Nitrogen Content Estimation of Apple Leaves Using Hyperspectral Analysis

Weihao Sun,1 Dongwei Wang,1 Ning Jin,2 Shusheng Xu,1 and Haoran Bai1

1School of Mechanical and Electrical Engineering, Qingdao Agricultural University, Qingdao 266109, China
2School of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang 110866, China

Correspondence should be addressed to Haoran Bai; baihaoran111@126.com

Received 4 June 2021; Revised 11 September 2021; Accepted 20 September 2021; Published 6 October 2021

Copyright © 2021 Weihao Sun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Leaf nitrogen content (LNC) is an important factor reflecting the growth quality of plants. We estimated the nitrogen content of apple leaves using hyperspectral wavelength analysis using the differential spectrum, differential spectrum transformation, and vegetation spectrum index with different derivative gaps. We then used the characteristic wavelengths extracted via the correlation coefficient method as the input vectors to the gradient boosting decision tree (GBDT) model for analysis and performed cross-validation to optimize the inversion model parameters. We analyzed the results with different input variables and loss functions and compared the GBDT model with other mainstream algorithm models. The results show that the $R^2$ value of the optimized GBDT inversion model is higher than that obtained using the random forest (RF) and support vector regression (SVR) models. Thus, the GBDT model is accurate, and the characteristic wavelength analysis is helpful for the tasks of real-time monitoring and detection of apple tree health.

1. Introduction

Nitrogen content is an important indicator of plant health and nutritional value, with the lack of nitrogen greatly reducing the photosynthetic yield of crops [1, 2]. Most traditional detection methods for leaf nitrogen content (LNC) use chemometric detection methods, such as the Kjeldahl method, but these conventional methods are time-consuming and complex. Recent advances in hyperspectral remote sensing technology make use of the principle that crops cause changes in reflectance spectra when subjected to nitrogen stress and have led to significant progress in rapid detection of LNC. Existing studies have examined the properties of different indices for wheat, rice, corn, and other crops [3–9].

Current LNC detection research generally uses multiple vegetation indices or hyperspectral sensitivity band reflectance as estimation factors, with the spectral bands concentrated in the visible near-infrared shortwave range (350–1100 nm) [10–14]. However, using the full-band original reflectance spectrum as an estimation factor reduces the generalization accuracy of the inversion model because the original reflectance spectrum usually includes background soil. Other researchers have shown that spectral derivative transformation effectively reduces soil background information and low-frequency noise, making the spectral estimation model more reliable [15].

At present, differential spectroscopy and the spectral index it constructs have been widely and successfully applied, and they are considered to be the best method to estimate plant physiological parameters [16, 17]. However, this method also greatly reduces the number of spectral input variables of the nitrogen content estimation model. In order to counteract the negative influence from the reduction of model input variables, researchers have performed secondary transformations on the constructed spectral index or differential spectrum, such as reciprocal, logarithm, first derivative of the reciprocal, and first derivative of the logarithm to build as many estimation factors as possible [1, 18, 19]. Others have selected an algorithm with a small dependence on the number of input variables for the nitrogen content estimation model. If the input variables are
representative and significant, these estimation models also achieve good results, as with partial least squares regression \cite{18, 20}, SVR \cite{19}, RF \cite{1, 21}, and GBDT \cite{22}. In this study, we perform differential processing on the original spectral data of apple leaves using different derivative gaps, construct the spectral parameters, and model them using the GBDT algorithm to achieve an accurate inversion of the nitrogen content of apple leaves.

2. Materials and Methods

2.1. Experimental Area. The research object of this paper is apple leaves. The sample collection site is located in Xicheng Town, Qixia City, Shandong Province, China (120°45' 24” east longitude, 37°19' 20” north latitude), which is located in a hilly mountainous area at an altitude of 210 meters. The site is within the temperate monsoon climate zone and has a mean annual precipitation level of 640 mm–846 mm, with a mean annual number of sunshine hours of 2659.9 hours and a mean annual temperature of 11.4°C. The average age of trees is 7–8 years, and the apple tree variety is Red Fuji at maturity.

