Xu F *et al.* (2024)
Notulae Botanicae Horti Agrobotanici Cluj-Napoca
Volume 52, Issue 1, Article number 13565
DOI:10.15835/nbha52113565
Research Article

Construction of cotton leaf nitrogen content estimation model based on the PROSPECT model

Feng XU[1a], Yiren DING[1b], Shizhe QIN[1], Hongyu WANG[1], Lu WANG[1], Yiru MA[1], Xin LV[1], Ze ZHANG[1*], Bing CHEN[2*]

[1]*Shihezi University College of Agriculture/ The Key Laboratory of Oasis Eco-agriculture, Xinjiang Production and Construction Group, Shihezi 832003, China; 20212012017@stu.shzu.edu.cn; dingyiren@stu.shzu.edu.cn; 20212012024@stu.shzu.edu.cn; wanghongyu@stu.shzu.edu.cn; wanglu98@stu.shzu.edu.cn; 20192012009@stu.shzu.edu.cn; lxshz@126.com; zhangze1227@shzu.edu.cn (*corresponding author)*

[2]*Xinjing Academy of Agricultural and reclamation science, Shihezi 832003, China; zyrcb@126.com (*corresponding author)*

[a,b]*These authors contributed equally to the work*

Abstract

Leaf nitrogen content (LNC) is an important index to measure the nitrogen deficiency in cotton. The rapid and accurate monitoring of LNC is of great significance for understanding the growth status of cotton and guiding precise fertilization in the field. At present, the hyperspectral technology monitoring of LNC is very mature, but it is interfered with by external factors such as shadow and soil, and data acquisition is still dependent on manpower. Therefore, on the basis of clarifying the correlation and quantitative relationship between physiological parameters and cotton LNC, the 400-2500 nm spectral curve was simulated based on PROSPECT-5 model. Combined with the measured spectra, the sensitive bands of leaf nitrogen content were screened, and four machine learning algorithms based on the reflectance of the sensitive bands were compared to construct a model for the estimation of LNC in cotton and determine the optimal model. The results show the following: (1) The parameter with the best correlation with nitrogen content was Cab, and the linear relationship was $y=0.3942x+12.521$, $R^2=0.81$, RMSE=12.87 g/kg. (2) The shuffled frog leaping algorithm (SFLA) and the successive projections algorithm (SPA) were used to screen the relevant bands sensitive to LNC. SFLA selected nine characteristic bands, mainly distributed between 700 and 750 nm. SPA screened seven characteristic bands, mainly distributed between 670 and 760 nm. The characteristic bands of both screening methods were distributed near the red edge. (3) Based on the sensitive bands, the four machine learning algorithms were compared. Among them, the band modeling of SFLA screening under the random forest (RF) algorithm was the best (modeling set $R^2=0.973$, RMSE=1.001 g/kg, rRMSE=3.41%, validation set $R^2=0.803$, RMSE=3.191 g/kg, rRMSE=10.85%). In summary, this study proposes an optimal estimation model of cotton leaf nitrogen content based on the radiative transfer model, which provides a theoretical basis for the dynamic, accurate, and non-destructive monitoring of cotton leaf nitrogen content.

Keywords: cotton; hyperspectral; leaf nitrogen content; machine learning; radiative transfer model

Received: 20 Dec 2023. Received in revised form: 12 Feb 2024. Accepted: 16 Feb 2024. Published online: 22 Feb 2024.

From Volume 49, Issue 1, 2021, Notulae Botanicae Horti Agrobotanici Cluj-Napoca journal uses article numbers in place of the traditional method of continuous pagination through the volume. The journal will continue to appear quarterly, as before, with four annual numbers.

Introduction

China is the world's largest consumer and importer of cotton, and the world's second-largest producer of raw cotton (Zhao *et al*., 2023). As the largest production base of high-quality commercial cotton in China (Kang *et al*., 2023), Xinjiang's output and product quality are particularly important for the development of national trade. The growth and development of cotton are inseparable from fertilizer application, and nitrogen, as the most important mineral nutrient element of cotton, is the key limiting factor for cotton growth and yield formation (Rehman *et al*., 2019). The unreasonable application of nitrogen fertilizer will not only reduce fertilizer use efficiency, but also lead to environmental pollution (Kaushal *et al*., 2011; Cui *et al*., 2020). Therefore, the real-time and accurate monitoring of leaf nitrogen content in the process of cotton growth is of great significance for formulating scientific and efficient nutrient management measures and optimizing cotton fertilization management. Traditional crop nitrogen content monitoring methods include morphological methods, physiological and biochemical analysis methods, etc. (Lemaire *et al*., 2019; Feng *et al*., 2020). Morphological diagnosis is based on visual observation only, which is prone to misdiagnosis and has poor timeliness; although the results of physiological and biochemical analysis and diagnosis are more accurate than those of morphological diagnosis, the destructive sampling of plants is needed, which is time-consuming, labor-consuming, and expensive.

Hyperspectral technology, as a non-contact and non-destructive method for monitoring crop biochemical information (Li *et al*., 2012; Feng *et al*., 2014; Chen *et al*., 2022), can obtain narrow and continuous fine spectral information from crops (Li *et al*., 2019) and can better establish the relationship between spectral data and crop nutrient biochemical information to build a high-precision monitoring model, which has been widely used in recent years. In previous studies, hyperspectral technology has achieved considerable success in crop nutrient estimation. Ryu *et al*. (2011) used hyperspectral technology to collect rice spectral data and nitrogen content information at the heading stage for three consecutive years, and established a hyperspectral prediction model for rice nitrogen content at the heading stage. The model R^2 reached 0.89. Liang *et al*. (2018) collected hyperspectral data from wheat leaves and established a new hyperspectral index based on LNC for estimation. The results showed that the accuracy of this index in estimating LNC was improved fivefold compared with the traditional hyperspectral index. Chen *et al*. (2022) established an LNC estimation model based on sensitive spectral bands and the LNC spectral index by studying the spectral information of cotton LNC. The results showed that the R^2 of the LNC estimation model based on the spectral index reached 0.72, but the stability of the model was greatly improved. The above research shows that hyperspectral technology can be used for crop nutrient monitoring, but at present, the acquisition of hyperspectral data is affected by factors such as light intensity, angle, soil, shadow, and human operation, and data acquisition is still dependent on manpower (Zhang *et al*., 2023).

