

Article

Estimating Soil Arsenic Contamination by Integrating Hyperspectral and Geochemical Data with PCA and Optimizing Inversion Models

Fei Guo ^{1,2,3}, Zhen Xu ^{4,*} , Honghong Ma ^{1,2,3} and Xiujiu Liu ^{1,2,3}

¹ Institute of Geophysical & Geochemical Exploration, Chinese Academy of Geological Sciences, Langfang 065000, China; flynn1991@sina.cn (F.G.); mahonghong@mail.cgs.gov.cn (H.M.); liuxiujiu@mail.cgs.gov.cn (X.L.)

² Key Laboratory of Geochemical Cycling of Carbon and Mercury in the Earth's Critical Zone, Chinese Academy of Geological Sciences, Langfang 065000, China

³ Geochemical Research Center of Soil Quality, China Geological Survey, Langfang 065000, China

⁴ College of Information Science and Engineering, Hohai University, Nanjing 211100, China

* Correspondence: xuzhen@iee.org

Highlights

What are the main findings?

- This study proposed a novel inversion method based on the fusion of geochemical data and PCA-dimensionality-reduced spectral data.
- This study demonstrated the superior performance of the Random Forest model on the fused data.

What is the implication of the main finding?

- This study innovatively integrates direct and precise geochemical data with macroscopic and continuous spectral data. By applying Principal Component Analysis to reduce the dimensionality and noise of the high-dimensional spectral data, core features are effectively extracted, resulting in a fused dataset with more comprehensive information. This method overcomes the limitations of using a single data source and significantly improves the inversion accuracy for mapping the spatial distribution of soil arsenic.
- Among various machine learning models, this study conclusively verifies the exceptional effectiveness of the Random Forest algorithm in processing the multi-source fused data. The model proficiently captures the complex non-linear relationships between arsenic content and multi-source environmental features. Its inherent resistance to overfitting and capability for feature importance assessment provide a reliable tool for high-precision arsenic inversion and pollution mechanism analysis.



Academic Editor: Somsubhra Chakraborty

Received: 29 September 2025

Revised: 2 November 2025

Accepted: 5 November 2025

Published: 10 November 2025

Citation: Guo, F.; Xu, Z.; Ma, H.; Liu, X. Estimating Soil Arsenic Contamination by Integrating Hyperspectral and Geochemical Data with PCA and Optimizing Inversion Models. *Sensors* **2025**, *25*, 6857. <https://doi.org/10.3390/s25226857>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract

Soil arsenic (As) contamination presents serious threats to ecosystems and human health, necessitating the development of accurate and efficient monitoring techniques. This study introduces a novel multi-source data fusion approach to enhance the hyperspectral inversion of soil arsenic concentrations by integrating dimensionality-reduced spectral data with soil components significantly correlated with arsenic (e.g., Cd, Cr, Cu, Ni, Pb, Zn, S, and total Fe₂O₃(T-Fe₂O₃)). Principal Component Analysis (PCA) was utilized to reduce the dimensionality of hyperspectral data, effectively addressing issues of collinearity and redundancy while preserving critical spectral information. The performances of three models, namely Partial Least Squares Regression (PLSR), Artificial Neural Networks (ANN), and

Random Forest (RF), were assessed under four input variable combinations: (1) original spectral data, (2) original spectral data with soil components, (3) PCA dimensionality-reduced spectral data, and (4) PCA dimensionality-reduced spectral data combined with soil components. The results demonstrated that the RF model, when applied to the multi-source data of PCA-reduced spectra and soil components, achieved the highest inversion accuracy with an R^2 value of 0.86, significantly outperforming the PLSR model ($R^2 = 0.75$). This study underscores the effectiveness of enhancing model performance and highlights the superior capability of the RF model in handling complex, high-dimensional datasets. The findings of soil arsenic estimation provide theoretical foundation for optimizing hyperspectral remote sensing technology in monitoring soil heavy metal contamination and establishing a robust framework for future research and practical applications in environmental science.

Keywords: hyperspectral; soil arsenic contamination; PCA; multi-source data fusion

1. Introduction

Arsenic (As), a toxic metalloid prevalent in natural environments, poses severe threats to ecosystems and human health through excessive accumulation [1,2]. Though it is not a true heavy metal [3,4], arsenic exhibits high toxicity, environmental mobility, and behavioral similarities to heavy metals [1]. Consequently, arsenic is generally termed a “pseudo-heavy metal” or “toxic element” and is systematically studied alongside heavy metals in environmental science and pollution management [5,6]. Soil arsenic contamination, driven by its persistence, bioaccumulation, and toxicity, poses significant risks to ecological integrity and public health [2,7]. From this point of view, accurate monitoring and estimation of soil arsenic concentrations are crucial for advancing environmental governance, pollution mitigation, and food safety protocols [8].

Traditional soil heavy metal detection employs field sampling with laboratory chemical analysis, supplemented by geostatistical interpolation to map spatial contamination patterns [9,10]. While offering high precision, these methods face practical constraints including labor intensity, time consumption, and elevated operational costs [11]. By contrast, Visible and Near-Infrared Reflectance (VNIR) hyperspectral spectroscopy presents a transformative alternative through its rapid, cost-effective, and non-destructive advantages [12]. Thus, the VNIR hyperspectral spectroscopy technique can capture high-resolution quasi-continuous spectral data, enabling comprehensive characterization of soil composition and facilitating efficient heavy metal detection [13,14]. Nevertheless, three inherent data challenges—such as high dimensionality, spectral collinearity, and information redundancy—often degrade model performance through overfitting and reduced inversion accuracy, hindering its practical application [15,16]. Therefore, developing optimized data processing workflows and enhanced modeling frameworks remains critical for advancing hyperspectral applications in soil pollution monitoring.

Beyond Principal Component Analysis (PCA), common dimensionality reduction methods include Linear Discriminant Analysis (LDA)—a supervised linear approach, Locally Linear Embedding (LLE), and t-distributed Stochastic Neighbor Embedding (t-SNE) for non-linear scenarios, Auto encoders (AE) driven by deep learning, and Kernel Principal Component Analysis (KPCA)—a non-linear extension of PCA. While each of these methods has its own strengths (e.g., LDA excels at classification tasks, t-SNE is adept at preserving local clustering structures, and AE can adapt to complex data distributions), they all suffer from notable drawbacks compared to PCA: generally, they face higher computational

complexity and stronger reliance on parameters or additional information (such as labels and large sample sizes). In contrast, PCA boasts advantages of linear efficiency, stability, and reliability, as well as strong interpretability, making it more versatile across various application scenarios [17]. Principal Component Analysis (PCA) demonstrates substantial advantages in hyperspectral inversion through effective dimensionality reduction [18,19]. By compressing high-dimensional spectral data into a few principal components (typically accounting for >95% cumulative variance), PCA eliminates inter-band collinearity, preserves critical spectral features, and suppresses noise, thereby enhancing model accuracy and computational efficiency [19,20]. Empirical studies validate PCA's utility in soil heavy metal inversion, where it reduces hundreds of spectral bands to 3–5 interpretable components while maintaining diagnostic spectral signatures [18,21,22]. Beyond terrestrial applications, PCA has proven effective in aquatic hyperspectral monitoring, particularly for quantifying suspended sediments and chlorophyll-a concentrations [23,24]. However, the linear transformation mechanism of PCA struggles with non-linear spectral response patterns inherent in complex soil matrices. Recent advancements propose synergistically integrating PCA with machine learning architectures and multi-source geospatial data to address this limitation and expand its operational scope [25].

While hyperspectral data can partially characterize soil properties, its inherent limitations constrain its ability to meet the requirements for high-precision inversion [26,27]. The VNIR data mainly captures the spectral response properties of the soil surface [28], whereas the distribution and concentration of soil heavy metals (such as arsenic, cadmium, and lead) are influenced by a variety of environmental factors, including soil components (such as organic matter, iron oxides, and clay minerals), pH value, redox conditions, and human activities [5,29]. Soil constituents not only provide a critical environmental context but also compensate for the shortcomings of spectral data in characterizing the internal chemical properties of soil [30]. For example, iron oxides (such as hematite and goethite) have a strong adsorption capacity for heavy metals like arsenic, and their content is closely related to heavy metal concentrations. Similarly, sulfides can form insoluble compounds with arsenic under reducing conditions, thereby affecting the mobility and bioavailability of arsenic [31,32]. In this view, exclusive reliance on spectral inputs risks omitting environmentally mediated mechanisms, ultimately compromising inversion accuracy.

The critical role of inversion models in predicting soil heavy metal concentrations necessitates judicious model selection and optimization to ensure prediction accuracy, reliability, and operational feasibility [18,33]. Different inversion models demonstrate distinct performances when it comes to processing hyperspectral data and estimating soil compositional parameters [26]. Linear inversion models like Partial Least Squares Regression (PLSR) excel in handling high-dimensional datasets through latent variable extraction, effectively mitigating spectral collinearity in strongly linear relationships [34,35]. However, their capacity to interpret complex nonlinear interactions remains constrained. Conversely, non-linear models, including but not limited to Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN), can better capture the complex non-linear relationships between spectral data and soil heavy metal concentrations [36–38]. Among the non-linear models, the RF model stands out in multi-source data fusion, capable of combining spectral data and soil component information to significantly improve inversion accuracy while being highly robust to noise and outliers [39]. The SVM leverages kernel functions to process nonlinear patterns, making it suitable for limited-sample scenarios [40]. The ANN, despite superior nonlinear fitting capabilities, demands high-quality training data, large sample sizes, and careful regularization to prevent overfitting [41,42]. Accordingly, optimal model selection must balance data characteristics (dimensionality, sample size, noise levels), research objectives, and computational resources.