2.2. Sample Collection and Measurement. To evaluate our approach, we collected apple samples on April 20 (floriculture stage), May 20 (new shoots flourishing stage), June 20 (spring shoots stop growing stage), and September 20 (autumn shoots stop growing stage), 2020. The sample included 158 apple trees from 4 apple orchards. We performed random sampling and included the leaves from different growth conditions as far as possible. We chose 4 mature healthy leaves of similar size and color from the periphery of the canopy, put them in a resealable bag immediately after removing the leaves, placed them in a foam box after sealing and numbering, and brought them back to the laboratory to measure the spectrum and LNC.

We performed spectral scanning on the samples using an AvaSpec-ULS2048 spectrometer (manufactured by the Dutch company AvaSpec) between 350 and 1100 nm, with a resolution of 3 nm and a sampling interval of about 1 nm. Before taking the measurement, we wiped the leaves to be measured with absorbent cotton. During the measurement itself, we placed a single-layer of blades flat on the black rubber surface, set the spectrometer’s field of view angle to 25°, and aimed the probe toward the middle of the measuring blades, with distance between the probe and blade as 6 cm. In order to reduce the effect of environmental changes, we measured each sample 10 times and calculated the average (mean) of the data as shown in Figure 1. Number 1 was the floriculture stage, number 2 was the new shoots flourishing stage, number 3 was the spring shoots stop growing stage, and number 4 was the autumn shoots stop growing stage. The experiment collected a total of 632 samples and randomly divided them into two groups: 70% of them, or 442 samples, were used as the training set; the other 30%, or 190 samples, were used as the prediction set. We also determined LNCs using the Kjeldahl method with dry samples.

2.3. Spectral Data Processing. Converting the original spectrum is an important measure for improving the accuracy of spectral diagnosis, reducing redundant spectrum interference, and improving the signal-to-noise ratio. In this study, we performed spectral conversions on the original spectrum of the leaves with methods such as one-stage differentiation of different derivative gaps, logarithmic, reciprocal, reciprocal derivative, and logarithmic derivative, to verify commonly used plant spectral indices.

The first transformation was the first-order differential with a differential gap of 1–30. The differential transformation formula is

\[
FD_i = \frac{R_{i+w} - R_{i-w}}{2w},
\]

where \(FD_i\) represents the first-order differential value at wavelength \(i\), \(R_i\) represents the hyperspectral reflectance value at wavelength \(i\), and \(w\) represents the derivative gap value. We then performed correlation analysis on the nitrogen content of apple leaves and the transformed first-order differential spectrum values. The analysis results are shown in Figure 2.

According to the correlation analysis results, we determined that five sensitive wavelengths are significantly related to the nitrogen content of apple trees under 30 derivative gaps, and we then constructed the spectral parameters of the nitrogen content of apple leaves. From high to low, they are FDW_{6, 806}, FDW_{5, 837}, FDW_{6, 813}, FDW_{17, 415}, and FDW_{17, 1001} (A_i/B and A_i represent the first-order differential value at derivative gap \(i\), and \(B\) represents the wavelength).

The second transformation applied reciprocal and logarithmic spectrum vector transformations to the original spectrum vector to obtain the reciprocal and logarithmic differential spectrum vectors. Figure 3 shows the correlation analysis between the spectrum vector and the nitrogen content. The variance of the correlation coefficient of the differential spectrum vector is large, and there is no
significant difference between the mean value of the absolute value of the correlation coefficient of the differential spectrum vector and the reciprocal and logarithmic spectrum vectors. Therefore, we also selected the spectral vector at the 775 nm wavelength of the reciprocal spectrum and the spectral vector at the 801 nm wavelength of the logarithmic spectrum as the characteristic vector for the study.

Finally, we selected six spectral indices with clear physical meaning and high recognition for comparative analysis. The calculation methods and literature sources of these indexes are shown in Table 1. According to existing research, the bands used by these indexes are in the visible and near-infrared range.

| Spectral index | Formula | References |
|----------------|---------|------------|
| NDVI705        | \( \frac{R_{750} - R_{705}}{R_{750} + R_{705}} \) | [10]        |
| mNDVI705       | \( \frac{R_{750} - R_{715}}{R_{750} + R_{715} + 2R_{445}} \) | [10]        |
| VOG3           | \( \frac{R_{734} - R_{747}}{R_{715} + R_{734}} \) | [13]        |
| PRI            | \( \frac{R_{531} - R_{570}}{R_{531} + R_{570}} \) | [23]        |
| NDCI           | \( \frac{R_{762} - R_{527}}{R_{762} + R_{527}} \) | [24]        |
| RVI3           | \( \frac{R_{810}}{R_{660}} \) | [25]        |

Table 1: Hyperspectral indices for nitrogen content estimation.