The development of the radiation transfer model provides a new idea for studying the interaction between solar radiation and plant leaves (Jacquemoud *et al*., 2009). The PROSPECT model, as a uniform leaf model, has been widely used to retrieve crop biochemical and structural variables. It can accurately simulate the surface reflectance of different types of vegetation with high simulation accuracy (Haboudane *et al*., 2002; Upreti *et al*., 2020). It can be applied to the data retrieval of different remote sensing sensors and vegetation retrieval in different geographical regions and has a wide range of applicability. FéRET et al. (2019) compared the PROSPECT model and machine learning model to retrieve leaf mass per area (LMA) and leaf equivalent water thickness (EWT) data from broad-leaved leaves. The results showed that the estimation errors of LMA and EWT were reduced by 55% and 33%, respectively, when using the PROSPECT model. Relevant research shows that the spectral curve of wheat is successfully simulated by the PROSPECT model, and the relationship between the simulated reflectance of PROSPECT and LNC can be established using partial least squares regression (PLSR) (Yang *et al*., 2015). Hao *et al*. (2023) proposed a hybrid inversion algorithm by combining the PROSPECT model with a multi-layer perceptron (MLP) algorithm to accurately retrieve carotenoids, and the established model RMSE reached 3.14 g/kg. Many scholars have combined the PROSPECT model with simple empirical models to provide an accurate method for estimating leaf nitrogen content and improve the estimation accuracy (Wang *et al*., 2015; DeSa *et al*., 2021). Li *et al*. (2018) Constructed n-prosail model based on prospect model and sail model canopy model to retrieve crop nitrogen status at leaf and canopy scales. FéRET et al. (2021) Successfully retrieved the content of leaf protein and other carbon-based components using the prospect model, with an accuracy of 0.9.

But at present, there are few studies on the estimation of cotton leaf nitrogen content. To sum up, the main objectives of this study are as follows: (1) On the basis of clarifying the correlation and quantitative relationship between physiological parameters and cotton LNC, the 400-2500 nm spectral curve was simulated based on PROSPECT-5 model; (2) to use the shuffled frog leaping algorithm (SFLA) and successive projections algorithm (SPA) to screen the bands sensitive to nitrogen content by comparing the measured and simulated spectra; (3) to use four machine learning algorithms to model and compare the selected characteristic bands, and determine the best model for estimating cotton LNC. Therefore, this study takes cotton under different nitrogen content treatments as the research object, analyzes the correlation between the physiological parameters and LNC, simulates leaf spectral reflectance based on the PROSPECT-5 model, and constructs a stable cotton LNC estimation model combined with machine learning algorithms, resulting in the monitoring of cotton LNC content being more convenient and providing a theoretical basis for the dynamic, accurate, and nondestructive monitoring of cotton leaf nitrogen content.

Materials and Methods

Test area overview

The experiment was carried out in Shihezi University's teaching experimental field in Shihezi (85°59′41″E, 44°19′54″N), Xinjiang, in 2022. The area has an altitude of 429 meters, a frost-free period of 168-171 days, an annual sunshine duration of 2721-2818 hours, and an active accumulated temperature of ≥ 10 °C of 3570-3729 °C. The test soil texture in this area is mainly loam, and the basic physical and chemical properties are: alkali hydrolyzable nitrogen 60.88 mg/kg, organic matter 19.90 g/kg, available potassium 134 mg/kg, available phosphorus 17.95 mg/kg, and the previous crop was cotton. The tested variety was Xinluzao 53 (a tube-shaped plant with a dark green leaf color). Five nitrogen application levels were designed, namely, N0 (0 kg/hm^2), N1 (120 kg/hm^2), N2 (240 kg/hm^2), N3 (360 kg/hm^2), and N4 (480 kg/hm^2). The layout of the test area is shown in Figure 1. The experiment adopted a randomized block design. Each treatment was repeated three times with a total of 15 communities. The area of each community was 150 m^2, and protective rows were set throughout the area. The planting mode was "one film, three tubes and six rows", and the plant spacing was 10 cm + 66 cm + 10 cm. In the whole growth period of the cotton, urea (nitrogen content 46%) was applied with a water drip, and the phosphorus and potassium (potassium dihydrogen phosphate) was 150 kg/hm^2. Other field management measures were carried out according to the local high-yield cultivation requirements.

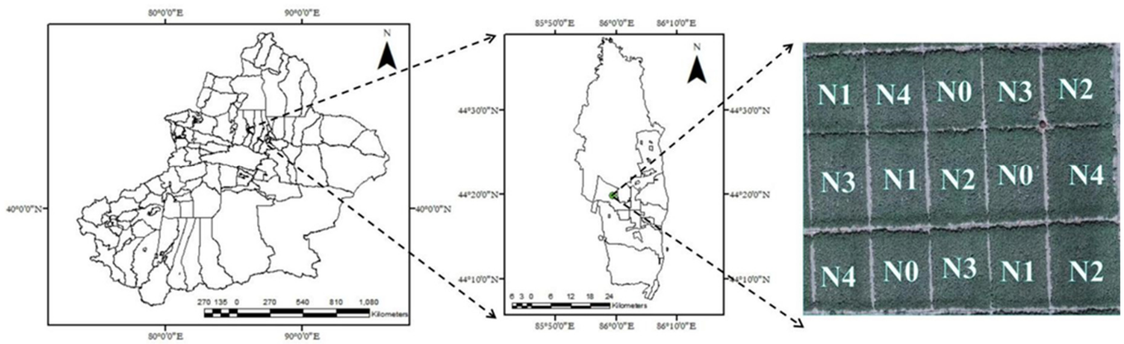

Figure 1. Layout of test area

Data acquisition

Reflectance measurement of cotton leaves

Leaf spectrum: the hyperspectral data instrument used for cotton leaves was the SR-3500 portable Spectral Evolution spectrometer from a spectral evolution company in the United States, with a spectral range of 350-2500 nm and an interval of 1 nm. The parameter data are shown in Table 1. The spectrometer was used for *in vitro* acquisition. During data acquisition, the veins of the cotton leaves were avoided. Three positions were measured for each leaf, the average value of three repetitions for each position was recorded as the spectral value of this position, and the average data of every three positions were taken as the spectral data for that leaf. Whiteboard correction was conducted before measuring the cotton leaves of different plants.

Table 1. Technical parameters of SR-3500 portable full-spectrum surface feature spectrometer

Technical Index	Technical Parameter	Technical Index	Technical Parameter
Spectral range	350-2500 nm	Spectral resolution	3.5 nm (350-1000 nm)
Number of spectral channels	2151		10 nm (1000-1900 nm)
Wavelength repeatability	0.1 nm		7 nm (1900-2500 nm)

Acquisition of agronomic parameters of cotton

All of the main stem leaves of the tested cotton plants were collected. Then, determination of cotton leaf area with Li-3100c instrument, the leaves were placed in an oven at 105 °C for 30 min, and baked at 80 °C to constant weight. The dry matter of the cotton leaves was weighed, and the total nitrogen content of the leaves was determined using Kjeldahl method after crushing (Bremner *et al*., 1960). Finally, the nitrogen content of the cotton leaves was calculated according to the dry matter weight.

