JARS Journal of Applied Remote Sensing

RESEARCH PAPER

Hyperspectral estimation for nitrogen and phosphorus content in *Camellia oleifera* leaves based on machine learning algorithms

Xuehai Tang,[a,b,*] Fan Kuang,[a,b] Genshen Fu,[a,b] Lipeng Yan,[a,b]
Qingfeng Huang,[a,b] Xinwen Wang,[c] Bin Wang,[a,b] and Qiangxin Ou[a,b]
[a]Anhui Agricultural University, School of Forestry and Landscape Architecture, Hefei, China
[b]Anhui Province Key Laboratory of Forest Resources and Silviculture, Hefei, China
[c]Huangshan Xian'ge Ecotourism Development Co. Ltd., Huangshan, China

ABSTRACT. Nitrogen and phosphorus are essential elements of plants, which play important roles in representing plant growth, physiological function regulation, fruit harvest, etc. Hyperspectral technology provides a nondestructive, rapid, highly accurate, and cost-efficient method for plant leaf nutrient content estimation. There are very limited studies on nutrient diagnosis of *Camellia oleifera* leaves using hyperspectral technology. In this work, 160 *Camellia oleifera* samples were used. Hyperspectral data were obtained using a full-band spectrometer. On the basis of preprocessing, the spectral response characteristics of leaf nitrogen content (LNC) and leaf phosphorus content (LPC) were revealed by comparing different combinations of spectral indices, and the spectral variables were further selected. The optimal LNC and LPC estimation models based on three machine learning algorithms [i.e., support vector machine (SVM), random forest (RF), and back propagation neural network (BPNN)] were constructed. The results showed that the spectral sensitive regions of leaf nitrogen and phosphorus content were mainly reflected in green band, followed by red band and the long-wave direction of short-wave infrared band. Savitzky–Golay first derivative (SGFD) pretreatment method was generally better than multiplicative scatter correction. The maximum correlation coefficients of the absolute values of LNC, LPC, and spectral transformation features were 0.56 and 0.49. The optimal LNC and LPC models were both SGFD-TBNDSI-BPNN, with R^2 of 0.81 and 0.79, and RMSEP of 0.55 and 0.06 g/kg, respectively. The research results can provide a reliable theoretical basis for large-scale optical remote sensing monitoring of nutrient content for *Camellia oleifera*.

[DOI: 10.1117/1.JRS.18.038502]

Keywords: ground hyperspectral technology; *Camellia oleifera*; nutrient content; machine learning; feature transformation

Paper 240008G received Jan. 2, 2024; revised Jun. 14, 2024; accepted Jul. 10, 2024; published Aug. 7, 2024.

1 Introduction

Nitrogen and phosphorus are essential elements for plant absorption, transport and assimilation from soil, maintaining all stages of growth and development.[1] Improving soil nutrient structure by adding nitrogen fertilizer, phosphate fertilizer, and compound fertilizer can create a good survival foundation for plants with fast growth, high quality, high yield, and stable yield. The traditional way of soil nutrient determination and balance (i.e., soil testing formula) is to understand the physical and chemical properties of the soil at the root of the plant through

[*] Address all correspondence to Xuehai Tang, tangxuehai@ahau.edu.cn

soil sampling and chemical analysis, so as to guide the supply balance of soil nutrients.[2] Although such technology has made significant contributions to alleviating the contradiction between plant fertilizer demand and soil fertilizer supply for a period of time, it is still impossible to determine the actual amount of nitrogen and phosphorus that plants actually absorb from soil. In plant physiology, the content of nitrogen and phosphorus in plants represents the real nutritional level of plants in real time, and leaves are the most sensitive organs to plant physiological and metabolic changes.[3,4] Therefore, the monitoring of nitrogen and phosphorus content in leaves has always been hot topics in long-term research both at home and abroad, and the quantitative diagnosis of nutrient profit and loss status in different growth stages of plants is the focus.[5] However, rigorous laboratory testing conditions, expensive analytical instruments, and technical bottlenecks, which are difficult to break through have greatly increased the difficulty of extending destructive monitoring to the determination of nitrogen and phosphorus content in leaves at large regional scales.

Nitrogen is involved in regulating plant photosynthesis. Nitrogen deficiency will have a serious impact on plant growth and yield, while excessive consumption of nitrogen fertilizer can lead to a series of environmental problems. Phosphorus content in plant leaves is second only to nitrogen, but it also plays a crucial role in plant growth and development. Compared with the research breadth and depth of nitrogen in the field of spectral analysis, there are relatively few reports on phosphorus. Phosphorus deficiency is easy to produce anthocyanins and change chlorophyll content and significantly change the arrangement of leaf cell structure and affect the synthesis of organic matter. Under the combined action of various factors, the change rule of phosphorus spectral characteristic curve is more complicated.[6,7] The spectral changes of different varieties of crops during the phenological period are generally similar, and the spectral reflectance decreases in the visible light band and increases in the near-infrared band with the increase of nitrogen application rate or the increase of leaf nitrogen content (LNC).[8,9] Due to the strong mobility of nitrogen, the spectral response of nitrogen in horizontal or vertical spatial heterogeneous distribution is different.[10,11] Nitrogen in leaf cells mostly exists in the form of protein, and some of it synthesizes chlorophyll. Under certain radiation levels, it causes the absorption and reflection of photosynthetic nitrogen or nonphotosynthetic nitrogen at specific wavelength positions and then produces significant differences in spectral reflectance.[12] Based on the highly correlated characteristics of nitrogen and chlorophyll content, many current studies focus on the visible and near-infrared short-wave ranges, ignoring the protein or nonphotosynthetic nitrogen components that respond to the near-infrared long-wave.[13] Ramoelo et al.[14] used *in situ* hyperspectral and environmental data to estimate the phosphorus content of grass and found that the sensitive regions were mostly located in the near-infrared band, which were susceptible to interference from proteins, sugars, and starches, showing low correlation similar to the visible light band. Guo et al.[15] found that 31 characteristic wavelengths can be selected in the range of $350 \sim 2500$ nm when monitoring the phosphorus content in rubber tree leaves, of which about 70% are distributed in the near-infrared band. Li et al. used *in situ* canopy spectra to monitor nitrogen and phosphorus in winter rapeseed, and obtained wavelengths related to phosphorus at 755, 832, 891, 999, 1196, and 1267 nm. They also pointed out that the spectral absorption characteristics of these regions may affect the final estimation accuracy due to differences in canopy or cell physical morphology.[6] Although the response difference in spectral response of leaf phosphorus has been confirmed to be related to internal biochemical components and cell structure, the previous research conclusions are mostly limited to short-term survey data. The effects of genetic conditions, external environment, and human interference cannot be completely excluded. Therefore, the use of spectral data to estimate leaf phosphorus content still has a large research space.

How to efficiently extract weak information of interest from redundant and complex hyperspectral data has always been a research difficulty, which has prompted researchers to start in-depth research on multivariate quantitative analysis techniques for weak hyperspectral signals in various fields. The process of hyperspectral data mining is essentially to solve typical high-dimensional problems, especially when the number of samples is far less than the dimension of spectral data, it may face "curse of dimensionality," which indirectly leads to the reduction of analysis accuracy. Therefore, to minimize or eliminate the influence of nontarget elements, it is very important to select appropriate spectral preprocessing and feature transformation to improve signal sensitivity when constructing hyperspectral estimation model. At present, the relevant processing methods are divided into two categories: spectral processing for spectral arrays and spectral array data processing combined with concentration array information.[16–20] Although so many preprocessing methods have been developed, there is still no preprocessing that can guarantee the complete removal of all irrelevant information independently. In addition, scholars have also fully learned from the successful experience of vegetation index in multispectral analysis and gradually used spectral index for hyperspectral analysis. By combining linear or nonlinear methods, potential effective information is mined from the full-band hyperspectral response characteristics of extremely narrow continuous intervals, and nonvegetation or irrelevant information is minimized, representing ratio nitrogen index, normalized area index and canopy chlorophyll concentration index.[21–25]

The objectives were: (1) to compare the accuracy between multivariate scattering correction (MSC) and Savitzky–Golay first derivative (SGFD) for spectral preprocessing; (2) to reveal sensitive bands or combinations of spectral response of leaf nitrogen and phosphorus content basis on two-band and three-band spectral indices; and (3) to construct the optimal estimation models of LNC and leaf phosphorus content (LPC) based on machine learning algorithms [support vector machine (SVM), random forest (RF), back propagation neural network (BPNN), etc.].

2 Materials and Methods

2.1 Overview of the Study Area

This paper takes the planting base of *Camellia oleifera* in Xian’ge Village, Xianyuan Town, Huangshan District, Huangshan City, Anhui Province, China, as the research scope. The coordinates of the central point are 118°10′13.448″E, 30°19′49.421″N. The whole area is located on the sunny slope, and the slope surface is inclined downward from northwest to southeast in the form of terraces. The maximum height difference between the upper and lower slopes is more than 50 m, and the average altitude is 290 m. The study area is located in the northern part of Huangshan District, facing adjacent to Taiping Lake Scenic Area in the north and Huangshan Mountain Scenic Area in the south. The climate characteristics are obvious, the four seasons are quite distinct, and the rain and heat are in the same period, which belongs to the typical subtropical monsoon humid climate.

2.2 Sample Tree Selection

From October 28 to November 2, 2021, 160 *Camellia oleifera* trees were selected as the research subjects in the study area, of which 120 trees were randomly sampled and 40 plants were sampled centrally. To reduce the blindness of ground sampling, this study uses an UAV to take low-altitude aerial photography of the study area to obtain high-resolution digital orthophoto map (DOM), and then uses ArcMap software to generate a fishnet map (3 m × 3 m) matching the DOM range of the study area. Thirty fishnet grids were randomly selected based on the spatial distribution of the *Camellia oleifera* trees, in which four plants were selected (i.e., four plants per grid × 30 grids = 120 plants). In addition, we set up a 15 m × 35 m square sample plot with 40 plants. Thus, a total of 160 plants were selected as research objects. All the *Camellia oleifera* trees tested in this study are Changlin series (i.e., Changlin No. 27, Changlin No. 40, and Changlin No. 53), which were planted in 2012. The investigation and fruit harvesting were carried out during the fruit ripening stage. There were no other operational measures, such as watering, fertilizing, dosing, etc., during the first seven months before investigation, except for necessary weeding.

