

HYPERSPECTRAL INVERSION OF SOIL TOTAL NITROGEN CONTENT USING SPECTRAL AND MACHINE LEARNING

ZHENYU DONG¹, NI WANG^{1*}, JIANCANG XIE¹ AND JUNGANG LUO¹

State Key Laboratory of Water Engineering Ecology and Environment in Arid Region, Xi'an University of Technology, Xi'an, 710048, China

Keywords: Soil Total Nitrogen, Hyperspectral, Machine Learning, Feature Extraction, Random Forest, Agro-pastoral Transitional zone

Abstract

The Agro-pastoral Transitional Ecotone of Northern Shaanxi is an ecologically fragile region. Soil nitrogen is a critical factor influencing crop productivity and regional ecological balance. Rapid and accurate monitoring of farmland soil total nitrogen (TN) content is therefore essential for implementing precision agriculture and ensuring sustainable land use in this area. This study focused on the cultivated soils of Dingbian County, a representative area within this ecotone. Field-collected soil samples underwent TN content determination and laboratory-based hyper spectral reflectance measurement. The original spectral reflectance (R) was first transformed using mathematical methods, including first-derivative (FDR), second-derivative (SDR), and logarithm of the reciprocal's first-derivative (LRFD) analysis, to enhance spectral information and suppress background noise. Subsequently, the successive projections algorithm (SPA) and correlation analysis (CA) were integrated to identify feature wavelengths highly correlated with soil TN from the transformed spectral data. Partial Least Squares Regression (PLSR), Decision Tree (DT), and Random Forest (RF) were developed to estimate soil TN content. Model performance was evaluated and compared using the coefficient of determination (R^2), root mean square error (RMSE), and ratio of performance to deviation (RPD). The results indicated that: (1) Spectral derivative transformations effectively highlighted subtle information, with LRFD showing the greatest sensitivity to TN. Models built using LRFD-derived features consistently outperformed those based on original reflectance. (2) Among the machine learning algorithms, the Random Forest (RF) model demonstrated superior predictive performance on the validation set ($R^2 = 0.85$, RMSE = 0.11 g/kg, RPD = 2.15), indicating the highest stability and generalization capability. The PLSR model performed moderately well, while the DT model was prone to over fitting and yielded the lowest accuracy. (3) This study confirms that the integration of hyper spectral technology and machine learning algorithms enables efficient and precise estimation of soil TN content in farmland within the Northern Shaanxi agro-pastoral ecotone, providing a reliable technical framework for soil fertility assessment and precision nutrient management in the region.

Introduction

Nitrogen (N) is an indispensable element for plant growth and development as its soil content directly affecting crop yield and quality. Soil total nitrogen (TN) serves as a fundamental indicator of soil N reserves and fertility status (Bao 2000). Traditional methods for quantifying soil TN, such as the Kjeldahl method, are accurate but involve labor-intensive procedures, high costs, and potential environmental pollution, making them unsuitable for large-scale, rapid soil nutrient monitoring (Viscarra Rossel and Behrens 2010, Liu *et al.* 2018). Consequently, developing rapid, non-destructive, and efficient techniques for estimating soil TN has become a priority in modern precision agriculture and smart farming research.

Hyperspectral remote sensing captures detailed reflectance information across dozens to hundreds of contiguous narrow bands. Its unique "image-spectrum merger" characteristic offers significant potential for the quantitative retrieval of soil properties (Ben-Dor *et al.* 1999, Wu *et al.* 2012). Soil spectral signatures are a composite reflection of its physical and chemical constituents.

*Author for correspondence: <wangni@xaut.edu.cn>. ¹Institute of Water Resources and Hydro-Electric Engineering, Xi'an University Of Technology, Xi'an, Shaanxi, 710048, China.

