Research on Hyperspectral Inversion of Soil Alkaline Hydrolysis Nitrogen Content and pH Value Based on DWD

Sijia Liu 1, 2a*, Zhiyun Xiao 1, 2b*

1 College of Electric Power, Inner Mongolia University of Technology, Hohhot 010080, 2 The Key Laboratory of Electromechanical Control, Inner Mongolia, Hohhot 010051, China

a*liuran5422@163.com, b* xiaozhiyun@imut.edu.cn

Abstract—Aiming at the shortcomings of traditional methods for detecting the content of Alkaline Hydrolysis Nitrogen (AHN) and pH value in soil, such as time-consuming and labor-consuming, this paper proposes a rapid quantitative inversion method based on hyperspectral analysis of AHN content and pH value. This method uses db4 discrete wavelet denoising (DWD) and wavelet denoising normalization (DWD-N) to carry out Pearson correlation analysis, and two methods, Ridge regression and Partial Least Squares Regression (PLSR), were used to compare the accuracy of hyperspectral inversion of soil AHN content and pH value. Experiments have demonstrated that in the inversion of the AHN content prediction model, Ridge regression has a good modeling effect under the DWD-N model, where $R^2=0.647$, RMSE=7.067mg/kg. PLSR has good prediction effect under DWD-N, where $R^2$ is the highest of 0.792, RMSE is 3.438mg/kg; in the model inversion of pH prediction, the full-band PLSR modeling effect of pH value under DWD pretreatment is the best, which modeling set and the prediction set of $R^2$ is 0.826 and 0.875, the RMSE is 0.217 mg/kg and 0.191 mg/kg respectively.

1. Introduction

Soil AHN content is also called available nitrogen. It includes inorganic nitrogen and organic nitrogen directly absorbed and utilized by crops, which can be absorbed by crops in the near future. Its content depends on the content of organic matter and the amount of nitrogen fertilizer. The material is extremely unstable and easily changed by the influence of hydrothermal environment and biological activities, but its content can reflect the recent soil nitrogen supply capacity. Due to the difference of topography and region, the measurement of soil pH has an important impact on the planting of crops, and after the pH is greater than 7.5, it will appear alkaline, which is unfavorable for the growth and development of crops. The use of hyperspectral technology can quickly and accurately obtain the AHN content and pH value of the soil, thereby greatly saving manpower and material resources, and providing a new approach and method for the study of soil properties.

In the study of soil properties inversion based on hyperspectrum, how to perform spectral preprocessing, selection of spectral characteristic bands, and comparison of inversion methods is an important direction to improve the accuracy of soil attribute prediction. Di Wupengyao et al. [2] summarized the methods of spectral preprocessing, including optimizing the number of factors of partial least squares (PLS) and the first derivative, second derivative, window parameters of SG smoothing, SG smoothing filter, the wavelet function and the decomposition scale of continuous wavelet transform, multivariate scattering correction, standard normal variable, centering, maximum
and minimum normalization, standardization... 120 kinds of preprocessing and their combinations are compared for PLS modeling, and finally point out the comparison of the advantages and disadvantages of the preprocessing methods under different data sets; Chen Zhikun [3] et al. compared the denoising effects of different wavelet bases and decomposition layers, and determined the best wavelet base db4 for selection. The 4-layer optimal wavelet decomposition has achieved good results. Luan Fuming [4] et al. used the first derivative spectrum of the logarithm of soil reflectance to separately decompose the four wavelet functions in multiple layers, and used the PLSR to establish the inversion of the soil AHN content model. Shen Congwang [5] et al. used three methods of principal component analysis regression, support vector regression and PLSR to establish the pH value and total potassium content models of purple soil and rice soil respectively, and performed hyperspectral analysis as well as comparative study of inversion accuracy. Lifei Wei et al. [6] chose the characteristic band based on different spectral index transformations and the Pearson correlation analysis method, in which the band with a correlation coefficient greater than 0.7 was taken as the characteristic band. Then, using ridge regression, nuclear ridge regression, Bayesian ridge regression and AdaBoost algorithm to establish the inversion model, it is found that the model effect established by the AdaBoost algorithm is compared with other methods, and the result is the best. Bao Qingling [7] used the random forest algorithm to model, then compare soil organic matter, and found that the combination of machine learning and wavelet decomposition can effectively improve the screening and prediction accuracy of the characteristic bands of organic matter while reducing noise.

