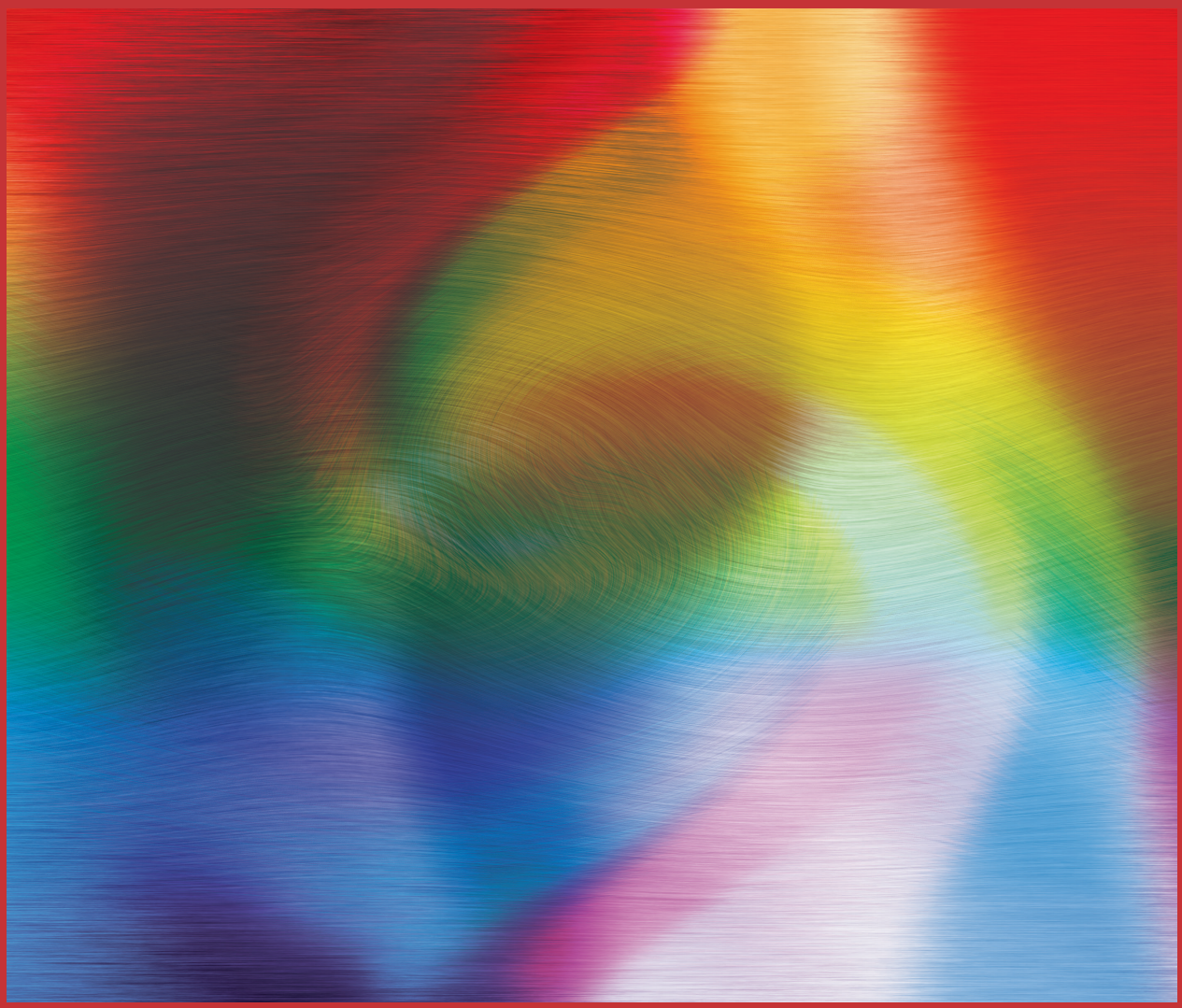
DECEMBER 2021
VOLUME 36 | NO. S12

Spectroscopy

SOLUTIONS FOR MATERIALS ANALYSIS



ADVANCES IN UV-VIS-NIR SPECTROSCOPY

A Peer-Reviewed Special Issue

SPECTROSCOPYONLINE.com

TAKING CENTER STAGE

RAPID • REPEATABLE • RELIABLE

(Did we mention super affordable?)

ENTER TO *WIN!*
UV-VIS Spectrometer
pasco.com/winuvvis

NEW!

UV-VIS
SPECTROMETER

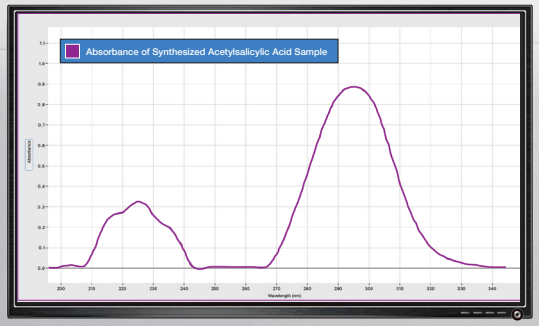
$2100
SE-3607

Spectrometry Software

Collect, analyze, and share data across devices with our free Spectrometry Software. Available for Mac®, Windows®, Chromebook™, iPad® & Android™ tablets.

Download on the App Store
GET IT ON Google Play

For Mac® and Windows® Computers go to pasco.com/**downloads**



PASCO UV-VIS Spectrometer

Get fast, accurate, and reliable performance in your undergraduate teaching labs with the new PASCO UV-VIS Spectrometer.

- Spectral scans from 175 to 1050 nm
- Intuitive, software-based operation
- One-click light and dark calibrations
- Isolated optic bench for consistent accuracy (±1 nm)
- Convenient tools for data export and printing

More Info: pasco.com/**UV-VIS**

PASCO scientific

Spectroscopy

MANUSCRIPTS: To discuss possible article topics or obtain manuscript preparation guidelines, contact the editorial director at: (732) 346-3020, e-mail: LBush@mmhgroup.com. Publishers assume no responsibility for safety of artwork, photographs, or manuscripts. Every caution is taken to ensure accuracy, but publishers cannot accept responsibility for the information supplied herein or for any opinion expressed.

SUBSCRIPTIONS: For subscription information: *Spectroscopy*, P.O. Box 457, Cranbury, NJ 08512-0457; email mmhinfo@mmhgroup.com. Delivery of *Spectroscopy* outside the U.S. is 3–14 days after printing.

CHANGE OF ADDRESS: Send change of address to *Spectroscopy*, P.O. Box 457, Cranbury, NJ 08512-0457; provide old mailing label as well as new address; include ZIP or postal code. Allow 4–6 weeks for change. Alternately, send change via e-mail to mmhinfo@mmhgroup.com for address changes or subscription renewal.

C.A.S.T. DATA AND LIST INFORMATION: Contact Melissa Stillwell, (218) 740-6831; e-mail: MStillwell@mmhgroup.com

Reprints: Contact Stephanie Shaffer, e-mail: SShaffer@mjhlifesciences.com

INTERNATIONAL LICENSING: Contact Kim Scaffidi, e-mail: KScaffidi@mjhlifesciences.com

CUSTOMER INQUIRIES: Customer inquiries can be forwarded directly to MJH Life Sciences, Attn: Subscriptions, 2 Clarke Drive, Suite 100, Cranbury, NJ 08512; e-mail: mmhinfo@mmhgroup.com

Alliance for Audited Media

50% Recycled Paper, 10-20% Post Consumer Waste

AN MJH life sciences™ BRAND

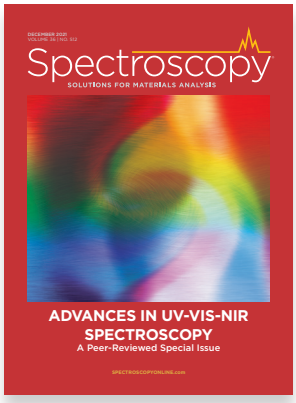
485F US Highway One South,
Suite 210
Iselin, NJ 08830
(609) 716-7777

Advances in UV-Vis-NIR Spectroscopy
December 2021

Spectroscopy

December 2021
Volume 36 | Number S12

PEER-REVIEWED RESEARCH



Cover Art: The UV-vis-NIR spectral regions reveal the details of electronic and molecular information

Cover image courtesy of Courtney Soden

Spectroscopy (ISSN 0887-6703 [print], ISSN 1939-1900 [digital]) is published monthly by MultiMedia Healthcare, LLC., 2 Clarke Drive, Suite 100, Cranbury, NJ 08512. Spectroscopy is distributed free of charge to users and specifiers of spectroscopic equipment in the United States. Spectroscopy is available on a paid subscription basis to nonqualified readers at the rate of: U.S. and possessions: 1 year (12 issues), $83.95; 2 years (24 issues), $151.20. Canada/Mexico: 1 year, $107.10; 2 years, $168.53. International: 1 year (12 issues), $157.50; 2 years (24 issues), $281.40. Periodicals postage paid at Trenton, NJ 08650 and additional mailing offices. POSTMASTER: Send address changes to Spectroscopy, P.O. Box 457, Cranbury NJ 08512-0457. PUBLICATIONS MAIL AGREEMENT NO. 40612608, Return Undeliverable Canadian Addresses to: IMEX Global Solutions, P. O. Box 25542, London, ON N6C 6B2, CANADA. Canadian GST number: R-124213133RT001. Printed in the U.S.A.


Spectroscopy
Follow us on social media for more updates on the field of spectroscopy
Join your colleagues in conversation and stay up-to-date on breaking news, research, and trends in the industry.
linkedin.com/company/spectroscopy-media
@SpectroscopyMagazine
@SpectroscopyMag
MJH life sciences

From the Editor

Jerome Workman, Jr.
is the senior technical editor of *LCGC North America* and *Spectroscopy*,
jworkman@mjhlifesciences.com

Advances in UV-Vis-NIR Spectroscopy: A Peer-Reviewed Special Issue

We present this special issue of *Spectroscopy* on ultraviolet (UV), visible (vis), and near-infrared (NIR) spectroscopy. This special issue features five specially selected and peer-reviewed papers that highlight exciting and important new developments in the potential of laboratory and remote sensing, combined with chemometric methods, to the application of UV-vis-NIR spectral data. Combining UV-vis-NIR spectra data and chemometrics provides a set of powerful analytical tools capable of discriminant analysis, classification, identification, and quantitative analysis for a variety of important applications. The UV-vis-NIR spectral regions reveal the details of electronic and molecular information for solid, liquid, and gas phases for natural and synthetic materials. The advantages of these spectral regions often include little or no sample preparation, capabilities for remote sensing, and rapid data acquisition and analysis. When multipurpose machine learning and other chemometric approaches are used for data analysis, the results can be surprising and dramatic. Many of these papers often include detailed chemometric terms and equations and we hope you will enjoy exploring these data analysis methods for your own use. The selected papers for this digital issue of *Spectroscopy* include the specific aspects of spectral data collection, data preprocessing, and chemometric model development.