To tackle the aforementioned challenges and limitations, this study introduces an innovative multi-source fusion framework integrating dimensionality-reduced hyperspectral features with arsenic-associated soil constituents. PCA-transformed spectral data and geochemically active soil components (e.g., iron oxides, organic matter) substantially enhance the accuracy and operational robustness of the arsenic inversion model. Systematic performance benchmarking evaluated three machine learning architectures in terms of PLSR, RF, and ANN under different input variable combinations to determine the optimal modeling strategy.

The main contributions of this study are threefold: (1) developing a multi-source data fusion framework to deal with hyperspectral data constraints in heavy metal inversion; (2) applying PCA for optimized spectral dimensionality reduction while retaining essential spectral information; and (3) comprehensively evaluating model performance through comprehensive metrics to identify the optimal model for inverting the soil arsenic concentration. This study not only provides a scientific basis and technical support for hyperspectral remote sensing monitoring of soil heavy metal pollution but also offers valuable insights into multi-source data fusion and model optimization. The proposed framework can be extended to similar toxic elements and other environmental monitoring scenarios, contributing to the advancement of remote sensing technology in the field of environmental science.

2. Materials and Methods

2.1. Description of the Study Area

The study area is located in Daye City, Hubei Province, whose topography exhibits a distinctive pattern: ascending southward, and descending northward, while maintaining a relatively flat and uniform terrain along the east–west axis [43]. Geographically, the area extends between 114°31′ E to 115°20′ E and 29°40′ N to 30°15′ N (Figure 1). The terrain consists primarily of hills, mountains, and plains, with elevations ranging from 120 to 200 m above sea level [18]. Daye City possesses abundant mineral resources with significant diversity. However, the extensive mining and smelting operations in the region have inevitably led to soil contamination in the surrounding areas [44,45].

2.2. Soil Sampling and Data Preprocessing

As shown in Figure 1, 56 soil samples were collected across the study area. The sampling locations were strategically distributed to ensure comprehensive coverage of diverse land use patterns and soil conditions. Surface soil samples (0–20 cm depth) were extracted at each sampling point using a standardized stainless steel soil sampler, with each sample weighing more than 1000 g [18,46]. Sample preparation followed a standardized protocol: air-drying, grinding, and sieving through a 2 mm (10-mesh) nylon screen [47]. Following preparation, each sample was divided into two portions: one designated for chemical analysis in the laboratory, and the other allocated for spectroscopic reflectance measurements.

2.2.1. Laboratory Analysis

Laboratory analysis was conducted on the collected soil samples to determine the actual concentrations of heavy metals and other soil elements, establishing ground truth data for model validation. The soil's arsenic concentrations were specifically measured using determination of arsenic, antimony, and bismuth by hydride atomic fluorescence spectrometry. The analytical methods for other elements (e.g., Cd, Cr, Cu, Hg...) are detailed in Table 1. In this study, quality control measurements, including the employment

of standard reference materials and duplicate samples, were implemented throughout the laboratory analysis process to ensure analytical accuracy and data reliability.

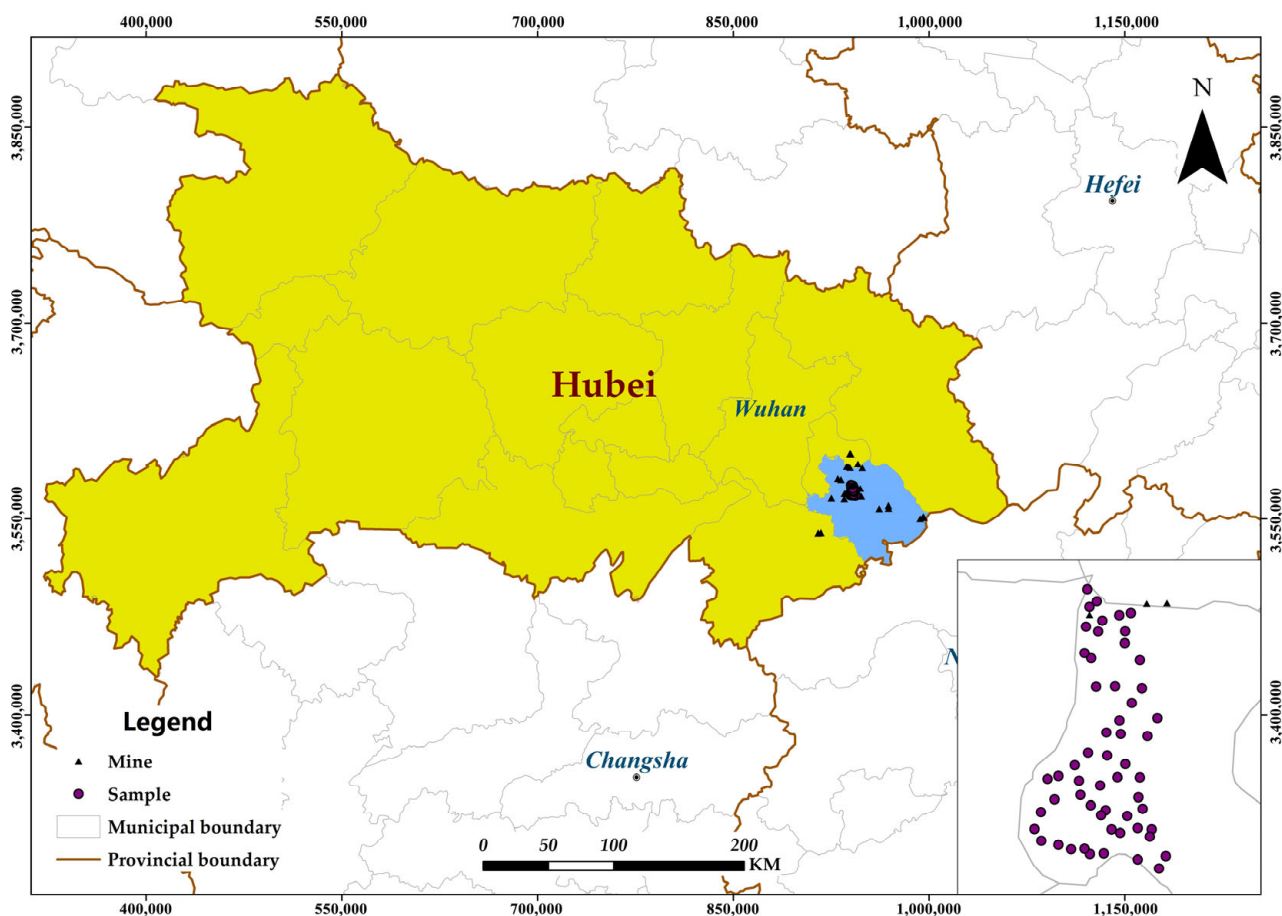


Figure 1. Overview of the study area and sampling points.

Table 1. The analytical methods of soil elemental composition.

Elements	Analytical Methods
As	Determination of arsenic, antimony, and bismuth by hydride atomic fluorescence spectrometry
Cd	Determination of 32 trace elements by plasma mass spectrometry
Cr	Determination of 34 primary, secondary, and trace elements by X-ray fluorescence spectrometry
Cu	Determination of 32 trace elements by plasma mass spectrometry
Hg	Determination of mercury by cold vapor atomic fluorescence spectrometry
Ni	Determination of 32 trace elements by plasma mass spectrometry
P	Determination of 34 primary, secondary, and trace elements by X-ray fluorescence spectrometry
Pb	Determination of 32 trace elements by plasma mass spectrometry
SiO ₂	Determination of 34 primary, secondary, and trace elements by X-ray fluorescence spectrometry
Al ₂ O ₃	Determination of 34 primary, secondary, and trace elements by X-ray fluorescence spectrometry
T-Fe ₂ O ₃	Determination of 34 primary, secondary, and trace elements by X-ray fluorescence spectrometry
MgO	Determination of 22 elements by plasma optical emission spectrometry
CaO	Determination of 34 primary, secondary, and trace elements by X-ray fluorescence spectrometry
Na ₂ O	Determination of 22 elements by plasma optical emission spectrometry
K ₂ O	Determination of 34 primary, secondary, and trace elements by X-ray fluorescence spectrometry
pH	Determination of pH value of forest soil
SOM	Determination of total carbon and organic carbon by high frequency combustion-infrared carbon sulfur meter

2.2.2. Hyperspectral Data Acquisition

The soil hyperspectral data were measured using an ASD FieldSpec4 spectrometer (Analytical Spectral Device, Inc., Longmont, CO, USA) with a spectral range of 350–2500 nm

(Visible and Near-Infrared Reflectance (VNIR) hyperspectral spectroscopy) [34]. The instrument features varying sampling intervals: 1.4 nm for the 350–1100 nm range and 2 nm for the 1000–2500 nm range [48]. Through resampling, the final output provided 2151 spectral bands at uniform 1 nm intervals.

The spectral measurements were conducted under controlled laboratory conditions. A 50 W halogen lamp served as the illumination source in a dark room environment. The samples were positioned at a 15° angle and 50 cm from the light source to prevent shadowing effects. The optical probe is mounted about 7 cm above the sample surface. To ensure the measurement stability, the spectroradiometer was preheated for 30 min before the data collection. Prior to the first measurement, the instrument was calibrated using a standard white BaSO₄ reference panel [49]. Each soil sample was homogeneously distributed in a Petri dish, from which 10 consecutive spectral measurements were acquired. Finally, the spectral reflectances with wavelength ranges of 350–399 nm and 2450–2500 nm were excluded from this analysis due to their low signal-to-noise ratio (SNR) [18,48,50].

As presented in Figure 2, the spectral reflectance curves exhibit three distinct absorption features centered at the wavelengths of approximately 1400 nm, 1900 nm, and 2200 nm. These absorption valleys, particularly pronounced after spectral preprocessing, are characteristic of soil clay mineral compositions. The presence and depth of these absorption features directly correlate with the soil's mineralogical properties [51].