Chlorophyll content: After obtaining the fresh weight of the leaves, a hole punch with a uniform diameter was used to remove the functional leaves from the leaf stems. The same number of discs were taken (with weights of approximately 0.1 g), an extraction solvent (anhydrous ethanol: acetone: water=4.5:4.5:1 mixture) was used as the blank for extraction, the absorbance values were measured under A470, A645, and A663 with a spectrophotometer, and the chlorophyll a, chlorophyll b, and other indicators were determined.

$$Ca = 12.21 * A663 - 2.81 * A645 \quad (1)$$

$$Cb = 20.13 * A645 - 5.03 * A663 \quad (2)$$

$$Cx = \frac{1000*A470-3.27*Ca-104*Cb}{229} \quad (3)$$

Note: Ca is the content of chlorophyll a (μ/cm^2), Cb is the content of chlorophyll b (μ/cm^2), and CX is the total content of chlorophyll (μ/cm^2).

The calculation of the equivalent water thickness (Cw) and dry matter content (Cm) of the blade is as follows:

$$Ewt\ \ (g/cm^2) = \frac{FW(g)\ -DW\ \ (g)}{LA\ \ (cm^2)} * 100\% \quad (4)$$

$$Dm\ \ (g/cm^2) = \frac{DW\ \ (g)}{LA\ \ (cm^2)} \quad (5)$$

where FW is the weight of the fresh leaves, DW is the weight of the dried leaves, and LA is the leaf area. EWT stands for equivalent water thickness, Dm stands for dry matter content.

Models and methods

PROSPECT-5 model

The PROSPECT-5 model (Feret *et al*., 2008) was used to simulate the spectral reflectance of the cotton leaves. The main input parameters include Cab, Car, Cw, Cm, and N. The basic assumption is that the blade is still regarded as a compact flat plate structure composed of N layers of the same material. In this model, the scattering part is represented by the refractive index n of light and the mesophyll structure parameter n, while the absorption part is mainly controlled by the concentration of chlorophyll, water content and other factors. The value range is shown in Table 2.

Table 2. Parameter ranges used in PROSPECT-5 model

Input Parameter	Name	Value Range	Unit
Cab	Chlorophyll content	10.0–70.0	μg/cm^2
Car	Carotenoid content	5-20	μg/cm^2
Cw	Water content	0.001–0.1	g/cm^2
Cm	Dry matter content	0.002–0.02	g/cm^2
N	Leaf structure parameter	1.0–3.0	

Spectral feature screening

For the simulated spectral curve, the shuffled frog leaping algorithm (SFLA) (Eusuff *et al*., 2003) and the successive projections algorithm (SPA) (Chen *et al*., 2023) were used to screen the characteristic bands, determine the characteristic wavelength of the parameter, calculate the correlation with LNC, and identify the characteristic bands sensitive to LNC.

Model building

In this study, four machine learning algorithms, adaptive boosting (AdaBoost) (Freund *et al*., 1997), random forest (RF) (Breiman, 2001), back propagation neural network (BPNN) (Rumelhart *et al*., 1986), and bootstrap aggregation (Bagging) (Breiman, 2001), were used to build the cotton LNC estimation model based on the radiative transfer model, and the optimal model was selected according to the model evaluation results. In this paper, 700 samples were randomly selected as the modeling set, and 300 samples were verified to establish the model.

Evaluation of the model

The determination coefficient (R^2), root-mean-square error (RMSE), and standardized root-mean-square error (rRMSE) were used as the accuracy evaluation indexes of the model. The closer the R^2 of the calculated validation set is to 1, the better the fitting of the model is. The smaller the RMSE and rRMSE are, the smaller the deviation between the simulated value and the measured value is, and the higher the accuracy of the model is. The above index calculation formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\bar{x}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \quad (8)$$

$$rRMSE = \frac{RMSE}{\bar{y}_i} * 100\% \quad (9)$$

where is the estimated nitrogen content (g/kg), is the measured nitrogen content (g/kg), and n is the number of samples. is the mean value of the estimated nitrogen content (g/kg), and is the mean value of the measured nitrogen content (g/kg).

Results and Discussion

Correlation analysis of agronomic parameters and nitrogen

In the whole growth period, 80 sample sizes were used to find the correlation between leaf nitrogen content and agronomic parameters, as shown in Figure 2. Chlorophyll, carotenoid, equivalent water thickness, and dry weight were used as the characterization parameters of cotton leaf size. It can be seen that the correlation between Cab, Car, Cw, and nitrogen content is significant at $p \leqq 0.01$; however, there was no significant correlation between Cm and nitrogen content. Among them, Cab had the best correlation with leaf nitrogen content (r=0.90), followed by carotenoid (r=0.76), and these two parameters were significantly positively correlated with nitrogen content. There was a significant negative correlation between equivalent water thickness and leaf nitrogen content (r=0.58). As shown in Figure 3, chlorophyll is in direct proportion to nitrogen, and the obtained equation is R^2=0.81, RMSE=12.87 g/kg.

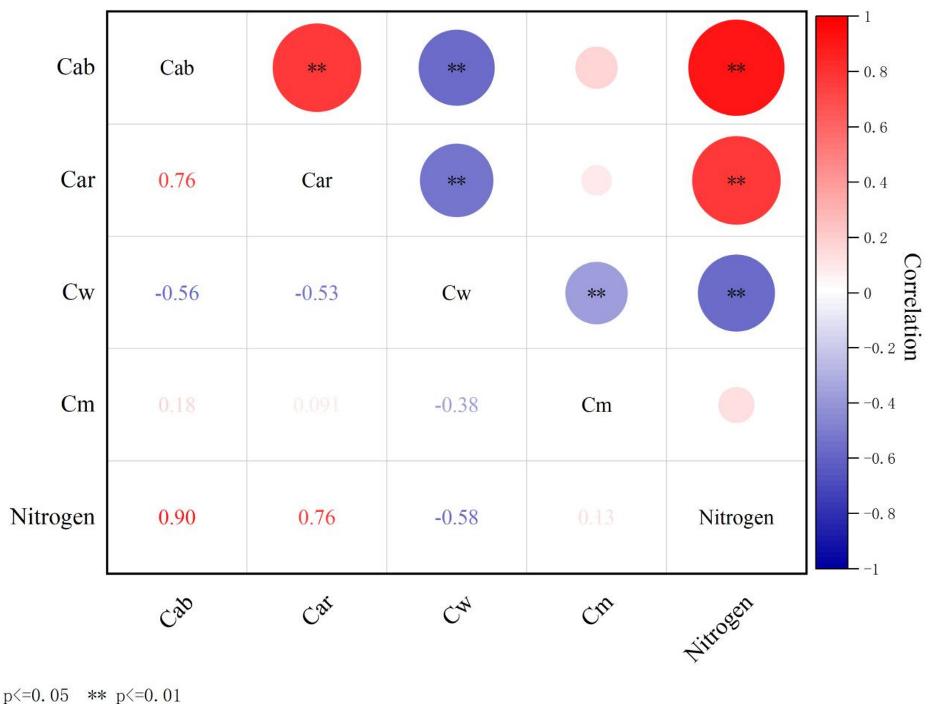

Figure 2. Correlation between nitrogen content and physiological parameters

Note: Cab (chlorophyll content, μg/cm²), Cw (equivalent water thickness, g/cm²), Cm (dry matter content, g/cm²), Nitrogen (leaf nitro gen content, g/kg).