The special topography of the Mu Us sandy land has resulted in the local desertified soil. Compared with the previous farmland, the soil water and fertilizer use efficiency is lower. Optimizing the use of fertilizers in the soil can improve the living environment of farmers and their living standards. Since there are few studies on the inversion of soil pH and AHN content in the Mu Us sandy area based on hyperspectral technology, this article intends to use Rigde and PLSR two methods to invert the local desertified soil’s AHN and pH value. Provide a basis for its rapid measurement.

2. Materials and method

2.1 Material
The measurement of soil properties and other nutrient elements will be carried out in the Mengda Industrial Park (38°5’N, 109°0’E) in Wushen Banner from 2019 to 2020. The first stubble of the test field was helianthus castor, hemp and other crops. The depth of soil collected was 0-20cm and 20-40cm respectively. The main types of soil were sandy soil.

2.2 Soil sample processing
Generally, in order to avoid the influence of moisture and soil roughness, the soil is collected first. The soil of the demonstration base depends on the different crops. Collect 0-20cm and 20-40cm soil under each crop, a total of 136 sets of samples, and then put the soil in a ziplock bag and bring it back indoors, let it dry naturally, pass through a 0.850 mm sieve to screen out other impurities in the soil, and then make duplicates, one is used for the analysis and determination of chemical physical and chemical properties, the other is used for hyperspectral analysis.

The AHN measurement method is mainly alkaline hydrolysis diffusion method [8]. The steps for determining the pH value are as follows: weigh the soil and a beaker, add deionized water, stir with an 85-2a thermostatic magnetic for 1 min to fully diffuse the soil particles, let stand for 1 hour, use a pH meter to measure the supernatant and make a record after the value is stabilized.

The study used the Finnish company Specim IQ handheld hyperspectrometer to measure. Specim IQ is a complete, portable, handheld hyperspectral camera. Whiteboard calibration should be done before and after each measurement. When measuring, the spectral range is 400-1000nm, the wavelength accuracy is 3nm, the interval is 7nm, and a total of 204 wavelength points are recorded.

The soil spectral picture is mainly taken indoors. Compared with the outdoor environment, the indoor environment and the measured spectrum is relatively stable, and the signal-to-noise ratio is
high [9]. Select the current environment to perform spectrum shooting in a darkroom environment. The laboratory temperature is 19°C and the humidity is 30hPa. In order to avoid external stray light interference, the laboratory uses two 125W halogen sources for non-contact measurement. The operation is as follows: The soil is placed in a petri dish with a diameter of 1cm and a diameter of 6cm. The soil sample is about 30cm away from the camera lens, and the two halogen lamps are about 15cm away from the camera. The angle of the light to the object is about 90 degrees, as shown in Figure 1.

\[ R = \frac{I - B}{W - B} \]  \hspace{1cm} (1)

Where \( W \) is the all-white calibration image obtained by the standard whiteboard, \( B \) is the all-black calibration image of the camera, and \( I \) is the original hyperspectral image. The mode of shooting at the same time as the whiteboard is used for data collection to eliminate the problem of environmental mismatch.

2.3 Description of specimens

Although the soil photos taken by the camera ensure the stray light interference from the external environment of the experiment, there will still be an army of noise at the beginning and the end of the spectrum, which has a great influence on the accuracy of the experimental results. In order to eliminate the influence of high-frequency noise on the experimental results, we use wavelet transform, using its unique advantages of local analysis in both the time domain and frequency domain [3] to denoise the spectrum.

The quality of DWD results is mainly composed of the following parts. One is the choice of wavelet base, the other is the selection of decomposition levels, and the third is the threshold function and the estimation method, and the threshold function is the most important link. There are usually 4 threshold estimation methods.

