Inversion Models for Orchard Soil Nutrient Content Using Near-Infrared Spectroscopy

Rongguang Sun[1], Chaojun Hou[1], Huawei Cui[1*]

[1] Zhongkai University of Agriculture and Engineering, Guangzhou, 510225, China

* cuihuawei820303@163.com

https://doi.org/10.70695/AA1202502A16

Abstract

Precise determination of available nitrogen (N), phosphorus (P) and potassium (K) concentrations in soil is essential for orchard fertilizer management. However, conventional chemical analytical methods are time-consuming and costly, making it difficult to meet the demand for rapid detection in modern agriculture. Near-infrared spectroscopy (NIRS) is a promising approach for rapidly detecting soil nutrient concentrations. This study aims to develop inversion models, including partial least squares regression (PLSR), support vector machine (SVM), and random forest (RF), to accurately determine the concentrations of available N, P, and K in soil using NIRS. To mitigate spectral interference and enhance prediction performance of inversion models. we systematically implemented multiple preprocessing methods including Savitzky-Golay smoothing (SG), multiplicative scatter correction (MSC), standard normal variate transformation (SNV), moving average (MA) and first-derivative transformation (FD). The results indicated that the RF model achieved superior predictive accuracy compared to the PLSR and SVR, with prediction performance for available N being R^2=0.539, RMSE=21.408 mg/kg, and RPD=1.490; for available P, R^2=0.536, RMSE=25.056 mg/kg, and RPD=1.484; and for available P, R^2=0.429, RMSE=42.452 mg/kg, and RPD=1.338. Compared with the spectral data without preprocessing, the performance of RF model improved by 8.53%, 3.17%, and 0.47% for the R^2 values for available N, P, and K, respectively. The comparison of the preprocessing methods in combination with the RF model revealed that MSC had the best model accuracy for the available N (R^2=0.585, RMSE=20.326 mg/kg, RPD=1.569), SG had the best model accuracy for the available P (R^2=0.553, RMSE=24.585 mg/kg, RPD=1.513), and MA had the best model accuracy for the available K (R^2=0.431, RMSE=42.383 mg/kg, RPD=1.340). It is noteworthy that the model accuracy is worst for the available K, possibly due to its inherently weaker spectral response characteristics and lower signal-to-noise ratio. This study shows that NIR spectroscopy using appropriate preprocessing methods and regression models can accurately predict soil nutrient concentrations, supporting precision fertilization management in orchard production systems.

Keywords Near-infrared Spectroscopy; Soil; Available Nutrient; Feature Selection; Random Forest

1 Introduction

Soil constitutes the fundamental resource underpinning agricultural production systems, with its physicochemical properties directly influencing crop developmental processes, yield performance and nutritional quality parameters [7]. Contemporary demographic expansion coupled with accelerated urbanization processes has intensified both quantitative and qualitative pressures on China's arable land resources, thereby rendering the preservation of the 1.8 billion mu (approximately 120 million hectares) agricultural land baseline critical for national food security and ecological stability maintenance [13].However, the sustained application of excessive chemical fertilizers has precipitated a cascade of soil degradation processes, including nutrient imbalances, structural deterioration, and environmental contamination pathways that collectively compromise soil integrity and ecosystem functionality [6]. Consequently, the development of rapid, accurate, and cost-effective soil nutrient assessment methodologies has emerged as a fundamental requirement for implementing precision fertilization strategies and establishing sustainable soil management frameworks [1].

Nitrogen (N), phosphorus (P), and potassium (K) constitute essential macronutrients that determine soil fertility status and crop productivity potential, functioning respectively to promote vegetative growth, root system development, and plant stress tolerance mechanisms [12]. Conventional analytical methodologies for soil nutrient determination, including Kjeldahl nitrogen analysis, molybdenum-

antimony colorimetric phosphorus detection, and flame photometric potassium measurement while providing high analytical accuracy, are characterized by procedural complexity, prolonged analytical timeframes, and substantial operational costs. These limitations prevent such methods from meeting the high-throughput analytical demands of modern precision agricultural systems [8].Near-infrared spectroscopy (NIRS) technology has emerged as a promising complementary analytical approach to traditional methodologies, distinguished by its capacity for rapid, simultaneous, multi-parameter analysis [10]. Operating within the 780-2526 nm spectral range, NIRS enables the quantification of soil chemical components through characteristic molecular absorption patterns, thereby significantly reducing analytical time and associated costs. This technology demonstrates particular suitability for applications requiring extensive spatial coverage and high temporal resolution, addressing critical limitations of conventional soil analysis methods [14].In the domain of soil nitrogen inversion methodologies, Zheng et al. (2005) demonstrated that integrating NIRS with partial least squares regression (PLSR) for soil nitrogen prediction resulted in R^2 values ranging from 0.86 to 0.92. Shao et al. (2011) applied least squares support vector machine (LS-SVM) models in Zhejiang Province, China, achieving R^2 values of 0.90. Liu et al. (2013) enhanced methodology by using Monte Carlo uninformative variable elimination (MC-UVE) with PLSR for critical wavelength selection, resulting in a validation correlation coefficient of 0.84.Regarding soil phosphorus inversion research trajectories, Maleki et al. (2006) used visible and near-infrared spectroscopy to develop calibration models for predicting soil phosphorus content based on spectral signatures. They applied Partial Least Squares Regression (PLSR) to create predictive models for soil phosphorus quantification within both spectral domains. The validation showed R^2 values of 0.75 and 0.73, respectively. Jia et al. (2015) expanded this research by investigating available phosphorus prediction using near-infrared spectroscopy (NIRS) and recursive partial least squares regression (PLSR), exploring correlations between phosphorus content and spectral features. Their framework achieved an R^2 of 0.85 and an RMSE of 0.14 mg/kg. This approach is useful for both laboratory and field-based rapid phosphorus content prediction, supporting precise phosphorus management and automated fertilization.

