Soil Heavy Metal Lead Inversion Through Combined Band Selection Methods: A Case Study in Gejiu City, China. In *Sensors*, 2025; Vol. 25.
9. Wang, J.; Cui, L.; Gao, W.; Shi, T.; Chen, Y.; Gao, Y., Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. *Geoderma* **2014,** 216, 1-9.
10. He, K.; Zhang, X.; Ren, S.; Sun, J., Deep Residual Learning for Image Recognition. In *IEEE conference on computer vision and pattern recognition*, 2016; pp 770-778.
11. Zeng, P.; Song, X.; Yang, H.; Wei, N.; Du, L., Digital Soil Mapping of Soil Organic Matter with Deep Learning Algorithms. *ISPRS International Journal of Geo-Information* **2022,** 11, (5).
12. Ng, W.; Minasny, B.; Mendes, W. d. S.; Demattê, J. A. M., The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *Soil* **2020,** 6, (2), 565-578.
13. Pan, S. J.; Yang, Q., A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **2010,** 22, (10), 1345-1359.
14. Yang, J.; Wang, X.; Wang, R.; Wang, H., Combination of Convolutional Neural Networks and Recurrent Neural Networks for predicting soil properties using Vis–NIR spectroscopy. *Geoderma* **2020,** 380, 114616.
15. Kok, M.; Sarjant, S.; Verweij, S.; Vaessen, S. F. C.; Ros, G. H., On-site soil analysis: A novel approach combining NIR spectroscopy, remote sensing and deep learning. *Geoderma* **2024,** 446, 116903.
16. Padarian, J.; Minasny, B.; Mcbratney, A. B., Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma* **2019,** 340, 279-288.
17. Zhong, X.; Chen, Z.; Li, Y.; Ding, K.; Liu, W.; Liu, Y.; Yuan, Y.; Zhang, M.; Baker, A. J. M.; Yang, W.; Fei, Y.; Wang, Y.; Chao, Y.; Qiu, R., Factors influencing heavy metal availability and risk assessment of soils at typical metal mines in Eastern China. *Journal of Hazardous Materials* **2020,** 400, 123289.
18. Luo, X.-s.; Yu, S.; Li, X.-d., The mobility, bioavailability, and human bioaccessibility of trace metals in urban soils of Hong Kong. *Applied Geochemistry* **2012,** 27, (5), 995-1004.
19. Caporale, A. G.; Violante, A., Chemical Processes Affecting the Mobility of Heavy Metals and Metalloids in Soil Environments. *Current Pollution Reports* **2016,** 2, (1), 15-27.
20. Ke, Z.; Ren, S.; Yin, L., Advancing soil property prediction with encoder-decoder structures integrating traditional deep learning methods in Vis-NIR spectroscopy. *Geoderma* **2024,** 449.
21. Haghi, R. K.; Pérez-Fernández, E.; Robertson, A. H. J., Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. *Geoderma* **2021,** 396, 115071.
22. Zhong, L.; Guo, X.; Ding, M.; Ye, Y.; Jiang, Y.; Zhu, Q.; Li, J., SHAP values accurately explain the difference in modeling accuracy of convolution neural network between soil full-spectrum and feature-spectrum. *Computers and Electronics in Agriculture* **2024,** 217.
23. Li, C.; Song, L.; Zheng, L.; Ji, R., DSCformer: Lightweight model for predicting soil nitrogen content using VNIR-SWIR spectroscopy. *Computers and Electronics in Agriculture* **2025,** 230, 109761.
24. Mkhatshwa, J.; Kavu, T.; Daramola, O., Analysing the Performance and Interpretability of CNN-Based Architectures for Plant Nutrient Deficiency Identification. In *Computation*, 2024; Vol. 12.
25. Albinet, F.; Peng, Y.; Eguchi, T.; Smolders, E.; Dercon, G., Prediction of exchangeable potassium in soil through mid-infrared spectroscopy and deep learning: From prediction to explainability. *Artificial Intelligence in Agriculture* **2022,** 6, 230-241.
26. Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O., LUCAS Soil, the largest expandable soil dataset for Europe: a review. *European Journal of Soil Science* **2017,** 69, (1), 140-153.
27. Panagos, P.; Meusburger, K.; Ballabio, C.; Borrelli, P.; Alewell, C., Soil erodibility in Europe: A high-resolution dataset based on LUCAS. *Science of The Total Environment* **2014,** 479-480, 189-200.