2.3 Data Acquisition

A full-band spectrometer (FieldSpec4 Wide-Res, Analytical Spectrum Devices Inc., United States) was used to collect the full-band (i.e., $350 \sim 2500$ nm) hyperspectral data. The armored optical fiber of the spectrometer was connected to the optical fiber extension cable through a fiber adapter and then the probe was lifted to 2.0 m above the center of the trees’ canopies using a custom bracket. Thus, a cone-shaped detection space with a diameter of more than 1.8 m is formed from the surface, and this detection field basically covers most of the top and middle layers of the canopy of a single tree. The measured hyperspectral data do not include background noise, such as soil and tree trunks.

In the outdoor environment, external factors, such as light intensity, cloud cover, wind speed, air temperature, and humidity, will affect the quality of hyperspectral data. Therefore, to minimize the interference of external factors, this experiment chose to carry out measurement operations under sunny, windless and cloudless weather conditions. Generally, hyperspectral data are obtained between 10 a.m. and 2 p.m. in Beijing time. Before collecting the spectra of each tree, one or more corrections were performed using a reference whiteboard with a reflectivity close to 100% in full band range. Finally, to reduce the random measurement error, 10 spectra were continuously collected for each single tree.

2.4 Determination of Nitrogen and Phosphorus Content in Leaves

According to the relevant literature,[26–28] in this experiment, a single plant of *Camellia oleifera* was harvested in the upper and middle parts of the canopy and scattered in 4 directions, with two leaves in each direction, totaling 16 leaves (mixed sampling of new and old leaves, 160 samples × 16 leaves in total). After picking the leaves, immediately put them into an envelope bag and take them back to the laboratory. Place them in an air-drying oven and sterilize them at 105°C for 30 min. Then, dry them at 60°C until they reach a constant weight. After drying, grind through a 60-mesh sieve, and the screened oil tea leaf powder is used for machine measurement of LNC on the packaging sample, while a portion is used for digestion and boiling to measure LPC. Nitric acid-perchloric acid mixing method was used for plant sample digestion. LNC was measured using an element analyzer (EA3000, Euro Vector, Inc., Italy), while LPC was measured using a continuous flow analyzer (AA3, SEAL Analytical, Inc., Germany). LNC and LPC corresponded to the contents of total nitrogen (TN) and total phosphorus (TP) in leaves, respectively, and the unit was converted into g/kg.

2.5 Data Preprocessing

2.5.1 *Hyperspectral data preprocessing*

Before preprocessing, it is necessary to visually select and delete the outliers within the 10 hyperspectral reflectance curves repeatedly collected from a single plant and then calculate the average reflectance of the remaining hyperspectral curves.

Elimination of water vapor bands. The absorption characteristics of water vapor to solar radiation span the entire ultraviolet-visible-near infrared-short wavelength infrared (SWIR) spectral range and change with time and space. The absorption intensity near 1400, 1800, and 2500 nm is almost 100%. Since the measurement process was carried out outdoors, the transpiration of *Camellia oleifera* leaves will increase with the increase of solar elevation angle, and a large amount of water vapor will evaporate upward in the canopies, which makes the reception signal of the spectrometer very weak in these three places, so it is very easy to be affected by random noise and produce large fluctuations. Therefore, according to the collected hyperspectral reflectance curve of *Camellia oleifera*, the abnormal range of water vapor absorption can be roughly judged. The band reflectance in the range of 1351 ~ 1440, 1796 ~ 2025, and 2331 ~ 2500 nm is directly removed, and the original spectra at this time is recorded as R.

Multiplicative scatter correction. Multiplicative scatter correction (MSC) is a spectral preprocessing method to deal with the influence of diffuse reflection and optical path change of rough surfaces. It corrects the baseline translation and offset of each spectral data by establishing an "ideal spectrum."[29] However, the average value of all spectral data is approximately replaced by the "ideal spectrum," since it cannot be directly collected. Equation (1) describes the calculation process of MSC. By performing a unary linear regression between the canopy spectrum of a single tree and the "ideal spectrum," the specific intercept and slope are obtained, that is, the baseline translation and offset corresponding to each spectrum. Then subtract the obtained intercept and divide it by the slope, and finally obtain the spectrum after scattering correction

$$\mathrm{MSC}_i = \frac{(R_i - b_i)}{k_i}, \tag{1}$$

where R_i is the i'th original spectral reflectance and b_i and k_i are the corresponding baseline translation and offset, respectively.

Savitzky–Golay first-order derivative. SGFD is a spectral preprocessing method for resolving absorption overlapping peaks, improving resolution, correcting baseline offset, and eliminating background noise.[30] It combines the characteristics of Savitzky–Golay convolution smoothing filter to eliminate weak high-frequency noise with the ability of the first-order derivative to correct the spectral baseline drift and has three variable parameters (i.e., polynomial order, smoothing window size, and derivative order) [Eq. (2)]. Different combinations of variable parameters can also achieve flexible changes in correction effects.[31] SGFD makes up for the lack of noise suppression ability of direct difference derivation and realizes the polynomial correction and first-order derivative transformation of spectral reflectance based on least squares method in a smooth window and then realizes the correction of full-band spectra by moving the window. When using this method, the combination of the three variable parameters is very important. For instance, if the width of the smoothing window is too large, the detail information will get lost, and if the window is too small, the noise will not be weakened. In this paper, the polynomial degree is set to two, and the smoothing window size is five

$$y_i' = \sum_{j=-m}^{j=m} \frac{C_j y_{i+j}}{N}, \tag{2}$$

where y_i' is the new value after the first derivative of Savitzky–Golay convolution, C_j is the convolution weight determined by the polynomial order and the size of $2m+1$ moving window, N is the normalization factor, and y_{i+j} is the measured original spectral reflectance.

2.6 Multivariate Spectral Index Method

Spectral index is a new index that combines the spectral reflectance of different bands through simple linear or nonlinear algebraic operations. There are three common algebraic operations: difference, ratio, and normalized value.[32]

2.6.1 *Two-band spectral indices*

The simple combination of reflectance values of a few bands can effectively enhance the linear measurement of hyperspectral features on the physical-chemical parameters of ground objects and weaken the influence of errors and uncertainties caused by differences of background factors. The spectral index of the two-band combination expands the spectral feature space of the one-dimensional wavelength index range to the two-dimensional index scale and fully combines the correlation between the spectra. In this part, three combinations of two-band reflectance values [i.e., difference spectral index (DSI), ratio spectral index (RSI), and normalized difference spectral index (NDSI)] are used, as shown in Eqs. (3) to (5). To reduce computational cost, a 5-nanometer resampling interval is set before the calculation

$$\mathrm{DSI}(i,j) = R_i - R_j, \tag{3}$$

$$\mathrm{RSI}(i,j) = \frac{R_i}{R_j}, \tag{4}$$

$$\mathrm{NDSI}(i,j) = \frac{R_i - R_j}{R_i + R_j}, \tag{5}$$

where R_i and R_j are the reflectance values at i nm and j nm, respectively.

2.6.2 *Three-band spectral indices*

The three-band spectral index introduces a new wavelength index dimension, and in this section, we still use three combinations of reflectance values [i.e., three-band difference spectral index

(TBDSI), three-band radio spectral index (TBRSI), and three-band normalized difference spectral index (TBNDSI)], as shown in Eqs. (6)–(8). To reduce computational cost, a 10-nm resampling interval is set before the calculation

$$\mathrm{TBDSI}(i,j,k)=(R_i-R_j)+(R_k-R_j), \tag{6}$$

$$\mathrm{TBRSI}(i,j,k)=\frac{R_i}{R_j+R_k}, \tag{7}$$

$$\mathrm{TBNDSI}(i,j,k)=\frac{R_i-R_j}{R_i+R_k}, \tag{8}$$

where R_i, R_j, and R_k are the reflectance values at i, j, and k nm, respectively.

2.7 Machine Learning Model Construction and Evaluation

The independent nitrogen and phosphorus estimation model dataset is constructed by combining the spectral characteristics of *Camellia oleifera* and the nitrogen and phosphorus content parameters of leaves, and the rank sample set partitioning based on joint X-Y distances (RANK-SPXY) algorithm is used to divide the calibration set and the prediction set at a fixed ratio (7:3). SVM, RF, and BPNN algorithms are used to train and construct models using the calibration set. The generalization ability of the nitrogen and phosphorus content estimation model of *Camellia oleifera* leaves was evaluated by using independent prediction sets to test the applicability of the model.

3 Results and Analysis

3.1 Descriptive Analysis of Nitrogen and Phosphorus Content in *Camellia oleifera* Leaves and Canopy Hyperspectral Reflectance

3.1.1 *Descriptive analysis of nitrogen and phosphorus content in* Camellia oleifera *leaves*

The RANK-SPXY method was used to divide the sample dataset of nitrogen and phosphorus content (i.e., LNC and LPC) of *Camellia oleifera* leaves with a capacity of 160 at according to the ratio of 7:3. The calibration set of 112 and the prediction set of 48 were established, respectively. The descriptive statistical results of the dataset are shown in Table 1. Among the 160 plants, the maximum N content was 15.33 g/kg, the minimum was 8.00 g/kg, the average was 11.15 g/kg, and the standard deviation was 1.31. The maximum P content was 1.12 g/kg, the minimum was 0.52 g/kg, the average was 0.83 g/kg, and the standard deviation was 0.12. We also performed a one-way analysis of variance test, all of which were greater than 0.05, indicating that there was no significant difference. The range of LNC in *Camellia oleifera* leaves reached 7.33 g/kg, and the range of LPC was 0.60 g/kg. From the coefficient of variation, it can be seen that the degree of variation of LPC is greater than that of LNC.

Table 1 Descriptive statistics of datasets and partition results.