Research on near-infrared spectroscopic soil potassium inversion has primarily focused on optimizing preprocessing techniques and modeling algorithms. Jin et al. (2020) compared 29 preprocessing methods and found that combining Savitzky-Golay smoothing, standard normal variate transformation and detrending techniques significantly improved model stability. Their study showed that AdaBoost models achieved R^2 values of 0.945 for low-concentration potassium prediction, while Gradient Boosting Regression Trees (GBRT) reached R^2 values of 0.947 for high-concentration potassium. Endut et al. (2023) developed potassium content prediction models using NIR data, achieving an outstanding R^2 of 0.9998 with RMSE of 0.0600, showcasing exceptional predictive accuracy for soil potassium quantification. Kone et al. (2018) used VIS-NIR spectroscopy on 877 soil samples from Mali to predict exchangeable potassium content. Their results showed that PLSR models (R^2=0.57) outperformed PCR models (R^2=0.50), indicating that while VIS-NIR spectroscopy with multivariate statistical methods can estimate soil potassium, there is still room for improving predictive accuracy.

This study focuses on orchard soils, using near-infrared spectroscopy with various preprocessing techniques to develop optimized models for quantifying available N, P, and K content. This integrated approach aims to offer strong technical support for precise fertilization and sustainable soil management in modern agriculture.

2 Methods

2.1 Sample Collection

The investigation took place in a 1,100-hectare orchard located at the southeastern end of Haizhu District, Guangzhou (113°18′45″E, 23°03′55″N), with sandy, clay and loam soil predominantly planted with Litchi chinensis Sonn, Dimocarpus longan Lour and Averrhoa carambola L. The research area has a subtropical monsoon climate, with a mean annual temperature of 20°C and an average annual precipitation of 1,500 mm.In September 2023 (Fig. 1), a stratified sampling grid based on remote sensing imagery was established, consisting of 67 sampling locations (39 in the central orchard and 28 in peripheral regions). Each location, depending on its spatial characteristics, yielded 1-2 soil samples through a five-point cross-sampling method, resulting in the collection of 89 samples from the 0-20 cm soil layer: 27 from Litchi zones, 49 from Dimocarpus zones, and 13 from Averrhoa zones. Each sample weighed approximately 500 g.The collected soil samples underwent a standardized preparation process:

air-drying under controlled conditions, manual removal of contaminants, thermal treatment at 45°C for 30 minutes, pulverization with agate mortars to minimize contamination and filtration through 2 mm nylon mesh to ensure uniformity. Each processed sample was split into two 50 g subsamples, one for chemical analysis and the other for near-infrared spectroscopic analysis. This multidisciplinary approach enabled a comprehensive pedological characterization of the orchard ecosystem while ensuring consistency to establish robust spectroscopic-chemical correlations within the complex soil-plant-atmosphere continuum.

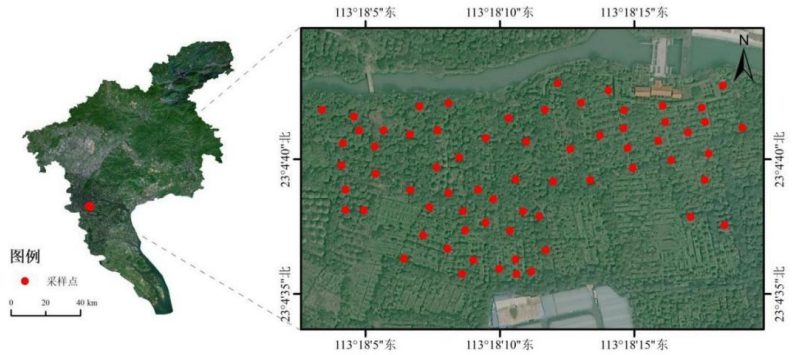


Fig. 1. Geographical location map of the study area

2.2 Spectral Data Acquisition

After filtration through a 2 mm nylon mesh, 5 g aliquots were extracted from each of the 89 soil samples, homogenized through mechanical agitation, and compressed using a ZS-24T hydraulic press under controlled pressure conditions (15 MPa) to create uniformly dimensioned soil tablets (25 mm diameter, 5.0-5.5 mm thickness) with optimized surface planarity for spectroscopic analysis. Near-infrared spectral acquisition was performed using a Specim FX17 hyperspectral imaging system (Specim Ltd., Finland), as shown in Fig. 2. The system has a spectral range of 935-1720 nm with an optical resolution of 8 nm, covering 224 spectral bands. It is equipped with a 50 W halogen illumination source and a motorized positioning platform for optimal spatial registration. To minimize systematic errors, the spectroscopic apparatus underwent a 30-minute thermal equilibration before use, reducing baseline drift. This was followed by rigorous radiometric calibration using standardized black and white reference panels, along with optimization of scanning velocity and integration time to maximize signal-to-noise ratio. Each sample was scanned twice (obverse and reverse orientations), generating a dataset of 178 spectral signatures. The signatures were processed in the ENVI remote sensing analysis environment, where circular regions of interest (ROI) with precise dimensions (18 mm diameter) were extracted from each image to eliminate peripheral artifacts, establishing a robust spectroscopic foundation for subsequent multivariate analysis of soil nutrient-spectral correlations.

Fig. 2. Data acquisition process of near-infrared spectrometer

The spectral dataset of 178 soil samples was divided using a stratified random allocation method in a 7:3 ratio, creating a calibration cohort (n=124) for model development and an independent validation cohort (n=54) for performance assessment. The partitioning method used a computationally-optimized stochastic sampling algorithm to ensure proper representation across the multivariate landscape, maintaining statistical homogeneity between cohorts while preserving pedological variability essential for model generalization. This rigorous partitioning procedure supported the development of spectroscopic prediction frameworks with better transferability across varied soil conditions, reducing spatial autocorrelation confounders and ensuring a valid foundation for comprehensive model evaluation which is critical for assessing the efficiency of hyperspectral methods in complex agronomic systems with significant spatial and temporal variability.