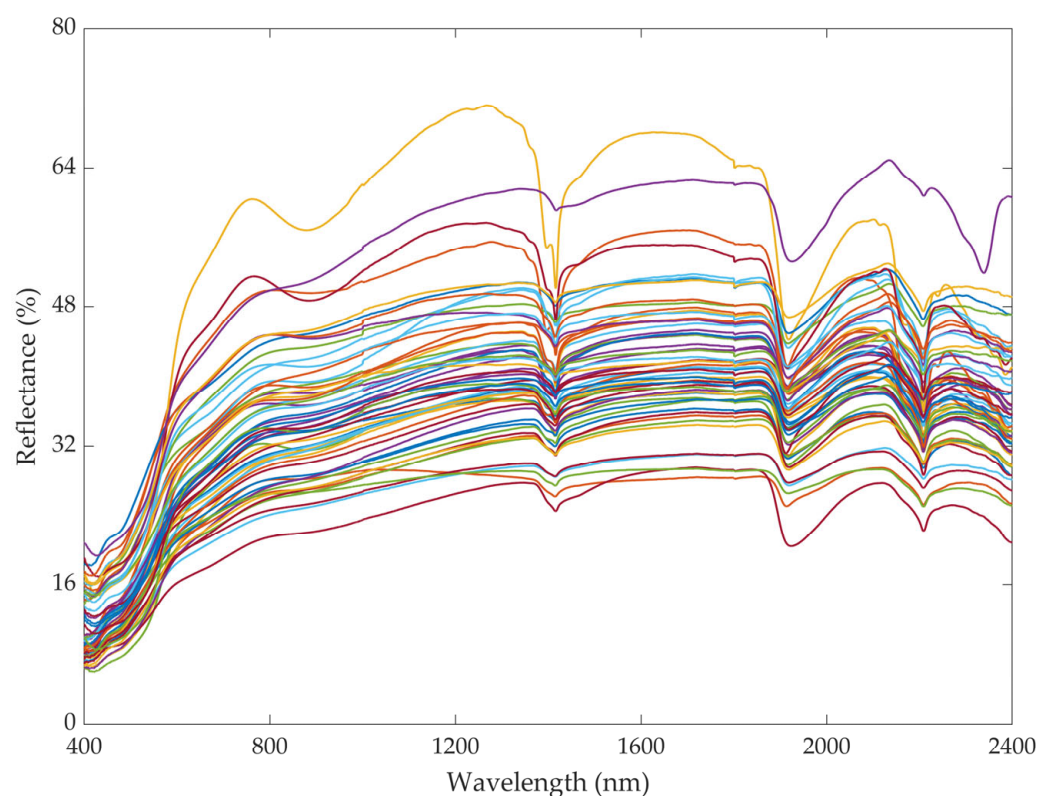


Figure 2. The original spectral reflectance.

2.3. Model Input Variable Filtering

2.3.1. Principal Component Analysis

PCA is an effective dimensionality reduction technique optimized for feature extraction and visualization of high-dimensional data [52,53]. The method employs a systematic process of data standardization, covariance matrix computation, eigenvalue decomposition, and projection to effectively reduce high-dimensional data into a lower-dimensional space while preserving essential information characteristics [54]. The fundamental principle

of PCA operates by identifying directions of maximum variance (principal components) within the dataset [55]. These principal components are orthogonal vectors arranged in descending order of explained variance. Generally, the first principal component captures the highest variance, followed by subsequent components with progressively decreasing variance contributions. Numerous studies have demonstrated that PCA plays a significant role in improving model inversion accuracy for heavy metal estimation based on hyperspectral reflectance data. The method's effectiveness in reducing data dimensionality while maintaining crucial spectral information has made it a fundamental processing step in hyperspectral data analysis [56,57].

2.3.2. Correlation Analysis

Correlation analysis is a statistical methodology for quantifying the magnitude and directionality of relationships across different variables [58]. Correlation analysis identifies linear relationships crucial for feature selection in modeling [59,60]. The metric used to measure the linear relationship between variables ranges from -1 to $+1$, where a value near -1 represents a strong negative correlation, while a value near 1 indicates a strong positive correlation [61]. The Pearson correlation coefficient, specifically, is widely used to identify features that exhibit strong correlations with target variables. This coefficient serves as a crucial tool for selecting input variables in regression analysis and classification models. In this research, we employ correlation analysis to identify soil properties that demonstrate significant relationships with arsenic contamination levels. By integrating these correlated soil properties with dimensionally reduced hyperspectral data, we aim to enhance both the prediction accuracy and practical applicability of the arsenic contamination model.

2.4. Model Construction and Validation

2.4.1. Modeling Method

Various inversion models were employed in this study, wherein the PLSR is a highly effective multivariate statistical modeling approach specifically designed for handling high-dimensional data, such as hyperspectral measurements, and addressing multicollinearity issues [62]. Originally developed by Wold et al. in 1983 [18], PLSR has become one of the most widely adopted linear modeling techniques. The method operates by extracting latent variables that link independent variables (hyperspectral bands) with dependent variables (heavy metal concentrations) to construct a linear regression model [18,63]. The fundamental principle of PLSR lies in its ability to maximize covariance between independent and dependent variables during dimensionality reduction, thereby efficiently capturing essential data patterns. PLSR demonstrates several key advantages: it effectively manages high-dimensional datasets, resolves multicollinearity issues, and maintains robustness against noise and outliers in the data structure [64]. In hyperspectral inversion applications, PLSR has proven particularly valuable for estimating soil heavy metal concentrations.

ANN has emerged as a powerful tool for predicting heavy metal concentrations from hyperspectral data due to its exceptional nonlinear modeling capabilities [65,66]. The architecture of ANN, comprising multiple neural layers and nonlinear activation functions (ReLU, Sigmoid), effectively captures nonlinear relationships between the hyperspectral signatures and heavy metal concentrations. The key strength of ANN lies in its adaptive learning mechanism and generalization ability. This allows for automatic feature extraction from large datasets and optimal parameter adjustment through training [67]. In hyperspectral applications, ANN processes spectral bands as input nodes and produces heavy metal concentrations as output, with nonlinear transformations occurring through hidden layers. ANN demonstrates superior performance compared to traditional linear models, particularly in handling nonlinear relationships and high-dimensional hyperspectral data.

The model's accuracy can be further enhanced by incorporating additional environmental variables, such as soil properties and topographic data. However, ANN implementation faces several challenges: it requires large datasets and careful tuning to prevent overfitting, and complex hyperparameter optimization. These limitations can be effectively addressed through cross-validation, regularization, and early stopping techniques [68]. Overall, ANN holds broad application prospects in the hyperspectral inversion of heavy metals, providing reliable technical support for environmental monitoring and pollution control.

RF represents an ensemble learning approach that has proven highly effective for heavy metal concentration estimation using hyperspectral data [69,70]. The method operates by constructing multiple decision trees and aggregating their predictions, demonstrating robust performance in handling high-dimensional spectral data [69]. The advantages of RF include automatic feature selection capability, efficient dimensionality reduction, and strong resilience to noise and outliers in hyperspectral measurements. The ensemble nature of RF enhances prediction accuracy and stability compared to single-model approaches, particularly in soil heavy metal estimation [70]. Additionally, the method requires minimal parameter tuning and offers straightforward implementation and interpretation. However, RF may face computational challenges with high-dimensional datasets and show limitations with small sample sizes, which can be addressed through strategic feature selection and hyperparameter optimization procedures. Overall, RF has broad application potential in the hyperspectral inversion of heavy metals, providing an efficient and reliable tool for environmental monitoring and pollution assessment.

2.4.2. Model Validation

The models are validated using cross-validation techniques to ensure their generalizability and robustness. The performance of the models is evaluated using metrics such as the coefficient of determination (R^2), root mean square error (RMSE), and Residual Prediction Deviation (RPD) [71].

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

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

$$RPD = \frac{S.D.}{RMSE} \quad (3)$$

where n represents the number of samples; $\bar{y}_i$ and $\hat{y}_i$ denote the measured and predicted values in the validation set, respectively; $\bar{y}_i$ is the mean of the measured values; and S.D. stands for the standard deviation.

A robust model is typically defined by high R^2 and RPD values coupled with a low RMSE. R^2 and RPD serve as key metrics to evaluate the accuracy of inversion performance. In contrast, RMSE is affected by the range of the measured values. The classification of RPD values is as follows: an RPD greater than 2.0 signifies excellent inversion performance; an RPD between 1.4 and 2.0 indicates the model's ability to distinguish between high and low values; while an RPD below 1.4 denotes poor inversion performance [48,72].

2.5. Data Treatment

When screening independent variables, the following steps were carried out: first, Principal Component Analysis (PCA) was used to reduce the dimensionality of the spectral data (Python 3.7.2 software, Wilmington, DE, USA) for extracting principal components; meanwhile, correlation analysis was applied to analyze the soil components related to the

target variable (SPSS 27.0), and variables with high correlation were selected as another part of the input variables for the model. Finally, PLSR, ANN, and RF models were used to estimate the content of arsenic (Python 3.7.2). In addition, the data statistical work involved in this study was completed using SPSS 27. All maps were generated using ArcMap 10.2 within the GIS software (ArcGIS 10.2, Esri, Redlands, CA, USA). The spectral curve graph was plotted using MATLAB Version 2013b (Matlab Inc., Natick, MA, USA).

3. Results

3.1. Arsenic Contamination of Soil

In this study, 56 soil samples were collected and divided into 38 calibration samples and 18 validation samples, maintaining an approximate 7:3 ratio [18]. Table 2 summarizes the arsenic concentrations across the samples, including the mean, maximum, minimum, standard deviation (SD), and coefficient of variation (CV). The results revealed that the average arsenic concentration in the study area was 23.01 mg/kg, with a maximum concentration reaching 57.08 mg/kg. Compared to the natural background value of 11.2 mg/kg for soil arsenic concentration, as published by the China National Environmental Monitoring Centre, 48 sample points exceeded this threshold. Notably, the maximum arsenic concentration was over four times the natural background value, indicating severe arsenic contamination in the region.

