Estimating Soil Arsenic Content with Visible and Near-Infrared Hyperspectral Reflectance

Lei Han 1,2,3,*, Rui Chen 2,4,*, Huili Zhu 5, Yonghua Zhao 1,2, Zhao Liu 1,2 and Hong Huo 2,4

1 School of Land Engineering, Chang'an University, Xi'an 710054, China; yonghuaz@chd.edu.cn (Y.Z.);
lz13@chd.edu.cn (Z.L.)
2 Shaanxi Key Laboratory of Land consolidation, Xi'an 710054, China; huohong717@163.com
3 Degraded and Unused Land Consolidation Engineering, The Ministry of Land and Resources,
   Xi’an 710054, China
4 College of Earth Sciences and Resources, Chang’an University, Xi’an 710054, China
5 College of Geological Engineering and Geomatics, Chang’an University, Xi’an 710054, China;
   zhuhuili@chd.edu.cn
* Correspondence: hanshuanglei@chd.edu.cn (L.H.); chenruigis@163.com (R.C.)

Received: 6 January 2020; Accepted: 13 February 2020; Published: 17 February 2020

Abstract: Soil arsenic (AS) contamination has attracted a great deal of attention because of its detrimental effects on environments and humans. AS and inorganic AS compounds have been classified as a class of carcinogens by the World Health Organization. In order to select a high-precision method for predicting the soil AS content using hyperspectral techniques, we collected 90 soil samples from six different land use types to obtain the soil AS content by chemical analysis and hyperspectral data based on an indoor hyperspectral experiment. A partial least squares regression (PLSR), a support vector regression (SVR), and a back propagation neural network (BPNN) were used to establish a relationship between the hyperspectral and the soil AS content to predict the soil AS content. In addition, the feasibility and modeling accuracy of different interval spectral resampling, different spectral pretreatment methods, feature bands, and full-band were compared and discussed to explore the best inversion method for estimating soil AS content by hyperspectral. The results show that 10 nm + second derivative (SD) + BPNN is the optimum method to predict soil AS content estimation; \( R^2 \) is 0.846 and residual predictive deviation (RPD) is 2.536. These results can expand the representativeness and practicability of the model to a certain extent and provide a scientific basis and technical reference for soil pollution monitoring.

Keywords: estimation mechanism; soil AS contents; hyperspectral; back propagation neural network

1. Introduction

Arsenic (AS) is a kind of metalloid that is widely found in nature [1]. Atmospheric deposition, industrial production, sewage irrigation, and soil arsenic-containing pesticides cause soil AS pollution [2,3]. AS has strong neurotoxicity and teratogenicity and, moreover, the fact that AS compounds are relatively stable and do not easily degrade in the natural environment cause its accumulation in soil. Excessive AS content in the soil will not only harm the local ecological environment, but also cause irreparable damage to human health [4–6]. Therefore, how to obtain soil AS pollution information quickly and accurately has received much attention in recent years.

The traditional monitoring methods for soil AS pollution are field sampling followed by chemical analysis and rapid on-site monitoring by elemental analysis instruments. Traditional monitoring methods collect soil samples in the field and take them back to the laboratory to determine the soil AS content. These methods can obtain accurate information but are time-consuming, labor-intensive,
cost-intensive, and inefficient, and they are subject to limited soil spatial heterogeneity and analyzing costs [7], and it is difficult to achieve repeated periodic sampling in a short time. Even though the on-site rapid monitoring method has the advantages of fast, continuous, and high-density information acquisition, it is mostly in the qualitative or semi-quantitative experimental stage and is susceptible to surrounding factors [8]. Therefore, both laboratory and on-site monitoring methods have some limitations in obtaining surface pollution characteristics quickly and accurately. In recent years, several spectral technologies have been developed using hyperspectral [9]. Spectral analysis refers to the analysis method using the principal and experimental method of spectroscopy to find out the chemical composition of a substance [10]. Visible/near-infrared [11], thermal infrared, and even ultraviolet spectroscopy can be used to estimate soil element contents because it is rapid, non-destructive, environmental friendly, and cost effective. Therefore, it can quickly obtain the soil AS pollution by establishing the soil AS pollution parameter model based on spectral analysis principles. Much research in recent years has focused on this. For example, Zheng et al. [12] suggested that it is feasible to predict AS element contents in soils using reflectance spectral, and the model results of 4 nm + multiplicative scatter correction (MSC) + partial least squares regression (PLSR) were the best ($R^2 = 0.711$, residual predictive deviation (RPD) = 1.827); Cheng et al. [13] reported that AS contents in surface soils were detectable using visible/near-infrared spectral, and Savitzky–Golay (SG)+PLSR had the best effect ($R^2 = 0.75$, RPD = 1.81). Many previous studies have investigated models for the estimation of AS content from visible and near infrared (VNIR) hyperspectral [12–20]. PLSR was usually chosen as the estimation model [20]. PLSR and support vector regression (SVR) are linear modeling methods, which can effectively solve the linear problem. However, there is a clear non-linear relationship between the reflectance spectrum and soil AS content. There may be some shortcomings in using a linear model to deal with problems with a nonlinear relationship. The disadvantages of these methods are the single land use type and the unrepresentative models. In addition, because of the nonlinearity between soil AS content and spectral reflectance, it is difficult to estimate the soil AS content accurately using linear regression methods.

In the present study, six different types of land use in a gold mine tailing in Shangluo City and a suburb in Weinan City, Shaanxi Province, China were considered as research areas. Through collecting soil samples in the field and indoor chemical analysis, statistics on soil AS content in a mining area and a suburb were collected, and the spectrum of the soil samples was determined. The PLSR, SVR, and back propagation neural network (BPNN) methods were used to establish the relationship between soil hyperspectral reflectance and soil AS content. This experiment illustrates that it is feasible to estimate soil AS content by BPNN and hyperspectral, and improve the accuracy and stability of the model. This study provides useful guidance for rapid and extensive monitoring of soil AS pollution and environmental management.