Our first paper demonstrates classification and identification of different wood species using a portable near-infrared spectrometer, combined with four spectral pretreatment methods and three pattern recognition methods. Additional chemometric tools are used for comprehensive evaluation of classification model accuracy and complexity.

In the second paper, an effective NIR spectroscopy method is described, using a series of machine learning approaches for wavelength variable selection, to rapidly discriminate grape seed oil adulteration. Grape seed oil has shown beneficial effects for consumers as a dietary supplement, and there is now an interest in measuring grape seed oil for oil quality and potential.

In our third selected paper, a vis-NIR spectral reflectance method is proposed that measures leaf chlorophyll using a wavelet analysis algorithm approach. Here, spectral reflectance is shown as a non-destructive method that is applicable to remote sensing for chlorophyll content in a crop. Measured chlorophyll content is a recognized indicator for photosynthetic capacity, growth cycles, and degrees of stress on plant ecosystems.

Our fourth paper proposes a vis-NIR remote sensing method to determine copper content in mining areas as well as inferred measurement of the environmental impact of surface mining methods. This analytical method is important due to depletion of modern mineral resources from continuous exploitation and utilization, making it economically necessary to quickly identify and locate sources of low-grade copper ore.

In the final paper of this edition, a highly accurate and optimized UV spectroscopy method is proposed to simultaneously monitor nitrate and nitrite for rapid determination and continuous monitoring in environmental water applications. Recent regulations have been imposed to set legal limits of nitrate and nitrite in water worldwide, making accurate measurements of these analytes an important water quality analysis requirement.

PEER-REVIEWED RESEARCH
Spectroscopy papers have undergone a double-blind peer review process and are available on our open-access website.

Rapid Identification of Wood Species Based on Portable Near-Infrared Spectrometry and Chemometrics Methods

Yong Hao, Qiming Wang, and Shumin Zhang

In this paper, a portable near-infrared (NIR) spectrometer, combined with chemometrics methods, was used for rapid identification of 20 wood species. Four spectral pretreatment methods, including Norris-Williams smooth (NWS), standard normal variate (SNV), multiplicative scatter correction (MSC) and Savitzky-Golay 1st derivative (SG 1st-Der), were adopted for noise reduction and information enhancement of near-infrared (NIR) spectra of wood. Three pattern recognition methods, including principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), and support vector machine (SVM), were used to cluster analysis of sample spectra. The competitive adaptive reweighted sampling (CARS) method was proposed to select effective wavelengths (EWs). The Bayesian information criteria (BIC) value was used for comprehensive evaluation of model accuracy and complexity. Compared with PLS-DA models, both the correction set and test set of the SG 1st-Der-SVM model and the SNV-SVM model have obtained 100% correct recognition rates (CRRs). The CARS method shows the SG 1st-Der-SVM model having the smallest BIC value, and the model was optimal.

Classification and identification of wood species is an important part of wood processing and trade. Different wood species have different physical or chemical properties, which is of great significance to entry-exit inspection, quarantine departments, and furniture enterprises. Conventional wood species identification methods include microscopic cell structure and surface characteristic analysis. The microscopic cell structure analysis method needs microscope and wood slices processing, and the analysis process is complex (1). The surface characteristic analysis method is mainly used to analyze the color and texture of wood surfaces by means of image and spectral analysis. Image analysis mainly includes image acquisition, image processing and image recognition, and the identification process requires expertise and is also complex (2). In the image acquisition stage, a high-resolution camera and light source are necessary to ensure the clarity of the image. Complex image pretreatment methods are needed to enhance and extract the image features. Image recognition needs a better template or model to ensure its accuracy.

In recent years, with the development of optical instruments, near-infrared (NIR) spectroscopy has been widely used in qualitative and quantitative analysis of physical and chemical properties of substances because of its fast, nondestructive, and simple operation. The NIR spectrum is electromagnetic energy with a wavelength range of 780–2500 nm. NIR spectra mainly detect overtones and combination bands of the substance; and different bonds, including C-O, O-H, C-H, S-H, and N-H, have different spectral absorption. NIR applications have increased in the agricultural and forest products industries (3,4). Thayna and associates used NIR and partial least squares (PLS) to predict total anthocyanins content (TAC) and total phenolic compounds (TPC) in intact wax jambu fruit (5). Chen and associates have used NIR and PLS to analysis main catechins contents in green tea (6). Luna and associates used NIR and multivariate classification to discriminate soybean oil samples into non-transgenic and transgenic types (7). Zhou and associates used NIR and chemometrics to separate a green hem-fir mix online (8).

TABLE I: The corresponding relationship between attribute values and wood species

Attribute Values	Wood Species	Attribute Values	Wood Species
1	*Pcricopsis elata*	11	*Hymcnaca spp*
2	*Prcca sitchcnsis*	12	*Canarium spp*
3	*Lovoa spp*	13	*Dalbcrgia melanoxylon*
4	*Aglaia spp*	14	*Mansonia altissima*
5	*Burckella spp*	15	*Ocotea rodiei*
6	*Guibourtia spp*	16	*Didelotia spp*
7	*Intsia spp*	17	*Khaya spp*
8	*Manikara spp*	18	*Brachystegia laurentii*
9	*Microberlinia spp*	19	*Swietenia spp*
10	*Afzelia africana*	20	*Tabebuia spp*

TABLE II: The results of classification for wood species based on the PLS-DA model

Methods	*LVs*	*CRR (%)*	
		Calibration Set	Test Set
Origin	**17**	**97.50**	**98.25**
NWS	16	96.75	92.75
SNV	16	96.42	98.75
MSC	17	97.17	98.50
SG 1st-Der	17	97.08	97.50

TABLE III: The results of classification for wood species based on the SVM model

Methods	*C/g*	*CRR (%)*	
		Calibration Set	Test Set
Origin	32.00/0.50	97.50	97.25
NWS	32.00/0.35	97.25	97.00
SNV	**22.63/1.41**	**100.00**	**100.00**
MSC	16.00/1.41	100.00	94.50
SG 1st-Der	**22.63/0.13**	**100.00**	**100.00**

The chemical composition of wood is very complex. Many experimental studies have shown that cellulose, hemicellulose, lignin, and other organic molecules (such as glucose, fructose, pinitol, sorbitol, and inositol) are contained in wood, and these substances all have spectral response in the NIR region (9–15).

In this paper, a portable NIR spectrometer combined with chemometrics methods is used for qualitative identification of 20 wood species. Different spectral pretreatment and pattern recognition methods are used to optimize the optimal recognition model, and predict attribution of an unknown wood species.

Materials and Methods

Samples

For this study, 20 wood species were collected from Zhang Jiagang Entry-Exit Inspection and Quarantine Bureau of China. There were 80 samples from 80 different trees per type of wood, and each sample was made with dimensions of 21 × 10 × 2 cm^3 (length × width × height) according to the national standard method (16) for spectral collection and analysis. Therefore, a total of 1600 samples were used in the experiment.

To build the discriminant model and evaluate the accuracy of the model, 80 samples of each kind of wood were divided into calibration sets and test sets, with the ratio of 3:1 using the Kennard Stone (KS) method (17,18). Thus, a total of 1600 wood samples, of which 1200 samples were used to build the model and validation model, and the remaining 400 samples were taken as a test set for external test. To establish a multi-classification model, the 20 wood species were named 1 to 20 in sequence. The corresponding relationship between attribute values and wood species is shown in Table I.

NIR Spectra Acquisition

The NIR spectra were acquired by a portable spectrometer (MicroNIRS, VIAVI Corporation) with the spectral region of 900–1700 nm. The MicroNIRS instrument consists of a linear variable filter (LVF) dispersing element focused directly onto a 128-pixel linear indium gallium arsenide (InGaAs) array detector, and two tungsten light bulbs as the sources (19).

The NIR spectra acquisition system was set up based on a computer (with self-developed software based on Matlab R2014a) and the MicroNIRS spectrometer. For each wood sample, three diffuse reflectance spectra were measured randomly at different locations, with a temperature of 25±2 °C and humidity of 57±5% RH. Then, the mean spectrum of each wood sample is calculated and stored for the subsequent spectral analysis and species classification.

Traditional detection methods require multiple treatments of the wood, including cutting, softening, slicing, dyeing, dehydrating, and transparency processing. Pictures of the processed wood samples are then taken, or the samples may be observed through a microscope, and the characteristics and microstructure of the wood samples are compared with the standard wood to

TABLE IV: The *BIC* values of the model calibration set and test set

Methods	EWs	C/g	CRR (%)			
			CRR (%)	BIC	CRR (%)	BIC
SG 1st-Der- CARS-SVM	**22**	**5.66/5.66**	**100**	**311.96**	**100**	**263.62**
SNV-CARS-SVM	48	22.63/4.00	100	680.65	100	575.18

determine the type of wood being studied. The proposed wood type identification system is only required to collect the spectrum of the wood, and determine the type of wood by using chemometric analysis.

Methods

NIR spectra are often accompanied by noise for some factors, mainly because of the instability of the light source or detector due to temperature and power supply fluctuations, spectral acquisition modes and sample state variations, and other factors. Therefore, spectral pretreatment is very important for NIR analysis. Four spectral pretreatment methods, including Norris-Williams smooth (NWS) (20), standard normal variate (SNV) (21), multiplicative scatter correction (MSC) (22) and Savitzky-Golay 1st derivative (SG 1st -Der) (23), are used for noise reduction and information enhancement of the NIR wood spectra.

The NIR spectra are highly overlapping, so it is necessary to use chemometrics methods, such as principal component analysis (PCA), partial least squares discriminant analysis (PLS–DA) and support vector machines (SVM), for spectral interpretation. PCA is a method for dimensionality reduction for high-dimensional data by decomposing linear combination of origin variables into a few principal components (24,25). PCA was used to observe the samples spectral spatial distribution. PLS–DA and SVM are used to build discriminant models.