Type	Dataset	Size	Mean	Maximum	Minimum	Standard deviation	Skewness	Kurtosis	Coefficient of variation/%
LNC	The whole sample	160	11.15	15.33	8.00	1.31	0.34	0.26	11.78
	Calibration set	112	11.13	15.33	8.00	1.34	0.25	0.13	12.03
	Prediction set	48	11.18	15.03	8.80	1.26	0.62	0.74	11.29
LPC	The whole sample	160	0.83	1.12	0.52	0.12	0.06	−0.31	14.34
	Calibration set	112	0.83	1.12	0.52	0.12	0.05	−0.20	14.26
	Prediction set	48	0.83	1.07	0.58	0.12	0.06	−0.46	14.65

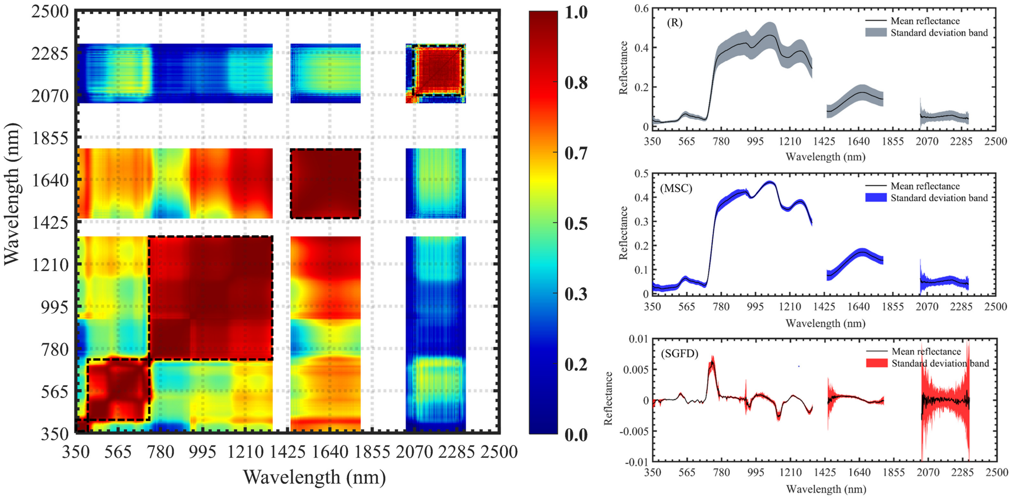
The RANK-SPXY method is used to divide the data set. The results show that the selection range of the correction set of LNC and LPC is well covered and greater than the numerical range of the prediction set. According to the skewness and kurtosis calculation values described in Table 1, the two types of datasets and the division results generally obey the normal distribution. This phenomenon indicates that the RANK-SPXY method can make the calibration set and the prediction set satisfy the homomorphic distribution that is approximately consistent with all the data and further consider the differences in spectral space and target properties and extract representative samples.

3.1.2 *Descriptive analysis of* Camellia oleifera *canopy hyperspectral reflectance*

Figure 1(a) shows the distribution of the Pearson correlation coefficient of hyperspectral reflectance for any adjacent wavelengths across the whole band of the *Camellia oleifera* canopy.

The regions with correlation coefficients above 0.8 are blocky, corresponding to UV, Vis, NIR, and SWIR bands, respectively. There are also small blocky regions with correlation coefficients close to 1 in the spectral region. These strong correlation regions are distributed in a specific wavelength range, with different widths and sizes. It shows that the same band has high redundancy and there is interference between adjacent bands. In addition, the correlation coefficient of different spectral regions in the joint region is generally lower than 0.8, indicating that different spectral regions have certain independence. The mutual interference and the redundancy between different spectral intervals are reduced, and the information load is relatively more obvious.

Figure 1(b) depicts the changes in the spectral curve characteristics of the raw hyperspectral reflectance (R) by MSC and SGFD pretreatments, respectively. Compared with R with typical characteristics of *Camellia oleifera* canopy, MSC pretreatment basically does not change the overall peak trend of its spectral curve, but reduces the standard deviation of reflectance of high reflection platform in the range of 780 ~ 1350 nm, and increases the standard deviation of reflectance in the range of 400 ~ 500 nm. It shows that MSC can weaken the scattering fluctuation caused by difference in leaf cell level or canopy level structure, and change the response characteristics of some Vis bands. Compared with the spectral curve characteristics of R and MSC, the SGFD pretreatment completely changed the original spectral presentation form, and amplifies the peak reflection and trough absorption characteristics of each band to varying degrees.



(a)

(b)

Fig. 1 Canopy hyperspectral reflectance feature analysis. (a) Correlation coefficient distribution diagram of reflectance at different wavelengths. (b) Characteristics of canopy hyperspectral reflectance curve under different pretreatments. Note: *r* is the absolute value of Pearson correlation coefficient.

3.2 Response Relationship Between Canopy Hyperspectral Reflectance and Leaf Nitrogen and Phosphorus Content of *Camellia oleifera*

3.2.1 *Response relationship between canopy hyperspectral reflectance and LNC of* Camellia oleifera

By comparing the response relationship of the average hyperspectral reflectance (R) of *Camellia oleifera* canopy in the five equal intervals of LNC [Fig. 2(a)], the results show that the overall trend of canopy hyperspectral reflectance at different nitrogen levels shows a certain rule. The amount of samples in five equal LNC intervals are 14, 54, 65, 23, and 4, respectively. The specific performance are as follows: with the increase of LNC, the reflectance of the visible spectral region gradually decreases, and the red-edge feature tends to shift towards the long wave direction, while the spectral reflectance in the SWIR spectral region has a monotonic downward tendency only in the range of 2026 ~ 2330 nm. Nitrogen plays a crucial role in regulating plant photosynthesis and is involved in the synthesis of chlorophyll. Phosphorus deficiency can lead to the production of anthocyanins, altering the chlorophyll content. Therefore, we have focused on exploring the differences in nitrogen and phosphorus levels and their correlation with spectral changes. It can also be seen from the figure that the spectral reflectance of the LNC values in the ranges of 10.93 ~ 12.40 g/kg and 12.40 ~ 13.86 g/kg are very close in the Vis band, and when LNC exceeds these two intervals, there is a significant change in reflectance. This indicates that there may be a transition stage in the regulation of nitrogen on the growth of *Camellia oleifera*. When the LNC increases or decreases beyond the critical value of the transition stage, there will be a clear direction of plant growth change.

Figure 2(b) shows the Pearson correlation coefficient between LNC and raw spectral reflectance, and the correlation coefficient between LNC and reflectance with MSC and SGFD pretreatment. It can be seen that there are three sensitive bands of canopy spectra (i.e., 515 ~ 660, 688 ~ 734, and 2078 ~ 2310 nm). The response degree has reached a very significant correlation level (p-value < 0.01), and the maximum correlation coefficient is at wavelength 705 nm ($r = 0.39$). After MSC pretreatment, the sensitive band changed significantly, only overlapped with raw reflectance in 698 ~ 740 nm. However, the correlation coefficients increased in the range of 352 ~ 520 nm, 658 ~ 688 nm, and 1025 ~ 1086 nm, and all reached a very significant

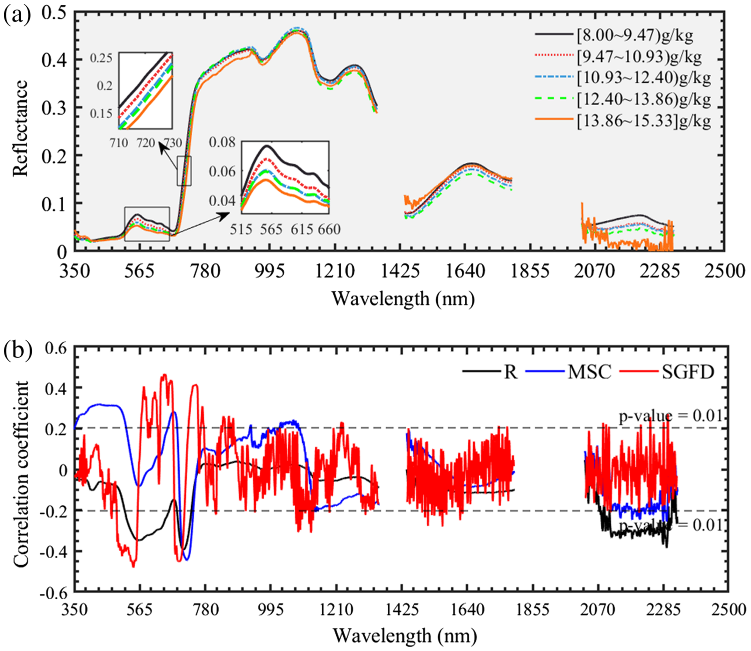


Fig. 2 Spectral response characteristics of LNC. (a) Response relationship of average canopy hyperspectral reflectance with different LNC gradients. (b) Correlation analysis of LNC and R, MSC, and SGFD spectral reflectance.

correlation level. The maximum correlation coefficient increased to 0.44, and the corresponding wavelength moved to 718 nm. The change of spectral reflectance by SGFD further enhanced the correlation with LNC, which was mainly reflected in the fact that most of the response relationships in the visible spectral region (490 ~ 775 nm) were at a very significant correlation level, and the sensitive intervals of NIR and SWIR1 bands were increased. The maximum correlation coefficient of SGFD was at a wavelength of 544 nm ($r = 0.48$). It is not difficult to find that the spectral reflectance of R, MSC, and SGFD reaches a very significant correlation level with LNC in the range of 698 ~ 734 nm, which just shows that the red-edge feature is a stable window for monitoring the changes of LNC. On the other hand, the correlation coefficient in the range of 2078 ~ 2310 nm decreased after pretreatment, among which the SGFD method decreased the most, which exactly coincides with the change relationship reflected in Fig. 2(b) (SGFD), indicating that both MSC and SGFD will reduce the signal-to-noise ratio of SWIR2 band.