2. Materials and Methods

2.1. Study Area

Shangluo City is located in the southeastern part of Shaanxi Province. It is between 108°34′20″−111°1′25″ E and 33°2′30″−34°2′40″ N. The total area is 19,851 square kilometers, which belongs to the monsoon climate zone. The soil type is yellow cinnamon soil. The Zhen’an Gold Mine operated in 1993. In 2006, the “4.30” dam failure occurred. After 13 years of treatment, it achieved certain results, but bare slag poses a serious threat to human health and was listed as one of the national key areas for heavy metal prevention and control, and after field soil sampling and analysis in July 2018, we found that the concentration of AS in the soil was seriously exceeded.

Weinan City is located in the middle reaches of the Yellow River, east of the Guanzhong Plain in Shaanxi Province, between 108°58′−110°35′ E and 34°13′−35°52′ N. Weinan City has a warm temperate semi-humid and semi-arid monsoon climate with four distinct seasons, sufficient sunshine and suitable rainfall. The main soil type is fluvo-aquic soil and yellow loess soil. Weinan has experienced rapid
urbanization over the past several decades. There are more than 20 types of polluting enterprises in the suburb, such as chemical plants, funeral homes, medical waste treatment plants, and so on. These polluting enterprises are also direct discharge areas of wastewater, waste gas, and waste slag in Weinan City. In recent years, heavy metal pollution caused by city expansion has drastically affects the soil ecological environment and human health.

2.2. Acquisition and Processing of Soil Data

The results of previous field investigations indicated that the soil AS content in the mining area is mainly affected by human activities and shows a large difference, and the suburban soil AS content is mainly affected by land use types and shows a small difference. Therefore, from November to December, 2018, 90 soil samples were collected in accordance with the uniform grid method with a higher sampling density in the mining area (50 m) and a lower density in the suburb (200 m). The 50 soil samples were collected from the mining area and the rest from the suburb. There were six members in the research group with unified training on the sampling method. First, sample points were selected on the satellite map, and then the six members were divided into two groups to finish the sampling job in each area. A real-time kinematic (RTK) was used to precisely locate every sampling location. Figure 1 shows the position of the sampling points. The sampling was conducted at a depth of 0–20 cm. Stones, plant residues, and other large debris were removed from each fresh sample, which were then mixed thoroughly and then stored in a labelled plastic bag. Each sample weighted 500 g. All samples were air dried at room temperature. Small stones and plant residuals were manually removed and the soil samples were then run through a 2 mm sieve. The samples were divided into three parts, one of which was used for chemical determination of pH and soil AS content; one was used to measure spectral reflectance; and one was sealed for backup. The content of AS in the soil was determined by the atomic fluorescence spectrometry method; the pH value of the soil was measured by the glass electrode method.

Figure 1. Study area and the distribution of soil sampling points; A–F denote six types of land use, (a) suburb, Weinan City, (b) mining area, Shangluo City.
2.3. Collection and Processing of Soil Spectral Data

The spectral reflectance of the soil samples was measured in a dark room using a FieldSpec4 field spectrum analyzer manufactured by American ASD Corporation. The measurements were conducted in a dark room with a 50 W halogen lamp as a light source, which was positioned 0.3 m away from the samples, with a zenith angle of 25°. The optical probe was perpendicular to the soil surface and about 10 cm away from the samples. The radiance of the standard white reflection panel was used to calibrate the instrument before collecting each of the five samples. The average of the 10 spectral curves for each soil sample was used as the reflectance spectrum of the soil samples.

Spectral pretreatment can effectively reduce the error caused by random influences in soil spectral data. In this paper, the 350–399 and 2401–2500 nm bands with low signal-noise ratio were excluded. The spectral data was resampled at intervals of 2 nm, 4 nm, 6 nm, 8 nm, 10 nm, 12 nm, and 14 nm, respectively. FD, SD, and MSC were performed on the spectral data to highlight the absorption and reflection characteristics of the spectral curve and eliminate redundant information between bands. Spectral resampling is one of the basic steps of spectral preprocessing, which has an important influence on the accuracy of the hyperspectral prediction model. With the increase of the spectral resampling interval, the spectral absorption band changes accordingly, which will affect the inversion accuracy [21]. The SG algorithm is a digital filtering algorithm that can smooth the spectral curve and improve the accuracy of the data without losing signal trends. Furthermore, for AS elements, spectral FD transformation can improve stability and accuracy requirements; SD transformation can enhance hidden information in the spectral data, amplify feature differences between bands, eliminate partial redundancy and noise, and reduce the random error caused by factors such as illumination and improper operation; MSC can correct the baseline offset effect caused by spectral scattering [21]. These spectral transformation methods have been widely used in hyperspectral vegetation and soil research, and have achieved good results [22–24].

2.4. Extraction Feature Bands

Since there are 2151 bands in the original spectral data, the high correlation between the bands leads to data redundancy, which brings great difficulties for spectral data analysis and processing. Therefore, it is necessary to reduce the dimensionality of the spectral data. Principal Component Analysis (PCA) is a commonly used method of dimensionality reduction [25]. Assuming that the original data is an m rows and n columns matrix X, the main steps of principal component analysis are:

Step 1: Zero-average each row of X;
Step 2: Calculate the covariance matrix:

\[ C = \frac{1}{m} (XX^T); \]  \hspace{1cm} (1)

Step 3: Calculate the eigenvalues of the covariance matrix and the corresponding eigenvectors;
Step 4: Arrange the feature vectors into a matrix from the largest to the smallest according to the corresponding feature value, and taking the first k rows to form a matrix P, Y = P × X is the data after dimension reduction to k dimensions.