Table 2. Statistical information for soil arsenic concentration (mg·kg^{−1}) in the study area.

Arsenic Contamination (mg·kg ^{−1})	Number	Mean	Max	Min	SD	CV
Calibration set	38	23.05	57.08	2.34	13.16	1.05
Validation set	18	22.95	54.57	6.14	12.26	0.53
Whole dataset	56	23.01	57.08	2.34	12.77	0.55

Additionally, Table 2 demonstrated significant spatial heterogeneity in the distribution of arsenic concentrations. While some areas showed minimal arsenic contamination, others were likely subjected to severe pollution. The coefficients of variation (CV) for all samples exceeded 0.36 [18], further corroborating the pronounced spatial variability of arsenic concentrations. This finding suggests the potential presence of point source pollution within the study area. Consequently, there is an urgent need to develop an accurate inversion model capable of monitoring arsenic pollution on a macro scale using hyperspectral remote sensing technology.

3.2. Statistical Results of Soil Properties

In this study, alongside measuring arsenic contamination in the 56 soil samples, other soil components and properties potentially associated with arsenic were analyzed. The statistical results of these measurements are summarized in Table 3. The findings revealed that the average soil pH in the study area was 6.05 ± 1.23 , with 66.1% of the sampling points exhibiting pH values below 6.5. This indicates that the surface soil in the study area is predominantly acidic. Additionally, the average soil organic matter (SOM) content was slightly lower than the average SOM level of surface soil in Hubei Province, which is 56.23 mg·kg^{−1} [73].

Table 3. Statistical results of soil properties (mg·kg^{−1}) in the study area.

Soil Properties	Mean	Max	Min	SD	CV
Cd	0.64	2.11	0.04	0.40	0.62
Cr	65.24	116.91	10.55	25.19	0.39

Table 3. Cont.

Soil Properties	Mean	Max	Min	SD	CV
Cu	89.04	320.86	21.68	64.48	0.72
Hg	0.11	0.41	0.02	0.07	0.63
Ni	24.77	47.98	5.70	10.46	0.42
Pb	65.28	592.44	17.30	75.66	1.16
Zn	135.47	401.09	47.43	65.64	0.48
P	850.30	3339.70	179.60	495.96	0.58
S	261.79	549.42	54.94	114.24	0.44
SiO ₂	65.63	76.62	7.45	9.83	0.15
Al ₂ O ₃	14.39	24.48	1.88	3.28	0.23
T-Fe ₂ O ₃	5.66	8.16	1.20	1.27	0.22
MgO	0.75	2.46	0.29	0.33	0.45
CaO	2.00	45.00	0.07	6.12	3.06
Na ₂ O	0.50	2.28	0.05	0.44	0.89
K ₂ O	1.88	2.78	0.31	0.41	0.22
SOM	38.66	75.97	6.66	18.57	0.48
pH	6.05	8.09	3.87	1.23	0.20

3.3. Filtering of Model Input Variables

3.3.1. Soil Components Associated with Arsenic

In this study, Pearson's correlation analysis was conducted to evaluate the relationships between arsenic concentrations and selected soil parameters, including heavy metals, pH, and potential arsenic-associated soil components. The objective was to identify key variables significantly influencing arsenic concentrations for inclusion as additional input variables in the inversion model. As shown in Figure 3, the analysis revealed strong correlations between arsenic and several heavy metals, specifically Cd, Cr, Cu, Ni, Pb, and Zn, with the exception of mercury (Hg). Additionally, significant correlations were identified between arsenic and sulfur (S) as well as total iron oxides (T-Fe₂O₃). These relationships are likely attributable to synergistic interactions among heavy metals, competitive adsorption processes, and the high adsorption capacity of iron oxides for arsenic species [74].

Figure 3 shows the pairwise correlations between elements. Green indicates a positive correlation, with darker shades representing a more significant positive correlation; yellow indicates a negative correlation, with darker shades representing a more significant negative correlation. Additionally, the size of the circle is proportional to the color intensity, both of which denote a stronger correlation. In the figure, all non-significant correlations between elements are marked with an "×". Based on these findings, the variables Cd, Cr, Cu, Ni, Pb, Zn, S, and T-Fe₂O₃ were incorporated into the model as auxiliary input parameters to improve the accuracy and reliability of arsenic concentration inversion. This approach effectively captures the complex interactions between arsenic and associated soil components while enhancing the data framework for hyperspectral remote sensing applications in monitoring soil heavy metal contamination. Consequently, this methodology contributes to advancing the technical robustness of geospatial pollution assessments.

3.3.2. Spectral Dimensionality Reduction Using PCA

In this study, PCA was employed to reduce the dimensionality of the original spectral data by selecting principal components with eigenvalues greater than 1. These components were subsequently used as selected input variables for the arsenic contamination inversion model. The dimensionality reduction results, detailed in Table 4, indicate that 7 principal components have eigenvalues exceeding 1, with a cumulative contribution rate of 99.9%. This high cumulative contribution rate demonstrates that these components effectively preserve the vast majority of the information contained in the original spectral data.

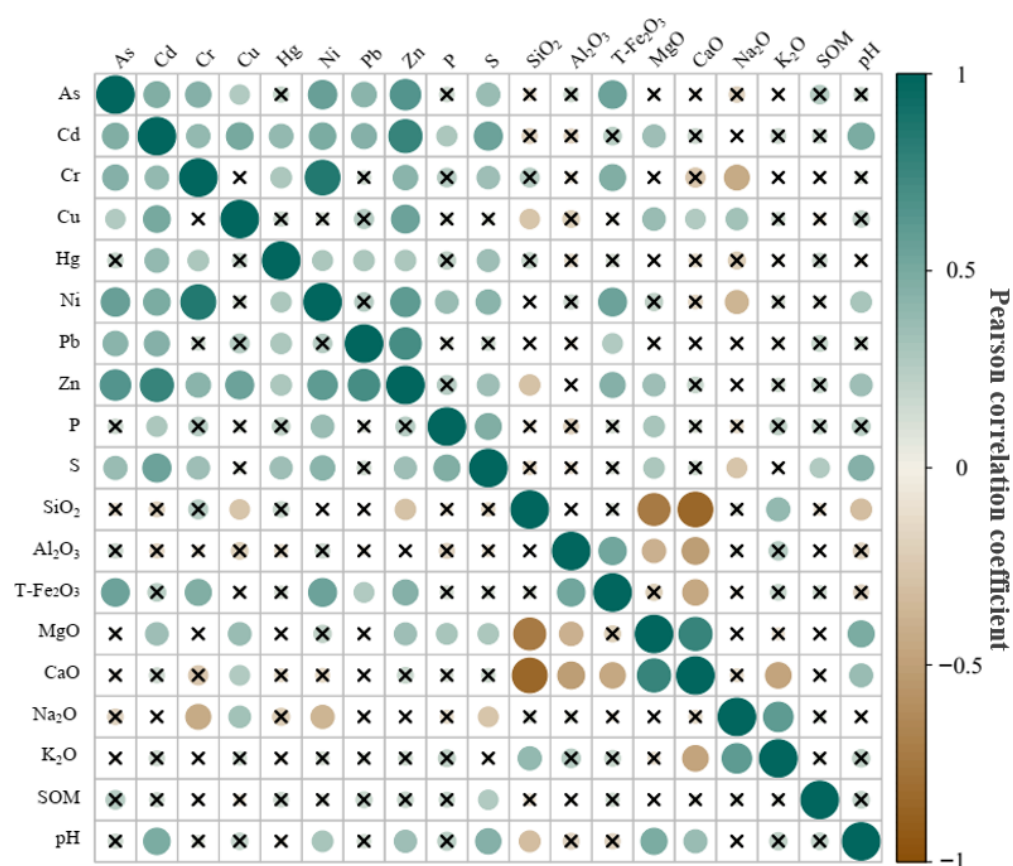


Figure 3. Results of correlation analysis of soil components related to arsenic elements.

Table 4. Results of principal component dimensionality reduction analysis.

Components	Eigenvalue	Variance (%)	Cumulative Contribution Rate (%)
1	1847.74	92.34	92.34
2	77.84	3.89	96.23
3	46.67	2.33	98.56
4	16.54	0.83	99.39
5	7.48	0.37	99.76
6	1.52	0.08	99.84
7	1.11	0.06	99.90

By implementing PCA, data dimensionality was significantly reduced, leading to a decrease in model complexity. Simultaneously, redundant information and noise in the spectral data were eliminated, thereby enhancing the model's computational efficiency and prediction accuracy. The ability of the selected principal components to adequately represent the features of the original spectra establishes a robust data foundation for subsequent arsenic contamination inversion modeling. These findings underscore the effectiveness of PCA in dimensionality reduction and information extraction from spectral data, providing essential support for the utilization of hyperspectral remote sensing technology in monitoring soil heavy metal pollution.

3.4. Performance of Inversion Models

In this study, four input variable combinations were designed for arsenic concentration inversion modeling: (1) original spectral data; (2) original spectral data combined with soil components significantly correlated with arsenic, as determined through correlation analysis; (3) principal components derived from dimensionality reduction; and (4) principal

components combined with soil components significantly correlated with arsenic, as identified through correlation analysis. To evaluate the performance of these combinations, both linear models and nonlinear models were employed for inversion modeling. In addition, a scatter plot was drawn using 18 samples from the validation set, which clearly reflects the correlation between the predicted values and the measured values.

By comparing the performance of different input variable combinations and modeling approaches, the study identified the optimal input variable combination and the most effective model for arsenic concentration inversion in the study area. This research provides valuable scientific evidence and technical support for the application of hyperspectral remote sensing technology in monitoring soil heavy metal contamination.

3.4.1. Modeling of Original Spectral Data

The original spectral data were used for hyperspectral inversion modeling of arsenic concentrations in the study area, employing both linear models (PLSR) and nonlinear models (ANN and RF). The scatter plots of the inversion results are presented in Figure 4. The results revealed that the inversion accuracies of the PLSR, ANN, and RF models were -0.14 , -0.22 , and -0.06 , respectively.