The correlation analysis results of three two-band spectral indices (i.e., DSI, RSI, and NDSI) constructed by LNC and R, MSC, and SGFD pretreatment of *Camellia oleifera* are shown in Fig. 3. Compared with R in Fig. 2(b), the maximum correlation coefficient of R-DSI spectral index increased from 0.39 to 0.47 and its sensitive interval was mainly distributed in the visible spectral region. The linear correlation between NIR band and LNC and SWIR band and LNC was weak. DSI in the range of 694 ~ 737 nm had a high correlation in the whole band, and the linear correlation is higher when the wavelength is more biased towards the visible spectral region. Compared with R-DSI, both R-RSI and R-NDSI increased the maximum correlation coefficient from 0.47 to 0.51, with an increase rate of 8.5%. The sensitive regions of the two in the visible spectral region decreased year-on-year, and the response degree was weakened. However, the linear correlation between the red edge band and the band combination in 740 ~ 1350 nm was significantly enhanced, and a high linear correlation region of 560 ~ 690 nm combined with 740 ~ 1350 nm was added. MSC-DSI and SGFD-DSI both

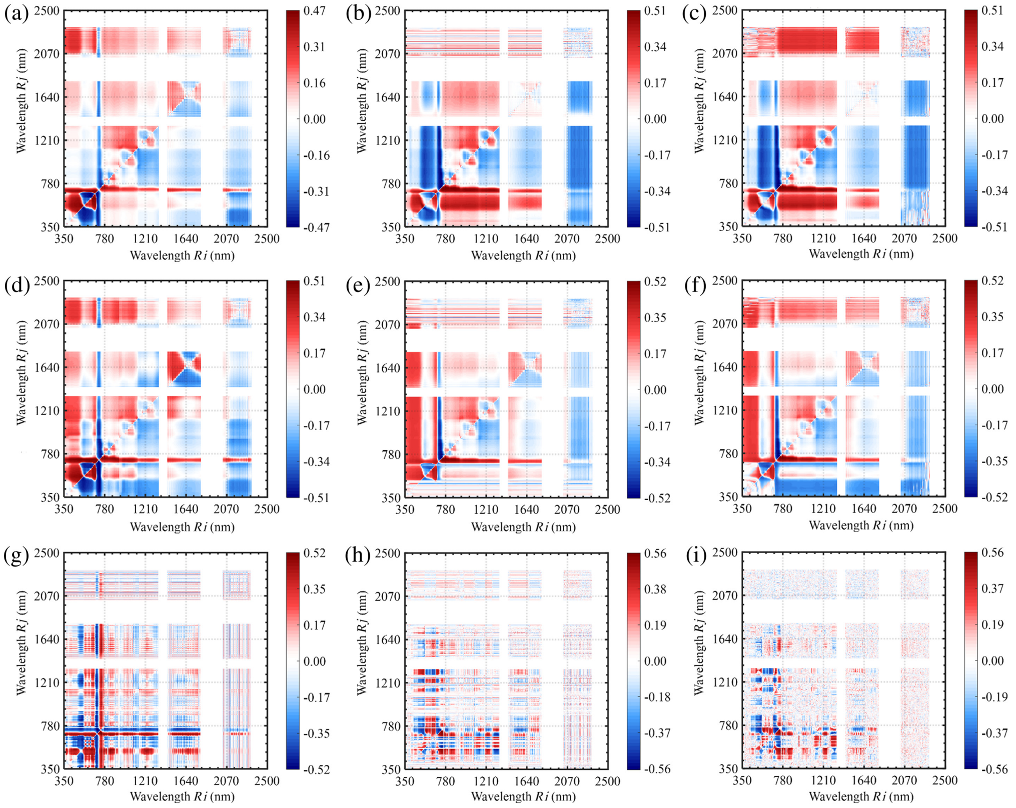


Fig. 3 Two-band spectral indices correlation analysis of LNC and R, MSC, and SGFD spectral reflectance. (a) R-DSI, (b) R-RSI, (c) R-NDSI, (d) MSC-DSI, (e) MSC-RSI, (f) MSC-NDSI, (g) SGFD-DSI, (h) SGFD-RSI, and (i) SGFD-NDSI.

increased the maximum correlation coefficient, which were 0.51 and 0.52, respectively. The sensitive interval of MSC-DSI was basically similar to that of R-DSI, while SGFD-DSI refined the response details, and the sensitive bands near 520, 690, and 760 nm were prominent. The correlation coefficients of SGFD-RSI and SGFD-NDSI both increased to 0.56. Although the sensitivity interval of SGFD-NDSI was similar to SGFD-RSI, the response region was more fragmented.

The correlation analysis results of three three-band spectral indices (i.e., TBDSI, TBRSI, and TBNDSI) constructed by LNC and R, MSC, and SGFD pretreatment of *Camellia oleifera* are shown in Fig. 4. It can be seen from Figs. 4(a)–4(c) that the maximum correlation coefficient of R-TBDSI increased to 0.49, and the correlation degree increased by 25.6% compared with R. Compared with R-DSI, the maximum correlation coefficient also increased by 0.02, and the maximum correlation coefficient of R-TBRSI and R-TBNDSI has improved, $r = 0.53$ and $r = 0.55$, respectively. This shows that the linear enhancement effect of the three-band spectral indices of R is obvious, and it is better than that of the two-band spectral indices. Among the three spectral index combinations, NDSI has the best improvement effect, followed by RSI and DSI. From Figs. 4(d)–4(f), it can be seen that the maximum correlation coefficients of the three three-band spectral indices of MSC are very close, all around 0.54, which is improved compared to R-TBDSI and R-TBRSI, but the enhancement effect is similar to R-TBNDSI. In Figs. 4(g)–4(i), the three-band spectral indices of SGFD were not significantly improved compared with the maximum correlation coefficient of its two-band spectral indices. The maximum correlation coefficients of SGFD-TBRSI and SGFD-TBNDSI are lower than those of SGFD-RSI and SGFD-NDSI. The linear enhancement effect of TBNDSI on SGFD is between TBRSI and TBDSI.

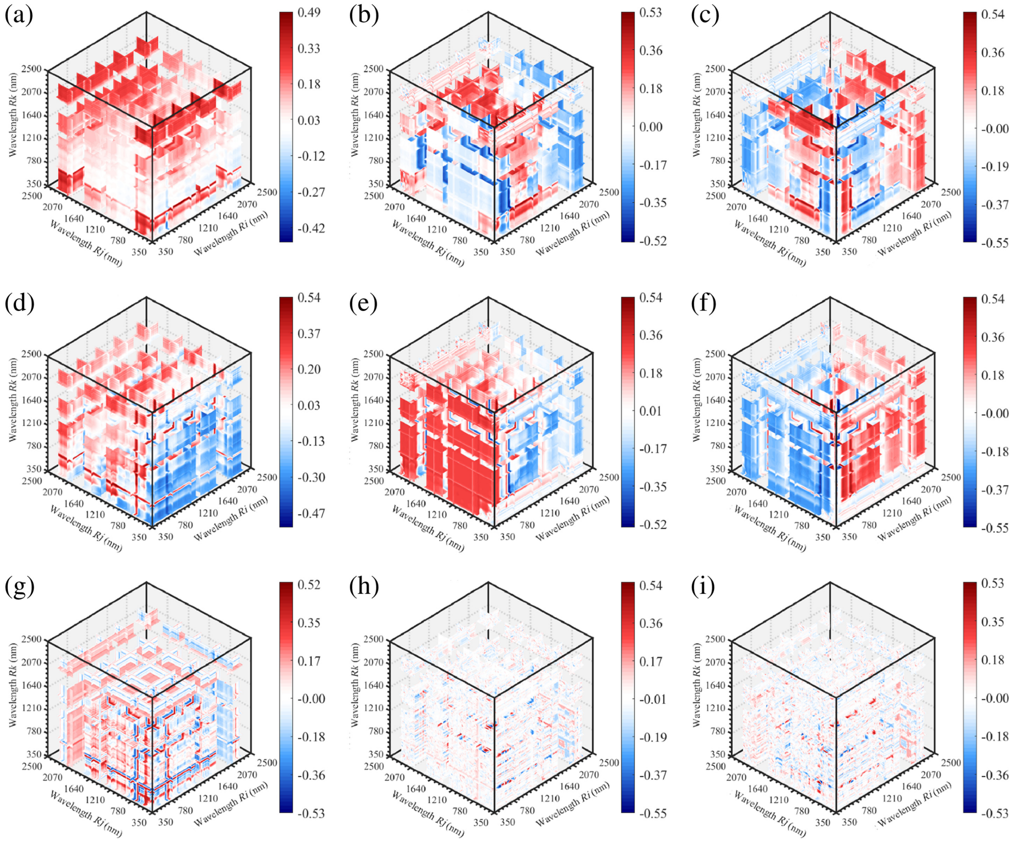


Fig. 4 Three-band spectral indices correlation analysis of LNC and R, MSC, and SGFD spectral reflectance. (a) R-TBDSI, (b) R-TBRSI, (c) R-TBNDSI, (c) MSC-TBDSI, (d) MSC-TBRSI, (e) MSC-TBNDSI, (f) SGFD-TBDSI, (g) SGFD-TBRSI, and (h) SGFD-TBNDSI.

3.2.2 *Response relationship between canopy hyperspectral reflectance and LPC of* Camellia oleifera

The amount of samples in five equal LPC intervals are 6, 41, 53, 47, and 13, respectively. By comparing the response relationship of the average hyperspectral reflectance (R) of *Camellia oleifera* canopy in the five equal intervals of LPC [Fig. 5(a)], the results show that the overall trend of canopy hyperspectral reflectance at different phosphorus levels shows a certain regulation. With the increase of LPC, the reflectance in the visible spectral region gradually decreases, and the red-edge feature will shift towards the long wave direction. The spectral reflectance in the range of 2026 ~ 2330 nm of SWIR2 has a more obvious monotonic downward trend. The difference between this and LNC is that the spectral reflectance of LPC in the ranges of 0.76 ~ 0.88 g/kg, 0.88 ~ 1.00 g/kg and 1.00 ~ 1.12 g/kg is close in Vis band. Only when LPC falls out of these three intervals, there is a significant change in reflectance. This indicates that the critical value of phosphorus regulating the growth change of *Camellia oleifera* into the transition stage is small and can be maintained in a wide concentration range.

Figure 5(b) shows the Pearson correlation coefficient between LPC and R, and the correlation coefficient between LPC and reflectance with MSC and SGFD pretreatment. It can be seen that the sensitive bands of canopy spectrum are 513 ~ 723 and 2083 ~ 2325 nm, and the response degree reaches a very significant correlation level. The maximum correlation coefficient is at 2295 nm ($r = 0.35$). After MSC pretreatment, the sensitive band changed significantly, only overlapped with R in 697 ~ 722 nm. The correlation coefficient in the range of 350 ~ 520 nm and 1021 ~ 1102 nm increased and reached a highly significant correlation level but the linear correlation degree in the range of 521 ~ 665 nm decreased. The maximum correlation coefficient was 0.33, and the corresponding wavelength moved to 386 nm. The maximum correlation coefficient after SGFD pretreatment is at 681 nm ($r = 0.40$). The spectral reflectance of R, MSC, and SGFD reached a significant correlation with LPC in 697 ~ 722 nm, indicating that the red-edge feature can also be used as a stable window for monitoring LPC changes.