Since most soil nitrogen is associated with organic matter, it lacks direct absorption features in the visible-near-infrared-shortwave infrared (VIS-NIR-SWIR, 350-2500 nm) range. Therefore, hyperspectral estimation of TN is essentially an indirect process achieved by detecting spectral features related to soil organic matter (SOM), such as absorption valleys near 570 nm, 660 nm, 1710 nm, and 2210 nm (Stenberg *et al.* 2010, Shi *et al.* 2015). Recent studies have extensively explored this field. For instance, Shi *et al.* (2015) systematically summarized the spectral characteristics of major soil types in China, confirming the utility of SOM-related bands for TN estimation. Research by Wu *et al.* (2012) also highlighted the importance of the visible and shortwave infrared regions for retrieving soil carbon and nitrogen content.

However, soil spectral signals are susceptible to interference from various factors like moisture, particle size, surface roughness, and parent material. This often results in weak TN-related spectral responses and complex nonlinear relationships with TN concentration (Zhang *et al.* 2016), posing challenges for processing hyperspectral data and extracting relevant information. To address this, various spectral preprocessing and feature selection techniques are employed. Common preprocessing methods include derivative analysis. First and second derivatives (FDR, SDR) effectively eliminate baseline drift and background noise, amplifying subtle spectral variations (Tsai and Philpot 1998). Further mathematical transformations, such as the logarithm of the reciprocal derivative (LRFD), can enhance absorption features related to constituent concentrations, adhering more closely to the principles of absorption spectroscopy and offering distinct advantages for soil spectra (Chu *et al.* 2004). For feature selection, beyond traditional correlation analysis (CA), the successive projections algorithm (SPA) is a forward-selection method designed to minimize variable collinearity, thereby identifying a minimal set of informative feature wavelengths, reducing model complexity, and improving computational efficiency (Araújo *et al.* 2001).

In terms of inversion modeling, machine learning algorithms have increasingly supplanted traditional linear models due to their superior ability to handle nonlinear relationships, becoming mainstream in hyperspectral soil property retrieval. Partial Least Squares Regression (PLSR) remains a prominent linear technique for handling high-dimensional, collinear data (Wold *et al.* 2001). However, simple linear models often fall short of capturing the complexity of soil systems. Decision Tree (DT) models are simple and interpretable but suffer from instability and overfitting (Breiman *et al.* 2017). Random Forest (RF), an ensemble learning method, enhances model robustness and predictive accuracy by constructing multiple decision trees and aggregating their predictions. RF has demonstrated excellent performance in estimating various soil properties like SOM, heavy metals, and salinity, establishing it as one of the most promising algorithms currently available (Breiman 2001, Xu *et al.* 2019). Research by Viscarra Rossel and Behrens (2010) also underscores the significant potential of data mining approaches like RF in interpreting soil spectra.

The agro-pastoral ecotone of Northern Shaanxi is a typical ecologically vulnerable zone in northern China, characterized by a transition between farming and pastoral cultures, environmental sensitivity, and significant soil degradation. Dingbian County, representative of this region, relies heavily on its agricultural productivity for local economic development and ecological security. Currently, research on hyperspectral inversion of soil TN specifically for this regional farmland is limited, particularly concerning feature variable optimization and comparative analysis of multiple machine learning models. Therefore, this study, using Dingbian County's farmland as a case study, aims to: (1) systematically analyze the correlations among original spectra, transformed spectra, and soil TN content to deepen the understanding of TN's spectral response mechanisms; (2) utilize SPA and CA algorithms to extract the optimal spectral feature variables for TN inversion, optimizing model inputs; (3) construct and compare the accuracy and stability of PLSR, DT, and RF inversion models, assessing their applicability in this

region; and (4) identify the optimal hyperspectral inversion model for soil TN content in the farmland of the Northern Shaanxi agro-pastoral ecotone, thereby providing a scientific basis for precise soil fertility evaluation and sustainable agricultural development in the area.