2.3 Spectral Data Thethods

Systematic noise artifacts, primarily in the peripheral spectral regions, led to the exclusion of edge-effect wavelength bands (935-956 nm and 1698-1720 nm) from the analytical framework, leaving 210 significant bands in the optimized range (956-1700 nm) for multivariate modelling. So as to improve spectroscopic signal quality and enhance nutrient-specific feature extraction in the complex soil matrix, this study conducted a comparative analysis of five spectral preprocessing methods, each chosen for its mathematical properties and proven ability to reduce spectral interference while amplifying chemometric signatures related to soil nutrients in the near-infrared spectrum. The integration of these preprocessing algorithms allowed for the systematic separation of overlapping spectral signatures, providing an optimized foundation for chemometric models that quantify the spatiotemporal dynamics of available nutrients in heterogeneous orchard soils with diverse biogeochemical characteristics：

Savitzky-Golay smoothing (SG)

The Savitzky-Golay (SG) preprocessing method represents a local polynomial least-squares fitting method for spectral smoothing, in which the spectral continuum was filtered using high-order polynomial regression within predefined sliding windows. SG effectively attenuates stochastic noise components while preserving critical inflection and absorption points contained in the original spectral architecture. SG can be formalized by a convolution operation,

$$x_{SG}(i) = \sum_{j=-m}^{m} c_j x(i+j) \qquad (1)$$

where, $x_{SG}(i)$ represents the value after SG smoothing, $x(i)$ is the original spectral value at the i-th band of the preprocessed spectrum, c_j is the coefficient of the j-th term in the polynomial fit within the window size of 2m + 1. The polynomial order was set to 3 and the window size was set to 9 for SG method employed in this paper.

Multiplicative Scatter Correction (MSC)

Multiplicative scatter correction (MSC) is a transformative mathematical framework that establishes a linear regression relationship between the individual spectrum and the referenced spectrum, effectively attenuating heterogeneous scattering phenomena caused by granulometric variations, surface microtopography and other physical matrix parameters that introduce non-chemical variance into the spectrum. The formula for MSC preprocessing is as follows,

$$x_{\mathrm{MSC}}(i) = \frac{x(i)-b_i}{a_i} \qquad (2)$$

where, a_i, b_i represent the coefficients obtained by regression of the preprocessed spectrum from the reference spectrum. The reference spectrum is usually the mean spectrum of the entire spectral dataset, $\bar{X}_{\mathrm{ref}} = \frac{1}{n}\sum_{k=1}^{n} X_k$, n is the total number of samples, $x_{\mathrm{MSC}}(i)$ is the value processed by MSC.

Standard Normal Variate transformation (SNV)

By calculating the mean and standard deviation of each spectrum, the multiplicative effects caused by factors such as particle size distribution and surface scattering between samples are eliminated, improving the comparability of spectra between different samples. The formula for SNV preprocessing is as follows,

$$x_{\mathrm{SNV}}(i) = \frac{x(i)-\bar{x}}{s} \qquad (3)$$

where, $\bar{x}$ is the mean of all bands of x, s is the standard deviation of the spectrum, m is the total number of spectral bands, $x_{SNV}(i)$ is the value processed by SNV.

Moving Average (MA)

The moving average preprocessing performs averaging smoothing over the original spectrum, effectively suppressing high-frequency stochastic noise components while improving the signal stability. MA is beneficial for spectral data that exhibits significant fluctuation patterns. The formula for MA is as follows,

$$x_{\mathrm{MA}}(i) = \frac{1}{n}\sum_{k=i-\frac{n-1}{2}}^{i+\frac{n-1}{2}} x(k)\frac{x(i)-\bar{x}}{s} \qquad (4)$$

where, n is the size of the sliding window (usually odd), $x_{MA}(i)$ is the average within the sliding window of the spectrum. In this paper, the size of the sliding window n is set to 5.

First Derivative (FD)

The first derivative transformation calculates the slope variations between neighboring points along the spectrum. This effectively emphasizes the peak-valley characteristics within the spectrum, while eliminating baseline drift influence and improving the detection capabilities for subtle absorption features. The formula for FD is as follows,

$$x_{\mathrm{FD}}(i) = x(i+1) - x(i) \qquad (5)$$

where, $x_{\mathrm{FD}}(i)$ is the value resulting from the difference between the neighboring band values of the processed spectrum.

2.4 Modeling methods

Random Forest (RF)

Random Forest (RF) is an advanced ensemble learning method that builds multiple decision trees and integrates their predictions, significantly improving model robustness and stability across various soil types. In soil nutrient inversion, RF excels at non-linear modelling, capturing the complex relationships between hyperspectral data and nutrient concentration gradients in diverse soil environments. This framework uses bootstrap aggregation (bagging) and stochastic feature selection, randomizing both training data and spectral variables during tree construction to reduce overfitting in high-dimensional spectroscopic datasets. The RF algorithm can effectively model complex variable interactions in hyperspectral data with high dimensionality, without requiring assumptions about data normality or homoscedasticity. Additionally, the ensemble structure is robust to spectral artifacts and outliers due to its averaging mechanism, providing a reliable framework for soil nutrient quantification that overcomes the limitations of traditional parametric methods.

Partial Least Squares Regression (PLSR)

Partial Least Squares Regression (PLSR) is a multivariate statistical method that combines principal component analysis for dimensionality reduction with multivariate linear regression for prediction. It is particularly effective for analyzing high-dimensional datasets with significant multicollinearity among predictors. In soil spectroscopic analysis, PLSR projects hyperspectral data into a latent variable space, extracting spectral features that are most strongly correlated with nutrient concentrations. This transformation enhances the covariance between the spectroscopic data and nutrient vectors, addressing multicollinearity between adjacent wavelength bands while reducing model complexity through efficient compression of the spectral data. By integrating multivariate statistics and chemometric theory, PLSR offers a powerful framework that bridges spectroscopy, soil science, and statistical learning.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) methodology is grounded in statistical learning theory and structural risk minimization, using optimized hyperplanes in multidimensional feature spaces to perform regression tasks. In soil nutrient inversion processes, SVM applies kernel transformations that project spectroscopic data into high-dimensional feature spaces, converting non-linear spectral-nutrient relationships into linear structures for analysis. This approach optimizes classification boundaries through support vector optimization, showing strong adaptability to high-dimensional datasets with limited observations. By calibrating penalty parameters (C) and kernel function coefficients, SVM achieves an optimal balance between model complexity and generalization, ensuring robustness against spectral artifacts and anomalies common in pedological spectroscopic datasets.

Figure 3 illustrates the comprehensive methodological framework of this investigation, outlining the research workflow from soil sample acquisition and nutrient quantification to spectral data collection, preprocessing optimization, feature extraction, and model construction. This methodological framework provides a structured foundation for near-infrared spectroscopic inversion of soil nutrient parameters, integrating classical pedology, spectroscopic techniques, and advanced machine learning.