The correlation analysis results of three two-band spectral indices (i.e., DSI, RSI, and NDSI) constructed by LPC and R, MSC, and SGFD pretreatment of *Camellia oleifera* are shown in Fig. 6. The sensitive interval of R-DSI spectral index is mainly distributed in the visible spectral region, and the wavelength around 700 nm has a high linear correlation with the DSI value in the whole band [Fig. 6(a)]. Compared with R-DSI, both R-RSI and R-NDSI increased the maximum

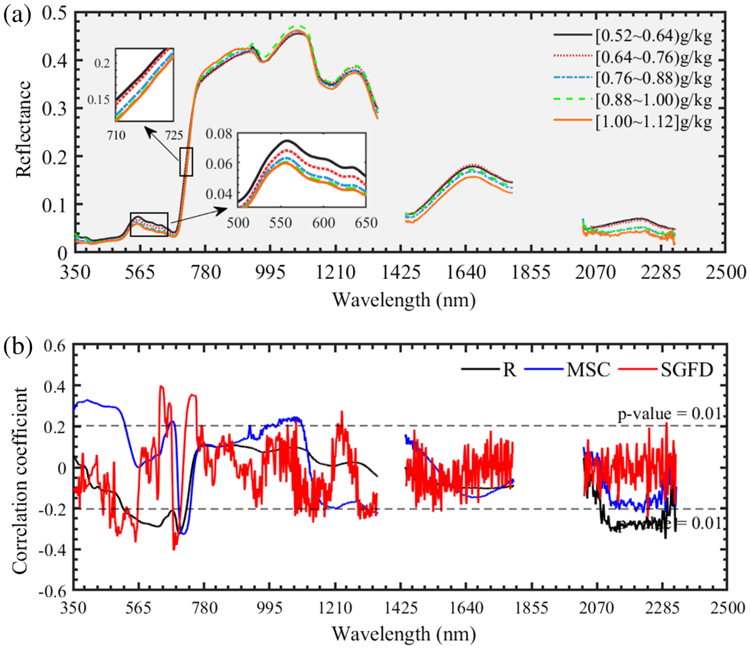


Fig. 5 Spectral response characteristics of LPC. (a) Response relationship of average canopy hyperspectral reflectance with different LPC gradients. (b) Correlation analysis of LPC and R, MSC and SGFD spectral reflectance.

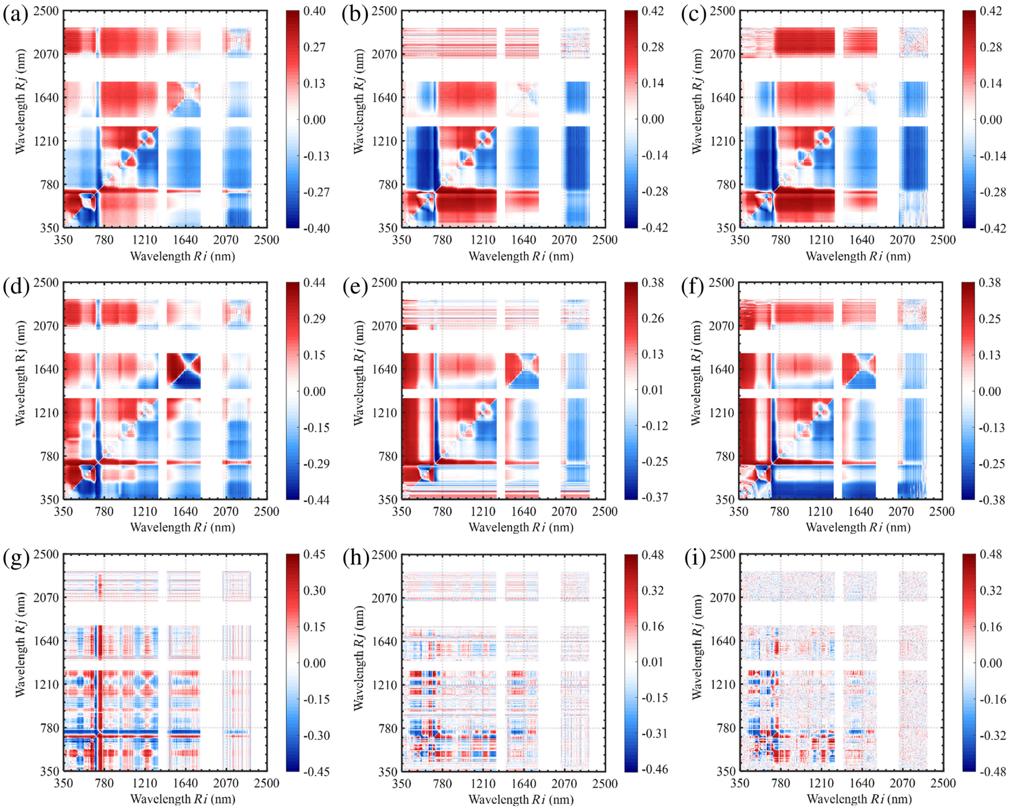


Fig. 6 Two-band spectral indices correlation analysis of LPC and R, MSC, and SGFD spectral reflectance. (a) R-DSI, (b) R-RSI, (c) R-NDSI, (d) MSC-DSI, (e) MSC-RSI, (f) MSC-NDSI, (g) SGFD-DSI, (h) SGFD-RSI, and (i) SGFD-NDSI.

correlation coefficient from 0.40 to 0.42, with an increase rate of 5%. MSC-DSI and SGFD-DSI increased the maximum correlation coefficients to 0.44 and 0.45, respectively. The sensitive range of MSC-DSI in the visible spectral region is basically similar to that of R-DSI, but it also enhances the response near 1640 nm. The combination of MSC-RSI and MSC-NDSI is not so good, and the maximum correlation coefficients of the two are significantly reduced to 0.38 [Figs. 6(e)–6(f)]. SGFD-RSI and SGFD-NDSI increase the maximum correlation coefficient to 0.48.

The correlation analysis results of three three-band spectral indices (i.e., TBDSI, TBRSI, and TBNDSI) constructed by LPC and R, MSC, and SGFD pretreatment of *Camellia oleifera* are shown in Fig. 7. It can be seen from Figs. 7(a)–7(c) that the maximum correlation coefficient of R-TBDSI increased to 0.43, which is 22.9% higher than that of R and is 0.03 higher than that of R-DSI. R-TBNDSI was basically the same as R-TBDSI, but the maximum correlation coefficient of R-TBRSI increased to 0.46. The maximum correlation coefficient of MSC with both TBDSI and TBRSI combination is 0.42 [Figs. 7(d)–7(f)]. Although it is lower than three-band spectral index combinations of R, the difference of MSC and the response degree normalized spectral index are not significantly weakened as the two-band spectral indices. The maximum correlation coefficient of the three-band spectral indices of SGFD did not increase significantly, and only the maximum correlation coefficient of SGFD-TBRSI reached 0.49 [Figs. 7(g)–7(i)].

3.3 Using VCPA-IRIV Algorithm to Select the Combination of Spectral Variables

3.3.1 *Results of canopy hyperspectral variable combination of* Camellia oleifera *LNC*

The variable selection combination scheme [i.e., variable combination population analysis-iteratively retained informative variables (VCPA-IRIV) algorithm] used in this paper reduced

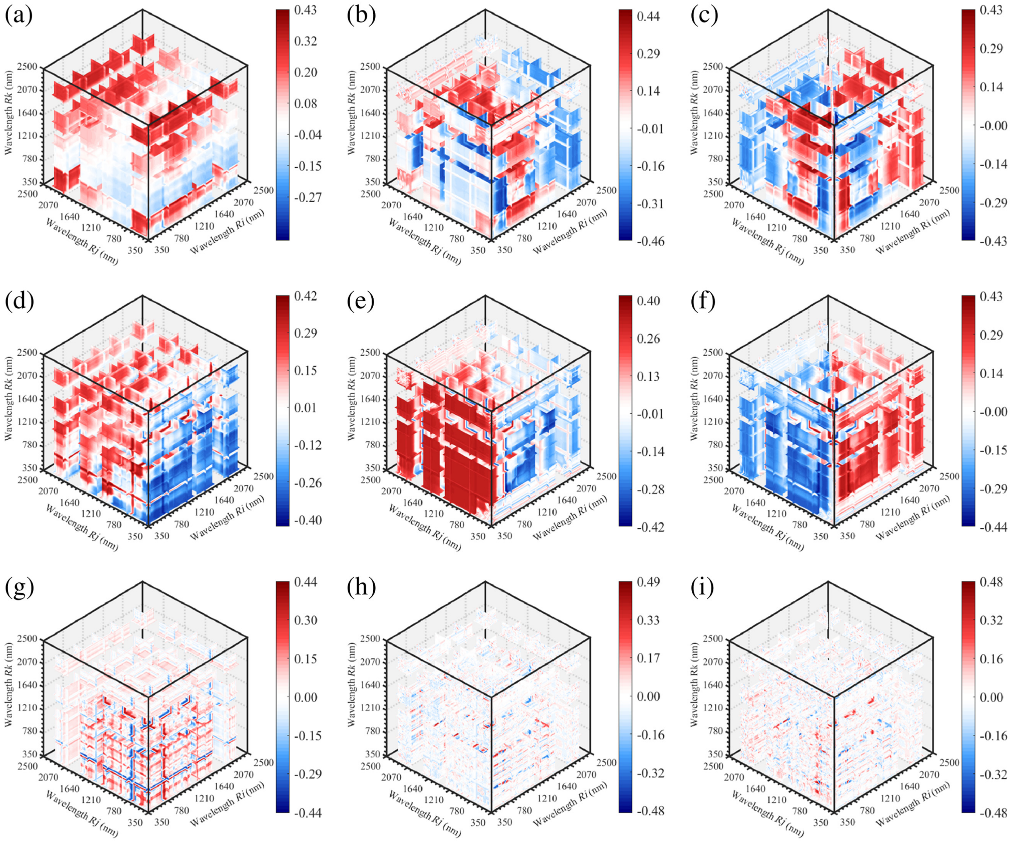


Fig. 7 Three-band spectral indices correlation analysis of LPC and R, MSC, and SGFD spectral reflectance. (a) R-TBDSI, (b) R-TBRSI, (c) R-TBNDSI, (c) MSC-TBDSI, (d) MSC-TBRSI, (e) MSC-TBNDSI, (f) SGFD-TBDSI, (g) SGFD-TBRSI, and (h) SGFD-TBNDSI.

the number of retained variables from 1661 (490 abnormal wavelength variables affected by water vapor have been eliminated) at the beginning to 4 with the increase of the number of iterations and the algorithm process after setting basic parameters, with a reduction rate of 99.76% (Fig. 8). After several of VCPA-IRIV analysis, the downward trend of variables showed the characteristics of fast first and then slow. The VCPA stage went through "fast selection" and "precise selection" in the EDF iteration process. Eight variables (i.e., 732, 733, 736, 757, 1101, 1226, 2029, 2296 nm) were retained. To reduce the multicollinearity between the variables, the correlation coefficients between the eight variables were calculated, and only one of the pairs of variables with a correlation coefficient higher than 0.9 was retained, which was highly correlated with LNC. The final subset of variable combinations is 732, 1101, 2029, and 2296 nm (Table 2).