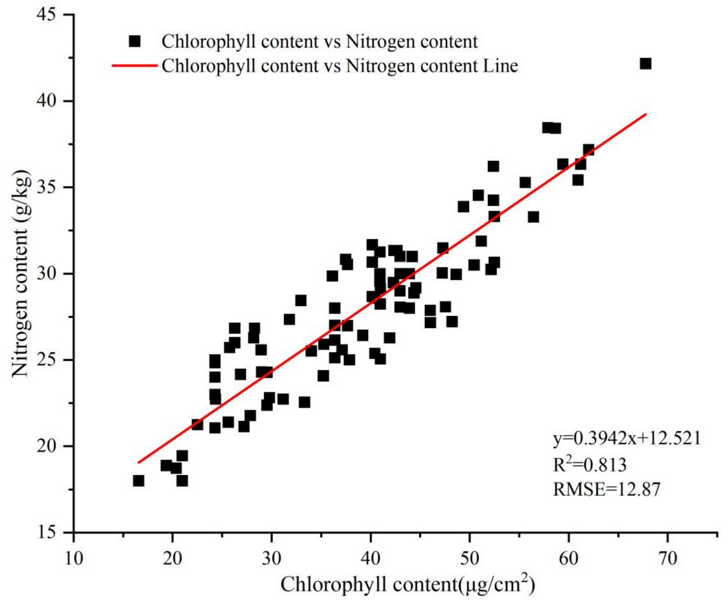

Figure 3. Linear relationship between chlorophyll and nitrogen content

Spectral simulation of cotton leaves based on PROSPECT-5

According to the measured parameters in this study and references, a total of 1000 sets of simulation parameters were generated using chlorophyll 17-67 (μg/cm^2) as the range to simulate the spectrum, and the model parameters of these sets were input into PROSPECT-5 to obtain a simulated canopy spectral dataset with a spectral interval of 1 nm. The spectral reflectance of 200 leaves collected from the 2022 field fertilizer experiment was used as the measured data, and 1000 simulated spectra and measured spectra were obtained. As shown in Figure 4, the spectral curves of the maximum, average, and minimum values were selected in the measured spectral set and compared with their simulated values, respectively. The maximum RMSE reached 0.0327, R^2=0.987; the average RMSE reached 0.024, R^2=0.993; and the minimum RMSE reached 0.035, R^2=0.987. When the PROSPECT model simulated single-leaf spectra, the trend of the simulated spectra was consistent with the measured spectra, and the reflectance was also consistent. The RMSE of the simulated spectra were all less than 0.05, indicating that the simulation accuracy was high.

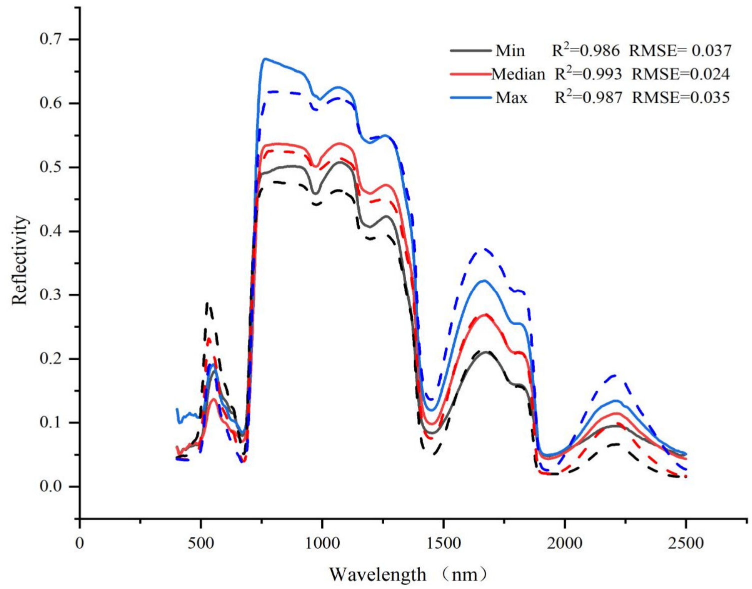

Figure 4. Comparison between measured spectra and PROSPECT-5-simulated spectra
Note: The solid line represents the measured spectrum, the dotted line represents the simulated spectrum, R^2 is the determination coefficient between the measured and simulated spectra, and RMSE is the root-mean-square error between the measured and simulated spectra.

Screening of characteristic bands

By calculating the correlation coefficient between the reflectance and chlorophyll content in each band, it can be seen that the absolute value of the correlation coefficient between the reflectance and chlorophyll content in the bands at 430-560 nm and 650-750 nm is greater than 0.9, indicating that the reflectance of the bands is strongly correlated with chlorophyll content in these two bands. Since there are 220 bands with an absolute value of correlation coefficient greater than 0.9, in order to reduce the calculation amount, feature bands were screened using a continuous projection algorithm. Seven feature bands were finally screened out, and nine feature bands were screened out using the random leapfrog algorithm. The screening results are shown in Table 3.

Table 3. Characteristic bands screened using different algorithms

Selection methods	Select band/nm	Number of bands
SPA	762、710、693、700、685、678、441	7
SFLA	747、743、742、707、518、751、746、761、754	9

Construction and verification of nitrogen content estimation model

The bands and nitrogen content selected based on the continuous algorithm were modeled. Each model established was validated, and the accuracy of each model was evaluated via computational modeling and predicting the RMSE. Four machine learning algorithms, AdaBoost, bagging, BPNN and RF, are used for analysis. Figure 5 shows a scatter plot of the predicted and true values of the model. It can be seen that the modeling accuracy of the RF algorithm is higher than the other three modeling methods; the modeling set R^2 reaches 0.965, RMSE=1.119 g/kg, rRMSE=3.77%, and the verification set R^2 reaches 0.782, RMSE=3.881 g/kg, rRMSE=14.50%. Secondly, the modeling set R^2 of BPNN algorithm reached 0.769, RMSE 2.603 g/kg, rRMSE 8.78%; Validation set R^2 reached 0.837, RMSE 2.974g/kg, rRMSE 10.03%. In the modeling of bands screened by the random leapfrog projection algorithm, as shown in Figure 6, the modeling accuracy of RF is higher than that of the other three modeling methods. The modeling set R^2 reaches 0.973, RMES=1.001 g/kg, rRMSE=3.41%, and the verification set R^2 reaches 0.803, RMSE=3.191 g/kg, rRMSE=10.85%. The method of band screening using random leapfrog is superior to the continuous projection algorithm on the verification set.