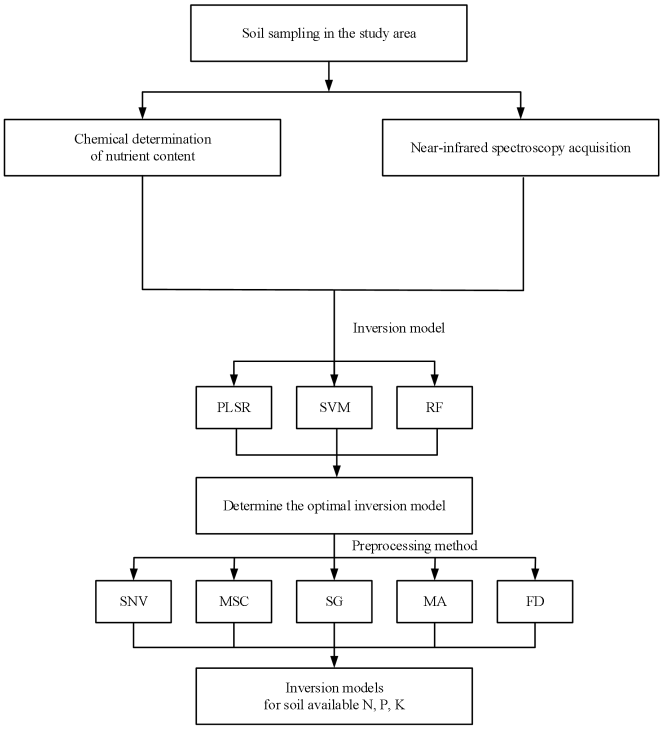


Fig. 3. Technical roadmap

2.5 Model Evaluation Methods

Regression coefficient of determination (R^2), root mean square error (RMSE) and relative prediction deviation (RPD) were employed as evaluative metrics for analytical assessment of modelling outcomes and validation efficiency. R^2 quantifies the model's explanatory capacity, RMSE measures the magnitude of deviation between predicted and observed values, while RPD provides further evaluation of predictive precision. The computational formulations are expressed as:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{7}$$

$$RPD = \frac{SD}{RMSE} \tag{8}$$

where, $\hat{y}_i \; and \; y_i$ represent the predicted and measured values, respectively, of the i-th sample. The mean of the measured values is denoted as $\bar{y}$, n represents the total number of samples, and SD is the standard deviation of the measured values. The correlation coefficient of the samples was also calculated. When the RPD exceeds 2.0, it indicates that the model possesses good predictive capability. When the RPD ranges between 1.4 and 2.0, it suggests that the model demonstrates acceptable predictive performance, although further improvement remains necessary. When the RPD falls below 1.4, it indicates that the model cannot effectively perform prediction tasks.

3 Results and Analysis

3.1 Statistical Analysis of Soil Nutrient Content

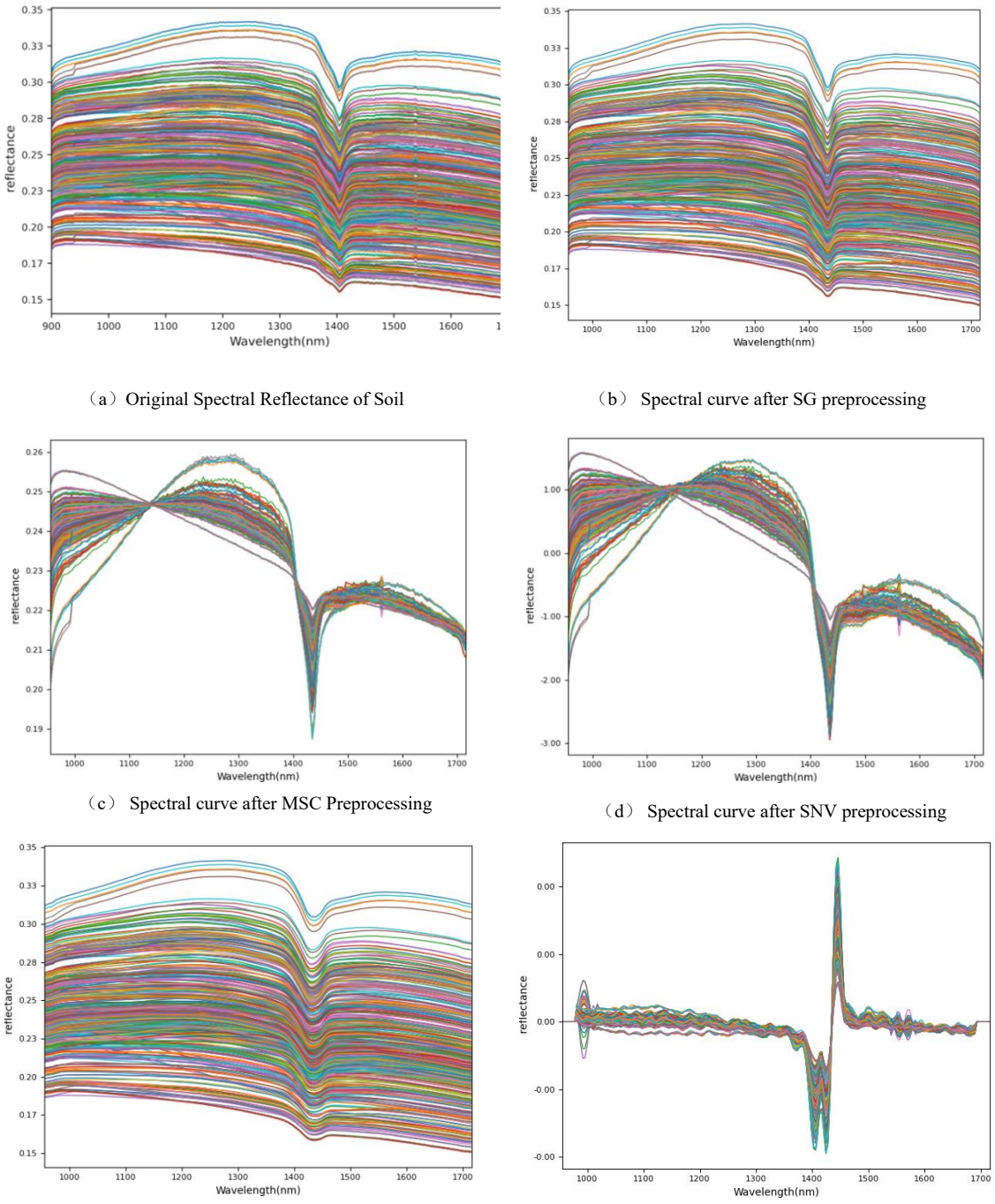
Spectral data were used to independently implement RF, SVM, and PLSR models for the inversion of available N, P, and K. As shown in Table 2, for available nitrogen content reconstruction, RF models demonstrated superior performance in the testing set with R^2=0.539, RMSE=21.408 mg/kg, and RPD=1.490, followed by SVM (R^2=0.589, RMSE=20.216 mg/kg, RPD=1.578) and PLSR (R^2=0.451, RMSE=23.365 mg/kg, RPD=1.365).For available phosphorus content prediction (Table 3), RF models again achieved the highest accuracy with R^2=0.536, RMSE=25.056 mg/kg, and RPD=1.484 in the testing set, while SVM showed moderate performance (R^2=0.405, RMSE=28.380 mg/kg, RPD=1.311) and PLSR exhibited the lowest accuracy (R^2=0.188, RMSE=33.137 mg/kg, RPD=1.122).Similarly, for available potassium content estimation (Table 4), RF models maintained their superior performance with R^2=0.429, RMSE=42.452 mg/kg, and RPD=1.338 in the testing set, outperforming SVM (R^2=0.307, RMSE=46.768 mg/kg, RPD=1.215) and PLSR (R^2=0.198, RMSE=50.299 mg/kg, RPD=1.129).These results consistently demonstrate that RF models excel at capturing non-linear relationships in spectroscopic data while maintaining robust performance despite noise interference. Although SVMs generally perform well with high-dimensional, limited-observation datasets, their underperformance in this study may be attributed to suboptimal kernel function parameter selection. As a linear approach, PLSR showed consistently lower predictive accuracy across all three nutrients, likely due to the complex non-linear soil-spectral relationships inherent in this system. These findings align with Zhou et al. (2020), who demonstrated that ensemble learning methods generally outperform traditional regression approaches in handling complex spectroscopic datasets.