Figure 4 shows a graph of the correlation coefficient between the hyperspectral index of apples and the nitrogen content of apple leaves for different derivative gaps. The graph shows that the correlation between the hyperspectral
index and the nitrogen content of apple leaves is irregular but with a low correlation coefficient when the derivative gap value is between 23 and 30. This indirectly shows that the derivative gap values higher than 30 have no practical use in finding sensitive spectral parameters. According to the analysis results, NDVI705_1, MNDVI_3, VOG3_23, PRI_1, NDCI_7, and RVI3_8 (A_B and A represent the vegetation spectral index, and B represents the derivative gap value) were selected as the spectral vector for nitrogen content estimation.

3. Results and Discussion

From the approach described above, we obtained a total of 13 feature vectors of the selected differential spectrum, the spectrum vector from the spectrum transformation, and the vegetation spectrum index and used them as the input vector of the gradient boosting (GBDT) model. GBDT [22] is a type of boosting algorithm consisting of three parts: DT (regression decision tree), GB (gradient boosting), and shrinkage. The algorithm decision result is composed of multiple decision trees. When constructing a subtree, the iterative decision tree uses the residuals formed after the results of the previous subtree construction as input data to construct the next subtree and predicts according to the order of subtree construction and then combines the prediction results to obtain the final results. The GBDT algorithm is suitable for low-dimensional data, handles nonlinear data, and supports some robust loss functions. It is also very robust to outliers. In this study, we optimized the maximum depth, loss function, number of iterations, and other model parameters that affect the estimation accuracy in GBDT through cross-validation, verified the rationality of feature selection, and combined the GBDT inversion model and other mainstream machine learning. We also compared and analyzed the prediction results of the algorithm.

The number of iterations represents the number of iterative enhancements to be executed, and the maximum depth represents the maximum depth of each regression estimator. We explored the influence of the number of iterations and the maximum depth in the range of 1–500 on the model, with the results shown in Figure 5. From these results, we set the number of iterations to 500 and the maximum depth to 5, with a final prediction evaluation index $R^2$ value of 0.88. In the setting of the number of iterations, although the test set error is small when the value is 100, the error on the training set is large with the $R^2$ value of only 0.60, indicating that the model is in an underfitting state. When the number of iterations is set to 350, the test error and training error reach a stable state, with the $R^2$ value of 0.86 and a difference with the final result of 0.02. However, considering the lower training cost of the algorithm model and the better prediction results, we set the number of iterations to 500. Values of the maximum depth run contrary to the number of iterations. When the maximum depth is set to 350, the test error and training error reach a stable state with an $R^2$ value of only 0.72. This is because the maximum depth represents the purity of the data. If the value is set too large, outliers affect the model and cause the model to overfit. Thus, we set the maximum depth to 5.

We considered four loss functions: least squares regression, least absolute deviation, a combination of least squares regression with least absolute deviation (Huber), and quantile regression. Figure 6 shows a comparison of the results for each function. The least squares regression had the highest degree of fit on the test set, with the $R^2$ value.
**Figure 5:** The influence of different number of iterations and the maximum depth values on the model.

**Figure 6:** Predictive analysis of apple leaf nitrogen content for different loss functions.
reaching 0.88 and the loss value of 0.033. Table 2 presents the
detailed prediction results of the model for each loss
function.

After selecting the model hyperparameters, we also
analyzed the importance of the input features in the GBDT
model to judge the rationality of the feature vector selection.
Those results are shown in Figure 7. The relative importance
of each feature differed between the training and test sets.

To further determine feature importance, we conducted
two sets of experiments. We separately imported the
differential spectral vector, the spectral vector obtained after spectral transformation, and the spectral index into the GBDT model as input features. The analysis results are shown in Figures 8–10. The detailed prediction results of the three are shown in Table 3. In terms of the prediction results, the $R^2$ value of the model using the differential spectrum vector alone as the input vector was the highest, but it was lower than the results using all three variables simultaneously. So, in this article, the feature selection method is feasible.

Finally, we compared the GDBT model with SVR and RFR. The RFR algorithm [21] is an integrated machine learning algorithm based on regression trees, using bagging ideas and employing majority voting to obtain the final prediction results. The SVR algorithm is a machine learning algorithm using the principle of structural risk minimization. This reduces the complexity of the system while ensuring the accuracy of the calibration model, making it generalizable and with high prediction accuracy [19].