For the PLS–DA method, the latent variable (LV) is an important optimization parameter (26,27), and a reasonable number of LV can make full use of spectral information and filter out noise (28). The Monte Carlo cross-validation (MCCV) method, proposed by Picard and Cook (29), is used to determine the number of LV. MCCV is a simple and effective method that can reduce the risk of model overfitting. The repeated MCCV criterion is defined in equation 1. When the MCCV is the smallest, the *m* is the number of LV:

$$MCCV_{n_v}(m) = \frac{1}{Nn_v}\sum_{i=1}^{N}||y_{s_{v(i)}} - \hat{y}_{s_{v(i)}}||^2 \quad [1]$$

where y and $\hat{y}$ are the true and predicted values of the samples, n_v is the size of the test sample, and repeat the procedure N times (i = 1, 2, ..., N).

SVM is based on the statistical learning theory and structural risk minimization. The basic principle of SVM is to find the optimal separation hyperplane to make the classification

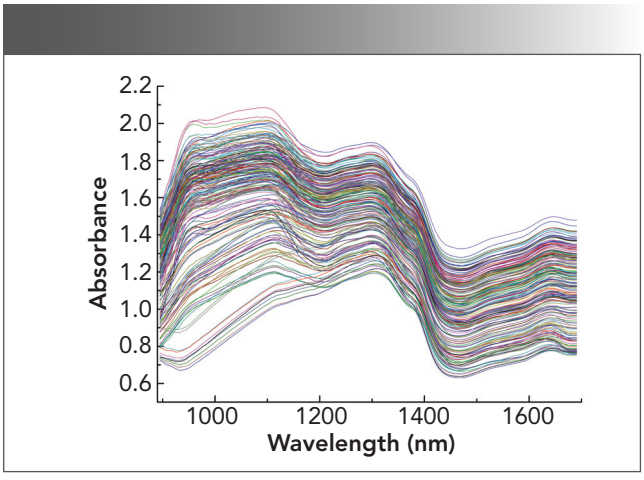


FIGURE 1: NIR absorbance spectra of wood samples.

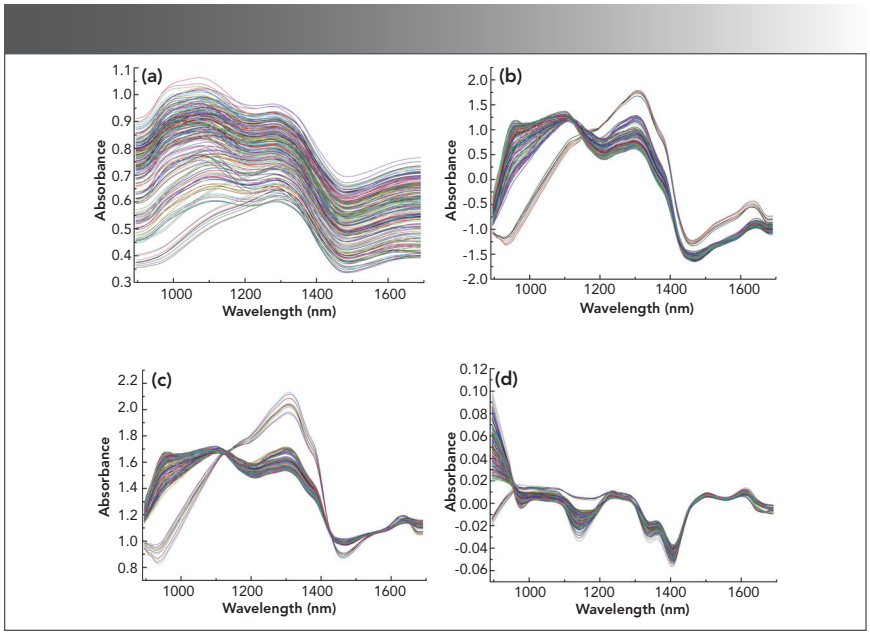


FIGURE 2: The NIR spectra of wood samples employing the different pretreatment methods: (a) NWS, (b) SNV, (c) MSC, and (d) SG 1st-derivative.

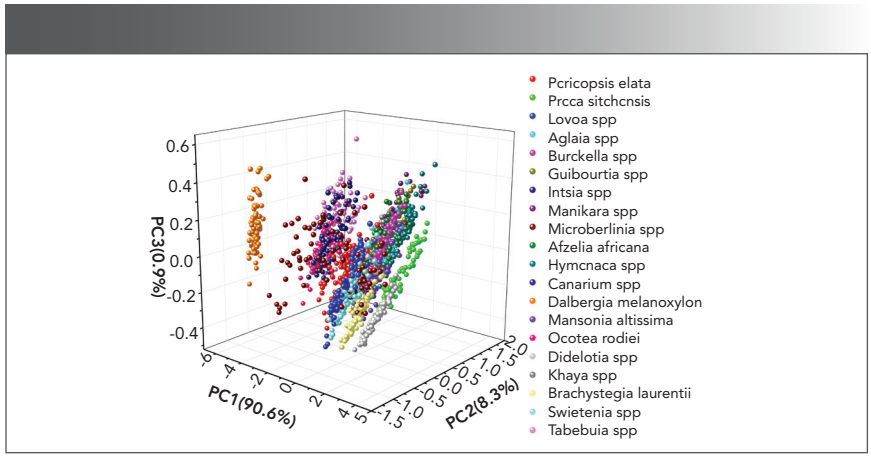


FIGURE 3: The distribution diagram of the first three principal components of the calibration set wood samples.

problem linearly separable (30). Assuming a sample set:

$$S = \{(x_i,y_i)_{i=1}^{n} | x_i \epsilon R^N, y_i \epsilon \{-1,1\}, i=1,2,...,l\} \quad [2]$$

Among them, x_i is the sample data, and y_i is the sample category. The optimization problem is written as:

$$\min \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\zeta_i \quad (\zeta_i \geq 0)$$
$$s.t. \begin{cases} y_i(wx_i+b) \geq 1-\zeta_i & (i=1,2,...,n) \\ C \geq 0 \end{cases} \quad [3]$$

where w represents the weight vector and b represents the bias vector. C is the penalized regression error. In order to ensure the accuracy of classification, introduce a relaxation factor $\xi_i \geq 0$.

As the classifier is a linear function, it is necessary to ensure that the classification hyperplane can accurately distinguish the two types of samples while also ensuring the maximum classification interval. For the case where the classifier is a nonlinear function, it is necessary to map the nonlinear separable problem in the low-dimensional space to the high-dimensional space by introducing the kernel function, and then find the optimal classification hyperplane in the high-dimensional space.

The radial basis function (RBF) is used as the kernel function, and expression of RBF is shown in equation 4 (31). Two parameters, penalized regression error (*C*) and gamma (*g*), need to be optimized to get the best analysis model, and the grid search (GS) method is employed to select optimal C and *g*. The expression of *g* is shown in equation 5:

$$k(x_i,x) = \exp(-\frac{||xi-x||}{2\sigma^2}) \quad [4]$$

$$g = \frac{1}{2\sigma^2} \quad [5]$$

where x_i is the training sample, x is the sample to be predicted, and σ is the width of the kernel function.

Competitive adaptive reweighted sampling (CARS) is a characteristic wavelength selection method based on Monte Carlo sampling and PLS regression coefficients (32). It can overcome the combinatorial explosion problem in variable selection to a certain extent, filter out an optimized subset of variables, and improve the predictive ability of the model. Establish the corresponding PLS model through the correction set samples selected by Monte Carlo sampling, and calculate the weight of the absolute value of the wavelength regression coefficient in this sampling. The larger the weight value, the greater the contribution of the variable to the establishment of the model. Remove the wavelength variables with small absolute values, and the number of variables is determined by the exponential decay function (EDF) (33). The remaining wavelength variables adopt adaptive reweighted sampling to select multiple subsets of wavelength variables to establish a PLS model. The subset of the model with the smallest root mean square error of cross-validation (RMSECV) is the selection of characteristic wavelength combinations. The calculation process is as follows:

$$T = XW \quad [6]$$

$$y = Tc + e = XWC + e = Xb + e \quad [7]$$

where X is the spectral data of the sample and y is the attribute value of each sample. T is the score matrix of X, which is a linear combination of X and W; c is the regression coefficient vector of y against T by least squares; and e is the prediction error.

The method used in the paper was run under Matlab R2014a.

Model Evaluation

The correct recognition rate (CRR) is used to evaluate the accuracy of the model. Good models have relatively higher CRR. The formula is as follows:

$$CRR = \frac{p}{t} \times 100\% \quad [8]$$

where p represents the number of correctly identified samples, and t represents the total number of samples.

To compare the best performing models, the Bayesian information criteria (BIC) (34) is used to determine a satisfactory compromise between model accuracy and model simplicity. The calculation formula of BIC is shown in equation 9:

$$BIC = -2n\ln CRR + \ln(n) \times 2k \quad [9]$$

In the formula, CRR is the accuracy of the model, n is the number of samples, and k is the number of variables to build the model. The model with the minimum BIC is optimal.

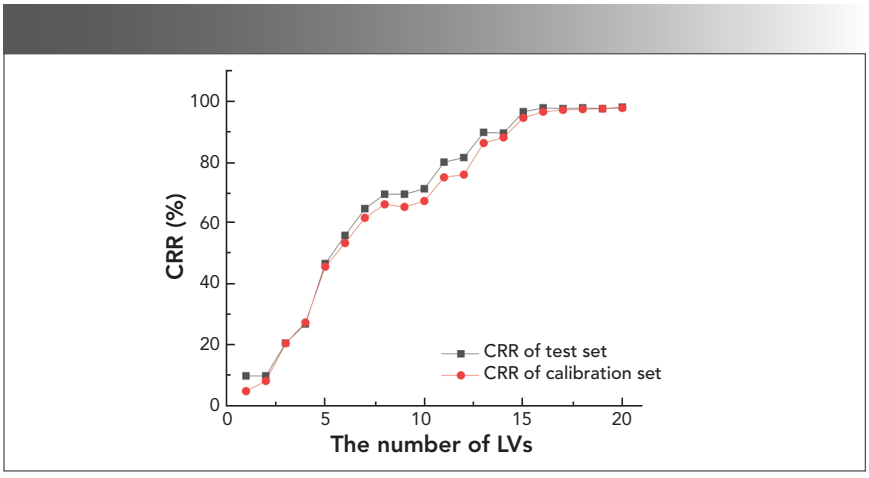


FIGURE 4: The trend chart of CRR changing with LVs in the calibration and test set.