Materials and Methods

The study area is Dingbian County (E107°15′–108°22′, N36°49′–37°53′), located in Yulin City, Shaanxi Province. It lies in the transition zone between the Loess Plateau of Northern Shaanxi and the Ordos Desert Steppe of Inner Mongolia, making it a typical area of the Northern Shaanxi agro-pastoral ecotone. The terrain is generally higher in the southwest and lower in the northeast, with an average elevation of approximately 1600 meters. The climate is characterized as a mid-temperate semi-arid continental monsoon, featuring cold, dry winters and warm, rainy summers. The average annual precipitation is about 316 mm, exhibiting significant temporal variability (annually and seasonally), while evaporation considerably exceeds precipitation. The predominant soil types are Huangmian soil and aeolian sandy soil, which are generally loose in structure, low in organic matter content, and have poor water and nutrient retention capabilities. Major crops include maize, potatoes, and buckwheat. Agricultural production in this region is highly dependent on soil fertility, particularly nitrogen levels, underscoring the importance of rapid soil TN monitoring.

In July 2022 (during the peak growing season), a total of 85 topsoil (0–20 cm) samples were collected from major cultivated areas within Dingbian County using either an "S" shaped sampling pattern or a grid pattern to ensure representativeness. Each sample was a composite mixture from five points around the sampling location, with a total weight of approximately 1 kg. The samples were placed in sealed bags and transported to the laboratory. After air-drying, removing stones and plant residues, and passing through a 2-mm nylon sieve, each sample was divided into two subsamples. One subsample was used for determining TN content via the standard Kjeldahl method (Bremner 1996). The other subsample was reserved for hyperspectral measurement. The measured TN content of the 85 samples ranged from 0.21 g/kg to 1.89 g/kg, with a mean value of 0.76 g/kg and a standard deviation of 0.38 g/kg, indicating sufficient variability for modeling purposes.

Hyperspectral measurements were conducted in a darkroom using an ASD FieldSpec®4 Hi-Res spectroradiometer (Analytical Spectral Devices, Inc., USA) covering the 350–2500 nm range. The sampling interval was 1.4 nm for 350–1000 nm and 2 nm for 1000–2500 nm, with data resampled to 1-nm intervals for output. During measurement, the sieved soil sample was placed in a Petri dish (approximately 9 cm diameter) and leveled smooth with a scraper. The spectrometer's fiber optic probe was held vertically approximately 15 cm above the sample surface. For each sample, 10 spectral scans were taken and averaged to obtain the original spectral reflectance (R). A white reference panel measurement was performed before and after each sample set for calibration. To minimize random noise from instrument error and sample heterogeneity, the original spectral data were smoothed using Savitzky-Golay filtering (window size: 9 points, polynomial order: 2). Subsequently, bands with low signal-to-noise ratio (350–399 nm and 2401–2500 nm) were removed, retaining 2001 bands from 400 nm to 2400 nm for subsequent analysis.

To enhance spectral information and accentuate features sensitive to TN, the smoothed original reflectance (R) was subjected to the following mathematical transformations:

First Derivative (FDR): Effective for removing linear background, approximating the slope of the spectral curve. Calculated as:

$$\text{FDR}(\lambda_i) = \frac{R(\lambda_{i+1}) - R(\lambda_{i-1})}{2\Delta\lambda}$$

Where λ_i is the central wavelength of the i -th band and $\Delta\lambda$ is the wavelength interval (1 nm). Second Derivative (SDR): Sensitive to the curvature of the spectrum, further amplifying absorption features. Calculated as:

$$\text{SDR}(\lambda_i) = \frac{R(\lambda_{i+1}) - 2R(\lambda_i) + R(\lambda_{i-1}))}{\Delta\lambda^2}$$

Logarithm of the Reciprocal First Derivative (LRFD): The first derivative was applied to log 10 (1/R). This transformation can enhance weak absorption features and tends to linearize the relationship with constituent concentrations (Chu *et al.* 2004).