After the principal component analysis of the spectral data, the first principal component variables with the largest explained variance can be calculated. According to each principal component variable, the correlation loadings corresponding to each band can be deduced. The greater the absolute value, the stronger the correlation between the band and the principal component, and the greater the contribution to the principal component. Therefore, a band having a large absolute value of correlation loadings can be selected as a feature band.
2.5. Model Calibration and Validation Methods

2.5.1. Partial Least Squares Regression

PLSR is a spectral analysis method that includes multiple linear regression, canonical correlation analysis, and principal factor analysis, and was proposed by Herman O. A. Wold [26]. The main research content of PLSR is the method of establishing a linear model for independent variables in the case of a large number of two sets of variables with high linear correlation in order to solve the problem that the number of samples is smaller than the number of variables and avoid over-fitting. The principle of PLSR is: Firstly, extract the mutually independent components $T_h (h = 1, 2, \ldots)$ from the independent variables $(x_1, x_2, \ldots, x_m)$, and the extracted principal components carry as many originals as possible; then extracting the independent components $U_h = (h = 1, 2, \ldots)$ from the dependent variable $(y_1, y_2, \ldots, y_m)$ requires that the covariance between $T_h$ and $U_h$ is maximized and establishes the regression equation between the extracted components and the dependent variable through the multivariate regression method. The basic models of partial least square regression are:

$$X = T_h P^T + E$$
$$Y = U_h Q^T + E$$

where $P$ and $Q$ are the orthogonal load matrices of $m \times h$, respectively, and $E$ and $F$ are the error terms, which are random variables obeying the normal distribution.

2.5.2. Support Vector Regression

Support Vector Machine (SVM) is a machine learning method proposed by Professor Vapnik et al. in 1995 [27], which has some advantages such as versatility, robustness, effectiveness, simple calculation, etc. The main idea of the SVM is to establish a classification hyperplane as the decision surface so that the margin edge between the positive and negative examples is maximized. The theoretical basis of the SVM is statistical learning theory, that is, risk minimization approximate implementation. SVM is originally used to solve classification problems. After continuous research and innovation, the theory of support vector regression (SVR) was gradually developed based on SVM theory [28]. The essence and structure of SVM and SVR are very similar, but the difference is that the value of SVM output in the classification problem is the discrete value between $[-1, 1]$, and the value of SVR output in the regression problem is any real number [29]. The type and parameters of the kernel function in SVR indirectly determine the distribution of samples in high-dimensional linear space, which affects the performance of SVR. In SVM theory, different kernel functions will lead to different SVR algorithms. Currently, there are four commonly used kernel functions:

- Linear kernel function: $K(x, x_i) = x^T x_i$
- Polynomial kernel function: $K(x, x_i) = (\gamma x^T x_i + r)^p$
- Radial basis function: $K(x, x_i) = \exp(-\gamma ||x - x_i||^2) \gamma > 0$
- Sigmoid function: $K(x, x_i) = \tan(\alpha (x x_i) + b)$

2.5.3. Back Propagation Neural Network

The BPNN is a multi-layer feedforward neural network [30], which is characterized by forward signal transmission and backward error propagation. In forward transfer, the input data is processed layer by layer from the input layer through the hidden layer to the output layer, and the state of each layer of neurons only affects the state of the next layer of neurons. If the output layer does not get the expected output value, it goes to the backpropagation and adjusts the network weight and threshold according to the prediction error, so that the BPNN prediction output is continuously approaching
the expected output until the accuracy meets the requirement. Figure 2 shows the topology structure of the BPNN, $X_1, X_2, \ldots, X_n$ are the input values of the BP neural network, that is, the independent variables of the function. $Y_1, Y_2, \ldots, Y_n$ are the predicted values, that is, the dependent variables of the function, and $\omega_{ij}$ and $\omega_{jk}$ are the network weights. The training process of the BPNN model is as follows:

Step 1: Network initialization. According to the system input and output sequence $(X, Y)$, determine the number of network input layer nodes, $n$, the number of hidden layer nodes, $l$, and the number of output layer nodes, $m$. Then initialize the connection weights, $\omega_{ij}$, between the input layer and the hidden layer, and $\omega_{jk}$, between the hidden layer and the output layer, and the thresholds $a$ and $b$ of the hidden layer and the output layer. The learning rate, $\eta$, and the neuron excitation function are given. Then determine the learning rate, $\eta$, and the excitation function, $f$.

Step 2: Calculate the hidden layer output. The hidden layer output, $H$, is calculated according to the input variable, $X$, the connection weight, $\omega_{ij}$, between the input layer and the hidden layer, and the hidden layer threshold, $a$.

$$H_j = f\left(\sum_{i=1}^{n} \omega_{ij}x_i - a_i\right) \quad j = 1, 2, \ldots, l$$

(8)

Step 3: Calculate the output layer output. The network prediction output, $O$, is calculated based on the hidden layer output, $H$, the connection weight, $\omega_{jk}$, and the output layer threshold, $b$.

$$O_k = \sum_{j=1}^{n} \omega_{jk}H_j - b_k \quad k = 1, 2, \ldots, m$$

(9)

Step 4: Calculate the error. The network prediction error, $e$, is calculated based on the network prediction output, $O$, and the expected output, $Y$.

$$e_k = Y_k - O_k \quad k = 1, 2, \ldots, m$$

(10)

Step 5: Update the weight. The network connection weights, $\omega_{ij}$ and $\omega_{jk}$, are updated according to the network prediction error, $e$.

$$\omega_{ij} = \omega_{ij} + \eta H_j(1-H_j)x(i) \sum_{i=1}^{n} \omega_{jk}e_k \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, l$$

(11)

$$\omega_{jk} = \omega_{ij} + \eta H_j e_k \quad j = 1, 2, \ldots, l; \quad k = 1, 2, \ldots, m$$

(12)

Step 6: Update the threshold. The network connection thresholds, $a$ and $b$, are updated according to the network prediction error, $e$.

$$a_j = a_j + \eta H_j(1-H_j) \sum_{j=1}^{n} \omega_{jk}e_k \quad j = 1, 2, \ldots, l$$

(13)

$$b_j = b_j + e_k \quad k = 1, 2, \ldots, m$$

(14)

Step 7: Determine whether the algorithm iteration ends. If not, return to step 2.