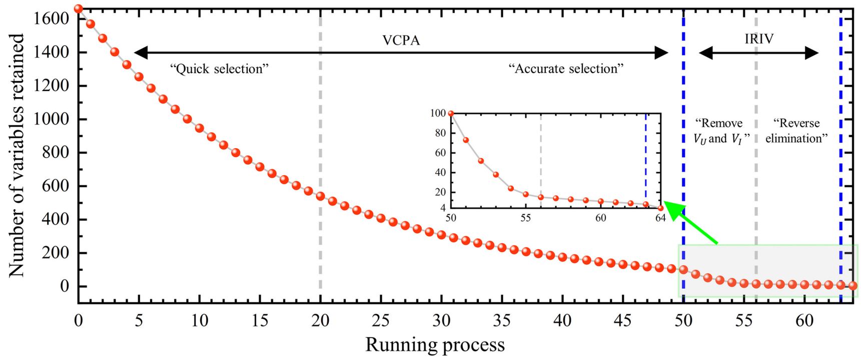


Fig. 8 Changes in the number of retained variables during the operation of VCPA-IRIV strategy.

Table 2 Optimal variable combination subset results of R, MSC, and SGFD of LNC.

Pretreatment type	Number of variables retained	Selected wavelength position/nm
R	4	732, 1101, 2029, and 2296
MSC	6	736, 758, 1254, 2030, 2107, and 2127
SGFD	7	736, 878, 913, 1145, 1192, 1328, and 1501

Table 2 shows that the number of retained variables after MSC and SGFD pretreatment increases, which indicates that the use of pretreatment methods can improve the ability of VCPA-IRIV to mine potential information. At the same time, from the perspective of the selected wavelength position, the wavelengths that make up the optimal subset are not all located in the sensitive interval with high linear correlation, such as 1101, 1254, etc., which further shows that the selection of spectral variables cannot simply take the wavelength where the maximum correlation coefficient is located as a single input variable. According to the results of the optimal variable combination subset selected by multiple iterations of VCPA-IRIV algorithm after combination of LNC and two and three band spectral indices, it can be seen that the two band spectral indices and three band spectral indices of LNC show the characteristics of combination of strong spectral signals (p-value < 0.01) and weak spectral signals (p-value ≥ 0.01) in the location distribution of selected wavelength combinations, and most of them are located in the joint region of visible spectral region and NIR, SWIR1, and SWIR2.

3.3.2 *Results of canopy hyperspectral variable combination of* Camellia oleifera *LPC*

From Table 3, it can be seen that the number of reserved variables after MSC and SGFD pretreatment increases sequentially, and the selected wavelength positions are more dispersed. Through the results of the optimal variable combination subset selected by multiple iterations of VCPA-IRIV algorithm after combining LPC with two and three band spectral indices, it is found that the two band spectral indices and three band spectral indices of LPC also have the characteristics of combining strong spectral signals with weak spectral signals in the location distribution of the selected wavelength combination. Similarly, the VCPA-IRIV algorithm maintains a high variable space compression ratio for high-dimensional data.

3.4 Estimation Models of Nitrogen and Phosphorus Content of *Camellia oleifera* Leaves Based on Canopy Hyperspectral Reflectance

3.4.1 *Evaluation of LNC estimation models based on canopy hyperspectral reflectance of* Camellia oleifera

Through the correlation analysis of LNC and R, MSC, and SGFD spectral reflectance, and the analysis combined with three two-band spectral index forms and three three-band spectral index forms, 63 machine learning models were constructed. The result showed that the R_C^2 is $0.35 \sim 0.82$, and RMSE_C is $0.56 \sim 1.08$ g/kg in calibration set. The R_P^2 is $0.30 \sim 0.81$,

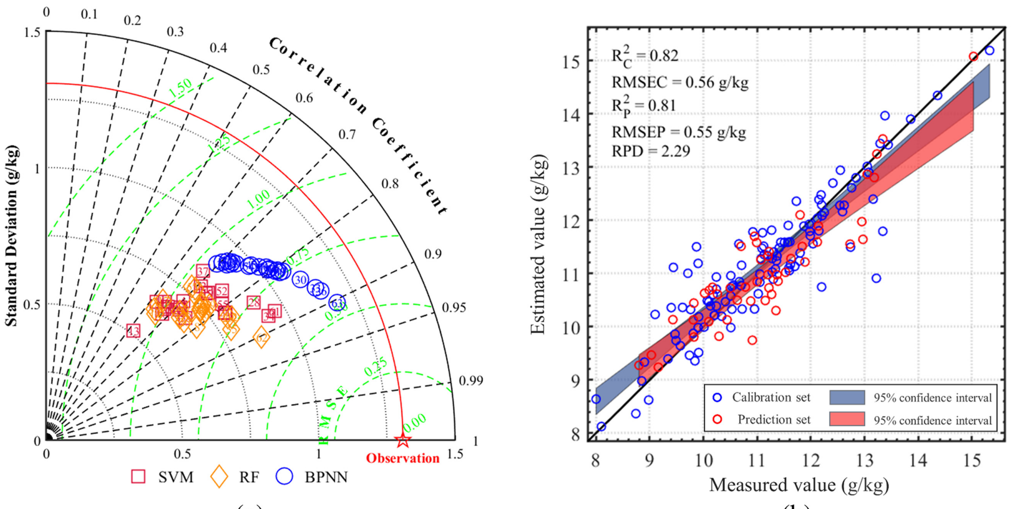
Table 3 Optimal variable combination subset results of *R*, MSC, and SGFD of LPC.

Pretreatment type	Number of variables retained	Selected wavelength position/nm
R	3	697, 926, and 1451
MSC	5	701, 1349, 1472, 2132, and 2300
SGFD	9	681, 734, 969, 1079, 1148, 1226, 1233, 1532, and 2236

RMSE_P is $0.55 \sim 1.05$ g/kg, and the RPD is $1.21 \sim 2.29$ in prediction set, which shows that there are obvious differences in the estimation performance of different combination types. In terms of overall performance of the three machine learning models, the R_C^2 of SVM model is $0.35 \sim 0.78$, and the R_P^2 is $0.30 \sim 0.77$. The R_C^2 of RF model is $0.38 \sim 0.76$, and the R_P^2 is $0.38 \sim 0.75$. The R_C^2 of BPNN model is $0.49 \sim 0.82$, and the R_P^2 is $0.45 \sim 0.81$. Through comprehensive analysis of the improvement effect of the three models after R, MSC, SGFD and two-band and three-band spectral indices processing, it can be concluded that the upper limit of LNC model estimation accuracy increases with the increase of spectral band dimensions. Overall, the average R^2 of R and its combination with the spectral index is 0.47, and the MSC and SGFD are 0.54 and 0.67, respectively, which shows that SGFD is superior to the overall estimated performance of LNC.

In Fig. 9(a), the accuracy results of 63 LNC candidate estimation models corresponding to the serial numbers are displayed intuitively and efficiently in Taylor diagram. It can be seen from the figure that the SGFD-TBNDSI-BPNN model (No. 63) is the closest to the straight-line distance from the reference point, so it has a high model fitting accuracy. Furthermore, the standard deviation of the estimated value is also more consistent with the numerical fluctuation of the measured value. Most correlation coefficients between the measured and estimated values of SVM, RF, and BPNN were between 0.7 and 0.8. Most of the corresponding RMSE ranged from 0.75 to 1.00 g/kg. In terms of numerical volatility, RF has a relatively smaller standard deviation, which is basically around 0.75 g/kg, representing that it can often form more stable results, while the standard deviation of BPNN is near 1.00 g/kg, and the corresponding estimated value changes are relatively unstable.

To further evaluate the changes between the estimated and measured value of the LNC optimal model (i.e., SGFD-TBNDSI-BPNN) in continuous concentration changes, the relationship between the two is presented in the scatter point plot [Fig. 9(b)]. The results showed that the LNC estimation model of SGFD-TBNDSI-BPNN had an R_C^2 of 0.82 and an RMSE_C of 0.56 g/kg in the calibration set, and had an R_P^2 of 0.81, and an RMSE_P of 0.55 g/kg in the prediction set. This explained 82% of the LNC in the training samples and 81% of the unknown LNC samples of *Camellia oleifera*. The RPD of 2.29 (>2.00) indicates that the model has a good estimation ability, which means that the model can better use the characteristic information of canopy hyperspectral data to accurately reflect the changes of LNC.



(a) (b)

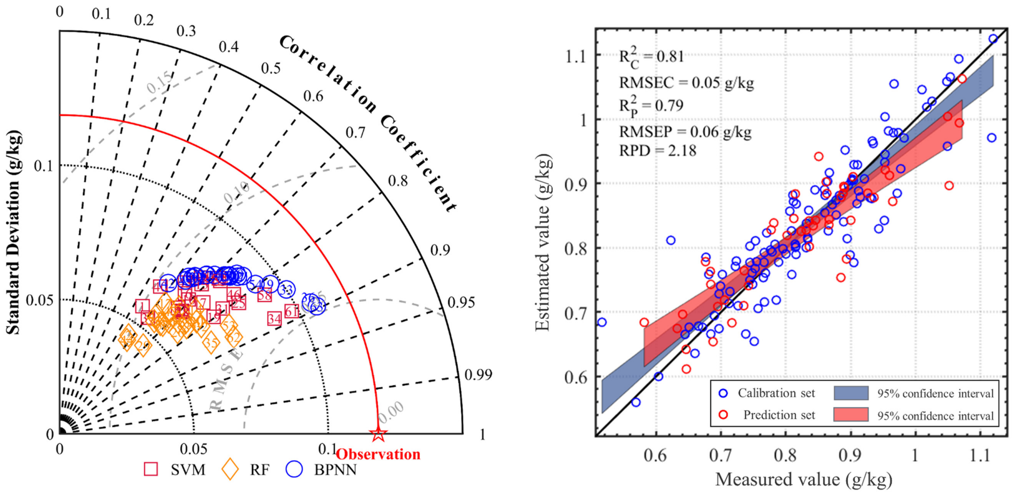
Fig. 9 LNC model performance evaluation. (a) The accuracy comparison results of 63 LNC candidate estimation models presented in Taylor diagram. (b) The relationship between the measured and estimated values of LNC given by SGFD-TBNDSI-BPNN model.