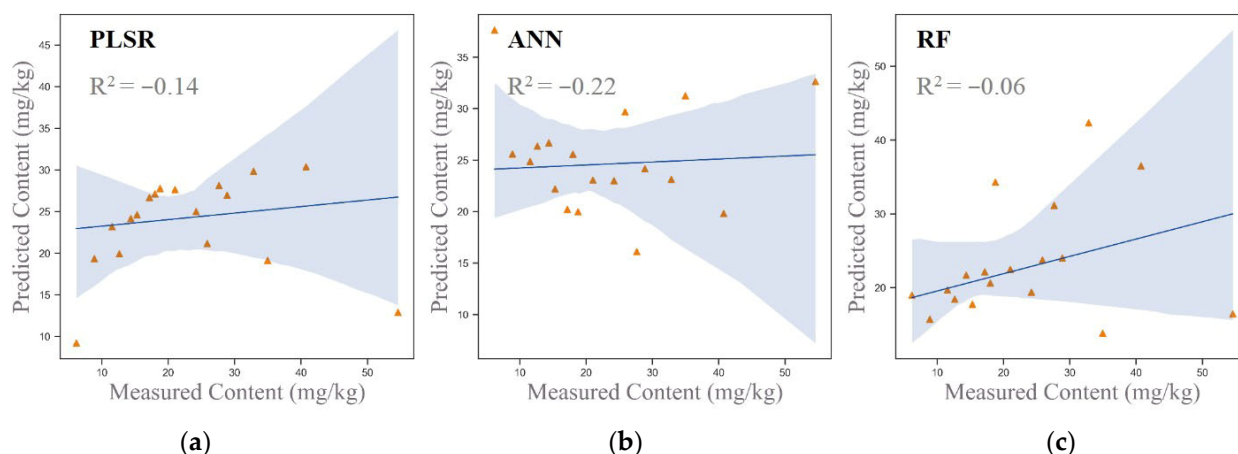


Figure 4. Scatter plots of original spectral data using three models.

All three models demonstrated negative accuracy values, indicating significant overfitting. It may be caused by the small number of training samples. This outcome suggests that using only the original spectral data is insufficient for accurately predicting soil arsenic concentrations. The underlying issues may include noise, redundant information, or unresolved nonlinear relationships within the spectral data. To address these limitations, further optimization of the input variables or the inclusion of additional auxiliary data is required to improve the models' generalization capabilities and prediction accuracy.

3.4.2. Modeling of Original Spectral Data Combined with Arsenic-Correlated Soil Components

To address the overfitting issue observed in the inversion of soil arsenic concentrations using only the original spectral data, this study integrated soil components significantly correlated with arsenic as additional input variables. These variables were combined with the original spectral data to enhance inversion accuracy. Using the combined dataset, inversion modeling was conducted with three models: PLSR, ANN, and RF. The scatter plots of the inversion results are shown in Figure 5.

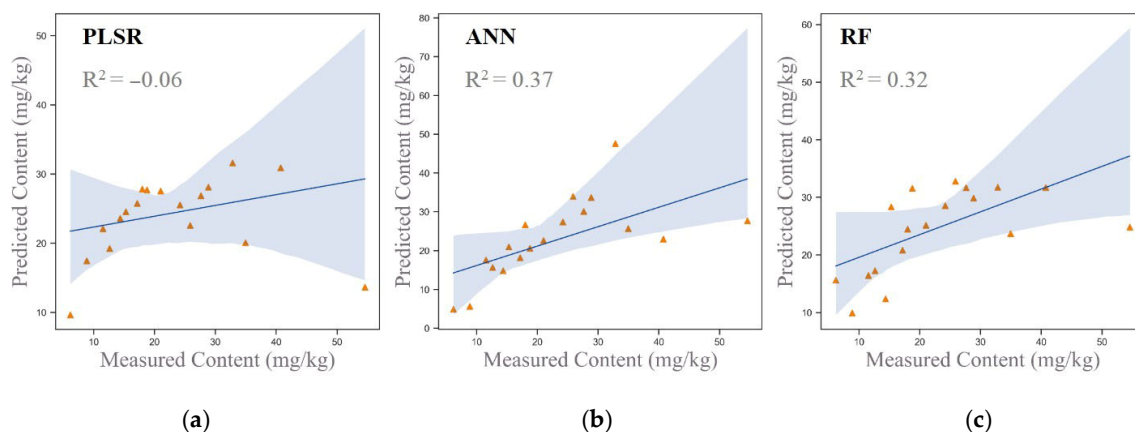


Figure 5. Scatter plots of original spectral data combined with arsenic-correlated soil components using three models.

The results revealed that the inversion accuracies of the PLSR, ANN, and RF models were -0.06 , 0.37 , and 0.32 , respectively. Compared to the results obtained using solely original spectral data, the ANN and RF models exhibited substantial improvements in inversion accuracy after incorporating arsenic-correlated soil components. This improvement suggests that the inclusion of soil components effectively enhances the models' predictive capabilities. The additional environmental information provided by the soil components compensates for the limitations of spectral data alone in characterizing arsenic concentrations. Such findings demonstrate the effectiveness of a multi-source data modeling approach that integrates spectral data and soil components, which provides valuable technical support for improving the accuracy and reliability of hyperspectral inversion of arsenic concentrations.

3.4.3. Modeling of the Principal Component by PCA

One of the primary reasons for the poor performance of original spectral data in inverting heavy metal concentrations is the presence of severe collinearity issues and an excessive number of independent variables, a challenge extensively validated in prior research. To address this, the study applied dimensionality reduction to the original spectral data, leveraging PCA to mitigate collinearity and remove redundant information. The dimensionality-reduced spectral data were then used as input variables for the models. Inversion modeling was performed using three models: PLSR, ANN, and RF, with the results presented in Figure 6.

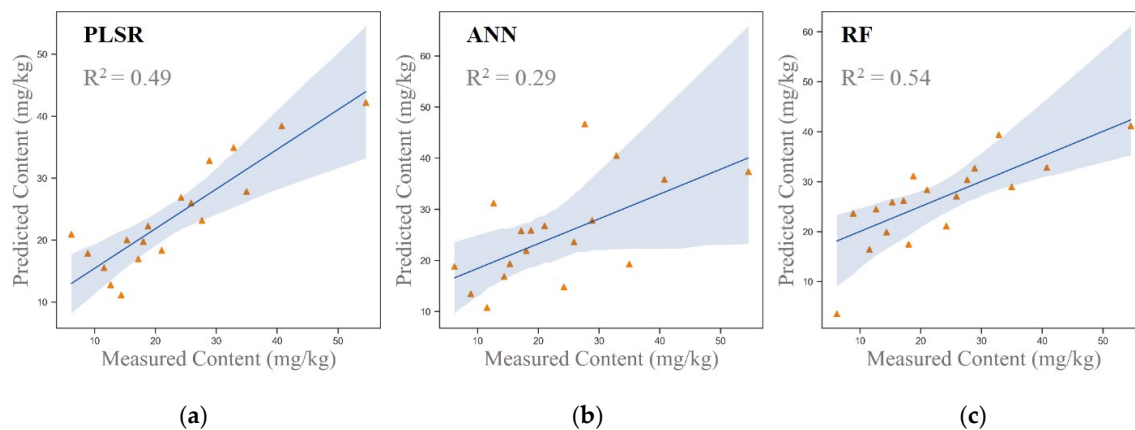


Figure 6. Scatter plots of principal components by PCA using three models.

Figure 6 shows that the inversion accuracies of the PLSR, ANN, and RF models were 0.49, 0.29, and 0.54, respectively, with the RF model achieving the highest accuracy. The overall performance ranking was RF > PLSR > ANN. These findings highlight the significant improvement in model performance after PCA-based dimensionality reduction. The reduction in data dimensionality and the elimination of noise and redundant information enhanced the models' generalization capabilities and prediction accuracy. This demonstrates the effectiveness of PCA in improving the reliability and robustness of inversion modeling for soil heavy metal concentration estimation.

3.4.4. Hybrid Modeling of Principal Components and Arsenic-Correlated Soil Components

By integrating soil components significantly correlated with arsenic and applying dimensionality reduction to spectral data, both of which have been shown to enhance model inversion performance, this study developed a new set of hybrid input variables. This approach reduces the dimensionality of the spectral data while retaining its essential spectral features and incorporates the additional environmental information provided by soil components. Together, these inputs were used to jointly invert soil arsenic concentrations. The scatter plots of the inversion results are shown in Figure 7.

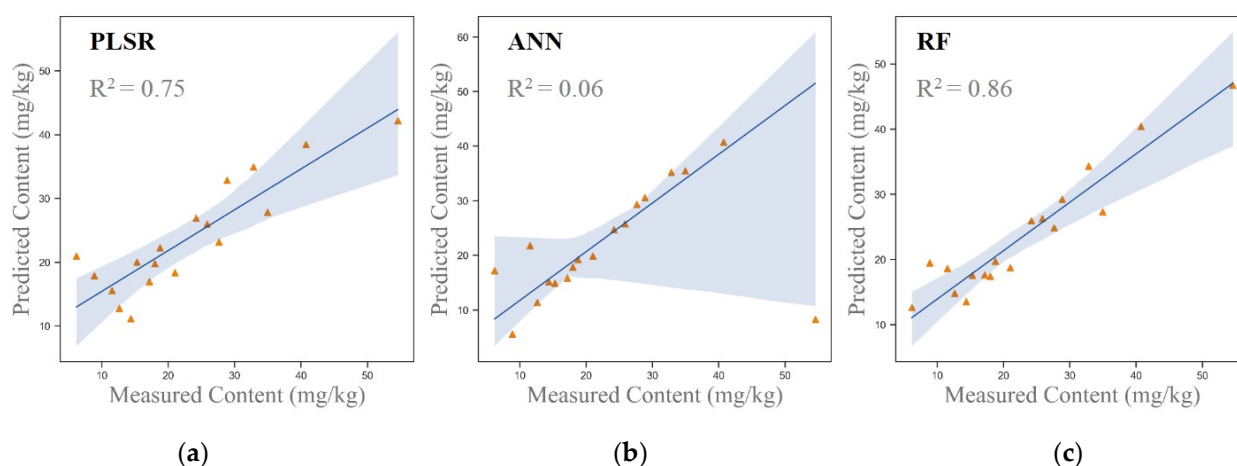


Figure 7. Scatter plots of PCs combined with arsenic-correlated soil components using three models.