In this study, the kernel function of the SVR model was a radial basis function. A three-layer BPNN with a single hidden layer was used to predict the soil AS content. The input layer of the network was composed of the spectral reflectance, and the output layer was the soil AS content. The number of hidden neurons is calculated by Equation (15); $n$ is the number of input neurons, and the activation function is a sigmoid function.

$$l = \log_2 n$$

(15)
The performance of models were assessed using the calibration set coefficient of determination \( R^2 \), the validation set coefficient of determination \( R^2_c \), the calibration set root mean square error (RMSEC) \([31]\), the validation set root mean square error (RMSEV), the calibration set mean absolute error (MAEC), the validation set mean absolute error (MAEV), and RPD \([32]\), according to the following equations:

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} \sum_{i=1}^{n} (y - \hat{y})^2}{\sum_{i=1}^{n} \sum_{i=1}^{n} (y - \bar{y})^2} \tag{16}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2} \tag{17}
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y - \hat{y}| \tag{18}
\]

\[
RPD = \frac{\text{standard deviation}}{\text{RMSEV}} \tag{19}
\]

where \( n \) is the number of samples and \( y, \hat{y} \) is measured and predicted values of AS content, respectively. The range of \( R^2 \) is \([0,1]\), and the larger the value of \( R^2 \), the stronger the linear relationship between the predicted value and the measured value, and the more stable the model is. The smaller the RMSEC and RMSEV values, the more accurate the model is, and the closer the values are, the more stable the model is and the stronger the prediction ability is. MAE is the average of the absolute values of the deviations of all individual observations from the arithmetic mean, which can accurately reflect the actual prediction error. When RPD > 3.0, the model has excellent predictive ability, 2.0 < RPD < 3.0 means the model has good predictive ability, RPD ≥ 2.0 means the model is reliable, 1.5 < RPD < 2.0 means the model is more reliable, but the reliability of the model can be improved by other methods, and RPD ≤ 1.5, means the model is not reliable.

Principal component analysis used IBM SPSS Statistics 19.0 software, spectral resampling, and spectral preprocessing, and model building used MATLAB R2015b software.

2.6. Evaluation Modeling Accuracy
3. Results

3.1. Descriptive Statistics

Since the study area includes some different types of land use, the background value of AS concentration is also different. Therefore, the study areas are divided into six areas; the pulp deposition area (A), the hillside (B), the remediation field (C), park area (D), residential area (E) and factory area (F). Table 1 shows the descriptive statistics of AS content and pH values of 90 sampling points. Table 1 indicates that the pH value of the soil in the whole study area is greater than 8.1, which is weakly alkaline. The average values of AS in the six regions are: 150.7, 35.21, 25.88, 12.58, 13.55, and 12.53 mg/kg, respectively. AS content in the A, B, and C areas exceeded the risk screen values for soil contamination of development land of the Soil Environment Quality Risk Control Standard for soil of development (agricultural) land of China (2018) [33,34] by 2.51, 1.41, and 1.03 times, respectively. The coefficient of variation can describe the degree of dispersion of each data value in the dataset. The larger the value, the more uneven the distribution of elements in space. The coefficient of variation of soil AS content in the six areas is ranked from large to small: B > A > C > F > E > D. Area B has the largest coefficient of variation and belongs to the strong variation type. It means that the AS content in area B is unevenly distributed, and the source might be human activities.

According to the systematic sampling method, the AS content of 90 samples were sorted in ascending order and divided into 30 groups. One sample in each group was randomly selected to compose the validation set. Thus, the validation set had a total of 30 values, and the remaining 60 values belonged to the calibration set. Table 2 shows the descriptive statistics of AS content in 90 sampling points in the calibration set and the validation set. It can be seen from Table 2 that the maximum, minimum, mean value and coefficient of variation for two sets of data are very close, indicating that the distribution of the two groups of data is nearly the same.

Table 1. Soil arsenic (AS) concentrations and pH of soil samples in different areas.

| Site Type | Item  | Max  | Min  | Mean  | Coefficient of Variation | Background Value | Ratio |
|-----------|-------|------|------|-------|--------------------------|------------------|-------|
| Mining Area | A     | AS (mg/kg) | 231.00 | 54.00 | 150.70 | 0.37 ** | 60 | 2.51 |
|           | pH    | 8.93 | 8.04 | 8.37 | | | | |
| B         | AS (mg/kg) | 100.00 | 13.30 | 35.21 | 0.71 ** | 25 | 1.41 |
|           | pH    | 8.56 | 5.33 | 8.19 | | | | |
| C         | AS (mg/kg) | 41.50 | 16.10 | 25.88 | 0.28 ** | 25 | 1.03 |
|           | pH    | 8.40 | 7.53 | 8.13 | | | | |
| Suburb    | D     | AS (mg/kg) | 14.30 | 11.30 | 12.58 | 0.07 | 60 | 0.21 |
|           | pH    | 9.13 | 8.40 | 8.81 | | | | |
| E         | AS (mg/kg) | 16.10 | 12.20 | 13.55 | 0.08 ** | 20 | 0.68 |
|           | pH    | 8.88 | 8.18 | 8.52 | | | | |
| F         | AS (mg/kg) | 14.90 | 9.40 | 12.53 | 0.12 * | 60 | 0.21 |
|           | pH    | 8.74 | 8.28 | 8.47 | | | | |

** means significant on 0.01 level, * means significant on 0.05 level. Take group D data as a control group.

Table 2. AS concentrations of soil samples in different data sets.