Correlation Analysis (CA): Pearson correlation coefficients (r) between the reflectance values of each band (for R, FDR, SDR, LRFD) and the measured TN content were calculated. Bands with a statistically significant correlation ($p < 0.01$) and an absolute correlation coefficient $|r| > 0.3$ were initially selected as candidate feature bands. To mitigate multicollinearity, within each contiguous interval of high correlation, only the band with the maximum $|r|$ was retained.

Successive Projections Algorithm (SPA): SPA is a variable selection technique that minimizes collinearity by iteratively selecting wavelengths with the least redundant information (Araújo *et al.* 2001). SPA was applied to the four types of transformed spectral data, with the maximum number of variables set to 30, to identify optimal feature subsets.

The feature wavelengths identified by CA and SPA were subsequently merged, and duplicates were removed to form the final set of input variables for constructing the machine learning models.

Partial Least Squares Regression (PLSR) models the relationship between the predictor variables (spectral data) and the response variable (TN content) by extracting latent factors that maximize the covariance between them. The optimal number of latent components for the PLSR model was determined using leave-one-out cross-validation to prevent overfitting.

Decision Tree (DT) models partition the data recursively based on feature values to create subsets with higher purity regarding the target variable. Key parameters, including the maximum tree depth (set to 10) and the minimum number of samples required at a leaf node (set to 5), were determined through preliminary experiments to prevent excessive model complexity.

Random Forest (RF) is an ensemble learner comprising multiple decision trees. It employs bootstrap aggregation (bagging) to create diverse training subsets and randomly selects a subset of features for splitting at each node in tree construction. The final prediction is the average (for regression) of the predictions from all individual trees. Key hyper-parameters of the RF model, including the number of trees (n -estimators: 50-500), the number of features considered for splitting, and the maximum depth of trees (max depth: 5-20), were optimized using grid search coupled with cross-validation.

The 85 samples were randomly split into a training set (64 samples, 75%) and a validation set (21 samples, 25%), ensuring that the statistical characteristics (mean, standard deviation) of TN content were similar between the two sets.

To objectively evaluate the performance of the developed soil TN inversion models, three common metrics were employed: the coefficient of determination (R^2), root mean square error (RMSE), and the ratio of performance to deviation (RPD). These metrics assess predictive accuracy and model robustness. Coefficient of Determination (R^2): Measures the proportion of variance in the dependent variable (TN content) explained by the model. Values closer to 1 indicate better fit. Root Mean Square Error (RMSE): Represents the average deviation between predicted and measured values. Lower values indicate higher prediction accuracy. Ratio of Performance to Deviation (RPD): Calculated as the ratio of the standard deviation (SD) of the measured values in the validation set to the RMSE. It is used to evaluate the predictive capability

of the model: $RPD > 2.0$ indicates excellent predictive ability; $1.4 < RPD < 2.0$ allows for approximate quantification; $RPD < 1.4$ indicates unreliable prediction (Chang *et al.* 2001).

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

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

$$RPD = \frac{RMSE}{SD}$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

Where, y_i is the measured value for the i -th sample, $\hat{y}_i$ is the predicted value for the i -th sample, $\bar{y}$ is the mean of the measured values, and n is the number of samples in the validation set. Where, SD is the standard deviation of the measured values in the validation set.

All model accuracy assessments were based on the validation set results. All data processing, analysis, and modeling were performed in Python 3.8 utilizing libraries such as scikit-learn, pandas, and numpy.

Results and Discussion

Descriptive statistics of soil TN content are presented in Table 1. The mean and standard deviation values were similar between the training and validation sets, indicating a reasonable and consistent distribution suitable for modeling and validation. The original spectral curves (Fig. 1) exhibited typical soil spectral characteristics, with reflectance generally increasing with wavelength and distinct water absorption valleys visible around 1400 nm, 1900 nm, and 2200 nm. Correlation analysis with TN content (Fig. 2) revealed generally weak correlations ($|r| < 0.4$) across the entire spectral range for the original reflectance (R). After applying derivative transformations, the correlations were significantly enhanced (Fig. 1c, d, e). Notably, the LRFD transformation produced several regions with $|r| > 0.6$, indicating very significant correlations, particularly in the ranges 520-580 nm, 780-820 nm, 1020-1100 nm, 1700-1750 nm, and 2200-2250 nm. These wavelength regions often correspond to characteristic absorption features associated with soil organic matter and clay minerals, suggesting that the LRFD transformation most effectively accentuated the spectral information related to TN.