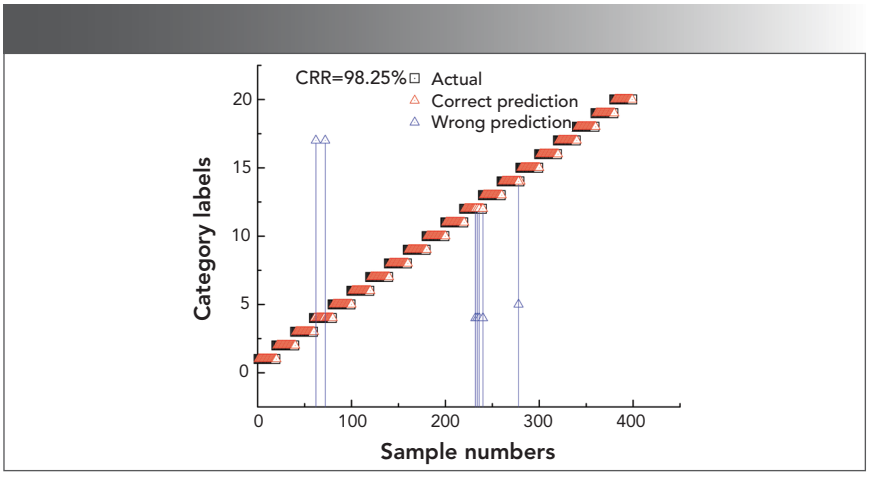


FIGURE 5: The scatter plot of the real attribute value and the predicted value of PLS-DA model for test set samples.

Results and Discussion

NIR Spectra of Wood Samples

Figure 1 shows the NIR spectra of a total of 200 wood samples randomly selected —10 samples from each type of wood. The absorbance spectra range was from 0.6 to 2.2 AU. The absorption peak appears at a wavelength of 1100 nm; this is because of the in-plane bending vibration of the aromatic C-H and the tensile vibration of the secondary alcohol C-O in the wood. Each spectrum is smooth, and the contour trends are consistent. However, when multiple wood spectra are plotted together, the spectra are interlaced and cannot be resolved intuitively. Further preprocessing and pattern recognition methods are required for spectral analysis.

Figure 2 shows the NIR spectra of wood samples employing the different pretreatment methods, including NWS, SNV, MSC, and SG 1st-Der. Different woods show different spectral characteristics after spectral pretreatment due to the fact that NIR spectra signature is strongly affected by the growth environmental conditions and texture of wood (35). Nevertheless, it is not possible to observe significant visual differences between the spectra.

PCA Result of Wood Species

PCA was applied to find the characteristics of each wood species according to the 1600 original spectra. Every wood has many characteristics that can be measured in spectra. PCA can select the important characteristics that can classify them among the 20 species of wood according the spectra, and reduce dimensionality based on decomposing linear combination of origin variables into a few principal components (36).

The distribution diagram of the first three principal components of the calibration wood samples is shown in Figure 3. The cumulative contribution rate of the first three principal components reached 99.8%. In Figure 3, we use different colored points to rep-

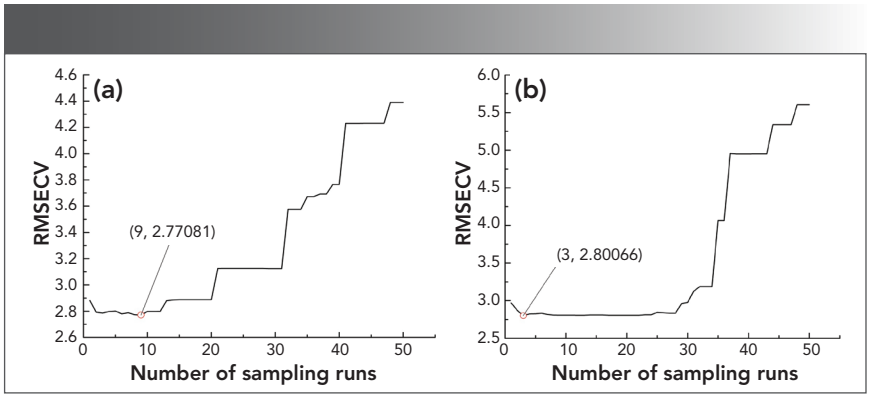


FIGURE 6: The relationship between RMSECV and the number of sampling runs in the CARS method; (a) the model of SG 1st-Der-CARS-SVM; and (b) the model of SNV-CARS-SVM.

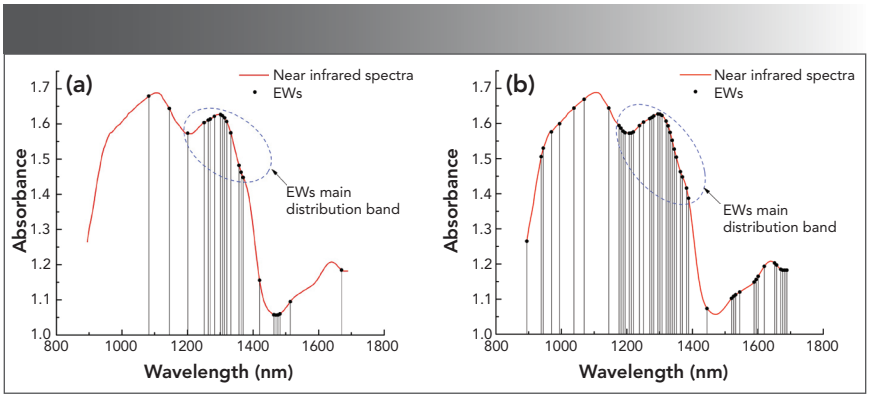


FIGURE 7: Distribution map of EW*s* selected by CARS; (a) The model of SG 1st-Der-CARS-SVM selected 22 EWs; (b) The model of SNV-CARS-SVM selected 48 EW*s*.

resent 20 types of wood, and there are 80 samples of each kind of wood. As shown in the figure, the sample points representing different wood species are interlaced with each other. It can be also be seen in the figure that PCA cannot clearly classify the 20 wood species in the form of a simple classification plane.

PLS–DA Model Analysis of Wood Species

PLS–DA models for identification of wood species were developed with origin and preprocessing spectra. MCCV was used to select the optimal number of LVs. For the MCCV method, 1200 samples were used for random sampling modeling, and 840 were selected for modeling each time, with the remaining 360 used for model verification. The number of random cycles was set to 500, and the model was evaluated using the CRR. The ordinate of Figure 4 is the average of CRR of 500 cycles. For origin spectra PLS–DA model, when LVs are 17, the curves of CRR tend to be flat, and the standard deviation (SD) of 500 operations is small for calibration and test sets. There is no risk of overfitting for models with approximately equal value of CRRs for both calibration and test sets. The results of classification for wood species based on PLS–DA models with origin and preprocessing spectra are shown in Table II, where it can be seen that the origin spectra PLS–DA model was optimal with 17 LVs, and the CRRs were 97.50% and 98.25% for calibration and test sets, respectively.

Figure 5 shows the scatter plot of the real attribute value and the predicted value of the origin spectra PLS–DA model for test set samples. The square and triangle represent the true attribute values of the test set and predictive attribute values of the optimal PLS–DA model, respectively. The coincidence degree of square and triangle reflects the precision of the model. The higher the coincidence degree in the scatter plot, the better the prediction accuracy of the model; the red triangle represents the correct prediction, and the blue triangle represents the wrong prediction. As shown in Table II, the value of CRR was 98.25%. It clearly observed that seven samples showed incorrect validation results in Figure 5. Among them, two samples of No. 4 (*Aglaia spp*) wood were wrongly predicted to be No. 17 (*Khaya spp*) sample, four samples of No. 12 (*Canarium spp*) wood were wrongly predicted to be No. 4 sample, and one sample of No. 14 (*Mansonia altissima*) wood was wrongly predicted to be No. 5 (*Burckella spp*) sample. As to reasons behind the wrong prediction of wood, a possible reason is that the absorption of the double frequency and the combined frequency of the hydrogen-containing groups of the two woods is similar.

SVM Model Analysis of Wood Species

The regularization constant *C* and the kernel function parameter *g* are the key parameters that affect the performance of the SVM. The basic idea of the GS method is an exhaustively search for optimization parameters; arrange and combine the possible values of each parameter, list all possible combinations to generate grid, and then each combination is brought into the model to verify its performance. Finally, the parameter values that make the model performance optimal are taken as the best parameters.

SVM models were developed with origin and pretreated spectra. Table III summarizes the modeling results. From Table III, it is possible to observe that SG 1st-Der and SNV pretreatment promoted the best SVM model, with 100% CRR for calibration set and test set. According to the prediction accuracy, the prediction result of the SVM model

was better than the PLS–DA model, proving that the kernel introduced by SVM was a good predictor of the nonlinear relationship between spectral data and wood species.

Establishment of CARS–SVM Model

To build a more parsimous discriminant model, a CARS (37,38) method was proposed to remove redundant variables with collinearity and select effective wavelengths. The number of repeated sampling of the CARS method was set to 50, and the five-fold cross-validation method was used to calculate the RMSECV. At the beginning of the EDF, as the number of sampling variables was eliminated, the RMSECV decreased, and then the number of sampling iterations continued to increase rapidly because of the decrease of effective wavelengths (EWs). At the end, interactive verification was used to select the smallest subset of RMSECV as the optimal variable combination. It can be concluded in Figure 6 that, when the number of sampling runs of the SG 1st-Der-CARS–SVM model was 9, the RMSECV reaches the minimum (Figure 6a), and 22 EWs were selected. When the number of sampling runs of the SNV–CARS–SVM model was 9, the RMSECV reaches the minimum (Figure 6b), and 48 EWs were selected. Figure 7a was the distribution map of 22 EWs selected by the SG 1st-Der-CARS–SVM model, and Figure 7b was the distribution map of the 48 EWs selected by the SNV–CARS–SVM model. The red line represents a spectrum, and the black circle represents EWs. From the figure, EWs are mainly distributed between 1200–1400 nm (The point inside the blue circle in Figure 7). In this band, the spectrometer introduces the least external interference in the process of collecting spectral data, and contains the most effective data.

In the process of establishing a discriminant model, choosing different combinations of variables would result in different models, and the values corresponding to the information criteria would also change. In this study, the calibration set and test set CRR of the SG 1st-Der-CARS–SVM model and SNV–CARS–SVM model were both 100% (Table IV), but the variable combinations of the two models were different. Therefore, the corresponding BIC was not the same. As shown in Table IV, it can be concluded that the BIC value of the calibration set and the test set of the SG 1st-Der-CARS–SVM model were the smallest. The SG 1st-Der-CARS–SVM model used fewer variables for modeling. While ensuring the accuracy of model prediction, the model was simplified to make the model better.