| Data Set   | Max (mg/kg) | Min (mg/kg) | Mean (mg/kg) | Coefficient of Variation |
|------------|-------------|-------------|--------------|--------------------------|
| Calibration Set | 231.00 | 9.40 | 45.61 | 1.28 ** |
| Validation Set   | 214.00 | 11.00 | 45.01 | 1.27 ** |

** means significant on 0.01 level.

3.2. Different Resampling Interval Modeling Results

PLSR models using soil samples with 2 nm, 4 nm, 6 nm, 8 nm, 10 nm, 12 nm, and 14 nm spectral interval resampling were built for AS content estimation. Scatter points of different resampling interval model results are showed in Figure 3. The closer the scatter points are to the 1:1 line, the higher the accuracy and stability of the models are. Table 3 compares the modeling accuracy of different
interval resampling spectrum using PLSR models. Different resampling intervals cause different accuracy; 10 nm interval resampling has the highest calibration and validation accuracy, indicating that spectral resampling at intervals of 10 nm can reduce spectral noise, remove redundant information, and maintain spectral data characteristics better. The bands referred to in the following sections, such as 690 nm, represent the arithmetic mean of 690–699 nm, 700 nm represents the arithmetic mean of 700–709 nm, and so on.

Figure 3. Scatter plots of the reference vs predicted arsenic contents for different resample intervals using partial least squares regression: (a) original, (b) 2 nm, (c) 4 nm, (d) 6 nm, (e) 8 nm, (f) 10 nm, (g) 12 nm, and (h) 14 nm. RMSE: root mean square error.
Table 3. The estimation results of different resample interval using partial least squares regression.

| k   | RMSEC | RMSEV | $R^2_C$ | $R^2_V$ | MAEC  | MAEV  | RPD    |
|-----|-------|-------|---------|---------|-------|-------|--------|
| Original | 5 | 26.40 | 30.72 | 0.7923 | 0.7282 | 16.9763 | 22.6354 | 1.6546 |
| 2 nm  | 4 | 26.52 | 30.49 | 0.7915 | 0.7238 | 16.8406 | 22.2215 | 1.6655 |
| 4 nm  | 3 | 26.45 | 30.21 | 0.7915 | 0.7288 | 16.5236 | 21.6547 | 1.6989 |
| 6 nm  | 3 | 26.51 | 30.00 | 0.7906 | 0.7383 | 15.4459 | 21.3645 | 1.6922 |
| 8 nm  | 3 | 26.46 | 30.21 | 0.7913 | 0.7455 | 15.4321 | 22.2156 | 1.6952 |
| 10 nm | 3 | 25.60 | 29.99 | 0.8046 | 0.7442 | 15.3981 | 20.5442 | 1.7253 |
| 12 nm | 3 | 26.37 | 30.50 | 0.7848 | 0.7334 | 16.2564 | 21.1965 | 1.6543 |
| 14 nm | 3 | 26.89 | 30.98 | 0.7846 | 0.7185 | 17.3365 | 22.0398 | 1.6312 |

k: the number of partial least squares regression (PLSR) factors used in each calibration. RMSEC: the root mean square error of the calibration set. RMSEV: the root mean square error of the validation set. $R^2_C$: the coefficient of determination of the calibration set. $R^2_V$: the coefficient of determination of the validation set. MAEC: the mean absolute error of the calibration set. MAEV: the mean absolute error of the validation set. RPD: residual predictive deviation.

3.3. Different Spectral Preprocessing Method Modeling Results

The curves of different spectral preprocessing methods are displayed in Figure 4. PLSR modeling was performed using four spectral data with different spectral pretreatments. The results of the different spectral preprocessing models are shown in Figure 5, which indicates that the position relationships between the scatter points and the 1:1 line are similar. However, the scatter point distribution of the SD model is closer to the 1:1 line. Table 4 shows the accuracy evaluation results of PLSR modeling after different preprocessing methods. It can be seen that the modeling accuracy of the four groups after different preprocessing methods is higher than that without the spectral preprocessing. After SD transformation, the modeling accuracy was the best, followed by SG, and MSC was the worst.

Figure 4. The spectral reflectance curves of the soil samples; (a) Savitzky–Golay (S-G), (b) First Derivation (FD), (c) Second Derivation (SD), and (d) Multiplicative Scatter Correction (MSC). (Each curve indicates the spectral curve of one soil sample).
3.4. Influence of Feature Band Extraction on Segment Modeling Accuracy

Principal component analysis was performed on the 10 nm resampled and SD-transformed spectral data. Table 5 shows that the cumulative explained variance of the first six principal components has reached 95%. Therefore, the 55 original bands with the largest absolute value of correlation loadings from the first six principal components can be selected as the characteristic bands (Table 6). The reflectance of 55 feature bands is the independent variable, and the AS content is the dependent variable. Figure 6 shows the modeling result of the PLSR method. The comparison between the characteristic band and the full band modeling accuracy is displayed in Table 7. It can be seen that, compared with the PLSR modeling results of the feature band extraction by principal component analysis, the RMSE of the full-band modeling is lower, and $R^2$ and RPD are higher, indicating that feature bands cannot improve the model verification accuracy in this study.
1. The initial learning rate, \( \eta \), was 0.0715 and the di…

Table 5. The accumulated explained variance of the first 10 principal components (PC) of soil reflectance spectral.

| Principal Component | PC1  | PC2  | PC3  | PC4  | PC5  | PC6  | PC7  | PC8  | PC9  | PC10 |
|---------------------|------|------|------|------|------|------|------|------|------|------|
| Cumulative (%)      | 53.6 | 72.8 | 85.8 | 92.5 | 94.1 | 95.6 | 96.6 | 97.3 | 97.8 | 98.2 |

Table 6. Feature band extraction results.

| Soil Properties | Feature Bands/\( \text{nm} \) |
|-----------------|-----------------------------|
| AS              | 430, 450, 470, 480, 530, 610, 620, 640–660, 1030–1060, 1080–1110, 1230, 1240, 1280, 1300–1320, 1360–1380, 1400, 1400, 1480, 1510, 1580, 1750, 1850–1930, 2050, 2120, 2140, 2160, 2170, 2190–2220, 2250, 2340–2360 |

Figure 6. Scatter plots of the reference vs. predicted soil arsenic contents after selection feature bands using partial least squares regression.