In the RFR algorithm, we set the maximum depth to 30 and trained the RFR model using 100 decision trees. The

| Input vector                          | $R^2$  | Mean square error |
|---------------------------------------|--------|-------------------|
| Differential spectrum vector          | 0.82   | 0.04              |
| Spectral vector after spectral analysis | 0.74   | 0.07              |
| Spectral index vector                 | 0.70   | 0.09              |
training results are shown in Figure 11. The $R^2$ value of the RFR inversion model was 0.83, and the loss value was 0.04.

We explored the SVR model using both the radial basis function (RBF) and the least square regression function as the kernel function. We optimized the kernel function parameter $g$ and penalty coefficient $C$ that affect the estimation accuracy via cross-validation. The prediction results of the SVR model are shown in Figure 11. The $R^2$ value of the SVR model was 0.82 using the radial basis function and 0.71 using the least square regression function.

In our experiments, GDBT, SVR, and RFR all attained $R^2$ values over 0.8, with the $R^2$ of the GBDT inversion model being the highest at 0.88. This shows that if the input variables are representative and important, our machine learning estimation model achieved good results. The results further confirmed the effectiveness of our feature wavelength extraction and the feasibility of using the GBDT model as an inversion model for apple leaf nitrogen measurement.

4. Conclusion

In our study, we have shown that using selected sensitivity bands achieves good results with GDBT, SVR, and RFR. We obtained the best results (i.e., peak $R^2$ value) with GBDT, making it preferable to RFR and SVR for estimating apple leaf nitrogen content. We also show that selection of the appropriate spectral vector is essential for constructing the inversion model, confirming the effectiveness of our characteristic wavelength extraction method. We also observe the variation in the $R^2$ value as a result of specific values of the maximum depth, the number of iterations, and loss function, which proves that optimizing the modeling algorithm is an important part of improving the estimation accuracy.

Data Availability

Since the data are being used to apply for a patent, if it must be disclosed, the funders hope that, after the patent is accepted, it can be available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Guangdong Province Key Field Research and Development Plan (Project no. 2018B020241003-09), Shandong Province Agricultural Machinery Equipment Research and Development Innovation Plan (Project no. 2018YZ002), National Peanut Industry Technology System (CARS-13-Mechanized Sowing and Field Management Positions), and National University Student Innovation and Entrepreneurship Training Plan (Project nos. S202010435013 and 201810435050X).

References

[1] L. Liang, L. Di, T. Huang et al., “Estimation of leaf nitrogen content in wheat using new hyperspectral indices and a random forest regression algorithm,” Remote Sensing, vol. 10, no. 12, p. 1940, 2018.
[2] J. G. P. W. Clevers and A. A. Gitelson, “Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3,” International Journal of Applied Earth Observation and Geoinformation, vol. 23, pp. 344–351, 2013.
[3] W. Feng, B.-B. Guo, Z.-J. Wang et al., “Measuring leaf nitrogen concentration in winter wheat using double-peak spectral reflection remote sensing data,” Field Crops Research, vol. 159, pp. 43–52, 2014.
[4] W. Feng, X. Yao, Y. Zhu, Y. C. Tian, and W. X. Cao, "Monitoring leaf nitrogen status with hyperspectral reflectance in wheat," *European Journal of Agronomy*, vol. 28, no. 3, pp. 394–404, 2008.

[5] L. Liang, M. Yang, and Z. Zang, "Determination of wheat canopy nitrogen content ratio by hyperspectral technology based on wavelet denoising and support vector regression," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 26, no. 12, pp. 248–253, 2010.

[6] C. Byu, M. Suguri, and M. Umeda, "Model for predicting the nitrogen content of rice at panicle initiation stage using data from airborne hyperspectral remote sensing," *Biosystems Engineering*, vol. 104, no. 4, pp. 465–475, 2009.

[7] L. Du, S. Shi, J. Yang, J. Sun, and W. Gong, "Using different regression methods to estimate leaf nitrogen content in rice by fusing hyperspectral LiDAR data and laser-induced chlorophyll fluorescence data," *Remote Sensing*, vol. 8, no. 6, p. 526, 2016.