Table 1. Descriptive statistics of soil nutrients

	Maximum mg/kg	Minimum mg/kg	Mean mg/kg	Median mg/kg	Standard Deviation mg/kg	Coefficient of Variation
Available N	198.20	19.40	75.99	69.00	34.56	45.48%
Available P	243.50	25.60	83.02	59.30	42.61	51.33%
Available K	378.00	46.90	278.46	288.70	70.37	25.27%

3.2 Spectral Curve Characteristic Analysis

Reflectance spectra of soil samples within the near-infrared domain (956-1716 nm) showed dynamic variations from 0.15 to 0.35, characterized by notable fluctuations and systematic trends (Ben-Dor et al., 2015). In the 900-1200 nm spectral region, reflectance increased gradually from 0.15 to 0.20, most likely due to reduced soil moisture content and enhanced light scattering from iron oxides and other minerals. The 1200-1400 nm interval showed increased spectral fluctuations, with a prominent absorption feature around 1400 nm, where reflectance dropped to 0.17. This decrease is mainly due to the strong absorption of OH groups and H_2O molecules in soil moisture, possibly enhanced by bound water in clay minerals like kaolinite or montmorillonite (Hillel et al., 2005). After the 1400 nm absorption feature, reflectance quickly increased to over 0.30 as wavelengths moved beyond the strong water absorption band, and scattering dominated again (as shown in Fig. 4(a)). These spectral features form the basis for analyzing soil moisture, mineral composition, and particle characteristics, laying the foundation for soil nutrient inversion methods using near-infrared data acquisition.



（a）Original Spectral Reflectance of Soil

（b） Spectral curve after SG preprocessing

（c） Spectral curve after MSC Preprocessing

（d） Spectral curve after SNV preprocessing

（e） Spectral curve after MA preprocessing

（f）Spectral curve after FD preprocessing

Fig. 4. Comparison of original and pretreatment results of soil spectral reflectance

3.3 Comparative Analysis of Preprocessed Spectral Curves

This section presents a comparative analysis of the effects of various preprocessing methods that is: SG, MSC, SNV, MA, and FD on soil near-infrared spectral signatures. These methods enhance spectral signal quality by suppressing noise, correcting scatter, or enhancing features, thereby providing a more reliable data foundation for the inversion modeling of nitrogen (N), phosphorus (P), and potassium (K). The following analysis systematically examines the spectral characteristics of each preprocessing method and their differential effects on nutrient-relevant absorption bands.Fig. 4(b) demonstrates that SG preprocessing generates substantially smoothed spectral signatures while preserving and enhancing the characteristic water absorption feature at approximately 1400 nm, effectively highlighting the spectral signals associated with OH functional groups and clay minerals within the soil matrix.Fig.

4(c)illustrates that MSC preprocessing effectively mitigates scattering effects, substantially accentuating the chemical compositional information within the 1400-1700 nm spectral domain and enhancing the comparative analysis of available nitrogen-associated spectral features.Fig. 4(d)demonstrates that SNV preprocessing substantially reduces spectral amplitude fluctuations across the complete wavelength domain, effectively standardizing the water and mineral absorption features within the 1200-1400 nm region while accentuating nutrient-relevant spectral signals.Fig. 4(e) shows that MA preprocessing effectively suppresses high-frequency noise components, substantially enhancing the smoothness of the absorption feature at approximately 1400 nm and accentuating the subtle spectral response characteristics associated with available potassium.Fig. 4(f)illustrates that FD preprocessing substantially amplifies peak-valley characteristics at approximately 1400 nm and 1600 nm, effectively highlighting subtle absorption variations associated with soil moisture and organic matter components, although potentially introducing noise artifacts.

The five preprocessing methods demonstrate distinct optimization effects on spectral signatures: SG and MA effectively suppress noise, enhancing water and nutrient absorption features around 1400 nm; MSC and SNV improve chemical compositional analysis in the 1400-1700 nm range through scatter correction and standardization; FD amplifies peak-valley characteristics but may introduce noise artifacts, limiting its applicability. These preprocessing effects form the basis for enhancing random forest (RF) model performance, especially in improving the prediction accuracy of available N and P. Future studies may explore combinatorial preprocessing approaches to better capture the subtle spectral response related to available K.