Table 7. The estimation results of the feature bands and all-bands using partial least squares regression.

| Feature Bands | \( R^2_C \) | \( R^2_V \) | MAEC | MAEV | RPD |
|---------------|-------------|-------------|------|------|-----|
| Feature Bands | 0.7850      | 0.7384      | 22.1084 | 14.6293 | 1.7389 |
| All-bands     | 0.8152      | 0.7701      | 24.5437 | 16.8899 | 1.8874 |

3.5. Different Inversion Model Modeling Results

The spectral reflectance with 10 nm resampling and SD transformation is independent, and the AS content is used as the dependent variable. The model is established by PLSR, SVR, and BPNN, respectively. In the back propagation neural network model, the number of input layers, hidden layers, and output layers of the BP neural network model was 1. The number of neurons in the input layer was 200, the number of hidden layer neurons was 7, and the number of output layers was 1. The initial learning rate, \( \eta \), was 0.11, the training frequency was 20, and the expected error was 0.0001. The error performance curve of BP neural network training is shown in Figure 7. After only six iterations, the network reached the expected error level, showing the fast convergence speed of the model. Figure 8 shows the modeling results. Most of the scatter points in BPNN are located close to the 1:1 line, and the trend is more consist with the 1:1 line. Table 8 compares the modeling accuracy of different modeling methods under the same interval spectral resampling and pretreatment conditions. In the BPNN modeling results, the difference of \( R^2 \) between the calibration set and the validation set was 0.0715 and the difference in accuracy was only 7.67%, which indicates that the established BPNN model has good applicability and there is not an over-fitting problem in the model. After analyzing and comparing the results of the three models, it can be seen that the BPNN model was the best model.
for predicting the AS content in soil. This means that the BPNN model can solve the problem of the number of training samples being insufficient to support modeling due to excessive differences between sample values. The superiority of the BPNN model was attributed to the optimization of the BPNN initial input parameters (thresholds and weights); this approach does not have the problem of low accuracy common in other models.

![Figure 7. Error of Back Propagation neural network.](image)

![Figure 8. Scatter plots of the reference vs. predicted soil arsenic contents using support vector regression (a) and back propagation neutral network (b) models.](image)

| Table 8. The estimation results of using partial least squares regression, support vector regression, and a back propagation neutral network. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Method          | RMSEC | RMSEV | $R^2_C$ | $R^2_V$ | MAEC | MAEV | RPD  |
|-----------------|-------|-------|--------|--------|------|------|------|
| PLSR            | 24.90 | 27.77 | 0.8152 | 0.7701 | 14.6293 | 16.8899 | 1.8874 |
| SVR             | 32.80 | 34.17 | 0.7570 | 0.7201 | 16.6159 | 20.3121 | 1.0313 |
| BPNN            | 16.59 | 20.29 | 0.9322 | 0.8607 | 11.2749 | 14.4226 | 2.5362 |

4. Discussion

4.1. Best Inversion Model Analysis

In this paper, PLSR, SVR, and BPNN methods are used to establish a hyperspectral prediction model for estimating soil AS content in mining and suburban soils form VNIR spectra. Some studies
regarding the models for estimating AS concentrations from VNIR spectrum are displayed in Table 9. By comparing the $R^2$ and RPD values, our estimation using the BPNN model for AS is better than most of them, and the validation set, $R^2$, and RPD are 0.861 and 2.536, respectively.

Table 9. Comparison of the estimation models of AS concentration in soils using visible and reflectance spectroscopy.

| Sampling Site          | Num. | Methods          | $R^2$ | RPD  | References |
|------------------------|------|------------------|-------|------|------------|
| Agricultural Soils     | 33   | MSC + PCR        | 0.3674|      | [14]       |
| Suburban Area          | 61   | FD + PLSR        | 0.7200| 1.900| [15]       |
| Suburban Area          | 93   | SG + PLSR        | 0.7500| 1.810| [13]       |
| Urban Area             | 161  | SG + PLSR        | 0.6700| 1.720| [16]       |
| Urban Area             | 97   | 4nm + MSC + PLSR | 0.7110| 1.827| [12]       |
| Urban Area             | 154  | BD + MLR         | 0.3900| 1.230| [17]       |
| Urban Area             | 96   | GA + PLSR        | 0.3500| 1.090| [18]       |
| Urban Area             | 90   | PSO + BPNN       | 0.8110|      | [19]       |
| Mining Area            | 45   | SD + PLSR        | 0.8400|      | [20]       |
| Mining and Suburban Areas | 90 | 10nm + SD + BPNN | 0.8607| 2.536| This work |

BD: band depth. GA: genetic algorithm. PSO: particle swarm optimization. RMSE: root mean square error. $R^2$: coefficient of determination.

From Figures 3, 5, 6 and 8a, it can be seen that the scatter points of the high soil AS content in the PLSR and SVR models are farer from the 1:1 line than in the BPNN model. This is because the number of the high soil AS content samples used for modeling are low and the extreme values between the sample values are too large. In the BPNN modeling results (Figure 8b), the scatter points of high soil AS content are relatively close to the 1:1 line, but there are still some points that are a long way from the line of equivalence. This maybe because the amount of data is too small and the model needs further optimization. Moreover, there are many factors affecting soil spectral, which may affect the modeling result. In general, BPNN has some strong abilities of adaptive learning, knowledge reasoning, and computational optimization. Therefore, the BPNN model can solve the nonlinear problem effectively between soil AS content and reflectance spectra, and the low modeling accuracy caused by insufficient data in other models can be avoided in order to improve model accuracy and stability simultaneously.