3.4.2 *Evaluation of LNC estimation models based on canopy hyperspectral reflectance of* Camellia oleifera

Through the correlation analysis of LPC and R, MSC, and SGFD spectral reflectance, and the analysis combined with three two-band spectral index forms and three three-band spectral index forms, 63 machine learning models were constructed. The result showed that the R_C^2 is $0.29 \sim 0.81$, and RMSE_C is $0.05 \sim 0.10$ g/kg in calibration set. The R_P^2 is $0.20 \sim 0.79$, RMSE_P is $0.06 \sim 0.11$ g/kg, and the RPD is $1.13 \sim 2.18$ in prediction set. The overall performance of the three machine learning models was evaluated separately. The R_C^2 of the SVM models is $0.29 \sim 0.80$, and the R_P^2 is $0.20 \sim 0.77$. The R_C^2 of the RF models is $0.30 \sim 0.70$, and the R_P^2 is $0.27 \sim 0.70$. The R_C^2 of the BPNN models is $0.38 \sim 0.81$, and the R_P^2 is $0.25 \sim 0.79$. It can be concluded that the upper limit of LPC model estimation accuracy will increase with the increase of spectral band dimension. Overall, the average value of R^2 in R and its combination with spectral index is 0.43, and the MSC and SGFD are 0.53 and 0.62, respectively, which shows that SGFD is better for the overall estimated performance of LPC.

The accuracy results of 63 LPC candidate estimation models are visually and efficiently presented in the form of Taylor plots in Fig. 10(a). It can be seen from the figure that the SGFD-TBNDSI-BPNN model (No. 63) has the shortest straight-line distance from the reference point. Thus, it has a high model fitting accuracy. Moreover, the standard deviation of the estimated value is also in line with the numerical fluctuation of the measured value. In general, most correlation coefficients between the measured and estimated values of SVM, RF and BPNN were between 0.6 and 0.8. Most of the corresponding RMSE ranged from 0.06 to 0.10 g/kg. In terms of numerical volatility, RF has a relatively smaller standard deviation, basically around 0.06 g/kg, which means that it can often form more stable results. In contrast, the standard deviation of BPNN is around 0.08 g/kg, and the corresponding estimated value changes are relatively unstable.

To further evaluate the changes between the estimated and measured value of the LPC optimal model (i.e., SGFD-TBNDSI-BPNN) in continuous concentration changes, the relationship between the two is presented in the scatter point plot [Fig. 10(b)]. The results showed that the LPC estimation model of SGFD-TBNDSI-BPNN had an R_C^2 of 0.81, and an RMSE_C of 0.55 g/kg in calibration set, and had an R_P^2 of 0.79, and an RMSE_P of 0.06 g/kg in the prediction set. This explained 81% of the LPC in the training samples and 79% of the unknown LPC samples of *Camellia oleifera*. The RPD of 2.18 (>2.00) implies that it has good estimation ability, which means that the model can better use the characteristic information of canopy hyperspectral data to accurately reflect the changes of LPC.



(a) (b)

Fig. 10 LPC model performance evaluation. (a) Taylor diagram presenting the accuracy comparison results of 63 LPC candidate estimation models. (b) The relationship between the measured value and the estimated value of LPC given by SGFD-TBNDSI-BPNN model.

4 Discussions

4.1 Response Relationship Between LNC and Spectral Transformation Characteristics of *Camellia oleifera*

According to the correlation analysis results between the raw spectral characteristics and LNC, the sensitive regions of leaf nitrogen at canopy scale are mainly located in the green, red, and SWIR2 bands of the visible spectral region. In the wavelength range of visible light and near-infrared light, chlorophyll content is the dominant factor affecting the spectral reflectance of vegetation, and nitrogen, as an important nutrient element in plants, is the main component of chlorophyll. These three response positions correspond to the strong reflection and absorption bands of chlorophyll and the reflection bands of water, revealing that the significant spectral response of nitrogen level in fresh leaves at canopy scale is dominated by chlorophyll and water, which is consistent with Wang et al.'s[33] view. The spectral response characteristics of canopy scale are affected by various factors, such as photosynthetic pigment concentration, canopy structure, and background radiation, thus the corresponding spectral characteristics associated with nitrogen components will be weakened to a certain extent in information expression. MSC and SGFD respectively corrected the scattering difference by linear methods and decomposed the spectral characteristics by differential techniques to weaken the influence of background noise from raw spectra with physicochemical and structural information of single plant canopy, and then highlighted the response sensitive regions of LNC. Both preprocessing methods increased the response in the visible spectral region, NIR, and SWIR1 bands, yet reduced the linear response of some bands of SWIR2 band. This indicates that background noise has a strong interference on the expression of information about photosynthetic pigments and canopy structure in spectral features. In addition, the linear gain effect of SGFD on the relationship between spectral characteristics and LNC response is better than that of MSC. The reason may be that the spectral transformation characteristics of SGFD can counteract the influence of soil background, while MSC is mainly targeted at the scattering interference of the surrounding environment. In the process of acquiring ground canopy hyperspectral data in this study, the influence of the soil background in the vertical observation field was significantly greater than that of the surrounding environment.[34]

The spectral indices have the characteristics of amplifying weak correlation between reflection information and minimizing the influence of external factors. The combination of MSC and SGFD with different algebraic operation forms of spectral indices can more effectively enhance the linear measurement of LNC with spectral transformation characteristics. Obviously, the correlation analysis results of the two-band spectral indices and three-band spectral indices of R all verify this point. The combination results of MSC and SGFD with spectral indices also show obvious synergistic gain effects. The combination of pretreatment methods and spectral indices increases the sensitivity of spectral transformation characteristics to the response of LNC. One important reason is that the constructed spectral index is insensitive to interference factors, forming complementary advantages of different methods.[35]

4.2 Response Relationship Between LPC and Spectral Transformation Characteristics of *Camellia oleifera*

In this paper, the sensitive areas of canopy-scale *Camellia oleifera* LPC on the R spectrum are basically similar to the LNC results, including the green, yellow, orange, red, and SWIR2 bands in the visible spectral region. The difference is the change of the correlation coefficient curve in the visible spectral region. The two low points of LNC fall at 561 and 710 nm, and those of LPC fall at 629 and 696 nm. There is no general rule about the difference in the specific location of the response wavelength, which involves the complex biochemical activities of phosphorus metabolism itself and its interaction with nitrogen. In addition, the canopy spectral response of LPC in this paper is weakly correlated, which may be the nutritional status of *Camellia oleifera* materials selected in this study. There is no typical effect of phosphorus deficiency, and only some individual phosphorus deficiency affects chlorophyll synthesis, which in turn weakens the response relationship with LNC overlap. The spectral transformation characteristics of MSC and SGFD and the response area of *Camellia oleifera* LPC are basically similar to LNC, indicating that the preprocessing method is to filter out the interference information with fixed functional attributes and to increase the information proportion of cell and canopy structure in NIR and SWIR1 bands. The combination effect of the two-band spectral indices (i.e., MSC-RSI and MSC-NDSI) of LPC will be worse than that of MSC-DSI, which is not consistent with the conclusion of LNC, indicating that the synergy between pretreatment and spectral index will also have a negative effect. It is possible that the difference in performance between MSC-DSI and MSC-RSI/MSC-NDSI is due to the calculation methods used. MSC-DSI is based on the calculation of difference spectra, which can highlight the differences between different bands and is highly sensitive to specific spectral features. On the other hand, MSC-RSI and MSC-NDSI are based on ratio spectra or normalized difference spectra, which analyze the ratios or normalized differences between different bands. In the case of LPC, the calculation of difference spectra in MSC-DSI may be more suitable for capturing information related to leaf pigment content because difference spectra can better reflect the absorption and reflection characteristics of different pigments in different bands. MSC-RSI and MSC-NDSI may be more sensitive to other targets, such as vegetation indices. Furthermore, this study also found that the effect of different spectral preprocessing strategies will vary with the increase of spectral index dimension, and the phenomenon of weakening with the increase of spectral dimension also appears in other research results.[36]

This study constructed LNC and LPC estimation models using three machine learning algorithms. The best overall performance was achieved by the SGFD-TBNDSI-BPNN model. The evaluation results of the model's generalization ability were LNC-$R_P^2 = 0.81$ and LPC-$R_P^2 = 0.79$. The differences in performance were consistent with the variations in the maximum linear metrics of the spectral responses to the target variables. This highlights that the upper limit of a model's learning capability often depends on the effective information carried by the data features. This implies that the accuracy of machine learning models is not solely determined by the complexity of the model but also directly related to the optimization of the spectral feature variables.

4.3 Response Mechanism of Plant Leaf Nutrients to Spectral Characteristics

The spectral reflectance can reflect the light absorption ability of vegetation. The more the absorption, the lower the spectral reflectance, the higher the light energy utilization rate of vegetation, and the spectral absorption ability of vegetation is closely related to the physiological and biochemical characteristics of leaves. The nutrient content in the leaves can effectively reflect the nutrient status of the whole plant, and the content of these nutrients determines the physiological condition of the plant. In particular, nitrogen and phosphorus content accounted for about 5% of the dry matter proportion of plant leaves, they have an impact on the synthesis of biological pigments such as chlorophyll, anthocyanins, and carotenoids in leaves, which in turn affect the physiological state of plants.[37] Nitrogen is involved in regulating the photosynthesis of plants, and nitrogen deficiency can seriously affect the growth and yield of plants, but excessive use of nitrogen fertilizers can cause a series of environmental problems. Phosphorus deficiency can easily lead to the production of anthocyanins, change the content of chlorophyll, and significantly affect the structural arrangement of leaf cells and the synthesis of organic matter. In addition, changes in nutrient content can be quickly reflected on spectral curves, so remote sensing technology can be used to quickly monitor and assess plant growth.

The canopy spectrum is affected by changes in the content of trace elements nitrogen and phosphorus, which are due to the influence of these elements on the light absorption and scattering processes. Nitrogen atoms have strong absorption bands in the ultraviolet band, while phosphorus atoms have absorption bands in the visible band, and nitrogen and phosphorus atoms can also scatter light, thus changing the intensity and shape of the canopy spectrum.[38] The absorption bands and scattering signatures of nitrogen and phosphorus may overlap, which makes it difficult to quantify their content separately, and the presence of background noise from other substances (e.g., chlorophyll, water) in the canopy spectrum can mask the weak signals of nitrogen and phosphorus. Atmospheric scattering and absorption can affect the shape and intensity of the canopy spectrum, introducing additional interferences. In order to mitigate and correct the effects of canopy spectral interference, we use multivariate scattering correction and SGFD, which can effectively mitigate and correct the effects of canopy spectral interference, thereby improving the accuracy and reliability of nitrogen and phosphorus content inversion.