Table 1. Descriptive statistics of soil total nitrogen content (Unit: g/kg).

Dataset	Sample Size	Minimum	Maximum	Mean	Std. Dev.	Skewness	Kurtosis
All Samples	85	0.21	1.89	0.76	0.38	0.85	0.42
Training Set	64	0.22	1.89	0.75	0.37	0.91	0.55
Validation Set	21	0.21	1.65	0.78	0.40	0.72	0.18

The feature wavelengths selected by CA and SPA from the four spectral transformations are exemplified for LRFD in Table 2. CA primarily selected individual bands with the highest correlation coefficients, whereas SPA tended to select a complementary set of wavelengths. For instance, from the LRFD data, CA selected 6 bands, while SPA selected 12 bands. After merging and removing duplicates, the final feature sets comprised 8, 11, 14, and 15 wavelengths for R ,

FDR, SDR, and LRFD, respectively. These selected features were predominantly located in the visible blue-green region (related to iron oxides/organic matter), the NIR plateau (related to general organic content), and the SWIR region (related to overtones and combinations of organic C-H, N-H bonds and clay mineral O-H bonds).

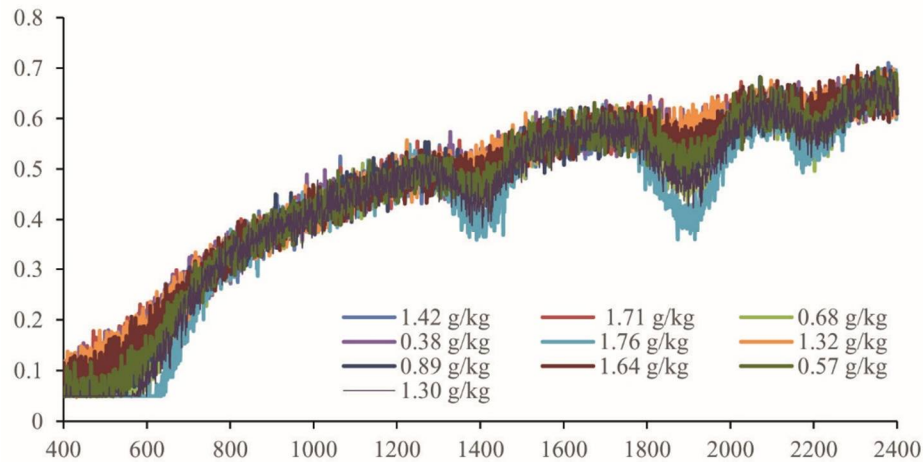


Fig. 1. Spectral Curve Characteristics with Different Total Nitrogen Contents.