4.2. Inversion Mechanism Analysis

Many previous studies have shown that there are several weak absorption peaks at 430, 530, and 650 nm in the original spectrum, which are mainly related to iron-manganese oxide [35]. The absorption peaks at 1400, 1900, and 2210 nm are related to soil organic matter, clay minerals, and hydroxyl groups in the water [36]. The weak absorption peak at 2250 nm is mainly related to soil organic matter [37]. The 10 nm interval spectral resampling data subjected to different pretreatment methods were divided into five groups according to the AS content (Figure 9). The results show that SD transformation can effectively amplify the differences between spectral reflectance of feature bands with different AS contents, compared with S–G smoothing, FD, and MSC. This is the reason why SD transformation can improve the accuracy and stability of the model. There are some different degrees of reflection (absorption) peaks at 430, 530, 1380, 1400, 1430, 1870, 1900, 2140, 2200, and 2340 nm, which can be considered as characteristic bands of soil AS content, consistent with the characteristic bands of Fe (450, 550, 1000, 1400, 1900, 2050, 2200, 2250, 2400, and 2470 nm) [38]. Moreover, studies have shown that soil AS content is highly correlated with soil Fe content [35]. These indicate that the inversion mechanism of soil AS content is indirectly constructed by the correlation between AS and iron-manganese oxide, organic matter, and clay minerals.
We are also actively exploring this work. In order to increase the range of AS content in soil samples, when selecting the study area, we selected a mining area with very high AS content and an urban suburb with low AS content. The 90 samples were divided into five groups: 0–20 mg/kg, 20–40 mg/kg, 40–80 mg/kg, 80–160 mg/kg, and 160–320 mg/kg, and the accuracy and stability of the BPNN model were, respectively, 45.34\%, 43.68\%, 42.41\%, 41.55\%, and 40.84\%. The differences in spectral characteristics of different AS content are not significant, and the detection method is difficult. In published research that estimate soil AS content based on hyperspectral, there is little research in estimating the AS content in different existing forms. We are also actively exploring this work.

4.3. Limitations and Future Work

The presence of AS in soils is an issue of concern. Relative to the total AS content, the existing forms of AS is of great significance for understanding the source, migration, transformation characteristics, and bioavailability of AS in the soil. At present, most research on soil AS existence focusses on the areas of bio-remediation and phytoremediation [40]. According to the National Standard Soil environmental quality Risk control standard for soil contamination of agricultural land [33] and Soil environmental quality Risk control standard for soil contamination of development land [34] of the People’s Republic of China, the determination method of AS in soil is total AS element content. In order to compare with national standards, the total AS content in this study was used as the research object. Obviously, the estimation of AS content in different existing forms is a question worthy of discussion. However, the differences in the spectral characteristics of different AS content are not significant, and the detection method is difficult. In published research that estimate soil AS content based on hyperspectral, there is little research in estimating the AS content in different existing forms based on hyperspectral, but it is of great importance to distinguish the spectral differences in the AS content of different existing forms. We are also actively exploring this work.

The extreme value of soil AS suitable for hyperspectral estimation is still unknown, which is also a question worth exploring. In order to increase the range of AS content in soil samples, when selecting...
the study area, we selected a mining area with very high AS content and an urban suburb with low AS content. The 90 samples were divided into five groups: 0–20 mg/kg, 20–40 mg/kg, 40–80 mg/kg, 0–160 mg/kg, and 160–320 mg/kg, and the accuracy and stability of the BPNN model were, respectively, calculated. The results are shown in Table 10. From the current sample data and test results, there is not much relationship between different AS content ranges and estimation accuracy, but it seems that the accuracy is higher when the amount of data is larger. Therefore, the study may be limited to individual sample points. In order to improve the estimation accuracy of soil AS content, it is necessary to collect more soil samples with high AS content in the future to optimize the soil AS content estimation models.

Table 10. The estimation results of different AS content ranges using back propagation neural networks.

| AS Content   | Number of Sample | RMSEC | RMSEV | $R^2_C$ | $R^2_V$ | MAEC  | MAEV  | RPD   |
|--------------|------------------|-------|-------|---------|---------|-------|-------|-------|
| 0–20 mg/kg   | 52               | 13.31 | 16.52 | 0.9254  | 0.8736  | 11.5896 | 13.6354 | 2.825 |
| 20–40 mg/kg  | 18               | 15.49 | 18.86 | 0.9109  | 0.8689  | 11.8614 | 14.2145 | 2.757 |
| 40–80 mg/kg  | 6                | 19.18 | 21.90 | 0.9852  | 0.8496  | 12.2016 | 14.249  | 2.365 |
| 80–160 mg/kg | 8                | 19.34 | 23.63 | 0.8761  | 0.8599  | 12.0254 | 13.9965 | 2.412 |
| 160–320 mg/kg| 8                | 17.09 | 21.36 | 0.8815  | 0.8606  | 12.6321 | 14.6254 | 2.613 |

Hyperspectral estimation of soil AS content is fast and cost saving and provides some convenience for monitoring soil AS pollution. However, soil spectral reflectance is affected by multiple components [41,42]; in order to obtain accurate spectral information, soil samples need to be carefully measured. Therefore, in areas with large differences in soil composition, the application of this model needs to be discussed. As the number of soil samples increases, we will continue to optimize the models.