1. Fixed threshold, where \( N \) is the number of bands.
2. Adaptive threshold selection based on Stein’s unbiased likelihood estimation principle (SURE): for a given threshold \( m \), get its likelihood estimation, and minimize the non-likelihood \( m \), then the selected threshold is obtained.
3. Heuristic threshold: It is a combination of the first two thresholds.
4. Maximum-minimum threshold.

The second and fourth threshold selection methods have more coefficients retained, while the other two coefficients are completely threshold denoising. All wavelet coefficients become zero. In summary, choose \( \text{dbN} \) wavelet basis, four layers as an example of wavelet decomposition, the heuristic threshold selection method is used to perform wavelet denoising from \( \text{db2} \) to \( \text{db6} \). See Table 1. The signal-to-noise ratio (SNR) and the minimum mean square error (MSE) are used to judge the denoised results, where the formulas of SNR and MSE are as follows, \( s(i) \) is the original spectrum, and \( s' (i) \) is the denoised spectrum signal. With the smaller the MSE, the less removed signal, and the larger SNR, the signal contains a little noise.
Finally, the wavelet base db4 is selected for 4-layer decomposition, as shown in Figure 2(b), after wavelet denoising, the spectral curve is smoother than the original spectral curve, and the high-dimensional, the non-zero variable background baseline and the interference of noise obtains the characteristic data corresponding to the analysis in the pure spectral data. Perform the normalization analysis on the spectrum of DWD data to make the spectrum data fall into a specific interval, and remove the interference of size difference and the information structure, then compare the inversion methods of soil AHN and pH content. The original signals and the corresponding signals after the noise reduction are shown in Figure 2. The wavelet denoising process is shown in Figure 3.

Table 1 db2-db6 wavelet denoising results

| Wavelet base | SNR   | MSE   |
|--------------|-------|-------|
| db2          | 131.18| 8.147*10^-7 |
| db3          | 129.68| 1.07*10^-6  |
| **db4**      | **133.31**| **2.6*10^-6** |
| db5          | 129.97| 1.42*10^-6  |
| db6          | 129.53| 1.33*10^-6  |

2.4 Regression prediction method and accuracy evaluation index

PLSR is a commonly used method of multi-linear regression model. By selecting the best principal component n, it will increase continuously. When the value $R^2$ is the largest, select the first n principal components for modeling analysis, and find the MSE and the coefficient of determination $R^2$. Ridge regression (Ridge), also known as Tikhonov regularization. It is different from the main transformation of linear regression is to add a regular term $\lambda$ to the loss function of multiple linear regression, the derivation process is as follows: the linear model $Y=XA$, $Y$ is the sample value, $XA$ is the calculated value of the model, that is the expected value. Because the calculated value and the sample value may not be 100% consistent, there is an error $B=Y-XA$, in order to find the most suitable model, it is necessary to solve the smallest E, which is the square of the error, the expression is as follows:

$$SNR = 10 \log\left(\frac{\sum_{i=1}^{n} s^2(i)}{\sum_{i=1}^{n} (s(i)-\hat{s}(i))^2}\right)$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (s(i)-\hat{s}(i))^2$$

**Fig. 2 Spectral curve**

(a) Original spectral curve  (b) Wavelet denoising  (c) Wavelet denoising normalization

**Fig. 3 Wavelet denoising process**

Import spectrum data
Perform wavelet decomposition
Set the denoising threshold and wavelet reconstruction
Output spectrum after denoising
After deriving the matrix E, obtain the minimum E as the corresponding A, the corresponding A when the derivative is 0, and the simplification of A is as follows:

\[ A = (X^T X)^{-1} X^T y \]  \hfill (5)

When the dimension of A is more than that of Y, the result will be very unstable. Adding a small disturbance \( \lambda I \) to the original least squares estimate of A, the situation where the generalized inverse cannot be obtained can be obtained. The generalized inverse makes the problem stable and can be solved, the expression is as follows:

\[ A(\lambda) = (X^T X + \lambda I)^{-1} X^T y \]  \hfill (6)

It can be seen that \( (X^T X + \lambda I)^{-1} \) becomes a full-rank matrix, and a stable inverse can be obtained. In the regression prediction model, set \( \lambda \) between 0.001 and 100 for modeling, and select the best \( \lambda \) to make the prediction model function accurate Highest.