Conclusion

In this study, 20 wood species were classified successfully based on the portable NIR spectra with chemometrics methods of PLS–DA and SVM. Compared to PLS–DA, the SG 1st-Der-SVM model and the SNV–SVM model both obtained 100% CRRs in the calibration set and test set for classifying 20 kinds of wood. The CARS method can effectively filter out EWs, reduce the invalid variables in the modeling, and simplify the SVM model. By comparing the calculated value of the models, the BIC value of the SG 1st-Der-CARS–SVM model was the smallest. The results show that the SG 1st-Der-CARS–SVM model is the best. A portable NIR spectrometer combined with SVM method can be used for fast identification of wood species with a higher recognition rate.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (21265006).

References

(1) G. Dou, G.S. Chen, and P. Zhao, *Spectrosc. Spectral Anal.* **36**(8), 2425–2429 (2016). Doi: 10.3964/j.issn.1000-0593(2016)08-2425-05

(2) K.P. Cui, X.R. Zhai, and H.J. Wang, *Advances in Forestry Letters* **2**(4), 61–66 (2013).

(3) Z.Y. Liu, F. Xu, J.B. Wen, and H.Z. Chen, *Chinese J. Anal. Lab.* **35**(10), 1117–1120 (2016).

(4) Y.N. Liu, Z. Yang, B. Lu, M.M. Zhang, and X.H. Wang, *Spectrosc. Spectral Anal.* **34**(3), 648–651 (2014).

(5) T.R. Viegas, A.L.M.L. Mata, M.M.L. Duarte, and K.M.G. Lima, *Food Chem.* **190**(1), 1–4 (2016). Doi: 10.1016/j.foodchem.2015.05.063.

(6) Q.S. Chen, J.W. Zhao, C. Sumpun, and Z.M. Guo, *Food Chem.* **113**(4), 1272–1277 (2009). Doi: 10.1016/j.foodchem.2008.08.042

(7) A.S. Luna, A.P. da Silva, J.S.A. Pinho, J. Feree, and R. Boque, *Spectrochim. Acta A* **100,** 115–119 (2013). Doi: 10.1016/j.saa.2012.02.085

(8) Z. Zhou, S. Sohrab, and S. Avramidis, *Eur. J. Wood Wood Prod.* **78**(1), 151–160 (2020) .Doi: 10.1007/s00107-019-01479-8

(9) E. Pecoraro, B. Pizzo, A. Alves, N. Macchioni, and J.C. Rodrigues, *Microchem. J.* **122,** 176–188 (2015).

(10) C.J.G. Colares, T.C.M. Pastore, V.T.R. Coradin, L.F. Marques, A.C.O. Moreira et al., *Microchem. J.* **124,** 356–363 (2016). Doi: 10.1016/j.microc.2015.09.022

(11) A. Sandak, J. Sandak, and M. Zborowska, *J. Archaeol Sci.* **37**(9), 2093–2101 (2010). Doi: 10.1016/j.jas.2010.02.005

(12) L. Tong and W.B. Zhang, *Appl. Spectrosc.* **70**(10), 1676–1684 (2016). Doi: 10.1177/0003702816644453

(13) X.S. Wang, Y.D. Sun, and M.G. Huang, *J. Northeast Forestry University* **43**(12), 82–85 (2015).

(14) B. Pizzo, E. Pecoraro, and N. Macchioni, *Appl. Spectrosc.* **67**(5), 553–562 (2013). Doi: 10.1366/12-06819

(15) K. Zhao, Y. Xiong, and M. Zhao, *Laser & Infrared* **41**(6), 649–652 (2011).

(16) National Timber Standardization Technical Committee, *General Method of Wood Identification: GB/T 29894-2013* (Standards Press of China, Beijing, People's Republic of China, 2013), pp. 1–12.

• Continued on Page 29

PEER-REVIEWED RESEARCH
Spectroscopy papers have undergone a double-blind peer review process and are available on our open-access website.

Rapid Quality Discrimination of Grape Seed Oil Using an Extreme Machine Learning Approach with Near-Infrared (NIR) Spectroscopy

Yang Li

In this paper, an effective identification method of wavelength variable selection to rapidly discriminate the grape seed oil adulteration by near-infrared (NIR) spectroscopy is investigated. The extreme learning machine (ELM) is employed to build a stable and accurate model, and a firefly algorithm combined with a successive projections algorithm (FA–SPA) is developed to eliminate redundant wavelengths (The model used throughout is called FA–SPA–ELM). The comparison among different models—the partial least squares discriminant analysis (PLS–DA) model, the support vector machine (SVM) model, the least squares support vector machine (LS–SVM), and the FA–SPA–ELM model—demonstrates that the wavelength number of the FA–SPA model can be effectively reduced with a wavelength variable of 17, and the model of FA–SPA–ELM presents the excellent predictive capability. The experimental results show that the proposed novel method could be used to identify adulterated grape seed oil quickly, effectively, and nondestructively.

Grape seed oil contains high content of lipids and bioactive compounds, such as vitamin E, linoleic acid, and proanthocyanidins, among other components (1,2). The total proanthocyanidins of grape seed oil have been demonstrated to offer benefits for antioxidant, anti-inflammatory, antihypertensive, and hypocholesterolemic activities (3,4). Additionally, grape seed oil has effect on anti-aging, radical-scavenging properties, and the protection against DNA damage (5). The results indicate that grape seed oil is a kind of senior health edible oil that has high nutrition health value and pharmacotherapy efficacy. However, adulterated grape seed oil is sometimes sold with other cheap edible oils by some producers so as to earn larger profits, thus disturbing the market and causing damage to the health of consumers. Therefore, it is of great value to employ rapid detecting techniques to discriminate the quality of grape seed oil.

Conventional methods detecting the quality of edible oils are mainly based on high performance liquid chromatography (HPLC), gas chromatography (LC), and thin-layer chromatography; however, these approaches are expensive, time-consuming, and require a plentiful supply of samples (6). Near-infrared (NIR) spectroscopy primarily measures the molecular vibrations of C-H, O-H, and N-H. Among existing techniques, the NIR spectroscopy technique has been widely applied to the nondestructive and rapid qualitative and quantitative detection of edible oils (7,8). Therefore, this research is intended to provide a potential reference method for detecting adulterated grape seed oil.

However, the full spectrum data measured by NIR contains too much redundant information between adjacent wavelength bands, causing some challenging problems regarding the identification of relevant and effective information, as well as the accuracy of the model. Wavelength selection techniques aim at eliminating the uninformative and interferential wavelength signals, while simultaneously obtaining an optimal subset of informative characteristic wavelength variables from the NIR spectrum (9). Among all different types of wavelength selection techniques, the *swarm intelligence op-*

timization algorithms are more interesting, because they simulate the social behavior of animals and insects to acquire the shortest path between a food source and their nests. Compared with the conventional optimization techniques, these techniques employ a stochastic, probabilistic, and crowd search process rather than a single solution. However, the number of the wavelength selected is still large, and it is easy to fall into a local optimal point through the swarm intelligence optimization algorithms. Hence, a genetic algorithm with a successive projections algorithm (GA–SPA) (10), a uninformative variable elimination with successive projections algorithm (UVE–SPA) (11), and a successive projections algorithm with particle swarm optimization (SPA–PSO) (12) are proposed to assist in rectifying these shortages, and to build a stable model using fewer wavelengths.

A high-quality, high-accuracy, and stable model is required for the qualitative analysis of adulteration by NIR spectroscopy. For the spectroscopy-based classification, chemometricians have developed many valuable algorithms, such as the partial least squares discriminant analysis (PLS–DA) (13), the support vector machine (SVM) (14), and the least squares support vector machines (LS–SVM) (15). Specifically, an algorithm for single-hidden layer feed-forward neural networks called an extreme learning machine (ELM) was proposed (16). Differing from the conventional learning algorithms that require adjusting input weights and hidden layer biases, the ELM arbitrarily assigns input weights and hidden layer biases, and calculates the output weights by a generalized inverse method (17). It is reported that ELM brings a higher learning rate, predictive accuracy, and generalization performance (18). As a rapidly developed technology, a large number of applications of the ELM have emerged in recent years (19,20). On the other hand, the ELM combined with NIR to analyze the adulterated grape seed oils has not been reported.

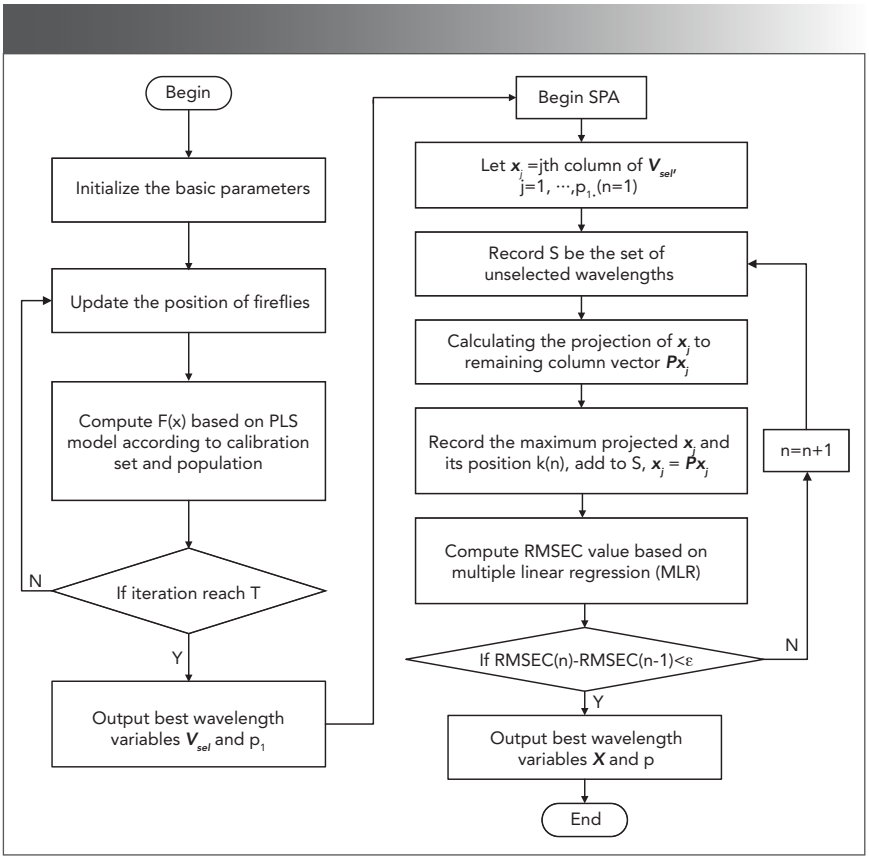


FIGURE 1: The flow chart of FA combined with SPA.