The results in Figure 7 indicate that the hybrid input variables significantly improved inversion accuracy. The PLSR and RF models achieved accuracies of 0.75 and 0.86, respectively, successfully inverting arsenic concentrations in the study area. However, the ANN model performed poorly, with an inversion accuracy of only 0.06. This underperformance is likely due to the ANN model's high sensitivity to outliers and errors in individual samples during the inversion process. These findings underscore the effectiveness of the multi-source data modeling approach, which combines spectral dimensionality reduction with soil components. This strategy markedly enhances the accuracy and reliability of arsenic concentration inversion, offering a robust framework for the application of hyperspectral remote sensing in soil heavy metal pollution monitoring.

4. Discussion

The arsenic (As) content in soil can be detected via hyperspectral data and is significantly influenced by various soil components. Specifically, under reducing conditions, sulfides react with arsenic to form insoluble arsenic sulfides (e.g., As_2S_3), which effectively immobilize arsenic. Iron oxides such as goethite and hematite, through adsorption or coprecipitation, also fix arsenic, leading to a significant correlation between their content and arsenic levels. Additionally, arsenic often coexists with heavy metals like cadmium

(Cd) and chromium (Cr) in mineral deposits, with their concentrations showing positive correlations due to co-enrichment during mineralization. These soil components thus serve as indirect indicators of arsenic variability, acting as effective auxiliary variables in hyperspectral inversion of soil arsenic contamination and playing a crucial role in enhancing inversion accuracy [75].

In previous studies on the hyperspectral inversion of arsenic elements (Table 5), the focus has mostly been on the processing and analysis of hyperspectral data itself to establish the relationship between arsenic content and spectral characteristics. For instance, some studies [12,18] have improved the correlation between spectra and arsenic content through preprocessing of hyperspectral data, such as first-order differentiation and multiplicative scatter correction, before developing inversion models. The distinctive feature of this study lies in the integration of soil component information into the input variables for inversion. Certain soil components, such as iron oxides, have absorption characteristic bands that may correlate with arsenic content. By incorporating such soil component information into the input variables, this study effectively compensates for the singularity of spectral information, providing the inversion model with more data relevant to the target variable (arsenic content). This approach is expected to enhance the accuracy and reliability of the inversion. Notably, this mechanism has rarely been considered in previous hyperspectral inversion studies of arsenic elements, making it an innovative research approach.

Table 5. Prediction of soil arsenic contents using the reflectance spectroscopy of soils.

Sampling Site	Content Range (mg/kg)	Model	R ²	Number of Samples	Authors
Agricultural regions	1.91–21.90	GA-PLSR	0.56–0.64	96	[76]
Agricultural area at mine	19.33–403.77	PLSR	0.58	33	[77]
Agricultural area at the Changjiang River Delta	6.13–13.30	PLSR	0.72	61	[78]
Agricultural area	10.25–133.36	GA-PLSR	0.42	94	[75]
Agricultural regions	2.34–57.08	RF	0.86	56	This study

PCA, whose effectiveness in spectral dimensionality reduction has been fully validated in previous studies [18,20], was applied to the original spectral data in this research. Seven principal components with eigenvalues > 1.0 were selected, accounting for a cumulative contribution rate of 99.9% (Table 4). These components retained key spectral information while significantly reducing data dimensionality, effectively resolving redundancy and severe collinearity inherent in spectral data. Additionally, soil components strongly correlated with arsenic (Cd, Cr, Cu, Ni, Pb, Zn, S, and T-Fe₂O₃) were incorporated as auxiliary input variables. This multi-source data fusion approach, integrating the merits of spectral dimensionality reduction and environmental information from soil components, notably enhanced the model's performance in arsenic concentration estimation. These findings offer robust technical support for promoting the application of hyperspectral remote sensing in soil heavy metal contamination monitoring.

To evaluate the estimation performance of different input variable combinations across three models (PLSR, ANN, and RF), this study designed four input variable combinations: Combination 1 (original spectral data), Combination 2 (original spectral data combined with soil components significantly correlated with As), Combination 3 (principal components derived from dimensionality reduction), and Combination 4 (principal components combined with soil components significantly correlated with As). The results, as shown in Figure 6, indicate that the PLSR and RF models achieved their best performance with Combination 4, obtaining inversion accuracies of 0.75 and 0.86, respectively, confirming that the multi-source data fusion approach combining dimensionality-reduced spectra and soil components significantly enhanced model performance. In contrast, the ANN model

exhibited poor performance across all combinations, with the highest inversion accuracy of only 0.32. This underperformance is likely due to the sensitivity of the ANN model to small sample sizes and its challenges in handling nonlinear relationships effectively. Overall, the RF model demonstrated superior performance in the context of multi-source data fusion, making it a reliable choice for hyperspectral inversion of arsenic concentrations.

The comprehensive comparison (Table 6) highlights that multi-source data fusion effectively addresses the issues of collinearity and redundancy in spectral data, leading to significant improvements in inversion accuracy. The outstanding performance of the PLSR and RF models with Combination 4 underscores the scientific value of integrating dimensionality-reduced spectral data with soil components. However, the limitations observed in the ANN model suggest the need for optimization of its network structure or the incorporation of regularization techniques to improve its performance.

Table 6. Compared the inversion performance of different input variables in the three models.

Input Variable	Model	R ²	RMSE	RPD
Combination 1	PLSR	−0.14	12.74	0.96
	ANN	−0.22	13.14	0.93
	RF	−0.06	12.30	1.00
Combination 2	PLSR	−0.06	12.29	1.00
	ANN	0.37	9.49	1.29
	RF	0.32	9.80	1.25
Combination 3	PLSR	0.49	8.52	1.44
	ANN	0.29	10.03	1.22
	RF	0.54	8.11	1.51
Combination 4	PLSR	0.75	5.91	2.07
	ANN	0.06	11.55	1.06
	RF	0.86	4.45	2.75

In conclusion, this study confirms that integrating multi-source data related to soil arsenic can enhance the accuracy of arsenic content prediction. It not only provides a solid scientific basis and technical support for the application of hyperspectral remote sensing technology in soil heavy metal contamination monitoring, but also offers valuable insights into the effectiveness of multi-source data fusion and the significance of selecting appropriate inversion models. Meanwhile, several geographic parameters, such as distance to cities, slope gradient, distance to roads, and distance from pollution sources, may affect the spatial distribution of heavy metals [79]. Therefore, it is recommended that further research be conducted on the integration of hyperspectral technology with geostatistical methods based on geographic information systems (GIS), so as to further improve the prediction accuracy of arsenic content in agricultural soils [80]. Furthermore, the application of remote sensing data from different platforms, such as airborne and spaceborne hyperspectral data, represents a potential future direction for estimating soil heavy metals.

5. Conclusions

This study enhanced the accuracy and reliability of soil arsenic concentration inversion using hyperspectral remote sensing through the integration of multi-source data fusion, model performance optimization, and dimensionality reduction techniques. The exceptional performance of the RF model in multi-source data fusion and the effectiveness of PCA for dimensionality reduction offer valuable theoretical and methodological support for advancing hyperspectral remote sensing applications in soil heavy metal pollution monitoring. The key conclusions drawn from this study are as follows:

(1) Significant Improvement in Inversion Accuracy through Multi-Source Data Fusion

Combining dimensionality-reduced spectral data with these soil components (Combination 4) yielded the best results in both the PLSR and RF models, with inversion accuracies of 0.75 and 0.86, respectively. This highlights the scientific importance and practical potential of multi-source data fusion methods in enhancing inversion accuracy.

(2) Optimal Performance of the RF Model

The RF model consistently achieved higher inversion accuracy across all input variable combinations, particularly when multi-source data fusion was applied. This establishes the RF model as highly suitable for hyperspectral inversion of soil arsenic concentrations, offering robust and reliable performance.

(3) Effectiveness of Spectral Dimensionality Reduction

Principal Component Analysis (PCA) effectively addressed the challenges of redundancy in hyperspectral data. By retaining key spectral information while reducing dimensionality, PCA significantly enhanced model performance, providing essential technical support for processing hyperspectral data in arsenic concentration inversion tasks.

The results underscore the value of integrating advanced data processing techniques, such as multi-source data fusion and PCA, with robust modeling approaches like RF, to advance hyperspectral remote sensing technology for monitoring and assessing soil heavy metal contamination.

Author Contributions: Conceptualization, F.G.; methodology, F.G.; software, F.G.; validation, H.M. and X.L.; formal analysis, F.G.; investigation, F.G.; resources, F.G.; data curation, Z.X.; writing—original draft preparation, F.G.; writing—review and editing, Z.X.; visualization, Z.X.; supervision, Z.X., H.M. and X.L.; project administration, F.G. and Z.X.; funding acquisition, F.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Geological Survey Project of the China Geological Survey under Grant DD20221770, in part by the Fundamental Research Funds for the Central Universities under Grant B250201238, in part by the Director Foundation of the Institute of Geophysical and Geochemical Exploration, Chinese Academy of Geological Sciences, under Grant AS2019J02, and in part by the National Science Foundation of China under Grant 42101398.