Two indicators, the coefficient of determination \( (R^2) \) and the root mean square relative error (RMSE) of prediction, were used to evaluate the two methods of Ridge and PLSR. The closer the \( R^2 \) is to 1, and the smaller the value of RMSE, the better the effect of the inversion model. The expressions of \( R^2 \) and RMSE are as follows:

\[ R^2 = \frac{\sum_i (y_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \]  \hfill (7)

\[ RMSE = \sqrt{\frac{1}{n \sum_i (y_i - \bar{y})^2}} \]  \hfill (8)

In the formula, \( n \) is the number of samples, \( y_i \) is the actual reference chemical value of the i-th sample, \( \bar{y} \) is the predicted value of the i-th sample in the specified sample set, and \( \bar{y} \) is the average value of the actual samples.

3. results and analysis

3.1 Soil spectral curve analysis

The spectrum curve is selected by extracting the average value of the soil in a certain range in ENVI5.3. The black and white correction of the camera is shown in 2(a) after pretreatment of the soil sample. Observation shows that the soil spectrum is in visible light and in the near-infrared light range, there are relatively gentle changes. The effect of DWD is shown in Figure 2(b), and the effect of DWD-N is shown in Figure 2(c).

3.2 Correlation Analysis of Soil AHN Content, pH Value and Spectral Reflectance

The Pearson correlation analysis method was performed on the soil sample reflectivity of the original spectral information (RSI), wavelet denoising (DWD), and wavelet denoising normalized (DWD-N) with the soil AHN content and pH value respectively. The correlation coefficient is analyzed with significance of P<0.01 or P<0.05, and the characteristic band within the correlation coefficient is selected. The results are shown in Figure 4 and Figure 5. The soil AHN content and the spectral curve show the strongest negative correlation. The correlation coefficient ranges from -0.478 to -0.128 under RSI, and -0.482 to -0.228 under DWD and DWD-N, among which 590nm-610nm the band shows a great negative correlation; the soil pH value shows a negative correlation with the spectral curve. Under RSI, the correlation coefficient ranges from -0.412 to 0.137, and the strongest negative correlation is shown at 768nm, 826nm and 941nm. The DWD ranges from -0.429 to -0.169, and the strongest negative correlation is shown at 720nm and 807nm. DWD-N is between -0.194 and 0.142, and the band at 892nm has a peak, showing a great negative correlation. All the characteristic bands
with \( P<0.05 \) are regarded as the characteristic bands of pH value, and the characteristic bands of AHN content are carried out with the characteristic bands of \( P<0.01 \) as the characteristic bands for inversion modeling. For pH, 42 characteristic wavebands were selected under RSI, 52 characteristic wavebands were selected under DWD, and 54 characteristic wavebands were selected under DWD-N. The full-band inversion modeling of AHN was performed under the three conditions.

![Fig.4 Correlation coefficient between spectral reflectance and AHN content](image1)

![Fig.5 Correlation coefficient between spectral reflectance and pH value](image2)

### 3.3 Division of the sample set

The SPXY algorithm is widely used in the division of sample sets, and is better than random sampling (RS), Kennard-Stone (KS) algorithm, two-way algorithm (duplex), etc. [10] Guo et al. [11], Zhu et al. [12], have achieved good results, indicating that this method can effectively improve the accuracy of the model. In this study, AHN or pH value \( y \) was used as the dependent variable, and near-infrared spectroscopy \( x \) was used as the variable. The soil samples were divided into 102 modeling set and 34 prediction set to establish the model. The statistical values of the samples are shown in Table 2.