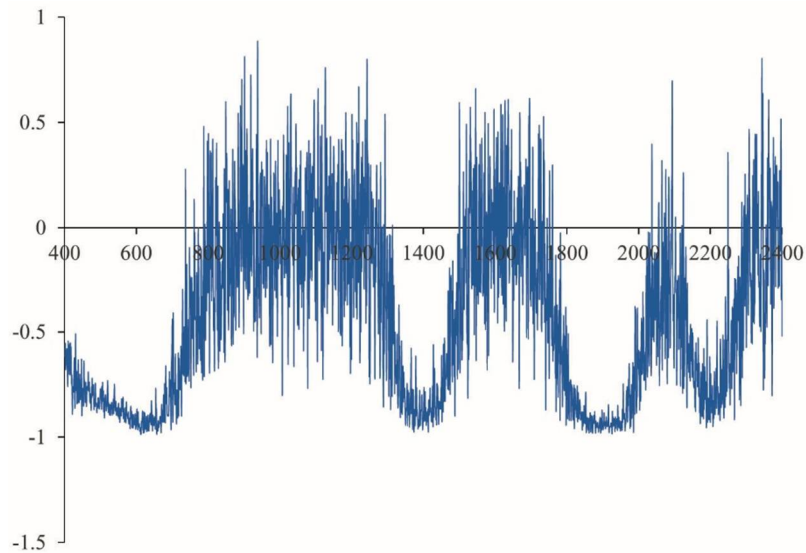


Fig. 2. Correlation Analysis Based on Original Spectral Data.

Table 2. Example of feature wavelength selection results (using LRFD data).

Selection Method	Selected Feature Wavelengths (nm)
CA	528, 798, 1025, 1082, 1721, 2248
SPA	435, 567, 632, 701, 798, 1025, 1150, 1321, 1721, 1850, 2015, 2248
Merged Set	435, 528, 567, 632, 701, 798, 1025, 1082, 1150, 1321, 1721, 1850, 2015, 2248

The performance metrics of the PLSR, DT, and RF models, built using the selected feature wavelengths, are compared in Table 3. Impact of Spectral Transformation: For all three machine learning models, those built using derivative-transformed spectra (FDR, SDR, LRFD) consistently achieved significantly higher accuracy than models based on the original reflectance (R). This demonstrates the effectiveness of derivative processing in suppressing noise and enhancing the relevant spectral signal. The LRFD transformation yielded the most stable and prominent improvements, with both PLSR and RF models achieving their highest accuracy when using LRFD features, confirming its advantage in strengthening concentration-related spectral features.

Table 3. Accuracy comparison of different model combinations on the validation set.

Spectral Data	Model	R ²	RMSE (g/kg)	RPD
R	PLSR	0.62	0.25	1.60
	DT	0.51	0.31	1.29
	RF	0.71	0.21	1.90
FDR	PLSR	0.68	0.22	1.82
	DT	0.58	0.28	1.43
	RF	0.78	0.18	2.22
SDR	PLSR	0.72	0.20	2.00
	DT	0.61	0.26	1.54
	RF	0.81	0.16	2.14
LRFD	PLSR	0.76	0.18	2.12
	DT	0.65	0.24	1.67
	RF	0.85	0.11	2.15

(2) Model Performance Comparison: PLSR Model: As a linear model, PLSR performed robustly. It achieved its best performance with LRFD features ($R^2=0.76$, $RPD=2.22$), indicating a reasonably strong linear relationship between TN and the selected LRFD features. DT Model: The DT model performed the poorest among the three. Its validation R^2 was consistently lower, and its RMSE higher, than both PLSR and RF models. While achieving very high R^2 on the training set, its performance dropped markedly on the validation set, exhibiting typical overfitting and poor generalization capability. RF Model: The RF model demonstrated the best overall performance. It achieved the highest R^2 and lowest RMSE across all four spectral data types. The RF model using LRFD features yielded a validation R^2 of 0.85, RMSE of 0.11 g/kg, and RPD of 2.15, indicating excellent predictive ability. The ensemble approach of RF effectively mitigated the overfitting tendency of individual decision trees, showcasing strong nonlinear fitting capability and robustness.

Optimal Model Identification: Based on the comprehensive comparison, the Random Forest model utilizing features from the LRFD transformation (RF-LRFD) was identified as the optimal model for inverting soil TN content in the farmland of Dingbian County. The scatter plot of predicted versus measured values for this model (Fig. 3) shows points closely distributed, further verifying its high accuracy and reliability.