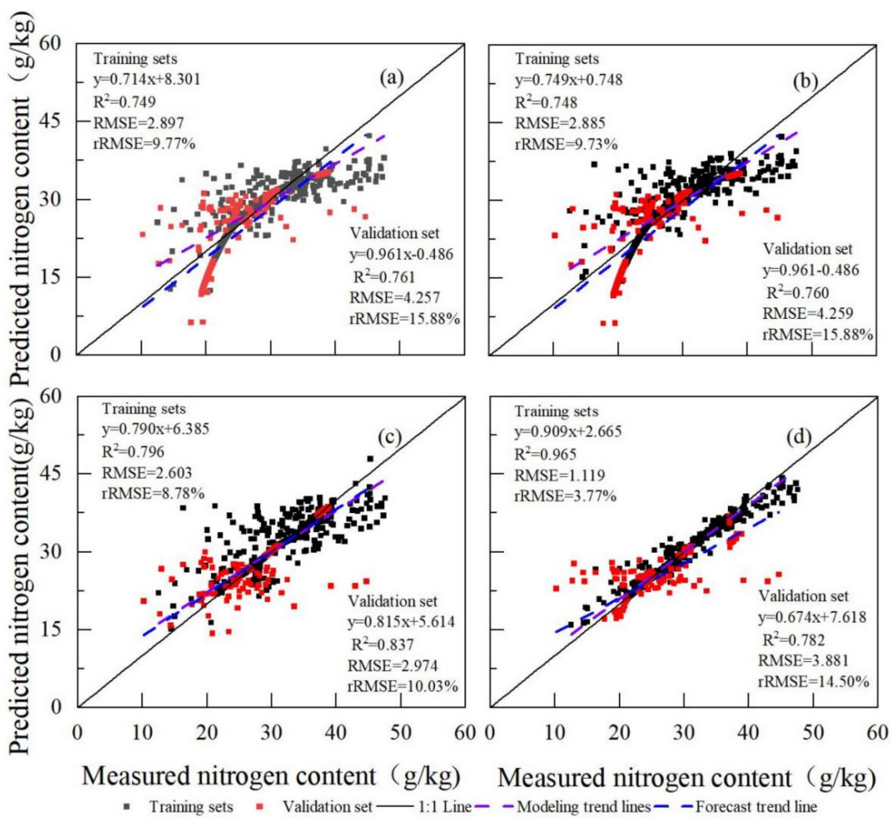

Figure 5. Comparison diagram of modeling and validation of different algorithms based on continuous projection

Note: (a) AdaBoost, (b) Bagging, (c) BPNN, (d) RF.

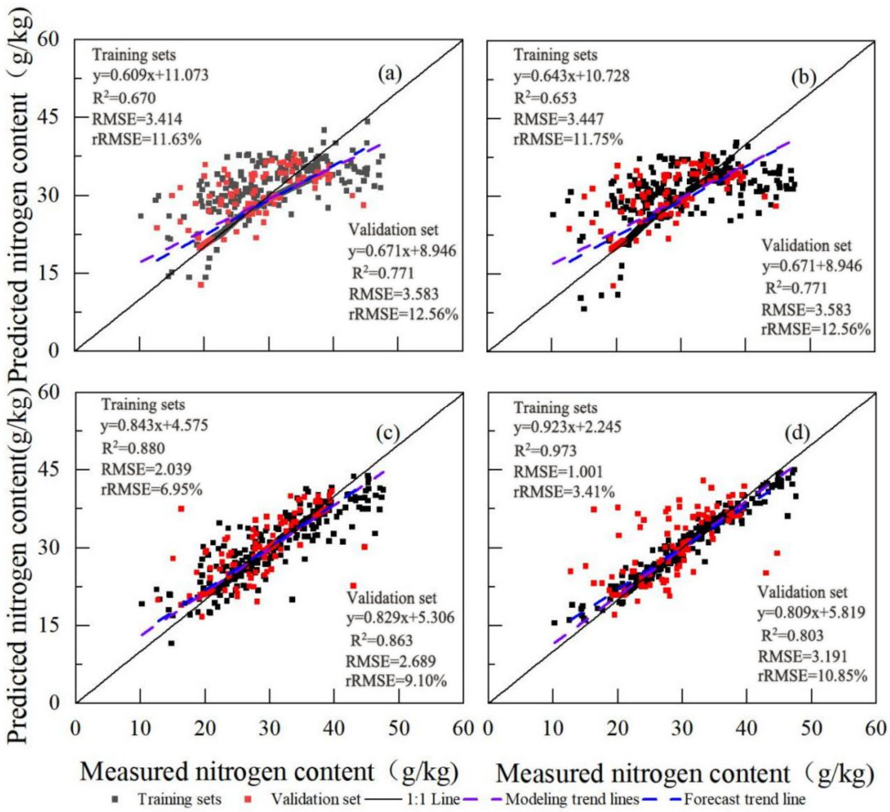

Figure 6. Comparison diagram of modeling and validation of different algorithms based on random frog jumps

Note: (a) AdaBoost, (b) Bagging, (c) BPNN, (d) RF.

Discussion

Based on the quantitative relationship between the input parameters in the PROSPECT-5 model and the nitrogen content in cotton leaves, this study selected spectral bands sensitive to nitrogen content changes in cotton leaves based on a continuous projection algorithm and random leapfrog algorithm. The estimation model of the nitrogen content in cotton leaves was constructed based on the selected bands and four machine-learning algorithms to provide a basis for nutrient management in cotton.

In terms of the relationship between chlorophyll and nitrogen content, nitrogen is a structural element of chlorophyll and the protein molecules that affect the formation of chloroplasts and the accumulation of chlorophyll in them. Many researchers have proven that there is a very close relationship between chlorophyll and nitrogen content (Evans, 1983; Sims and Gamon, 2002; Tucker, 2022). The total nitrogen necessary in leaves is 50% to 70% for chloroplast formation (Bohman *et al*., 2021), and the content of chlorophyll increases with the increase in nitrogen application (Tremblay *et al*., 2012). Relevant studies have shown that the nitrogen content in rice leaves is positively correlated with chlorophyll content (Wang *et al*., 2022). Based on this fact, a large number of studies have explored the feasibility of estimating crop nitrogen content by leaf chlorophyll content and achieved good definite research results (Eugenio *et al*., 2023). Therefore, measuring the chlorophyll content can indicate the real-time nitrogen nutrition status of plants (Knapp and Carter, 2023), and these results are consistent with the research results in this paper.