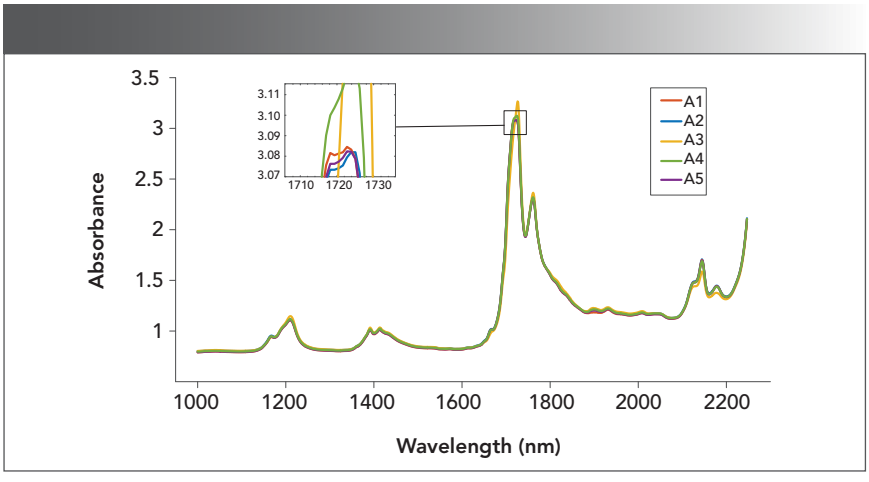


FIGURE 2: The NIR spectrum of the adulterated grape seed oil samples; insert is a close-up of 1710 to 1730 nm peak.

TABLE I: Comparison of different wavelength selection methods

Method	The Number of Wavelengths	RMSEC	RMSEP
FA	710	0.093	0.146
SPA	120	0.112	0.231
FA-SPA	17	0.382	0.264

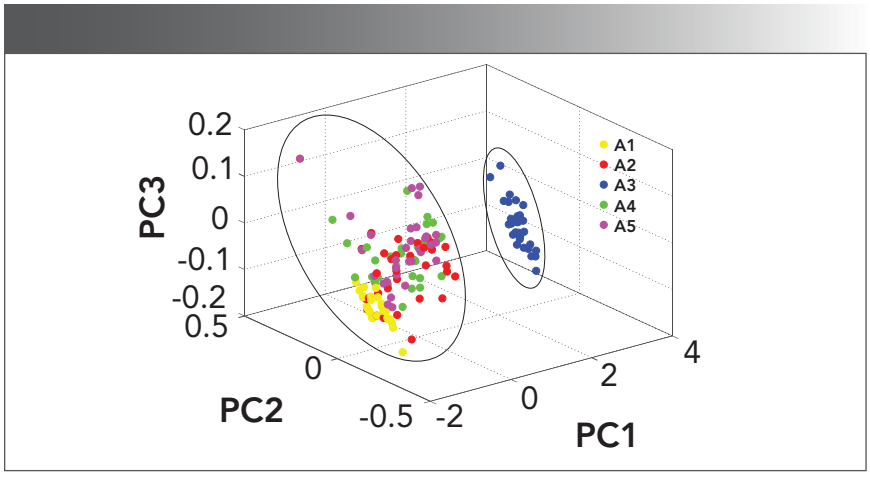


FIGURE 3: PCA score plots (PC1, PC2, and PC3) for adulterated grape seed oil. Groups A1 through A5 are shown in different colored dots.

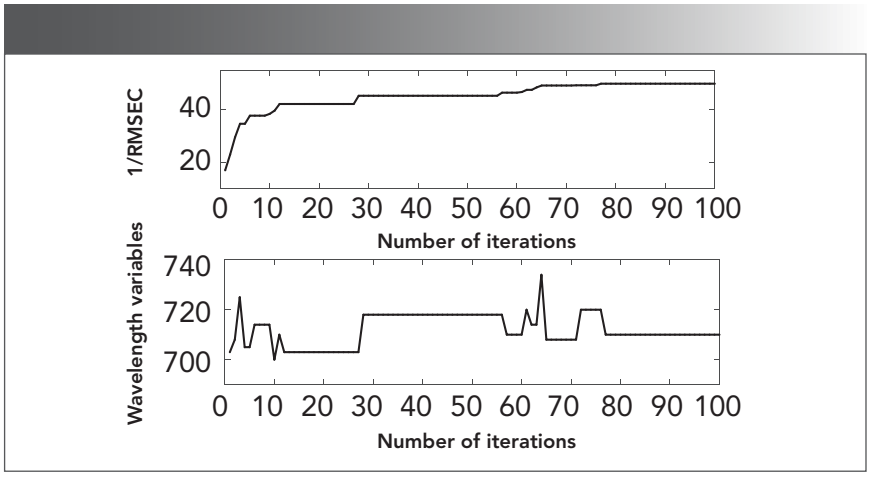


FIGURE 4: Illustration of iteration process plots of FA algorithm.

In this paper, an ELM approach is proposed to demonstrate the feasibility of the combination of NIR spectroscopy and the firefly algorithm combined with a successive projections algorithm ELM (FA–SPA–ELM) model in the quality of grape seed oil. NIR spectra is collected from grape seed oil blended with four kinds of different edible oils. FA–SPA is applied to optimize the characteristic wavelength, and then the ELM is built to determine the adulteration of grape seed oil. Moreover, the predictive performance of ELM is evaluated by a comparison with the conventional modeling methods, including the PLS–DA model, the SVM model, and the LS–SVM model.

This paper also formulates the oil samples, the spectral data collection, and the methods used throughout, as well as provides the experimental comparison and results to verify the effectiveness of the FA–SPA–ELM model.

Materials and Methods

In this section, the mixing of oil samples, the collection of NIR spectra, and the partition of the samples are described in detail. The flow chart of wavelength selection, and the structure and application of the ELM neural network in the research, are illustrated as well.

Oil Samples

The grape seed oil, soybean oil, peanut oil, corn oil, and sunflower oil employed in this study were purchased at the local market. The grape seed oil was mixed with different amounts of soybean oil, peanut oil, corn oil, and sunflower oil to obtain the calibration and prediction set of 31 adulteration samples, respectively. These samples were prepared in volumetric flasks of 200 mL. In the samples of the adulterated grape seed oil, the volume content of the other four oils ranged from 0 mL to 200 mL, with increments of 5 mL, respectively.

Spectral Collection

The spectra data were acquired by the measurement system consisting of oil samples, the Fourier transform near-infrared (FT-NIR) Antaris II spectrometer, and a laptop. The oil samples were put into the spectrometer to measure spectral information. The spectral data were obtained by the spectrum information acquisition software, and saved as a .csv file (including wavelength values and corresponding absorbance data). Spectral measurement were executed at 25 °C and 60% relative air humidity. The experimental samples were packaged in 10 mm quartz cuvettes. The spectra were scanned in the range of 10,000–4000 cm^{-1}, with the resolution of 16 cm^{-1} and with 32 replicate scans every time. The spectrum of each sample were analyzed in triplicate, and the average value was taken as the NIR absorption spectrum of the sample.

In this study, 31 samples for each category were divided into a calibration set and a prediction set through the Kennard-Stone (21) algorithm. The calibration set of 21 samples for each category was employed to build the classification model, and the remaining 11 samples were used as a prediction set to evaluate the prediction capability of the model. In the following sections, A1, A2, A3, A4, and A5 represent NIR data of pure grape seed oil, grape seed oil blended with different levels of soybean oil, peanut oil, corn oil, and sunflower oil, respectively.

Selection of Characteristic Wavelength Variables

The full spectral data was weak, owing to the redundant wavelength information. The redundant wavelength information reduced computation speed and accuracy of prediction modeling. Therefore, a firefly algorithm (FA) was applied to screen out characteristic wavelength variables from the NIR spectrum in this work. FA is a swarm intelligence algorithm originally proposed by Yang (22) simulating the social behavior of fireflies that use light to attract mates (23). Yang (24) formulates the following three idealized rules:

(i) All fireflies are unisex, and attract each other;

(ii) Attractiveness is related to their brightness; for any two flashing fireflies, the less brighter one will move towards the brighter one. However, the brightness can decrease as their distance increases. If no one is brighter than a particular firefly, it moves randomly; and

(iii) The brightness of a firefly is determined by the objective function.

The fitness function is defined as:

$$f = \frac{1}{\mathrm{RMSEC}} = \frac{1}{\sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}} \quad [1]$$

where RMSEC is root mean square error of calibration based on partial least squares (PLS) regression, N is the number of calibration set samples, y_j, $1 \le j \le N$ is category labels of the oil samples in calibration set, and $\hat{y}_j$, $1 \le j \le N$ is predictive category labels through the PLS model.

However, the optimized characteristic wavelengths by FA are still vast, and contain collinear interference. These wavelengths are time-consuming to obtain, and unstable when used to establish classification models. The successive projections algorithm (SPA) is a forward selection method that uses vector projection analysis in the spectral matrix to minimize variable collinearity. SPA is used to extract efficient wavelengths further. The basic principle of SPA starts with one wavelength incorporates another wavelength at each iteration, until a specified number of wavelengths is reached. For each iteration, the combination of wavelength information is selected for constructing multiple linear regression models and calculating the RMSEC. When the value of RMSEC reaches a minimum and tends to be stable, the corresponding preferred wavelength number is the optimal wavelength combination.

Figure 1 shows the flow of the characteristic wavelength selection by FA combined with SPA, where p_0 is the dimension of the spectral, V_{sel} is the wavelength variable matrix for the N calibration set and p_1 wavelengths selected by FA, and X is the wavelength variable matrix for N calibration set and p wavelengths selected by SPA.