3.4 Comparative Analysis of Inversion Models

Spectral data were used to independently implement RF, SVM, and PLSR models for the inversion of available N, P, and K. As shown in Table 2, RF models consistently outperformed both SVM and PLSR. Without preprocessing, RF models showed superior performance metrics for available N (R^2=0.539, RMSE=21.408 mg/kg, RPD=1.490), available P (R^2=0.536, RMSE=25.056 mg/kg, RPD=1.484), and available K (R^2=0.429, RMSE=42.452 mg/kg, RPD=1.338). These results confirm that RF models excel at capturing non-linear relationships in spectroscopic data while maintaining strong performance despite noise. Although SVMs generally perform well with high-dimensional, limited-observation datasets, their underperformance here may be due to suboptimal kernel function parameter selection. As a linear approach, PLSR showed lower predictive accuracy, likely due to the complex non-linear soil-spectral relationships in this system. These findings align with Zhou et al. (2020), who showed that ensemble learning methods generally outperform traditional regression approaches in handling complex spectroscopic datasets.

Table 2. Reconstruction results of soil available nitrogen content using three models

Model	Training Set			Testing Set		
	R^2	RMSE(mg/kg)	RPD	R^2	RMSE(mg)	RPD
SVM	0.698	18.205	1.825	0.589	20.216	1.578
PLSR	0.480	23.891	1.391	0.451	23.365	1.365
RF	0.728	17.295	1.921	0.539	21.408	1.490

Table 3. Reconstruction results of soil available phosphorus content using three models

Model	Training Set			Testing Set		
	R^2	RMSE(mg/kg)	RPD	R^2	RMSE(mg)	RPD
SVM	0.513	27.679	1.437	0.405	28.380	1.311
PLSR	0.392	30.923	1.286	0.188	33.137	1.122
RF	0.751	19.803	2.008	0.536	25.056	1.484

Table 4. Reconstruction results of soil available potassium content using three models

Model	Training Set			Testing Set		
	R^2	RMSE(mg/kg)	RPD	R^2	RMSE(mg/kg)	RPD
SVM	0.259	48.988	1.165	0.307	46.768	1.215
PLSR	0.431	42.941	1.329	0.198	50.299	1.129
RF	0.648	33.740	1.691	0.429	42.452	1.338

3.5 Impact of Preprocessing Methods on Model Performance

To eliminate spectral noise and other distortions, MSC, SG, MA, FD, and SNV preprocessing methods were applied to the spectral dataset, followed by the independent construction of inversion models for available nitrogen (N), phosphorus (P), and potassium (K) using the random forest (RF) algorithm. The comparative evaluation results for the preprocessing methods are presented in Table 3. MSC preprocessing yielded optimal performance for predicting available N, with metrics of R^2=0.585, RMSE=20.326 mg/kg, and RPD=1.569, outperforming other methods and unprocessed models (R^2 improvement of 0.02-0.05, RMSE reduction of about 5%). SG smoothing achieved optimal performance for predicting available P, with metrics of R^2=0.553, RMSE=24.585 mg/kg, and RPD=1.513. MA preprocessing showed optimal performance for predicting available K, with metrics of R^2=0.431, RMSE=42.383 mg/kg, and RPD=1.340. FD preprocessing consistently underperformed in all nutrient prediction models (available N: R^2=0.398, RMSE=24.468 mg/kg; available P: R^2=0.323, RMSE=30.261 mg/kg; available K: R^2=0.381, RMSE=44.178 mg/kg).

The results show that suitable preprocessing methods significantly improve model prediction accuracy. MSC enhances available N prediction by correcting scatter interference, consistent with Luo et al. (2019); SG smoothing optimization improves available P prediction, indicating its spectral response is particularly vulnerable to noise; and MA enhances available K prediction by effectively suppressing random noise. The suboptimal performance of FD preprocessing likely results from first-derivative operations amplifying noise artifacts and reducing signal-to-noise ratios. Optimal preprocessing methods vary across nutrient elements: MSC performs best for available N, SG for available P, and MA for available K. This differentiation reflects the diverse states and spectral response mechanisms of nutrient elements in soil matrices. Compared to unprocessed modelling approaches, preprocessing improved R^2 values by 0.02-0.05 and reduced RMSE by about 5%. Although these improvements are modest, they can lead to significant economic and environmental benefits in precise agriculture. This study highlights the complexity of soil nutrient spectral responses and the importance of optimizing preprocessing methods. Future research should consider: 1) extending the spectral range into visible and mid-infrared regions to capture signals for available P and K; 2) expanding sample size to improve model generalization; and 3) exploring combined preprocessing approaches to optimize inversion accuracy.

Table 5. Comparison results of several pretreatment methods of different models of available potassium

	N			P			K		
	RMSE	R^2	RPD	RMSE	R^2	RPD	RMSE	R^2	RPD
SG	21.480	0.539	1.489	24.585	0.553	1.513	42.916	0.416	1.323
MSC	20.326	0.585	1.569	29.592	0.353	1.257	42.619	0.424	1.333
SNV	22.872	0.474	1.394	29.955	0.337	1.242	43.757	0.393	1.298
MA	21.8	0.522	1.463	24.66	0.55	1.508	42.383	0.431	1.34
FD	24.468	0.398	1.303	30.261	0.323	1.229	44.178	0.381	1.286

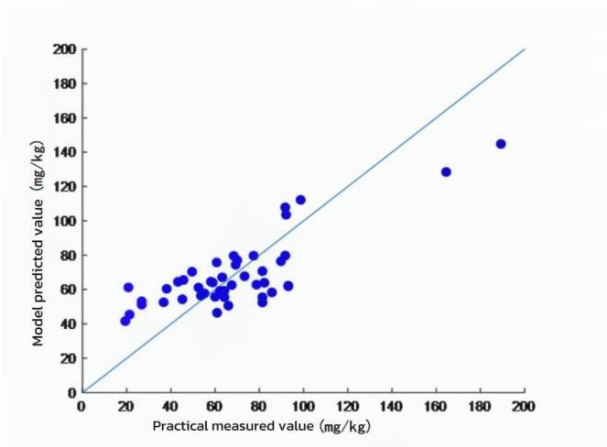


Fig. 5. Residual prediction of available soil nitrogen using a random forest model