In terms of the simulation spectrum of the PROSPECT-5 model, PROSPECT-5 can simulate the spectra of leaves with different biochemical content and mesophyte structure based on the radiation transfer model within the leaves. On the one hand, it can provide a large number of datasets and improve the robustness of modeling. On the other hand, the input parameters of the model are controllable, which reduces the interference factor of the reflection spectrum and can be used for the spectral inversion of leaf structure parameters and chlorophyll and other chromatosomes (Berger *et al*., 2018). Yi *et al*. (2014) estimated the water content of cotton leaves and canopy based on the hyperspectral index and radiative transfer model, and the results showed that the estimation of cotton leaf spectral index combined with the radiative transfer model had high accuracy. Yu *et al*. (2019) simulated wheat canopy reflectance with a radiative transfer model, and the research results showed that the method of combining wavelet coefficients and partial least squares based on the radiative transfer model to invert chlorophyll had the highest accuracy. Wang *et al*. (2015) used PROSPECT-5 model results to invert leaf traits and indirectly estimate leaf nitrogen content. In this paper, when the PROSPECT-5 model was used to simulate cotton spectra, the RMSE was less than 0.05, the simulated spectra at 650-750 nm and 1450-2000 nm were consistent with the measured spectral reflectance, and the simulation effect in the near-infrared band was not good. However, with the development of the growth period, the changes in the hormones in the leaves are more complex, and the simulation accuracy and stability are reduced. There is a saturation effect similar to the previous empirical statistical model on the estimation of physiological and biochemical parameters such as chlorophyll and equivalent water thickness, which is represented by an underestimation of reflectivity. This conclusion is similar to that of a previous study on the spectral simulation of silage maize and winter wheat using a radiative transfer model (Danner *et al*., 2019).

In terms of band screening, previous studies have shown that the absorption band center of nitrogen in the spectral curve is mainly 430, 460, 640, 660, 910, 1510, and 2350 nm, including the red edge spectrum and SWIR spectrum range (Duanr *et al*., 2019). Due to the influence of environmental conditions and the phenological changes of the vegetation itself, the characteristic bands of nitrogen absorption often shift (Zeng *et al*., 2022). Fava *et al*. (2009) collected the hyperspectral high-resolution spectral reflectance data of grassland in the Mediterranean area and found that the ratio index of the near-infrared band (775-820 nm) and red edge spectrum range (740-770 nm) had the best performance in predicting nitrogen concentration. Feng et al. (2009) monitored nitrogen accumulation in wheat leaves using the red edge characteristic parameter. In this study, the feature bands were mainly screened near the red edge, similar to the NNI reference ratio index of winter wheat obtained by Wang *et al*. (2014), which also indicated the close relationship between the red edge parameters and agronomic components (Zheng *et al*., 2018). In this paper, sensitive bands for leaf nitrogen content are screened based on the simulated spectrum, and the influence of some environmental factors on the measured spectral curve is eliminated to find a more stable spectral curve and sensitive band for leaf nitrogen content.

Some studies have shown that the RF algorithm has the best prediction effect in predicting nitrogen content in cotton (Abdel-Rahman *et al*., 2012; Yin *et al*., 2022; Zhou *et al*., 2023). This paper shows that RF has the best modeling accuracy no matter which band is selected by the continuous projection algorithm or the band selected by the random leapfrog algorithm. RF is a combinatorial classification algorithm belonging to ensemble learning; the core idea of ensemble learning is to combine several weak (base) classifiers to obtain a strong classifier with significantly superior classification performance. It performs well and has great advantages compared with other algorithms. For unbalanced datasets, random forest can balance out errors. When there is a classification imbalance, random forest can provide an effective method to balance the errors of datasets. The random forest algorithm has a strong anti-interference ability, and its anti-overfitting ability is also strong.

In summary, this study was based on PROSPECT-5 simulation data to expand the amount of data and avoid the measured spectra affected by the external influences of inaccurate measurement, small modeling data, and other problems, to screen the characteristic bands and construct a cotton leaf nitrogen content estimation model. The verification effect was good. However, this study only conducted spectral simulation based on the quantitative relationship between chlorophyll content and leaf nitrogen content. In future studies, the effects of other physiological and environmental factors and noise on spectral reflectance, as well as the quantitative relationship with nitrogen nutrition, will be considered to improve the accuracy of the simulation spectrum.

Conclusions

In this study, a combined feature screening and machine learning method based on the PROSPECT-5 model was established to construct a mixed model for cotton leaf nitrogen estimation. The conclusions are as follows:

(1) It was determined that the best correlation between physiological parameters and cotton LNC was cab, and the linear relationship was established ($y=0.3942x+12.521$ $R^2=0.81$, RMSE=12.87 g/kg). Based on the prospect-5 model, the spectral curve of 400-2500 nm was simulated. The simulation results showed that the reflectance of measured and simulated leaves was sensitive to the nitrogen content of leaves near the red edge, but not sensitive to chlorophyll.

(2) SFLA and SPA were used to screen the relevant bands sensitive to LNC. SFLA screened out nine characteristic bands, mainly distributed in the range of 700-750 nm. SPA screened out seven characteristic bands, mainly distributed between 670-760 nm. The characteristic bands of the two screening methods were distributed near the red edge to predict the leaf nitrogen content.

(3) In this paper, four machine learning algorithms were used to model the selected feature bands, and the conclusion is drawn that the RF algorithm is the most accurate in modeling the selected feature bands regardless of the continuous projection algorithm or the random leapfrog algorithm. Among them, the band selected by the random jumping frog algorithm had the best modeling effect (modeling set $R^2=0.97$, RMSE=1 g/kg, rRMSE=3%; verification set $R^2=0.80$, RMSE=3.20 g/kg, rRMSE=10%).

Authors' Contributions

FX and YRD were responsible for determining some indicators and writing this manuscript. SZQ, LW and YHW conducted experiments and collected all data sets. SZQ and RYM are responsible for editing and typesetting. XL is responsible for revision. Corresponding authors ZZ and BC are responsible for the revision and quality control of the paper. All authors read and approved the final manuscript.

Ethical approval (for researches involving animals or humans)

Not applicable.

Acknowledgements

This work was supported by the National Natural Science Foundation of China, China (Grant No. 42061058), "Tianshan Talents" Training Program, China, Key Scientific and Technological Research Program of XPCC, China (Grant No. 2023AA008), The Corps "Strong Green" Scientific and Technological Innovation Backbone Personnel Project, China (Grant No. 2022CB002-001), Eight Division Shihezi City Key Areas of Innovation Team Plan, China (Grant No. 2023TD01), and XPCC Agricultural Innovation Project Special Fund Project(NCG202304).

Conflict of Interests

The authors declare that there are no conflicts of interest related to this article.