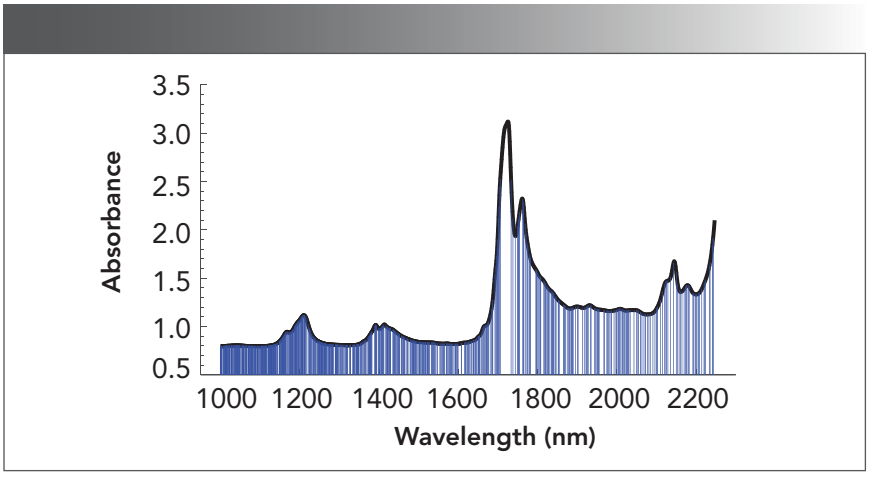


FIGURE 5: A plot of the 710 wavelengths selected by the FA algorithm.

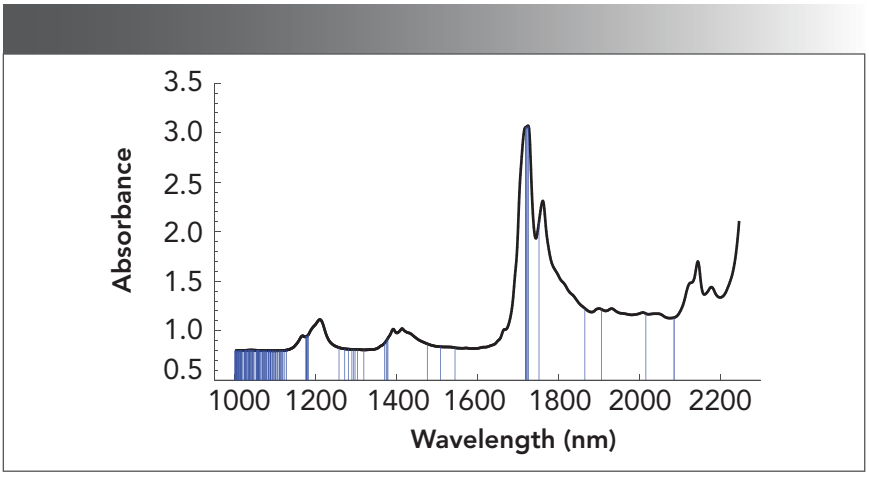


FIGURE 6: A plot of the 120 wavelengths selected by SPA algorithm.

The ELM Classification Model

The ELM classification model is a simple and practical single-hidden layer feedforward neural network proposed by Guangbin in 2006 (25). ELM can effectively overcome the issue of traditional neural networks, such as the complexity of training parameters, and the problem of local optimum. Here, there is N training spectra samples $\{\boldsymbol{X},\boldsymbol{Y}\} = \{\boldsymbol{x}_j, \boldsymbol{y}_j\}_{j=1}^{N}$ to be employed to establish the classification model between the spectra and category labels, where x_j is the measured spectra absorbance of the jth sample, and $x_j = [x_{j1}, x_{j2},\ldots,x_{jp}]^T$, x_{jk}, and $k = 1,\ldots p$ are the selected

TABLE II: Discriminant results of different wavelength selection based on the ELM model

Samples		ELM-FULL	FA-ELM	FA-SPA-ELM
A1	Calibration set	95%	95%	100%
	Prediction set	100%	100%	100%
A2	Calibration set	100%	100%	100%
	Prediction set	91%	100%	100%
A3	Calibration set	100%	100%	100%
	Prediction set	91%	100%	100%
A4	Calibration set	100%	100%	100%
	Prediction set	100%	100%	100%
A5	Calibration set	95%	95%	100%
	Prediction set	100%	100%	100%

TABLE III: Discriminant results of the four classification models based on FA-SPA

Samples		ELM	PLS-DA	SVM	LS-SVM
A1	Calibration set	100%	85%	95%	95%
	Prediction set	100%	100%	100%	100%
A2	Calibration set	100%	100%	100%	100%
	Prediction set	100%	91%	91%	100%
A3	Calibration set	100%	100%	100%	100%
	Prediction set	100%	100%	100%	100%
A4	Calibration set	100%	100%	100%	100%
	Prediction set	100%	100%	100%	100%
A5	Calibration set	100%	95%	95%	95%
	Prediction set	100%	100%	100%	100%

optimal wavelength by the firefly algorithm combined with successive projections algorithm FA–SPA; $y_j = [y_{j1}, y_{j2},...,y_{jq}]^T$ is the category labels of the jth sample, where q is the dimension of the category labels, $y_{jk} \in \{0,1\}$, $y_{jk} = 1$ means the spectral data is labeled as Class k.

For the jth sample, the ELM classification model can be mathematically modeled as follows:

$$\sum_{i=1}^{m} \beta_{ik} g(\boldsymbol{w}_i^T \boldsymbol{x}_j + b_i) = y_{jk}, k = 1,...q; j = 1,...,N \quad [2]$$

where m is the number of the hidden nodes and $g(x)$ is the activation function determined with sigmoid function, i = 1, 2,…, m, w_j is the input weight vector connecting input nodes with the hide node, b_i is the hidden layer vector bias corresponding to its hidden node, and β_{ik} is the output weight vector connecting the output nodes with the hide node; w and b are assigned arbitrarily.

For N samples, the equation [2] can be abbreviated as:

$$\boldsymbol{H\beta} = \boldsymbol{T} \quad [3]$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_m^T]^T$; $\boldsymbol{\beta}_i = [\beta_{i1} \cdots \beta_{iq}]^T$; $\boldsymbol{T} = [\boldsymbol{y}_1 \cdots \boldsymbol{y}_N]^T$ and

$$\boldsymbol{H} = \begin{bmatrix} g(\boldsymbol{w}_1^T \boldsymbol{x}_1 + b_1) & \cdots & g(\boldsymbol{w}_m^T \boldsymbol{x}_1 + b_m) \\ \vdots & \cdots & \vdots \\ g(\boldsymbol{w}_1^T \boldsymbol{x}_N + b_1) & \cdots & g(\boldsymbol{w}_m^T \boldsymbol{x}_N + b_m) \end{bmatrix} \quad [4]$$

For given number of the hidden m, input weight w_1, and the hidden layer biases b_i is to estimate β such that the output least-squares error of the model is minimized. It is directly equivalent to solve the following optimization problem:

$$\|\boldsymbol{H}(\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_m,b_1,b_2,...,b_m)\hat{\boldsymbol{\beta}} - \boldsymbol{T}\| = \min_{\boldsymbol{\beta}} \|\boldsymbol{H}(\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_m,b_1,b_2,...,b_m)\boldsymbol{\beta} - \boldsymbol{T}\| \quad [5]$$

where $\hat{\beta}$ is the estimated value of β and its solution is:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{H}^{\dagger}\boldsymbol{T} \quad [6]$$

$H^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix H.

Given a calibration set, the sigmoid function $g(x)$, and the hidden node number m (details see results and discussion), the ELM algorithm can be realized according to the following procedure:

Step 1: Generate input weight w_i and the hidden layer biases b_i randomly, i=1, 2, … , m.

Step 2: Calculate the hidden layer output matrix H according to equation [4].

Step 3: Calculate the output weight according to equation [6].

Results and Discussion

Spectral Analysis

Spectral data in the wavelength range from 1000 to 2300 nm are taken as the analytic spectral data. The spectrum of grape seed oil mixed with different vegetable oils and pure grape seed oil are shown in Figure 2. It can be observed that the differences between the spectra are extremely small, and it is hard to make distinctions directly. However, when enlarged at a local position, for example, and the major spectral bands presented in 1660–1820 nm and 2100–2200 nm, there is a difference in the spectrum of all samples. The reason is that vegetable oils contain some unsaturated fatty acids, such as oleic acid, linoleic acid, and some saturated fatty acids. The spectral absorption peak in the 1660–1820 nm region is mainly assigned to the bending vibrations of the -CH_2 and -CH_3 groups, and the functional groups of CH-CH. The intensity of the absorption peak near 2100–2200 nm corresponds to the cis-double bond stretching vibration of the -CH groups.

The spectral data are collected under the background of the same instrument parameters. The multivariate scatter correction is employed to preprocess the spectrum to eliminate background noise and physical interference, and then the subsequent experiments are carried out.

Principal component analysis (PCA) (26) is the most commonly used unsupervised pattern recognition method. This method provides a visual representation of the relationship between samples and variables. In this work, PCA is employed to reduce the dimension to achieve the visualization of the raw features. Figure 3 displays the distribution of the five-class oil samples in PC1-PC2-PC3 space. The results show that the first three PCs explain 99.56% of the variability in spectral data (PC1 = 97.46%, PC2 = 1.39%, and PC3 = 0.72%). It can be seen that the samples in A3—and in A1, A2, A4, A5—are respectively classified into two distinct clusters. However, A1, A2, A4, and A5 are superposed in the PC space, and are not clearly discriminated or classified in the PC space, as shown. Therefore, a new method is needed to extract useful information and overcome the disadvantage of the redundant wavelength information.

Optimal the Characteristic Wavelength by FA–SPA

The FA–SPA is developed to overcome the disadvantage of the redundant wavelengths, and to optimize it. Moreover, the ELM model is applied to discriminate the adulterated grape seed oil in this work.

To further illustrate the efficiency of the proposed method, FA, SPA, and FA–SPA are respectively employed to make a comparison. The whole spectra contain 1441 spectrum wavelengths.

In FA, the relevant parameters are set as follows: The maximum fluorescence intensity of firefly is 1, the number of step size factor is 0.7, the number of initial firefly population is 50, and the maximum number of iterations is 100. The calibration set of 100 samples are set as input matrix (100 × 1441) to begin the wavelength search through FA.

The iterative process about extracting the characteristic wavelength with FA algorithm is shown in Figure 4. It can be seen that the fitness value increases with the number of iterations increasing, and finally tends to be stable. It means that FA converges to the optimal value successfully. Although the fitness value is increasing, the number of the optimal wavelength does not decrease accordingly. Correspondingly, the number of selected wavelengths constantly rises and falls during the iteration, and finally stabilizes at around 710. As the fitness function value converges to a steady state, the best wavelength combination is extracted, and shown in Figure 5, where the vertical lines represent the selected wavelength. Clearly, the dimension of the selected wavelength is still high, because the superposed spectral absorption peak does not extract well. SPA is then applied to reduce the dimensionality further.

With respect to the SPA, the number of the selected wavelengths is 120, while the optimum value of RMSEC falls to 0.112. Figure 6 shows this optimal wavelength distribution, where the vertical lines represent the selected wavelengths.