Data Availability Statement: The author's institution has strict requirements for data. If data is needed, please contact the author and provide it for use only after obtaining permission from the institution.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fatoki, J.O.; Badmus, J.A. Arsenic as an environmental and human health antagonist: A review of its toxicity and disease initiation. *J. Hazard. Mater. Adv.* **2022**, *5*, 100052. [\[CrossRef\]](#)
2. Jomova, K.; Jenisova, Z.; Feszterova, M.; Baros, S.; Liska, J.; Hudecova, D.; Rhodes, C.J.; Valko, M. Arsenic: Toxicity, oxidative stress and human disease. *J. Appl. Toxicol.* **2011**, *31*, 95–107. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Rae, I.D. Arsenic: Its chemistry, its occurrence in the earth and its release into industry and the environment. *ChemTexts* **2020**, *6*, 25. [\[CrossRef\]](#)
4. Tamás, M.; Sharma, S.; Ibstedt, S.; Jacobson, T.; Christen, P. Heavy Metals and Metalloids As a Cause for Protein Misfolding and Aggregation. *Biomolecules* **2014**, *4*, 252–267. [\[CrossRef\]](#)
5. Xu, W.; Jin, Y.; Zeng, G. Introduction of heavy metals contamination in the water and soil: A review on source, toxicity and remediation methods. *Green Chem. Lett. Rev.* **2024**, *17*, 2404235. [\[CrossRef\]](#)
6. Moreno-Jiménez, E.; Manzano, R.; Esteban, E.; Peñalosa, J. The fate of arsenic in soils adjacent to an old mine site (Bustarviejo, Spain): Mobility and transfer to native flora. *J. Soils Sediments* **2010**, *10*, 301–312. [\[CrossRef\]](#)
7. Smith, E.; Naidu, R.; Alston, A.M. Arsenic in the Soil Environment: A Review. *Adv. Agron.* **1998**, *64*, 149–195.
8. Li, W.; Higgins, P. Controlling Local Environmental Performance: An analysis of three national environmental management programs in the context of regional disparities in China. *J. Contemp. China* **2013**, *22*, 409–427. [\[CrossRef\]](#)

9. Sun, W.; Zhang, X.; Sun, X.; Sun, Y.; Cen, Y. Predicting nickel concentration in soil using reflectance spectroscopy associated with organic matter and clay minerals. *Geoderma* **2018**, *327*, 25–35. [\[CrossRef\]](#)
10. Ma, C.L.; Zhou, J.M.; Wang, H.Y.; Du, C.W.; Huang, B. Methods for assessment of heavy metal pollution in cropland soils—A case study of Changshu. *J. Ecol. Rural. Environ.* **2006**, *22*, 48–53.
11. Guo, F.; Xu, Z.; Ma, H.; Liu, X.; Gao, L. On Optimizing Hyperspectral Inversion of Soil Copper Content by Kernel Principal Component Analysis. *Remote Sens.* **2024**, *16*, 2914. [\[CrossRef\]](#)
12. Delaney, J.K.; Pezzati, L.; Salimbeni, R.; Zeibel, J.G.; Thoury, M.; Littleton, R.; Morales, K.M.; Palmer, M.; de la Rie, E.R. Visible and infrared reflectance imaging spectroscopy of paintings: Pigment mapping and improved infrared reflectography. In Proceedings of the O3a: Optics for Arts, Architecture, & Archaeology II, Munich, Germany, 14–18 June 2009; p. 739103.
13. Chen, X.; Warner, T.A.; Campagna, D.J. Integrating visible, near-infrared and short-wave infrared hyperspectral and multispectral thermal imagery for geological mapping at Cuprite, Nevada: A rule-based system. *Int. J. Remote Sens.* **2010**, *31*, 1733–1752. [\[CrossRef\]](#)
14. Bilgili, A.V.; Akbas, F.; Es, H.M.V. Combined use of hyperspectral VNIR reflectance spectroscopy and kriging to predict soil variables spatially. *Precis. Agric.* **2011**, *12*, 395–420. [\[CrossRef\]](#)
15. Morales, G.; Sheppard, J.W.; Logan, R.D.; Shaw, J.A. Hyperspectral Dimensionality Reduction Based on Inter-Band Redundancy Analysis and Greedy Spectral Selection. *Remote Sens.* **2021**, *13*, 3649. [\[CrossRef\]](#)
16. Aishwarya, G.; Kumar, B.L.N.P.; Syamala, D.; Sreya, N.S. Dimensionality Reduction Technique for Hyperspectral Remote Sensing Image Classification. In Proceedings of the 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 1–3 June 2023; pp. 1694–1698.
17. Anowar, F.; Sadaoui, S.; Selim, B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput. Sci. Rev.* **2021**, *40*, 100378. [\[CrossRef\]](#)
18. Guo, F.; Xu, Z.; Ma, H.; Liu, X.; Tang, S.; Yang, Z.; Zhang, L.; Liu, F.; Peng, M.; Li, K. Estimating chromium concentration in arable soil based on the optimal principal components by hyperspectral data. *Ecol. Indic.* **2021**, *133*, 108400. [\[CrossRef\]](#)
19. Salem, N.; Hussein, S. Data dimensional reduction and principal components analysis. *Procedia Comput. Sci.* **2019**, *163*, 292–299. [\[CrossRef\]](#)
20. Zabalza, J.; Ren, J.; Yang, M.; Zhang, Y.; Wang, J.; Marshall, S.; Han, J. Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2014**, *93*, 112–122. [\[CrossRef\]](#)
21. Howley, T.; Madden, M.G.; O’Connell, M.-L.; Ryder, A.G. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowl.-Based Syst.* **2006**, *19*, 363–370. [\[CrossRef\]](#)
22. Rodarmel, C.; Shan, J. Principal Component Analysis for Hyperspectral Image Classification. *Surv. Land Inf. Sci.* **2002**, *62*, 115–122.
23. Wallace, J.; Champagne, P.; Hall, G. Multivariate statistical analysis of water chemistry conditions in three wastewater stabilization ponds with algae blooms and pH fluctuations. *Water Res.* **2016**, *96*, 155–165. [\[CrossRef\]](#)
24. Haji Gholizadeh, M.; Melesse, A.M.; Reddi, L. Water quality assessment and apportionment of pollution sources using APCS-MLR and PMF receptor modeling techniques in three major rivers of South Florida. *Sci. Total Environ.* **2016**, *566–567*, 1552–1567. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Rodionova, O.; Kucheryavskiy, S.; Pomerantsev, A. Efficient tools for principal component analysis of complex data—A tutorial. *Chemom. Intell. Lab. Syst.* **2021**, *213*, 104304. [\[CrossRef\]](#)
26. Wang, Y.; Zou, B.; Li, S.; Tian, R.; Zhang, B.; Feng, H.; Tang, Y. A hierarchical residual correction-based hyperspectral inversion method for soil heavy metals considering spatial heterogeneity. *J. Hazard. Mater.* **2024**, *479*, 135699. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Wang, S.; Guan, K.; Zhang, C.; Lee, D.; Margenot, A.J.; Ge, Y.; Peng, J.; Zhou, W.; Zhou, Q.; Huang, Y. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing. *Remote Sens. Environ.* **2022**, *271*, 112914. [\[CrossRef\]](#)
28. Dwivedi, R.S. Spectral Reflectance of Soils. In *Remote Sensing of Soils*; Ravi Shankar, D., Ed.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 267–303.
29. Wan, Y.; Liu, J.; Zhuang, Z.; Wang, Q.; Li, H. Heavy Metals in Agricultural Soils: Sources, Influencing Factors, and Remediation Strategies. *Toxics* **2024**, *12*, 63. [\[CrossRef\]](#)
30. Sharma, V.; Chauhan, R.; Kumar, R. Spectral characteristics of organic soil matter: A comprehensive review. *Microchem. J.* **2021**, *171*, 106836. [\[CrossRef\]](#)
31. Shi, M.; Min, X.; Ke, Y.; Lin, Z.; Yang, Z.; Wang, S.; Peng, N.; Yan, X.; Luo, S.; Wu, J.; et al. Recent progress in understanding the mechanism of heavy metals retention by iron (oxyhydr)oxides. *Sci. Total Environ.* **2021**, *752*, 141930. [\[CrossRef\]](#)
32. Chen, T.; Wen, X.; Zhou, J.; Lu, Z.; Li, X.; Yan, B. A critical review on the migration and transformation processes of heavy metal contamination in lead-zinc tailings of China. *Environ. Pollut.* **2023**, *338*, 122667. [\[CrossRef\]](#)