[8] K. S. Song, B. Zhang, L. Yu, and Z. M. Wang, "Study on the hyperspectral remote sensing estimation model for above-ground biomass of corn with canopy reflectance," *System Sciences and Comprehensive Studies in Agriculture*, vol. 1, 2005.

[9] M. N. Tahir, J. Li, B. Liu, G. Zhao, Y. Fuqi, and C. Chengfeng, "Hyperspectral estimation model for nitrogen contents of summer corn leaves under rainfed conditions," *Pakistan Journal of Botany*, vol. 45, no. 5, pp. 1623–1630, 2013.

[10] D. A. Sims and J. A. Gamon, "Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages," *Remote Sensing of Environment*, vol. 81, no. 2-3, pp. 337–354, 2002.

[11] A. Gitelson and M. N. Merzlyak, "Spectral reflectance changes associated with autumn senescence of Aesculus hippocastanum L. and Acer platanoides L. leaves. Spectral features and relation to chlorophyll estimation," *Journal of Plant Physiology*, vol. 143, no. 3, pp. 286–292, 1994.

[12] B. Datt, "A new reflectance index for remote sensing of chlorophyll content in higher plants: tests using Eucalyptus leaves," *Journal of Plant Physiology*, vol. 154, no. 1, pp. 30–36, 1999.

[13] J. E. Vogelmann, B. N. Rock, and D. M. Moss, "Red edge spectral measurements from sugar maple leaves," *International Journal of Remote Sensing*, vol. 14, no. 8, pp. 1563–1575, 1993.

[14] M. H. Yang, L. Y. Liu, T. J. Liu, W. J. Huang, and C. J. Zhao, "Research on a method to retrieve biophysical and biochemical parameters of wheat canopy with hyperspectral remote sensing," *Acta Geodaetica et Cartographica Sinica*, vol. 31, no. 4, pp. 316–321, 2002.

[15] T. H. Demetriades-Shah, M. D. Steven, and J. A. Clark, "High resolution derivative spectra in remote sensing," *Remote Sensing of Environment*, vol. 33, no. 1, pp. 55-64, 1990.

[16] A. Tillack, A. Clasen, B. Kleinschmit, and M. Förster, "Estimation of the seasonal leaf area index in an alluvial forest using high-resolution satellite-based vegetation indices," *Remote Sensing of Environment*, vol. 141, pp. 52–63, 2014.

[17] L. Liang, L. Di, L. Zhang et al., "Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method," *Remote Sensing of Environment*, vol. 165, pp. 123–134, 2015.

[18] M. A. Cho, A. Skidmore, F. Corsi, S. E. Van Wieren, and I. Sobhan, "Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression," *International Journal of Applied Earth Observation and Geoinformation*, vol. 9, no. 4, pp. 414–424, 2007.

[19] F. Wang, J. Huang, Y. Wang, Z. Liu, and F. Zhang, "Estimating nitrogen concentration in rape from hyperspectral data at canopy level using support vector machines," *Precision Agriculture*, vol. 14, no. 2, pp. 172–183, 2013.

[20] F. Li, B. Mistele, Y. Hu, X. Chen, and U. Schmidhalter, "Reflectance estimation of canopy nitrogen content in winter wheat using optimised hyperspectral spectral indices and partial least squares regression," *European Journal of Agronomy*, vol. 52, pp. 198–209, 2014.

[21] L. Prado Osco, A. P. Marques Ramos, D. Roberto Pereira et al., "Predicting canopy nitrogen content in citrus-trees using random forest algorithm associated to spectral vegetation indices from UAV-imagery," *Remote Sensing*, vol. 11, no. 24, p. 2925, 2019.

[22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[23] J. A. Gamon, J. Peñuelas, and C. B. Field, "A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency," *Remote Sensing of Environment*, vol. 41, no. 1, pp. 35–44, 1992.

[24] A. Marshak, Y. Knyazikhin, A. B. Davis, W. J. Wiscombe, and P. Pilewskie, "Cloud-vegetation interaction: use of normalized difference cloud index for estimation of cloud optical thickness," *Geophysical Research Letters*, vol. 27, no. 12, pp. 1695–1698, 2000.

[25] Y. Zhu, X. Yao, Y. C. Tian, X. J. Liu, and W. X. Cao, "Analysis of common canopy vegetation indices for indicating leaf nitrogen accumulations in wheat and rice," *International Journal of Applied Earth Observation & Geoinformation*, vol. 10, no. 1, pp. 1–10, 2008.