References

Abdel-Rahman EM, Ahmed FB, Ismail R (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. International Journal of Remote Sensing 34(2):712-728. *http://dx.doi.org/10.1080/01431161.2012.713142*

Berger K, Atzberger C, Danner M, Wocher M, Mauser W, Hank T (2018). Model-based optimization of spectral sampling for the retrieval of crop variables with the PROSAIL model. Remote Sensing 10(12):2063. *https://doi.org/10.3390/rs10122063*

Bohman BJ, Rosen CJ, Mulla DJ (2021). Relating nitrogen use efficiency to nitrogen nutrition index for evaluation of agronomic and environmental outcomes in potato. Field Crops Research 262:108041. *https://doi.org/10.1016/j.fcr.2020.108041*

Breiman L (1996). Bagging predictors. Machine Learning 24:123-140. *https://doi.org/10.1007/bf00058655*

Breiman L (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199-231. *https://doi.org/10.1214/ss/1009213726*

Bremner JM (1960). Determination of nitrogen in soil by the Kjeldahl method. The Journal of Agricultural Science 55(1):11-33. *https://doi.org/10.1017/s0021859600021572*

Chen X, Li F, Chang Q (2023). Combination of continuous wavelet transform and successive projection algorithm for the estimation of winter wheat plant nitrogen concentration. Remote Sensing 15(4):997. *https://doi.org/10.3390/rs15040997*

Chen X, Lv X, Ma L, Chen A, Zhang Q, Zhang Z (2022). Optimization and validation of hyperspectral estimation capability of cotton leaf nitrogen based on SPA and RF. Remote Sensing 14(20):5201. *https://doi.org/10.3390/rs14205201*

Cui M, Zeng L, Qin W, Feng J (2020). Measures for reducing nitrate leaching in orchards: A review. Environmental Pollution 263(Pt B):114553. *https://doi.10.1016/j.envpol.2020.114553*

Danner M, Berger K, Wocher M, Mauser W, Hank T (2019). Fitted PROSAIL parameterization of leaf inclinations, water content and brown pigment content for winter wheat and maize canopies. Remote Sensing 11(10):1150. *https://doi.org/10.3390/rs11101150*

de Sa NC, Baratchi M, Hauser LT, van Bodegom P (2021). Exploring the impact of noise on hybrid inversion of prosail rtm on sentinel-2 data[J]. Remote Sensing 13(4):648. *https://doi.org/10.3390/rs13040648*

Duan DD, Zhao CJ, Li ZH, Yang GJ, Yang WD (2019). Estimating total leaf nitrogen concentration in winter wheat by canopy hyperspectral data and nitrogen vertical distribution. Journal of Integrative Agriculture 18(7):1562-1570. *https://doi.org/10.1016/s2095-3119(19)62686-9*

Eugenio FC, Grohs M, Schuh M, Venancio LP, Schons C, Badin TL, ... Fantinel RA (2023). Estimated flooded rice grain yield and nitrogen content in leaves based on RPAS images and machine learning. Field Crops Research 292:108823. *https://doi.org/10.1016/j.fcr.2023.108823*

Eusuff MM, Lansey KE (2003). Optimization of water distribution network design using the shuffled frog leaping algorithm. Journal of Water Resources Planning and Management 129(3):210-225. *https://doi.org/10.1061/(ASCE)0733-9496(2003)129:3(210)*

Evans JR (1983). Nitrogen and photosynthesis in the flag leaf of wheat (*Triticum aestivum* L.). Plant Physiology 72(2):297-302. *https://doi.org/10.1104/pp.72.2.297*

Fava F, Colombo R, Bocchi S, Meroni M, Sitzia M, Fois N, Zucca C (2009). Identification of hyperspectral vegetation indices for Mediterranean pasture characterization. International Journal of Applied Earth Observation and Geoinformation 11(4):233-243. *https://doi.org/10.1016/j.jag.2009.02.003*

Feng D, Xu W, He Z, Zhao W, Yang M (2020). Advances in plant nutrition diagnosis based on remote sensing and computer application. Neural Computing and Applications 32:16833-16842. *https://doi.org/10.1007/s00521-018-3932-0*

Feng W, Guo BB, Wang ZJ, He L, Song X, Wang YH, Guo TC (2014). Measuring leaf nitrogen concentration in winter wheat using double-peak spectral reflection remote sensing data. Field Crops Research 159:43-52. *https://doi.org/10.1016/j.fcr.2014.01.010*

Feng W, Zhu Y, Yao X, Tian Y, Guo T, Cao W (2009). Monitoring nitrogen accumulation in wheat leaf with red edge characteristics parameters. Transactions of the Chinese Society of Agricultural Engineering 25(11):194-201. *https://doi.org/10.3969/j.issn.1002-6819.2009.11.035*

Feret JB, François C, Asner GP, Gitelson AA, Martin RE, Bidel LP, ... Jacquemoud S (2008). PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments. Remote Sensing of Environment 112(6):3030-3043. *https://doi.org/10.1016/j.rse.2008.02.012*

Féret JB, Berger K, De Boissieu F, Malenovský Z (2021) PROSPECT-PRO for estimating content of nitrogen-containing leaf proteins and other carbon-based constituents. Remote Sensing of Environment 252:112173. *https://doi.org/10.1016/j.rse.2020.112173*

Féret JB, Le Maire G, Jay S, Berveiller D, Bendoula R, Hmimina G, ... Lefèvre-Fonollosa MJ (2019). Estimating leaf mass per area and equivalent water thickness based on leaf optical properties: Potential and limitations of physical modeling and machine learning. Remote Sensing of Environment 231:110959. *https://doi.org/10.1016/j.rse.2018.11.002*

Freund Y, Schapire RE (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1):119-139. *https://doi.org/10.1006/jcss.1997.1504*

Haboudane D, Miller JR, Tremblay N, Zarco-Tejada PJ, Dextraze L (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. Remote Sensing of Environment 81(2-3):416-426. *https://doi.org/10.1016/s0034-4257(02)00018-4*

Hao W, Sun J, Zhang Z, Zhang K, Qiu F, Xu J (2023). Novel hybrid model to estimate leaf carotenoids using multilayer perceptron and PROSPECT simulations. Remote Sensing 15(20):4997. *https://doi.org/10.3390/rs15204997*

Jacquemoud S, Verhoef W, Baret F, Bacour C, Zarco-Tejada PJ, Asner GP, ... Ustin SL (2009). PROSPECT+SAIL models: A review of use for vegetation characterization. Remote Sensing of Environment 113:S56-S66. *https://doi.org/10.1016/j.rse.2008.01.026*

Kaushal SS, Groffman PM, Band LE, Elliott EM, Shields CA, Kendall C (2011). Tracking nonpoint source nitrogen pollution in human-impacted watersheds. Environmental Science & Technology 45(19):8225-8232. *https://doi.org/10.1021/es200779e*

Knapp AK, Carter GA (2001). Leaf optical properties in higher plants: lin Cing spectral characteristics to stress and chlorophyll concentration. American Journal of Botany 88(4):677e684. *https://doi.org/10.2307/2657068*