33. Boisson, J.; Rutters, A.; Mench, M.; Vangronsveld, J. Evaluation of hydroxyapatite as a metal immobilizing soil additive for the remediation of polluted soils. Part 1. Influence of hydroxyapatite on metal exchangeability in soil, plant growth and plant metal accumulation. *Environ. Pollut.* **1999**, *104*, 225–233. [\[CrossRef\]](#)
34. Zhang, S.; Shen, Q.; Nie, C.; Huang, Y.; Wang, J.; Hu, Q.; Ding, X.; Zhou, Y.; Chen, Y. Hyperspectral inversion of heavy metal content in reclaimed soil from a mining wasteland based on different spectral transformation and modeling methods. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *211*, 393–400. [\[CrossRef\]](#)
35. Tan, K.; Ye, Y.-Y.; Du, P.-J.; Zhang, Q.-Q. Estimation of Heavy Metal Concentrations in Reclaimed Mining Soils Using Reflectance Spectroscopy. *Spectrosc. Spectr. Anal.* **2014**, *34*, 3317–3322.
36. Zhou, W.; Yang, H.; Xie, L.; Li, H.; Yue, T. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. *Catena* **2021**, *202*, 105222. [\[CrossRef\]](#)
37. Chen, Y.; Shi, W.; Aihemaitijiang, G.; Zhang, F.; Zhang, J.; Zhang, Y.; Pan, D.; Li, J. Hyperspectral inversion of heavy metal content in farmland soil under conservation tillage of black soils. *Sci. Rep.* **2025**, *15*, 354. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Xiang, C.; Xiao, H.; He, F.; Dai, Z.; Huang, W.; Zhu, B.; Liu, S. Prediction of soil heavy metal content around mine tailings using multiple methods combined with transformed hyperspectral reflectance data. *Ore Energy Resour. Geol.* **2024**, *18*, 100072. [\[CrossRef\]](#)
39. Wang, W.; Liu, K.; Liu, C. Multi-source power data fusion method based on deep learning. In Proceedings of the Second International Conference on Energy, Power, and Electrical Technology (ICEPET 2023), Kuala Lumpur, Malaysia, 10–12 March 2023; pp. 903–908.
40. Zareapoor, M.; Shamsolmoali, P.; Kumar Jain, D.; Wang, H.; Yang, J. Kernelized support vector machine with deep learning: An efficient approach for extreme multiclass dataset. *Pattern Recognit. Lett.* **2018**, *115*, 4–13. [\[CrossRef\]](#)
41. Almeida, J.S. Predictive non-linear modeling of complex data by artificial neural networks. *Curr. Opin. Biotechnol.* **2002**, *13*, 72–76. [\[CrossRef\]](#)
42. Tealab, A.; Hefny, H.; Badr, A. Forecasting of nonlinear time series using ANN. *Future Comput. Inform. J.* **2017**, *2*, 39–47. [\[CrossRef\]](#)
43. Jiang, N.; Zhou, C.; Lu, S.; Zhang, Z. Effect of Underground Mine Blast Vibrations on Overlaying Open Pit Slopes: A Case Study for Daye Iron Mine in China. *Geotech. Geol. Eng.* **2018**, *36*, 1475–1489. [\[CrossRef\]](#)
44. Du, P.; Xie, Y.; Wang, S.; Zhao, H.; Zhang, Z.; Wu, B.; Li, F. Potential sources of and ecological risks from heavy metals in agricultural soils, Daye City, China. *Environ. Sci. Pollut. Res.* **2015**, *22*, 3498–3507. [\[CrossRef\]](#)
45. Xi, X.; Wang, S.; Yao, L.; Zhang, Y.; Niu, R.; Zhou, Y. Evaluation on geological environment carrying capacity of mining city—A case study in Huangshi City, Hubei Province, China. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102410. [\[CrossRef\]](#)
46. Hu, B.; Wang, J.; Jin, B.; Li, Y.; Shi, Z. Assessment of the potential health risks of heavy metals in soils in a coastal industrial region of the Yangtze River Delta. *Environ. Sci. Pollut. Res.* **2017**, *24*, 19816–19826. [\[CrossRef\]](#)
47. Malmir, M.; Tahmasbian, I.; Xu, Z.; Farrar, M.B.; Bai, S.H. Prediction of soil macro- and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique. *Geoderma* **2019**, *340*, 70–80. [\[CrossRef\]](#)
48. Sun, W.; Zhang, X. Estimating soil zinc concentrations using reflectance spectroscopy. *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *58*, 126–133. [\[CrossRef\]](#)
49. Cheng, H.; Shen, R.; Chen, Y.; Wan, Q.; Shi, T.; Wang, J.; Wan, Y.; Hong, Y.; Li, X. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. *Geoderma* **2019**, *336*, 59–67. [\[CrossRef\]](#)
50. Shen, Q.; Xia, K.; Zhang, S.; Kong, C.; Hu, Q.; Yang, S. Hyperspectral indirect inversion of heavy-metal copper in reclaimed soil of iron ore area. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *222*, 117191. [\[CrossRef\]](#)
51. Zhang, X.; Sun, W.; Cen, Y.; Zhang, L.; Wang, N. Predicting cadmium concentration in soils using laboratory and field reflectance spectroscopy. *Sci. Total Environ.* **2019**, *650*, 321–334. [\[CrossRef\]](#)
52. Raiko, T.; Ilin, A.; Karhunen, J. Principal Component Analysis for Sparse High-Dimensional Data. In Proceedings of the Neural Information Processing, Iconip, Kitakyushu, Japan, 13 November 2007; pp. 566–575.
53. 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*, 59–75. [\[CrossRef\]](#)
54. Van Wieringen, W.N.; Peeters, C.F.W. Ridge estimation of inverse covariance matrices from high-dimensional data. *Comput. Stat. Data Anal.* **2016**, *103*, 284–303. [\[CrossRef\]](#)
55. Bolcárová, P.; Kološta, S. Assessment of sustainable development in the EU 27 using aggregated SD index. *Ecol. Indic.* **2015**, *48*, 699–705. [\[CrossRef\]](#)
56. Lever, J.; Krzywinski, M.; Altman, N. Principal component analysis. *Nat. Methods* **2017**, *14*, 641–642. [\[CrossRef\]](#)
57. Maćkiewicz, A.; Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342. [\[CrossRef\]](#)
58. Reimann, C.; Filzmoser, P.; Hron, K.; Kynčlová, P.; Garrett, R.G. A new method for correlation analysis of compositional (environmental) data—A worked example. *Sci. Total Environ.* **2017**, *607–608*, 965–971. [\[CrossRef\]](#) [\[PubMed\]](#)
59. Gupta, B. Correlation and Regression. In *Interview Questions in Business Analytics*; Gupta, B., Ed.; Apress: Berkeley, CA, USA, 2016; pp. 45–55.

60. Harrington, P.d.B.; Urbas, A.; Tandler, P.J. Two-dimensional correlation analysis. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 149–174. [[CrossRef](#)]
61. Ratner, B. The correlation coefficient: Its values range between +1/−1, or do they? *J. Target. Meas. Anal. Mark.* **2009**, *17*, 139–142. [[CrossRef](#)]
62. Guebel, D.V.; Torres, N.V. Partial Least-Squares Regression (PLSR). In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013; pp. 1646–1648.
63. Cheng, J.-H.; Sun, D.-W. Partial Least Squares Regression (PLSR) Applied to NIR and HSI Spectral Data Modeling to Predict Chemical Properties of Fish Muscle. *Food Eng. Rev.* **2017**, *9*, 36–49. [[CrossRef](#)]
64. Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [[CrossRef](#)]
65. Raj, A.S.; Srinivas, Y.; Oliver, D.H.; Muthuraj, D. A novel and generalized approach in the inversion of geoelectrical resistivity data using Artificial Neural Networks (ANN). *J. Earth Syst. Sci.* **2014**, *123*, 395–411. [[CrossRef](#)]
66. Landi, A.; Piaggi, P.; Laurino, M.; Menicucci, D. Artificial Neural Networks for nonlinear regression and classification. In Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, Cairo, Egypt, 29 November–1 December 2010; pp. 115–120.
67. Demarchi, L.; Canters, F.; Cariou, C.; Licciardi, G.; Chan, J.C.-W. Assessing the performance of two unsupervised dimensionality reduction techniques on hyperspectral APEX data for high resolution urban land-cover mapping. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 166–179. [[CrossRef](#)]
68. Lee, K.Y.; Chung, N.; Hwang, S. Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. *Ecol. Inform.* **2016**, *36*, 172–180. [[CrossRef](#)]
69. Wang, Q.; Nguyen, T.-T.; Huang, J.Z.; Nguyen, T.T. An efficient random forests algorithm for high dimensional data classification. *Adv. Data Anal. Classif.* **2018**, *12*, 953–972. [[CrossRef](#)]
70. Fawagreh, K.; Gaber, M.M.; Elyan, E. Random forests: From early developments to recent advancements. *Syst. Sci. Control Eng.* **2014**, *2*, 602–609. [[CrossRef](#)]
71. Saeys, W.; Mouazen, A.M.; Ramon, H. Potential for Onsite and Online Analysis of Pig Manure using Visible and Near Infrared Reflectance Spectroscopy. *Biosyst. Eng.* **2005**, *91*, 393–402. [[CrossRef](#)]
72. Sawut, R.; Kasim, N.; Abliz, A.; Hu, L.; Yalkun, A.; Maihemuti, B.; Qingdong, S. Possibility of optimized indices for the assessment of heavy metal contents in soil around an open pit coal mine area. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *73*, 14–25. [[CrossRef](#)]
73. State Environmental Protection Administration; China National Environmental Monitoring Centre. *Background Values of Soil Elements in China*; China Environmental Science Press: Beijing, China, 1990.
74. Maliki, A.A.; Bruce, D.; Owens, G. Spatial distribution of Pb in urban soil from Port Pirie, South Australia. *Environ. Technol. Innov.* **2015**, *4*, 123–136. [[CrossRef](#)]
75. Shi, T.; Wang, J.; Chen, Y.; Wu, G. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 95–103. [[CrossRef](#)]
76. 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. [[CrossRef](#)]
77. Ren, H.-Y.; Zhuang, D.-F.; Singh, A.N.; Pan, J.-J.; Qiu, D.-S.; Shi, R.-H. Estimation of As and Cu Contamination in Agricultural Soils Around a Mining Area by Reflectance Spectroscopy: A Case Study. *Pedosphere* **2009**, *19*, 719–726. [[CrossRef](#)]
78. Wu, Y.; Chen, J.; Ji, J.; Gong, P.; Liao, Q.; Tian, Q.; Ma, H. A Mechanism Study of Reflectance Spectroscopy for Investigating Heavy Metals in Soils. *Soil Sci. Soc. Am. J.* **2007**, *71*, 918–926. [[CrossRef](#)]
79. Dragović, R.; Gajić, B.; Dragović, S.; Đorđević, M.; Đorđević, M.; Mihailović, N.; Onjia, A. Assessment of the impact of geographical factors on the spatial distribution of heavy metals in soils around the steel production facility in Smederevo (Serbia). *J. Clean. Prod.* **2014**, *84*, 550–562. [[CrossRef](#)]
80. Lee, C.S.-I.; Li, X.; Shi, W.; Cheung, S.C.-n.; Thornton, I. Metal contamination in urban, suburban, and country park soils of Hong Kong: A study based on GIS and multivariate statistics. *Sci. Total Environ.* **2006**, *356*, 45–61. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.