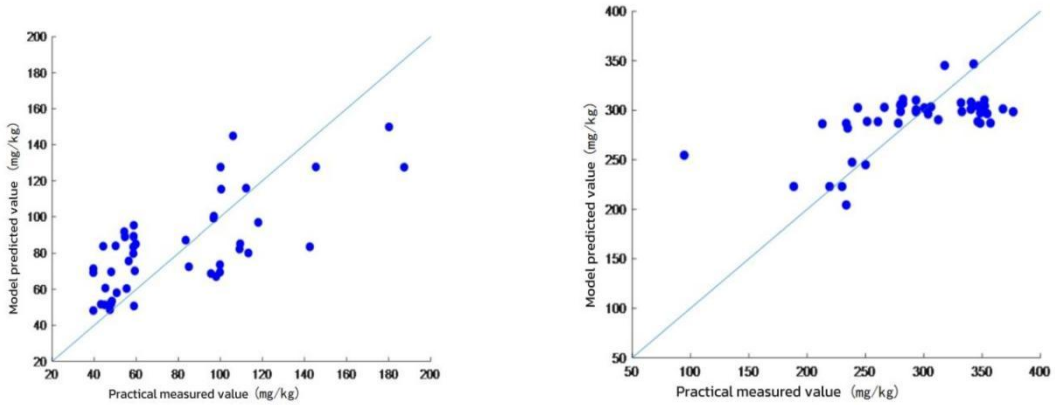


Fig. 6. Residual prediction of available soil phosphorus using a random forest model

Fig. 7. Residual diagram of random forest model prediction for soil available potassium

4 Conclusions

This study focused on orchard systems in Haizhu District, Guangzhou, using NIRS technology to analyze the near-infrared spectral signatures of soil samples for available nitrogen (N), phosphorus (P), and potassium (K), while constructing optimized inversion models. Results showed that RF-constructed inversion models outperformed SVM and PLSR methods in predictive accuracy, with available nitrogen (N) parameters of R^2=0.539, RMSE=21.408 mg/kg, RPD=1.490; available phosphorus (P) parameters of R^2=0.536, RMSE=25.056 mg/kg, RPD=1.484; and available potassium (K) parameters of R^2=0.429, RMSE=42.452 mg/kg, RPD=1.338. A comparative analysis of SG, MSC, SNV, MA, and FD preprocessing methods revealed that MSC preprocessing performed best for available N prediction (R^2=0.585, RMSE=20.326 mg/kg, RPD=1.569), SG preprocessing excelled in available P prediction (R^2=0.553, RMSE=24.585 mg/kg, RPD=1.513) and MA preprocessing showed the highest performance for available K prediction (R^2=0.431, RMSE=42.383 mg/kg, RPD=1.340). In conclusion, NIRS-based inversion modeling exhibited the highest accuracy for available N, followed by available P, while available K showed comparatively lower accuracy.

This study confirms the effectiveness of integrating NIRS with RF and optimized preprocessing methods for soil nutrient inversion, significantly supporting precision fertilization in orchard systems. However, the reduced predictive accuracy for available K may be due to its weak spectral response and limitations in laboratory data acquisition. Future research could focus on extending spectral bandwidth, implementing field measurements, integrating deep learning methods and developing portable devices to enhance the application of NIRS in precision agriculture.

Acknowledgement

1. The authors acknowledge support from the Natural Science Foundation of Guangdong Province, China (Grant No. 2021A1515010824).
2. This paper is by supported the teaching reform project of "Seed Testing" based on "student-centered" approach and "micro-video" technology (KA25YY081).

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Chen, X. H., Wang, W. (2019) Recent advances in rapid detection techniques for soil nutrients. Acta Pedologica Sinica., 56:521-530.
2. Du, X., Chen, H., Xie, J., Li, L., Cai, K., Meng, F. (2025) Quantitative analysis of soil potassium by near-infrared (NIR) spectroscopy combined with a three-step progressive hybrid variable selection strategy. Spectrochim. Acta. A, 324:124998. https://doi.org/10.1016/j.saa.2024.124998.
3. Endut, R., Sabri, M. S. A., Aljunid, S. A., Ali, N., Laili, A. R., Laili, M. H. (2023) Prediction of potassium (K) content in soil analysis utilizing near-infrared (NIR) spectroscopy. J. Adv. Res. Appl. Sci. Eng. Technol., 33:92-101. https://doi.org/10.37934/araset.33.1.92101.
4. Jia, S., Yang, X., Li, G., Zhang, J. (2015) Quantitatively determination of available phosphorus and available potassium in soil by near infrared spectroscopy combining with recursive partial least squares. Guang pu xue yu Guang pu fen xi= Guang pu, 35:2516-2520.
5. Jin, X., Li, S., Zhang, W., Zhu, J., Sun, J. (2020) Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. Appl. Sci., 10:1520. https://doi.org/10.3390/app10041520.
6. Kone, Y. J., Sogoba, B., Dembele, L., Ballo, A. Traore, K. (2018) Capability of visible-near infrared spectroscopy in estimating soils chemical properties in Mali. Open J. Soil Sci., 8:185-198.
7. Lin, F., Zhang, W. (2018) Application of near-infrared spectroscopy in rapid detection of soil nutrients. Acta Pedologica Sinica., 55:1057-1068.
8. Liu, Y., Zheng, X., Wang, Y., Cao, Z., Li, Y., Wu, W., Liu, Z., Liu, H., Li, R. (2018) Land consolidation and modern agriculture: From soil particles to agricultural systems: A case study of Yulin city, Shaanxi province. J. Geogr., 28:1896-1906. https://doi.org/10.1007/s11442-018-1570-1.
9. Luo, J., Wang, Y., Lou, W., Zhou, X., Tian, Y. (2020) Rapid, nondestructive and simultaneous predictions of soil content in Wuling mountain area using near infrared spectroscopy. Appl. Ecol. Environ. Res., 18:889-899. https://doi.org/10.15666/aeer/1801_889899.
10. Maleki, M., Van Holm, L., Ramon, H., Merckx, R., De Baerdemaeker, J., Mouazen, A. (2006) Phosphorus sensing for fresh soils using visible and near infrared spectroscopy. Biosyst. Eng., 95:425-436. https://doi.org/10.1016/j.biosystemseng.2006.07.015.
11. Rossel, R. V., Walvoort, D., McBratney, A., Janik, L. J., Skjemstad, J. (2006) Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 131:59-75. https://doi.org/10.1016/j.geoderma.2005.03.007.
12. Saberioon, M., Gholizadeh, A., Ghaznavi, A., Chabrillat, S., Khosravi, V. (2024) Enhancing soil organic carbon prediction of LUCAS soil database using deep learning and deep feature selection. Comput. Electron. Agric., 227:109494. https://doi.org/10.1016/j.compag.2024.109494.
13. Shao, Y., He, Y., Liu, W., Zhang, L. Wu, G. (2011) Application of near-infrared spectroscopy to predict soil nitrogen content in Zhejiang Province, China. J. Zhejiang Univ., 12:73-79.
14. Shepherd, K. D., Walsh, M. G. (2010) Near infrared spectroscopy for rapid soil analysis. Soil Sci. Soc. Am. J., 74:135-142.
15. Shibusawa, S., Anom, S. W. Sato, H. P. (2012) Spectral analysis for soil nutrient detection. Precis. Agric., 13:400-415.
16. Wang, Y., Li, M., Ji, R., Wang, M., Zheng, L. (2020) Comparison of soil total nitrogen content prediction models based on Vis-NIR spectroscopy. Sensors, 20:7078. https://doi.org/10.3390/s20247078.
17. Zhang, J. J., Li, Y. N. (2019) Research on influencing factors of land value-added income distribution in agricultural land non-agriculturalization—based on interpretative structural model. China Real Estate., 15:10-17.doi:10.13562/j.china.real.estate.2019.15.003.
18. Zhang, X., Wen, J., Zhao, D. (2010) Band selection method for retrieving soil lead content with hyperspectral remote sensing data. In: Earth Resources and Environmental Remote Sensing/GIS Applications. Washington, PP.377-383. https://doi.org/10.1117/12.864425.
19. Zhao, W., Wu, Z., Yin, Z., Li, D. (2023) Reducing moisture effects on soil organic carbon content estimation in vis-NIR spectra with a deep learning algorithm. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens., 16:7733-7748. https://doi.org/10.1109/JSTARS.2023.3287583.
20. Zheng, L. H., Mang, Y. T., Song, T. Zhang, J. C. (2005) Near infrared spectroscopy for rapid soil analysis in Zhejiang soils. J. Soil Sci., 45:200-210.
21. Zhou, Y. F., Li, Q. (2022) Current situation and development trend of soil nutrient detection technology. Soil Science., 57:521-530.