Kang X, Huang C, Chen JM, Lv X, Wang J, Zhong T, ... Tong Q (2023). The 10-m cotton maps in Xinjiang, China during 2018–2021. Scientific Data 10(1):688. *https://doi.org/10.1038/s41597-023-02584-3*

Lemaire G, Sinclair T, Sadras V, Bélanger G (2019). Allometric approach to crop nutrition and implications for crop diagnosis and phenotyping. A review. Agronomy for Sustainable Development 39:1-17. *https://doi.org/10.1007/s13593-019-0570-6*

Li F, Mistele B, Hu Y, Yue X, Yue S, Miao Y, ... Schmidhalter U (2012). Remotely estimating aerial N status of phenologically differing winter wheat cultivars grown in contrasting climatic and geographic zones in China and Germany. Field Crops Research 138:21-32. *https://doi.org/10.1016/j.fcr.2012.09.002*

Li F, Wang L, Liu J, Wang Y, Chang Q (2019). Evaluation of leaf N concentration in winter wheat based on discrete wavelet transform analysis. Remote Sensing 11(11):1331. *https://doi.org/10.3390/rs11111331*

Li Z, Jin X, Yang G, Drummond J, Yang H, Clark B, ... Zhao C (2018) Remote sensing of leaf and canopy nitrogen status in winter wheat (*Triticum aestivum* L.) based on N-PROSAIL model[J]. Remote Sensing 10(9):1463. *https://doi.org/10.3390/rs10091463*

Liang L, Di L, Huang T, Wang J, Lin L, Wang L, Yang M (2018). Estimation of leaf nitrogen content in wheat using new hyperspectral indices and a random forest regression algorithm. Remote Sensing 10(12):1940. *https://doi.org/10.3390/rs10121940*

Ma C, Liu M, Ding F, Li C, Cui Y, Chen W, Wang Y (2022). Wheat growth monitoring and yield estimation based on remote sensing data assimilation into the SAFY crop growth model. Scientific Reports 12(1):5473.

Rehman A, Farooq M (2019). Morphology. Physiology and Ecology of cotton. Cotton Production 23-46. *https://doi.org/10.1002/9781119385523.ch2*

Rumelhart DE, Hinton GE, Williams RJ (1986). Learning representations by back-propagating errors. Nature 323(6088):533-536. *https://doi.org/10.1038/323533a0*

Ryu C, Suguri M, Umeda M (2011). Multivariate analysis of nitrogen content for rice at the heading stage using reflectance of airborne hyperspectral remote sensing. Field Crops Research 122(3):214-224. *https://doi.org/10.1016/j.fcr.2011.03.013*

Sims DA, Gamon J A (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sensing of Environment 81(2-3):337-354. *https://doi.org/10.1016/s0034-4257(02)00010-x*

Tremblay N, Wang Z, Cerovic ZG (2012). Sensing crop nitrogen status with fluorescence indicators. A review. Agronomy for Sustainable Development 32:451-464. *https://doi.org/10.1007/s13593-011-0041-1*

Tucker M (2004). Primary nutrients and plant growth. In: Essential Plant Nutrients SCRIBD. Dept. of Agriculture, North Carolina, USA.

Upreti D, Pignatti S, Pascucci S, Tolomio M, Huang W, Casa R (2020). Bayesian calibration of the Aquacrop-OS Model for durum wheat by assimilation of canopy cover retrieved from VENμS satellite data. Remote Sensing 12(16):2666. *https://doi.org/10.3390/rs12162666*

Wang BF, Yu ZY, Cheng JP (2022). Research progress on effects of nitrogen on yield and quality formation of rice. Journal of Huazhong Agricultural University 41(01):76-83. *https://doi.org/10.13300/j.cnki.hnlkxb.2022.01.007*

Wang R, Song X, Li Z, Yang G, Guo W, Tan C, Chen L (2014). Estimation of winter wheat nitrogen nutrition index using hyperspectral remote sensing. Transactions of the Chinese Society of Agricultural Engineering 30(19):191-198. *https://doi.org/10.3969/j.issn.1002-6819.2014.19.023*

Wang Z, Skidmore AK, Darvishzadeh R, Heiden U, Heurich M, Wang T (2015). Leaf nitrogen content indirectly estimated by leaf traits derived from the PROSPECT model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8(6):3172-3182. *https://doi.org/10.1109/JSTARS.2015.2422734*

Yang G, Zhao C, Pu R, Feng H, Li Z, Li H, Sun C (2015). Leaf nitrogen spectral reflectance model of winter wheat (*Triticum aestivum*) based on PROSPECT: simulation and inversion. Journal of Applied Remote Sensing 9(1):095976-095976. *https://doi.org/10.1117/1.JRS.9.095976*

Yi Q, Wang F, Bao A, Jiapaer G (2014). Leaf and canopy water content estimation in cotton using hyperspectral indices and radiative transfer models. International Journal of Applied Earth Observation and Geoinformation 33:67-75. *https://doi.org/10.1016/j.jag.2014.04.019*

Yin C, Lv X, Zhang L, Ma L, Wang H, Zhang L, Zhang Z (2022). Hyperspectral UAV images at different altitudes for monitoring the leaf nitrogen content in cotton crops. Remote Sensing 14(11):2576. *https://doi.org/10.3390/rs14112576*

Yu KH, Yang GJ, Wang CC (2019). Winter wheat chlorophyll inversion using ground hyperspectral and PROSAIL models. Science of Surveying and Mapping, 44(11):96-102. *https://doi.org/10.16251/j.cnki.1009-2307.2019.11.014*

Zhao H, Chen Y, Liu J, Wang Z, Li F, Ge X (2023). Recent advances and future perspectives in early-maturing cotton research. New Phytologist 237(4):1100-1114. *https://doi.org/10.1111/nph.18611*

Zeng Y, Hao D, Huete A, Dechant B, Berry J, Chen JM, … Chen M (2022). Optical vegetation indices for monitoring terrestrial ecosystems globally. Nature Reviews Earth & Environment 3(7):477-493. *https://doi.org/10.1038/s43017-022-00298-5*

Zhang Y, Xiao J, Yan K, Lu X, Li W, Tian H, … Lan Y (2023). Advances and developments in monitoring and inversion of the biochemical information of crop nutrients based on hyperspectral technology. Agronomy 13(8):2163. *https://doi.org/10.3390/agronomy13082163*

Zheng T, Liu N, Wu L, Li M, Sun H, Zhang Q, Wu J (2018). Estimation of chlorophyll content in potato leaves based on spectral red edge position. IFAC-PapersOnLine 51(17):602-606. *https://doi.org/10.1016/j.ifacol.2018.08.131*

Zhou X, Yang M, Chen X, Ma L, Yin C, Qin S, … Zhang Z (2023). Estimation of cotton nitrogen content based on multi-angle hyperspectral data and machine learning models. Remote Sensing 15(4):955. *https://doi.org/10.3390/rs15040955*