5. Conclusions

In this paper, we explored the feasibility and best estimation method for estimating soil AS content by BPNN and VNIR hyperspectral. The main conclusions are as follows: (1) The results of the PLSR models modeling using the original spectrum, 2, 4, 6, 8, 10, 12, and 14 nm interval resampling spectrum indicated that the estimation accuracy of 10 nm interval resampling is the best ($R^2_C = 0.744$ and RPD = 1.725). (2) S–G smoothing was performed on the spectral data after 10 nm interval resampling, then FD, SD, and MSC transformation were performed, and the estimation model was established by the PLSR method. The results indicated that the SD transformation had the best modeling accuracy ($R^2_C = 0.770$ and RPD = 1.887). (3) Comparing the PLSR modeling accuracy with the full-bands and 55 feature bands extracted by principal component analysis, the results suggested that the feature band extraction cannot improve the model validation accuracy. (4) The independent variable was the original spectral reflectance after 10 nm re-sampling and SD transform, and the dependent variable was AS content. The estimation models were established using PLSR, SVR, and BPNN algorithms, respectively. The results showed that BPNN had the best modeling accuracy ($R^2_C = 0.861$ and RPD = 2.536). In summary, using BPNN and hyperspectral data to estimate soil AS content is feasible, and the best estimation method is: 10 nm + SD + BPNN.

Author Contributions: research conceptualization, L.H.; methodology, L.H., H.Z., and Y.Z.; investigation, L.H., R.C., Z.L., and H.H.; resources, H.Z.; writing, R.C. and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Opening fund of the Key Laboratory of Degraded and Unused Land Consolidation Engineering, the Ministry of Land and Resources (Program No. SXDJ2017-9), the Opening fund of Shaanxi Key Laboratory of Land consolidation (Program No. 2018-ZZ03, 2018-JC03), the National Natural Science Foundation of China (Program No. 41871190), and the National Science Basic Research Plan in Shaanxi Province of China (Program No. 2018QJ4027). The sponsors had no role in the design, execution, interpretation, or writing of the study.

Conflicts of Interest: The authors declare no conflict of interest.
24. Curran, P.J.; Dungan, J.L.; Peterson, D.L. Estimating the foliar biochemical concentration of leaves with reflectance spectrometry. Remote Sens. Environ. 2001, 76, 349–359. [CrossRef]
25. Fukunaga, K.; Koontz, W.L.G. Representation of random processes using the finite Karhunen-Loève expansion. Inf. Control 1970, 16, 85–101. [CrossRef]
26. Wang, H.W. Partial Least Squares Regression Method and its Application; National Defense Industry Press: Beijing, China, 1999; pp. 1–3.
27. Hsu, C.; Lin, C. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 2008, 13, 415–425.
28. Liu, X.B. Analysis and Research of Forecast Model on PM2.5 Using Support Vector Regression. Master’s Thesis, Southwestern University Of Finance And Economics, Chengdu, China, 2016.
29. Lian, C.Y. Support Vector Regression Based on Genetic Algorithm and its Application in Phase Separation Procedure with Salts. Master’s Thesis, Hebei University Of Technology, Wuhan, China, 2015.
30. Zhou, Z.H. Neural Network and its Application; Tsinghua University Press: Beijing, China, 2004.
31. Esbensen, K.H.; Guyot, D.; Westad, F.; Lars, P.H. Multivariate Data Analysis: An Introduction to Multivariate Data Analysis and Experimental Design, 5th ed.; CAMO Software: Oslo, Norway, 2002; p. 598.
32. Si, H.Q.; Yao, Y.M.; Qang, D.Y.; Liu, Y. Hyperspectral prediction of soil organic matter contents under different soil moisture contents. Trans. Chin. Soc. Agric. Eng. 2015, 31, 114–120.
33. GB 15618-2018. Soil Environmental Quality—Risk Control Standard for Soil Contamination of Agricultural Land; Ministry of Ecology and Environment of the People’s Republic of China: Beijing, China, 2018. Available online: http://www.mee.gov.cn/ywgz/fgbz/bzwb/trhj/201807/t20180703_446029.shtml. (accessed on 2 November 2019).
34. GB 36600-2018. Soil Environmental Quality—Risk Control Standard for Soil Contamination of a Development Land; Ministry of Ecology and Environment of the People’s Republic of China: Beijing, China, 2018. Available online: http://www.mee.gov.cn/ywgz/fgbz/bzwb/trhj/ (accessed on 3 November 2019).
35. Wu, Y.Z.; Chen, J.; Wu, X.M.; Tian, Q.J.; Ji, J.F.; Qin, Z.H. Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. Appl. Geochem. 2005, 20, 0–1059. [CrossRef]
36. Nayak, P.S.; Singh, B.K. Instrumental characterization of clay by XRF, XRD and FTIR. Bull. Mat. Sci. 2007, 30, 235–238. [CrossRef]
37. Ben-Dor, E.; Inbar, Y.; Chen, Y. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. Remote Sens. Environ. 1997, 61, 1–15. [CrossRef]
38. Thomas, K.; Stefan, S. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environ. Sci. Technol. 2002, 36, 2742–2747.
39. Song, L.; Jian, J.; Tan, D.J.; Xie, H.B.; Luo, Z.F.; Gao, B. Estimation of soil’s heavy metal concentration (As, Cd and Zn) in Wansheng mining area with geochemistry and field spectroscopy. Spectrosc. Spectr. Anal. 2014, 34, 812–817.
40. Zhao, F.J.; McGrath, S.P.; Meharg, A.A. Arsenic as a food chain contaminant: Mechanisms of plant uptake and metabolism and mitigation strategies. Annu. Rev. Plant Biol. 2010, 61, 535–559. [CrossRef] [PubMed]
41. Gao, P.; Fu, T.G.; Wang, K.L.; Chen, H.Y.; Zeng, F.P. Spatial heterogeneity of surface soil mineral components in a small catchment in Karst peakcluster depression area, South China. Chin. J. Appl. Ecol. 2013, 24, 3179–3184.
42. Liu, H.J.; Wang, X.; Li, H.X.; Meng, X.T.; Jiang, B.W.; Zhang, X.L.; Yu, Z.Y. Effect mechanism of soil minerals on spectral characteristics of main soil classes in songnen plain. Spectrosc. Spectr. Anal. 2018, 38, 3238–3244.

© 2020 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 (http://creativecommons.org/licenses/by/4.0/).