Biographies

1. **Sun Rongguang** graduated from Zhongkai University of Agriculture and Engineering.

2. **Hou Chaojun** is affiliated with Zhongkai University of Agriculture and Engineering, holding the academic title of Associate Professor. His major achievements are as follows:
   Chaojun Hou, Jiajun Zhuang, Yu Tang*, Yong He, Aimin Miao, Huasheng Huang, Shaoming Luo. "Recognition of Early Blight and Late Blight Diseases on Potato Leaves Based on Graph Cut Segmentation" [J]. Journal of Agriculture and Food Research, 2021, 5: 100154.
   C.J. Hou, Y. Tang*, S.M. Luo, J.T. Lin, Y. He, J.J. Zhuang, W.F. Huang. "Optimization of Control Parameters of Droplet Density in Citrus Trees Using UAVs and the Taguchi Method" [J]. International Journal of Agricultural and Biological Engineering, 2019, 12(4): 1–9
3. **Cui Huawei** is affiliated with Zhongkai University of Agriculture and Engineering, holding the academic title of Lecturer. His major achievements include "Prediction of Maize Seed Vigor Based on First-Order Difference Characteristics of Hyperspectral Data" (first author, published in Agronomy in 2022) and "Locality-Preserving Data Modelling and Its Application in Fault Classification" (co-corresponding author, published in The Canadian Journal of Chemical Engineering in 2021).

基於近紅外光譜的果園土壤養分含量反演模型研究

孫榮光[1]，侯超鈞[1]，崔華威[1]

[1] 仲愷農業工程學院，廣州，中國，510225

摘要：精準測定土壤速效氮（N）、磷（P）、鉀（K）含量對果園施肥管理至關重要。傳統化學分析方法耗時、成本高，難以滿足現代農業快速檢測需求。近紅外光譜（NIR）技術以其快速檢測的特點為土壤養分測定提供了新途徑。本研究基於 NIR 光譜數據，採用偏最小二乘回歸（PLSR）、支援向量機（SVM）和隨機森林（RF）構建土壤速效 N、P、K 含量反演模型。為了消除光譜數據採集的干擾因素，提高反演模型性能，分別應用了 Savitzky-Golay平滑（SG）、多元散射校正（MSC）、標準正態變量變換（SNV）、移動平均（MA）和一階導數（FD）等光譜預處理方法。結果表明，RF 模型在三種模型中的反演精度最高，速效氮的預測精度為 R^2=0.5390，RMSE=21.4083 mg/kg，RPD=1.4901；速效磷的預測精度為 R^2=0.5361，RMSE=25.0564 mg/kg，RPD=1.4841；速效鉀的預測精度為 R^2=0.4293，RMSE=35.0889 mg/kg，RPD=1.3383。通過對比不同預處理方法，其中 MSC 預處理方法對速效N 預測精度最高（R^2=0.585，RMSE=20.326 mg/kg，RPD=1.569）；SG 預處理方法對速效 P 的精度最高（R^2=0.553，RMSE=24.585 mg/kg，RPD=1.513）；MA 預處理方法對速效 K 的反演精度最高（R^2=0.431，RMSE=42.383 mg/kg，RPD=1.340）。相比無預處理方法，模型反演性能得到改善，具體表現為：速效氮 R^2提升 8.53%，速效磷提升 3.17%，速效鉀提升 0.47%。其中速效氮和磷的預測精度顯著提高，而速效鉀因光譜響應較弱，優化效果有限。研究證實，NIR 光譜結合適當預處理方法可有效預測土壤養分含量，為果園精準施肥提供技術支援，具有重要的應用價值。

關鍵詞：土壤；近紅外光譜；速效養分反演；特徵選擇；隨機森林

1. 孫榮光，畢業於仲愷農業工程學院；
2. 侯超鈞，目前在仲愷農業工程學院工作，副教授，以第一作者與通訊作者發表學術論文30余篇，其中第一作者發表SCI論文4篇，共同第一作者發表SCI論文5篇。授權實用新型專利10余件，登記計算機軟件著作權20余項；
3. 崔華威，目前在仲愷農業工程學院工作，講師。