#### Table 2 Descriptive statistics of alkaline hydrolysis nitrogen and pH content of soil samples

| Sample set              | Number of samples | AHN (mg/kg) | pH |
|-------------------------|-------------------|-------------|----|
|                         |                   | max         | min | mean | std | max | min | mean | std |
| Total sample            |                   | 136         | 44.88 | 1.45 | 20.95 | 9.79 | 8.96 | 7.00 | 7.69 | 0.54 |
| modeling set            | 102               | 44.88       | 1.45 | 21.37 | 10.35 | 8.96 | 7.00 | 7.78 | 0.52 |
| prediction set          | 34                | 34.02       | 7.84 | 19.70 | 7.53  | 8.52 | 7.00 | 7.42 | 0.50 |

### 3.4 Model building and comparison

#### 3.4.1 Establishment and comparison of AHN model

The inversion modeling of AHN is shown in Table 3, the Ridge method predicts a relatively high \( R^2 \) under RSI and DWD, and a relatively low RMSE, with values of 0.7, 0.690 and 4.425, 4.519 respectively. Select the principal component as 8, the \( R^2 \) of the predicted value of DWD-N under the PLSR method is the highest of 0.792, RMSE is 3.438, and DWD-N has the best modeling set data under the Ridge method, \( R^2=0.647 \) and RMSE= 7.067, \( \lambda \) has made a slight change in the optimization of the data modeling set, and improved the collinearity problem of the data set.

#### Table 3 Prediction and fitting effect of soil AHN

| Sample set | Rigde modeling set \( R^2 \) | RMSE | Prediction set \( R^2 \) | RMSE | \( \lambda \) | Principal component | PLSR modeling set \( R^2 \) | RMSE | Prediction set \( R^2 \) | RMSE |
|------------|-----------------------------|------|--------------------------|------|-------------|---------------------|---------------------------|------|--------------------------|------|
|            |                             |      |                          |      |             |                     |                           |      |                          |      |
3.4.2 Establishment of pH model

In the model fitting effect of pH value, the prediction and fitting effect of pH value are shown in Table 4, the RSI\(_{42}\) prediction set R\(^2\) obtains the maximum value (0.883) when the principal component 11 is selected under the PLSR, the RMSE modeling set obtains the minimum value (0.194), and the recommendation is obtained when the principal component 11 is selected under the PLSR. The modulus R\(^2\) is the maximum value (0.826), and the RMSE modeling set is the minimum value (0.217). Compared with the original band data set, both are improved. The feature selection of spectral bands did not improve the overall model prediction effect of Ridge regression. Under the PLSR method, DWD has played a great role in it, and after the prediction set under DWD-N\(_{54}\) is optimized in the characteristic band, the data volume is reduced from the original 204 bands to 52 bands, which reduces the complexity of the calculation. The overall accuracy rate has not dropped significantly.

### Table 4 Prediction and fitting effect of pH value

| Method    | RSI 42 | RSI | DWD | DWD-N | DWD-N\(_{54}\) |
|-----------|--------|-----|-----|-------|----------------|
| R\(^2\) modeling set | 0.701  | 0.493 | 0.692 | 0.792  | 0.676          |
| RMSE modeling set | 0.284  | 0.370 | 0.289 | 0.235  | 0.340          |
| R\(^2\) prediction set | 0.771  | 0.558 | 0.756 | 0.770  | 0.700          |
| RMSE prediction set | 0.257  | 0.376 | 0.266 | 0.242  | 0.302          |

### 4. Conclusion

The establishment of the inversion model of soil properties is inseparable from the rapid and accurate prediction of hyperspectral technology. In order to solve the problem of noise and redundancy in hyperspectral soil information, the DWD method has achieved remarkable results. It not only effectively reduces the noise and highlights the difference in spectral information, but also increases the correlation between the soil and the AHN and pH, which is a soil hyperspectral method. The improvement of inversion accuracy provides conditions. In the inversion of Ridge and PLSR methods, the following conclusions are drawn: (1) In the inversion of soil AHN content, in general, the prediction set effect of PLSR is better than that of Ridge method, but WTD-N later, the modeling set of the Ridge method is better, and the data is normalized to achieve good results in the accuracy of the two methods. (2) In the inversion of soil pH value, the accuracy of the feature extraction effect under the Ridge method is poor. DWD-N adopts the Ridge method in the full band relatively optimal, and the R\(^2\) of the modeling set and prediction set are 0.792 and 0.770, RMSE is 0.235 and 0.242. Under DWD, the model inversion of PLSR is improved under RSI. The R\(^2\) of the modeling set and prediction set are 0.826 and 0.875, and the RMSE is 0.191 and 0.217. Through feature selection, the accuracy difference under DWD-N\(_{54}\) is small, and the amount of original information retained is the least and optimal, which makes the extraction of spectral features more effective. In summary, these two methods are feasible for the prediction of soil AHN content and pH value.

### Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.61661042), Inner Mongolia Natural Science Foundation (2021MS06020), Inner Mongolia Autonomous Region Science and Technology Plan Project.
References

[1] Hou Le, Li Xinju, Yang Yunhui, Zhou Jingjing. Inversion of soil heavy metal content in mining area based on hyperspectral [A]. Professional Committee of Land Reclamation and Ecological Restoration of China Coal Society. Professional Committee of Land Reclamation and Ecological Restoration of China Coal Society Exchange materials of the 8th National Academic Conference on Land Reclamation and Ecological Restoration in Mining Areas[C]. Professional Committee of Land Reclamation and Ecological Restoration of China Coal Society: China Coal Society, 2019: 9

[2] Di Wupengyao, Bian Xihui, Wang Zifang et.al. Study on the selection of spectral pretreatment methods[J]. Spectroscopy and Spectral Analysis, 2019, 39(09): 2800-2806.

[3] Chen Zhikun, Zhang Hanjie, Wang Yutian, et.al. Mineral oil fluorescence spectrum data processing method based on wavelet transform[J]. Laser Journal, 2016, 37(10): 78-81.

[4] Luan Fuming, Xiong Heigang, Wang Fang, et.al. Hyperspectral inversion of soil alkali-hydrolyzable nitrogen content based on wavelet analysis[J]. Spectroscopy and Spectral Analysis, 2013, 33(10): 2828-2832.

[5] Shen Congwang, Xu Lihua. Comparison of hyperspectral inversion methods for soil pH and total potassium content[J]. Jiangsu Journal of Agriculture, 2020, 36(01): 92-98.

[6] Wei, L.; Yuan, Z.; Wang, Z.; Zhao, L.; Zhang, Y.; Lu, X.; Cao, L. Hyperspectral Inversion of Soil Organic Matter Content Based on a Combined Spectral Index Model. Sensors 2020, 20, 2777.

[7] Bao Qingling, Ding Jianli, Wang Jingzhe, et.al. Hyperspectral detection of soil organic matter content based on random forest algorithm[J]. Arid Land Geography, 2019, 42(06): 1404-1414

[8] LY-T 1229-1999, Determination of Hydrolyzable Nitrogen in Forest Soil. [S]

[9] Specim-IQ-user-manual [M]. Finland.2019

[10] Wang Shifang, Han Ping, Cui Guanglu, et.al. SPXY algorithm based near-infrared spectroscopy detection of watermelon soluble solids[J]. Spectroscopy and Spectral Analysis, 2019,39(03):738-742.

[11] GuoWC, ShangL, ZhuX H,etal. Food & Bioprocess Technology, 2015, 8(5):1126

[12] Zhu X H, Fang L J, Gu J S, et.al. Food AnalyticalMethods,2015, 9 (6):1

[13] Yun Y H, Li H D, Deng B C, Cao D S. Trends Anal. Chem. , 2019, 113: 102-115

[14] Jin Huining, Zhang Xinle, Liu Huanjun, et al. Research on prediction model method of soil water content based on spectral absorption characteristics [J] . Acta Pedologica Sinica, 2016, 53(3): 627-635.

[15] Guo Fei, Xu Zhen, Ma Honghong, Liu Xiujin, et.al. Comparative study of hyperspectral inversion models of soil Cd content based on PCA[J]. Spectroscopy and Spectral Analysis, 2021, 41(05):1625-1630.