The sensitive wavelengths identified in this study hold clear physical significance. The band near 528 nm is associated with strong absorption by organic matter in the blue-green region. Bands like 798 nm and 1025 nm in the NIR reflect the overall abundance of organic materials. The wavelengths 1721 nm and 2248 nm correspond to the first overtone of C-H bonds in organic matter and the combination bands of O-H bonds in clay minerals, respectively (Stenberg *et al.* 2010, Clark and Roush 1984). The successful identification of these bands confirms the feasibility

of indirectly estimating TN via its association with organic matter. The superior performance of the LRFD transformation is attributed to its ability to amplify weak absorption signals, thereby making these OM-related (and hence TN-related) features more pronounced and improving model sensitivity.

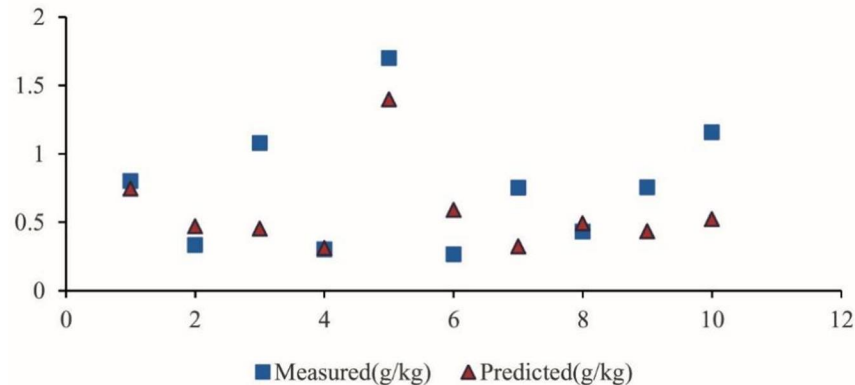


Fig. 3. Scatter plot of SOC optimal inversion model.

This study found that Random Forest (RF) outperformed both the traditional linear model (PLSR) and the simpler nonlinear model (DT) when dealing with the high-dimensional, nonlinear relationships inherent in soil hyperspectral data. This superiority stems from several factors:

(1) RF inherently handles multicollinearity without requiring explicit dimensionality reduction like PLSR, preserving more original information;

(2) The use of bagging and random feature selection effectively reduces overfitting and enhances generalization;

(3) RF is relatively insensitive to outliers, contributing to its stability. The DT model's limitations arise from its instability (high variance) and susceptibility to overfitting, especially with small datasets or without strict pruning. Therefore, ensemble algorithms like RF (and potentially others like Gradient Boosting) are highly recommended for similar soil property inversion studies involving hyperspectral data.

This study has some limitations. Firstly, the sample size ($n=85$), while sufficient for initial model development, could be expanded in future work to enhance model generalizability across a wider range of conditions. Secondly, soil moisture remains a significant confounding factor affecting spectra. Although derivative transformations can partially mitigate its effect, explicitly incorporating soil moisture content as an auxiliary variable in the models could be a valuable future direction. Furthermore, this research was based on laboratory-controlled spectral measurements. Translating the developed models to in-situ field measurements or to airborne/spaceborne hyperspectral remote sensing applications presents a significant challenge, involving complexities such as atmospheric correction and scaling effects, which need to be addressed in future research.

This study systematically investigated the inversion of soil total nitrogen content in the farmland of Dingbian County, Northern Shaanxi agro-pastoral ecotone, using hyperspectral techniques combined with machine learning. The main conclusions are as follows: Spectral derivative transformations, particularly the LRFD, significantly enhanced the correlation between soil spectra and TN content, effectively extracting sensitive feature bands located in the visible, NIR, and SWIR regions. Among the three machine learning models evaluated, the Random Forest