5 Conclusions

This paper focused on the influence of the combination of two-band and three-band spectral indices on the response relationships and the differences in model calibration and prediction accuracy. It is clear that VCPA-IRIV strategy can efficiently extract spectral transformation features and effectively improve model estimation accuracy. The main conclusions are as follows.

1. The response relationship between LNC, LPC, and raw spectra (R) was similar. The sensitive intervals were mainly concentrated in the green and red bands of the visible spectral regions related to chlorophyll and the short-wave infrared long-wave region related to moisture. After MSC and SGFD preprocessing, the background noise interference in the spectral information was significantly reduced, and the overall processing effect of SGFD was better than that of MSC. The combined effect of different pretreatment methods and spectral indices had different performances in LNC and LPC with the increase of spectral dimensions. The upper limits of Pearson correlation coefficient can reach 0.56 and 0.49, respectively in all treatment combinations.
2. The VCPA-IRIV variable selection strategy had a very high variable space compression rate for spectral transformation characteristics and can fully consider the interaction between variables. The selected spectral variables contained both strong and weak information variables, which were beneficial to describe the nonlinear relationship between canopy hyperspectral data and LNC. The sequence ranked by model stability was as follows: RF > SVM > BPNN, but the BPNN model had the highest accuracy.

Disclosures

The authors declare that they have no conflict of interest.

Code and Data Availability

The data and code that support the findings of this article are not publicly available because the project is still undergoing. They can be obtained from the corresponding author at tangxuehai@ahau.edu.cn by reasonable request.

Acknowledgments

The Xianyuan Xian'ge Oil Tea Professional Cooperative provides us access to *Camellia oleifera* trees. We thank Weijing Song and Cong Lv from Anhui Agricultural University for their help in field work. This paper was financially supported by the National Natural Science Foundation of China (Grant No. 32171783).

References

1. Y. Cai, *Plant Physiology*, pp. 38–66, China Agricultural University Press, Beijing (2014).
2. H. Wu, J. Li, and Y. Ge, "Ambiguity preference, social learning and adoption of soil testing and formula fertilization technology," *Technol. Forecast. Soc. Change* **184**, 122037 (2022).
3. S. Peanusaha et al., "Nitrogen retrieval in grapevine (Vitis vinifera L.) leaves by hyperspectral sensing," *Remote Sens. Environ.* **302**, 113966 (2024).
4. J. Zhu et al., "Diagnoses of rice nitrogen status based on characteristics of scanning leaf," *Spectrosc. Spect. Anal.* **29**(8), 2171–2175 (2009).
5. N. Liu et al., "Multi-year hyperspectral remote sensing of a comprehensive set of crop foliar nutrients in cranberries," *ISPRS J. Photogramm. Remote Sens.* **205**, 136–146 (2023).
6. L. Li et al., "Ability of models with effective wavelengths to monitor nitrogen and phosphorus status of winter oilseed rape leaves using in situ canopy spectroscopy," *Field Crops Res.* **215**, 173–186 (2018).
7. S. L. Osborne et al., "Detection of phosphorus and nitrogen deficiencies in corn using spectral radiance measurements," *Agron. J.* **94**(6), 1215–1221 (2002).
8. J. Li et al., "Integrating UAV hyperspectral data and radiative transfer model simulation to quantitatively estimate maize leaf and canopy nitrogen content," *Int. J. Appl. Earth Obs. Geoinf.* **129**, 103817 (2024).
9. J. Wang et al., "Inversion of winter wheat foliage vertical distribution based on canopy reflected spectrum by partial least squares regression method," *Spectrosc. Spect. Anal.* **27**(7), 1319–1322 (2007).

10. K. Berger et al., “Crop nitrogen monitoring: recent progress and principal developments in the context of imaging spectroscopy missions,” *Remote Sens. Environ.* **242**, 111758 (2020).
11. K. Klem et al., “Changes in vertical distribution of spectral reflectance within spring barley canopy as an indicator of nitrogen nutrition, canopy structure and yield parameters,” *Agriculture* **60**(2), 50–59 (2014).
12. P. J. Curran, “Remote sensing of foliar chemistry,” *Remote Sens. Environ.* **30**(3), 271–278 (1989).
13. D. Li et al., “Estimating leaf nitrogen content by coupling a nitrogen allocation model with canopy reflectance,” *Remote Sens. Environ.* **283**, 113314 (2022).
14. A. Ramoelo et al., “Non-linear partial least square regression increases the estimation accuracy of grass nitrogen and phosphorus using in situ hyperspectral and environmental data,” *ISPRS J. Photogramm. Remote Sens.* **82**, 27–40 (2013).
15. P. Guo et al., “A robust method to estimate foliar phosphorus of rubber trees with hyperspectral reflectance,” *Ind. Crops Prod.* **126**, 1–12 (2018).
16. L. Li et al., “Accurate modeling of vertical leaf nitrogen distribution in summer maize using in situ leaf spectroscopy via CWT and PLS-based approaches,” *Eur. J. Agron.* **140**, 126607 (2022).
17. X. Tang et al., “Hyperspectral prediction of mangrove leaf stoichiometries in different restoration areas based on machine learning models,” *J. Appl. Remote Sens.* **16**(3), 034525 (2022).
18. L. Xie, M. Hong, and Z. Yu, “A wavelength selection method combing direct orthogonal signal correction and Monte Carlo,” *Spectrosc. Spect. Anal.* **42**(2), 440–445 (2022).
19. T. Zhang et al., “Non-destructive analysis of germination percentage, germination energy and simple vigour index on wheat seeds during storage by Vis/NIR and SWIR hyperspectral imaging,” *Spectrochim. Acta A: Mol. Biomol. Spectrosc.* **239**, 118488 (2020).
20. R. Zhao et al., “Deep learning assisted continuous wavelet transform-based spectrogram for the detection of chlorophyll content in potato leaves,” *Comput. Electron. Agric.* **195**, 106802 (2022).
21. F. Carmona et al., “Vegetation index to estimate chlorophyll content from multispectral remote sensing data,” *Eur. J. Remote Sens.* **48**(1), 319–326 (2015).
22. Y. Inoue et al., “Diagnostic mapping of canopy nitrogen content in rice based on hyperspectral measurements,” *Remote Sens. Environ.* **126**, 210–221 (2012).
23. Y. Tian et al., “Assessing newly developed and published vegetation indices for estimating rice leaf nitrogen concentration with ground- and space-based hyperspectral reflectance,” *Field Crops Res.* **120**(2), 299–310 (2011).
24. Q. Yasir et al., “Spectral indices for tracing leaf water status with hyperspectral reflectance data,” *J. Appl. Remote Sens.* **17**(1), 1931–3195 (2023).
25. K. Zhang et al., “A new canopy chlorophyll index-based paddy rice critical nitrogen dilution curve in eastern China,” *Field Crops Res.* **266**, 108139 (2021).
26. L. Chen et al., “Latitudinal patterns of leaf carbon, nitrogen, and phosphorus stoichiometry in Phyllostachys propinqua McClure across Northern China,” *Forests* **14**(11), 2243 (2023).
27. J. Sun et al., “The morphology and nutrient content drive the leaf carbon capture and economic trait variations in subtropical bamboo forest,” *Front. Plant Sci.* **14**, 1137487 (2023).
28. Y. Zhang et al., “Environmental drivers of the leaf nitrogen and phosphorus stoichiometry characteristics of critically endangered Acer catalpifolium,” *Front. Plant Sci.* **13**, 1052565 (2022).
29. D. Xiao et al., “Inversion study of cadmium content in soil based on reflection spectroscopy and MSC-ELM model,” *Spectrochim. Acta A: Mol. Biomol. Spectrosc.* **283**, 121696 (2022).
30. A. Savitzky and M. J. E. Golay, “Smoothing and differentiation of data by simplified least squares procedures,” *Anal. Chem.* **36**(8), 1627–1639 (1964).
31. O. Abdel-Aziz et al., “Application of Savitzky-Golay differentiation filters and Fourier functions to simultaneous determination of cefepime and the co-administered drug, levofloxacin, in spiked human plasma,” *Spectrochim. Acta A: Mol. Biomol. Spectrosc.* **139**, 1627–1639 (2015).
32. P. Fu et al., “A new three-band spectral and metal element index for estimating soil arsenic content around the mining area,” *Process Saf. Environ. Prot.* **157**, 27–36 (2022).
33. R. Wang et al., “Estimation of winter wheat nitrogen nutrition index hyperspectral remote sensing,” *Trans. Chin. Soc. Agric. Eng.* **30**(19), 191–198 (2014).
34. A. K. Seema Ghosh, B. S. Das, and N. Reddy, “Application of VIS-NIR spectroscopy for estimation of soil organic carbon using different spectral preprocessing techniques and multivariate methods in the middle Indo-Gangetic plains of India,” *Geoderm. Reg.* **23**, e00349 (2020).
35. L. Chen et al., “Development of a soil heavy metal estimation method based on a spectral index: combining fractional-order derivative pretreatment and the absorption mechanism,” *Sci. Total Environ.* **813**, 151882 (2021).
36. Z. Zhang et al., “Quantitative estimation of soil organic matter content using three-dimensional spectral index: a case study of the Ebinur lake basin in Xinjiang,” *Spectrosc. Spect. Anal.* **40**(5), 1514–1522 (2020).
37. M. L. Clark and D. A. Roberts, “Species-level differences in hyperspectral metrics among tropical rainforest trees as determined by a tree-based classifier,” *Remote Sens.* **4**(6), 1820–1855 (2012).

38. L. He et al., "Improved remote sensing of leaf nitrogen concentration in winter wheat using multi-angular hyperspectral data," *Remote Sens. Environ.* **174**, 122–133 (2016).

Xuehai Tang graduated from Beijing Forestry University, Beijing, in 2011 and received his PhD in forestry equipment engineering. He now works at Anhui Agricultural University in Anhui Province, China. His research interests are quantitative estimation and evaluation of hyperspectral remote sensing, dynamic monitoring of forest resources, forestry 3S technology application, etc. He was a visiting scholar at Colorado State University in 2017. He has published about 20 papers.

Biographies of the other authors are not available.