 doi:10.20944/preprints202505.2402.v1

28. Liu, B.; Guo, B.; Zhuo, R.; Dai, F., Estimation of soil organic carbon in LUCAS soil database using Vis-NIR spectroscopy based on hybrid kernel Gaussian process regression. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2024**, 124687.
29. Li, H.; Yao, J.; Min, N.; Duran, R., Comprehensive assessment of environmental and health risks of metal(loid)s pollution from non-ferrous metal mining and smelting activities. *Journal of Cleaner Production* **2022,** 375.
30. Cheng, X.; Chen, Z.; Zhou, X.; Huang, Q.; Shen, J.; Chen, Y.; Hou, M.; Xiong, J., Evaluation of Contamination and Ecological and Health Risk in Surface Soil and Crops Contaminated with Metalloids and Heavy Metals in Datun, China, "The World Tin Capital". *Soil and Sediment Contamination: An International Journal* **2024**, 1-19.
31. Lundberg, S. M.; Lee, S.-I., A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 2017; pp 4765–4774.
32. Wu, Y.; Chen, J.; Wu, X.; Tian, Q.; Ji, J.; Qin, Z., Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. *Applied Geochemistry* **2005,** 20, (6), 1051-1059.
33. Khosravi, V.; Doulati Ardejani, F.; Yousefi, S.; Aryafar, A., Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. *Geoderma* **2018,** 318, 29-41.
34. Viscarra Rossel, R. A.; Walvoort, D. J. J.; McBratney, A. B.; Janik, L. J.; Skjemstad, J. O., Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006,** 131, (1), 59-75.
35. Tan, K.; Ma, W.; Chen, L.; Wang, H.; Du, Q.; Du, P.; Yan, B.; Liu, R.; Li, H., Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. *Journal of Hazardous Materials* **2021,** 401, 123288.
36. Arif, M.; Qi, Y.; Dong, Z.; Wei, H., Rapid retrieval of cadmium and lead content from urban greenbelt zones using hyperspectral characteristic bands. *Journal of Cleaner Production* **2022,** 374, 133922.
37. Rathod, P. H.; Rossiter, D. G.; Noomen, M. F.; van der Meer, F. D., Proximal Spectral Sensing to Monitor Phytoremediation of Metal-Contaminated Soils. *International Journal of Phytoremediation* **2013,** 15, (5), 405-426.
38. Zhou, M.; Zou, B.; Tu, Y.; Feng, H.; He, C.; Ma, X.; Ning, J., Spectral response feature bands extracted from near standard soil samples for estimating soil Pb in a mining area. *Geocarto International* **2022,** 37, (26), 13248-13267.
39. Liu, J.; Li, X.; Zhang, P.; Zhu, Q.; Lu, W.; Yang, Y.; Li, Y.; Zhou, J.; Wu, L.; Zhang, N.; Christie, P., Contamination levels of and potential risks from metal(loid)s in soil-crop systems in high geological background areas. *Science of The Total Environment* **2023,** 881, 163405.
40. Ashraf, M. A.; Maah, M. J.; Yusoff, I., Heavy metals accumulation in plants growing in ex tin mining catchment. *International Journal of Environmental Science & Technology* **2011,** 8, (2), 401-416.
41. Vega, F. A.; Covelo, E. F.; Andrade, M. L., Competitive sorption and desorption of heavy metals in mine soils: Influence of mine soil characteristics. *Journal of Colloid and Interface Science* **2006,** 298, (2), 582-592.
42. Jiang, H.; Li, T.; Han, X.; Yang, X.; He, Z., Effects of pH and low molecular weight organic acids on competitive adsorption and desorption of cadmium and lead in paddy soils. *Environmental Monitoring and Assessment* **2012,** 184, (10), 6325-6335.
43. Lu, Q.; Wang, S.; Bai, X.; Liu, F.; Wang, M.; Wang, J.; Tian, S., Rapid inversion of heavy metal concentration in karst grain producing areas based on hyperspectral bands associated with soil components. *Microchemical Journal* **2019,** 148, 404-411.
44. Wang, Y.; Zou, B.; Chai, L.; Lin, Z.; Feng, H.; Tang, Y.; Tian, R.; Tu, Y.; Zhang, B.; Zou, H., Monitoring of soil heavy metals based on hyperspectral remote sensing: A review. *Earth-Science Reviews* **2024,** 254, 104814.
45. Wang, Y.; Zhang, X.; Sun, W.; Wang, J.; Ding, S.; Liu, S., Effects of hyperspectral data with different spectral resolutions on the estimation of soil heavy metal content: From ground-based and airborne data to satellite-simulated data. *Science of The Total Environment* **2022,** 838, 156129.

 doi:10.20944/preprints202505.2402.v1

46. Hong, Y.; Shen, R.; Cheng, H.; Chen, Y.; Zhang, Y.; Liu, Y.; Zhou, M.; Yu, L.; Liu, Y.; Liu, Y., Estimating lead and zinc concentrations in peri-urban agricultural soils through reflectance spectroscopy: Effects of fractional-order derivative and random forest. *Science of The Total Environment* **2019,** 651, 1969-1982.