(RF) model exhibited the best predictive performance and stability, demonstrating clear advantages in handling nonlinearity and resisting overfitting. The PLSR model performed robustly as a linear method, while the Decision Tree (DT) model was prone to overfitting and showed the lowest accuracy. The combined RF-LRFD model was identified as the optimal approach for estimating soil TN content in the study area, achieving a validation set R^2 of 0.85, RMSE of 0.11 g/kg, and RPD of 2.15, meeting the requirements for practical application. The successful integration of hyperspectral analysis and machine learning algorithms provides an effective technical pathway for the rapid, non-destructive monitoring of soil nitrogen status in the ecologically vulnerable agro-pastoral ecotone of Northern Shaanxi, offering significant positive implications for guiding precision fertilization and promoting sustainable agricultural development in the region.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No.51979221), the Research Fund of the State Key Laboratory of Eco-hydraulics in Northwest Arid Region, Xi'an University of Technology (Grant No.2019KJCXTD-5) and Xi'an University of Technology Doctoral Dissertation Innovation Fund (104-252072305). Furthermore, we thank the anonymous reviewers and editor for their helpful comments.

References

- Araújo MCU, Saldanha TCB, Galvão RKH, Yoneyama T, Chame HC and Visani V 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometr. Intell. Lab.* **57**(2): 65-73.
- Bao SD. 2000. *Soil Agro-Chemical Analysis*; China Agric. Press: Beijing, China.
- Ben-Dor E, Irons JR and Epema GF 1999. Soil reflectance. In: *Remote Sensing of the Earth Sciences: Manual of Remote Sensing*. 3rd ed.; Rencz, AN (Ed). John Wiley and Sons: New York, NY, USA. **3**: 111-188.
- Breiman L, Friedman JH, Olshen RA and Stone CJ 2017. *Classification and Regression Trees*; Routledge, New York, NY, USA.
- Breiman L 2001. Random forests. *Mach. Learn.* **45**(1): 5-32.
- Bremner JM 1996. Nitrogen-total. In: *Methods of Soil Analysis, Part 3: Chemical Methods*. Sparks DL (Ed). SSSA Book Series No. 5. Soil Science Society of America, Madison, WI, USA. pp. 1085-1121.
- Chang CW, Laird DA, Mausbach MJ and Hurburgh CR 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. *Soil Sci. Soc. Am. J.* **65**(2): 480-490.
- Chu XL, Yuan HF and Lu WZ 2004. Progress and application of spectral pretreatment and wavelength selection methods in near-infrared spectroscopy. *Chem. J. Chin. Univ.* **16**(4): 528-542.
- Clark RN and Roush TL 1984. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *J. Geophys. Res.* **89**(B7): 6329-6340.
- Liu HJ, Zhang XL, Li LX, Peng J, Wang SZ and Liu DS 2018. Research progress on hyperspectral estimation model of soil nutrients. *Acta Pedol. Sin.* **55**(4): 781-794.
- Shi Z, Wang QL, Peng J, Ji WJ, Liu HJ and Li X 2015. Development of soil spectral libraries and quantitative analysis of soil properties in China. *Spectrosc. Spect. Anal.* **35**(11): 3083-3088.
- Stenberg B, Viscarra Rossel RA, Mouazen AM and Wetterlind J 2010. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* **107**: 163-215.
- Tsai F and Philpot W 1998. Derivative analysis of hyperspectral data. *Remote Sens. Environ.* **66**(1): 41-51.
- Viscarra Rossel RA and Behrens T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*. **158**(1-2): 46-54.

- Wu W, Shen YY, Liu XN, Li SJ and Wang JD 2012. Characterization of visible and near-infrared spectral properties of Chinese soils. *Spectrosc. Spect. Anal.* **32**(9): 2393-2398.
- Wold S, Sjöström M and Eriksson L 2001. PLS-regression: A basic tool of chemometrics. *Chemomet. Intell. Lab.* **58**(2): 109-130.
- Xu XB, Lyu JS and Wu QY 2019. Spatial simulation of soil salinity in the Yellow River Delta based on random forest. *Acta Pedol. Sin.* **56**(3): 565-576.
- Zhang J, Shi Z and Li FL 2016. Research progress on methods for eliminating moisture effects in visible and near-infrared spectroscopy for soil prediction. *Spectrosc. Spect. Anal.* **36**(8): 2480-2485.

(Manuscript received on 20 August, 2026; revised on 22 September, 2026)