Although the wavelengths are reduced from 1441 to 710 by the FA algorithm, there are still some collinear interferences in the preferred wavelengths. Moreover, FA easily converges into the local optimum. Therefore, to improve the wavelength selection, the FA–SPA for wavelength selection is proposed. Firstly, FA is used to select 710 informative wavelengths, and then SPA is followed to select 17 wavelengths with minimum redundant information from the 710 informative wavelengths. The selected 17 characteristic wavelength variables are 1002, 1004, 1012, 1013, 1016, 1036, 1076, 1124, 1165, 1261, 1378, 1381, 1471, 1546, 1901, 2084, and 2103 nm, respectively.

The results of three wavelength selection methods are summarized in Table I, where RMSEC is the root mean square error of calibration set and RMSEP is the root mean square error of prediction set. The wavelength variables are greatly reduced by FA–SPA. The RMSEC and RMSEP of FA–SPA are 0.382 and 0.264, respectively. Compared with the results of FA and SPA, although the precision of FA–SPA algorithm is a little bit larger, the redundant wavelengths decrease significantly, and the model is greatly simplified. In conclusion, the wavelength selected by FA–SPA is effective.

To summarize, the ELM classification model based on the 17 optimal wavelengths selected by FA–SPA can achieve accurate prediction results. Therefore, although the RMSEC value rises to 0.382, the 17 optimal wavelengths selected by FA–SPA are still efficient and effective.

The Contrast of ELM Model Based on the Three Wavelength Selection Methods

The classification model of ELM is adopted to classify the adulterated grape seed oils. The model shows high discrimination accuracy and stability when using a sigmoid function as the activation function of the single hidden layer. It is critical to select the single hidden layer neurons in ELM modeling analysis. In order to obtain the optimal number of hidden layer neurons, the initial number of neurons in the hidden layer starts from 5, and gradually iterates up to 150 with 5 steps.

In this paper, the ELM model with the full spectrum of 1441 wavelengths (denoted as ELM–FULL), the ELM model with 710 characteristic wavelengths selected by FA (denoted as FA–ELM), and the ELM model with 17 characteristic wavelengths selected by FA–SPA (denoted as FA–SPA–ELM) are constructed from the 100 samples of the calibration set, respectively, and then the models are used to predict the 55 samples from the prediction sets. After several experiments, the performance of the models are optimized when the number of hid-

den layer neurons in ELM–FULL, FA–ELM, and FA–SPA–ELM model are set to be 100, 70, and 30, respectively. The distinguishing results on the samples of the calibration set and the prediction set by the three models are summarized in Table II.

According to Table II, the FA–SPA–ELM model achieves the best predictive performance with the identification rates of 100% for both the calibration set and prediction set. In fact, FA can eliminate the redundant information between wavelengths but there is strong collinearity between the extracting adjacent wavelengths, and SPA can effectively eliminate collinear information. The comparative results demonstrate that although the RMSEC value of FA–SPA in the preceding paragraph rises to 0.382, the selected 17 optimal wavelengths are still effective with the predictive performance of the ELM model. Therefore, it is noteworthy to combine these two algorithms to improve the prediction accuracy of the classification model.

Comparison of the ELM Model and Other Discriminant Models Based on FA–SPA

To compare with the performance of the ELM model, the PLS–DA model, the SVM model, and the LS–SVM model are further built. We set 17 wavelengths optimized by FA–SPA as input variables to establish the four models. The optimal number of PLS–DA variables is set as 15, and the penalty parameters and kernel function parameters of SVM are 90 and 150, respectively; these were obtained by a grid search. The samples of the calibration set and the prediction set are tested by the above four classification models, and the different results can be observed in Table III.

Table III reveals the prediction performance of the four classification models. Compared to the LS–SVM, SVM, and PLS–DA model, the ELM model presents excellent predictive ability. Actually, classification with ELM can be implemented via parallel computations because of its network structure, so it has more potential for real-time applications with a comparable accuracy. The established four models are based on the selected characteristic wavelengths, indicating that the optimization of characteristic wavelength is beneficial to simplifying the model. Overall, the analyzing and comparison of all models indicated the greater ability of the FA-SPA-ELM model to discriminate the adulterated grape seed oil.

Conclusions

In this work, to rapidly and efficiently discriminate the adulterated grape seed oils, the optimized characteristic wavelengths by FA–SPA are developed to establish the ELM discriminant model based on the NIR spectroscopy data at 1000–2300 nm. The results show that FA-SPA can greatly reduce the wavelength variables, with the number of wavelength variables decreased from 1441 to 17. The accuracy, stability, and generalization of the ELM model are further improved based on the selected wavelength variables. This developed FA–SPA–ELM algorithm is effective and promising in identifying the adulterated grape seed oils based on NIR spectroscopy.

References

(1) M.M. Ozcan and F.Y. Al Juhaimi, *Chem. Nat. Compd.* **53**(1), 132–134 (2017).

(2) F.B. Shinagawa, F.C. Santana, E. Araujo et al., *Food Sci. Technol.* **38**(1), 164–171 (2018).

(3) M.D.L.L. Cádiz-Gurrea, I. Borrás-J. Lozano-Sánchez et al., *Int. J. Mol. Sci.* **18**(2), 376 (2017).

(4) D. Praud, M. Parpinel, V. Guercio, et al., *Cancer, Causes Control Pap.Symp.* **29**(2), 261-268 (2018).

(5) Y. Kim, Y. Choi, H. Ham et al., *Food Chem.* **137**(1–4), 136–141 (2013).

(6) J.J. Yuan, C.Z. Wang, H.X. Chen, et al., *Int. J. Food Prop.* **19**(2), 300–313 (2016).

(7) H. Azizian, M.M. Mossoba, A.R. Fardin-Kia, et al., *Lipids* **50**(7), 705–718 (2015).

(8) Z. Li, J. Wang, Y. Xiong, et al., *Vib. Spectrosc.* **84,** 24–29 (2016).

(9) A.M. Rady and D.E. Guyer, *Postharvest Biol. Technol.* **103,** 17–26 (2015).

(10) J.H. Cheng, D.W. Sun, and H. Pu, *Food Chem.* **197,** 855–863 (2016).

(11) S. Ye, D. Wang, and S. Min, *Chemom. Intell. Lab. Syst.* **91**(2), 194–199 (2008).

(12) X. Li, S. Wang, W. Shi, et al., *Food Analytical Methods* **9**(6), 1713–1718 (2016).

(13) P. Bai, J.Wang, H. Yin, et al. *Anal. Lett.* **50**(2), 379–388 (2016).

(14) C.J.C. Burges, *Data Mining and Knowledge Discovery* **2,** 121–167 (1998).

(15) Y. Liu et al., *J. Near Infrared Spectrosc.* **26**(1), 34–43 (2018).

(16) G.B. Huang, *Int. J. Mach. Learn. Cybern.* **2**(2), 107–122 (2011).

(17) Q.Y. Zhu, A.K. Qin, P.N. Suganthan, et al., *Pattern Recognit.* **38**(10), 1759–1763 (2005).

(18) G.B. Huang, Q.Y. Zhu, and C.K.Siew, *Neurocomputing* **70,** 489–501 (2006).

(19) X. Bian, S. Li, M. Fan, et al., *Anal. Meth.* **8**(23), 4674–4679 (2016).

(20) R. Moreno, F. Corona, A. Lendasse et al., *Neurocomputing* **128,** 207–216 (2014).

(21) L. Zhang, G. Li, M. Sun, et al. *Infrared Phys. Technol.* **86,** 116–119 (2017).

(22) X.S. Yang, "Firefly Algorithms for Multimodal Optimization," in *Stochastic Algorithms: Foundations and Applications. SAGA 2009. Lecture Notes in Computer Science*, O. Watanabe and T. Zeugmann, Eds.(Springer, Berlin, Germany, vol. 5792).

(23) M. Goodarzi and L. dos Santos Coelho, *Anal. Chim. Acta* **852,** 20–27 (2014).

(24) S.X. Yang, *Int. J.of Bio Inspired Computation* **2**(2), 78–84 (2010).

(25) G.B. Huang , Q.Y. Zhu, and C.K. Siew, *Neurocomputing,* **70**(1–3), 489-501 (2006).

(26) L. Gál, M. Oravec, P. Gemeiner P , et al. *Forensic Sci. Int.* **257,** 285–292 (2015).

Yang Li is with Concord University College-Fujian Normal University, in Fuzhou, Fujian, China. Direct correspondence to: 61580907@qq.com●

PEER-REVIEWED RESEARCH
Spectroscopy papers have undergone a double-blind peer review process and are available on our open-access website.

Model for Retrieving Leaf Chlorophyll Using the Wavelet Analysis Algorithm with the Prospect Radiative Transfer Model and Vis-NIR Spectra

Feifei Xie, Lin Sun, Jie Wang, and Fengzhu Liu

We proposed a new retrieving model (named CAB1) for estimating the chlorophyll content at leaf level using the Prospect radiative transfer (RT) model. CAB1 is built based on multiband information derived from the radiative transfer model. Specifically, a continuous wavelet analysis algorithm was used to find the wavelet coefficients for those spectral characteristics that can highlight the chlorophyll content. Then, the simulated data sets generated by the Prospect model were used to evaluate the sensitivity of CAB1 in chlorophyll and its stability to other biochemical components (leaf structural parameters, water, carotenoids, brown pigment, and dry matter). Finally, CAB1 was verified using the Lopex93 and Angers experimental spectral data sets. Quantitative and qualitative results revealed that the retrieving model was not only more accurate than the traditional spectral index (the highest R^2 of the inversion value and the measured value is 0.9), but also more stable.

Chlorophyll is the dominant factor affecting vegetation photosynthesis, and it is closely related to other biochemical parameters such as protein, nitrogen, lignin, and water (1). The chlorophyll content in a crop can be used to indicate the photosynthetic capacity, growth cycles, and degrees of stress (such as disease, insect pests, and heavy metal stress) (2,3). Traditional biochemical methods for measuring the chlorophyll content (such as spectrophotometry, in which samples must be pretreated in the laboratory) are destructive to the vegetation as well as time-consuming and laborious; thus, spectral reflectance is an attractive alternative (4,5).

During the past few decades, many studies have proposed the spectral indices required to estimate chlorophyll contents. By using one of the existing spectral indices or the sensitive bands of chlorophyll, a single or multivariate analysis model can be established. Several studies have successfully estimated the chlorophyll content in vegetation using visible ratios (6), vis-NIR ratios (7–9), red-edge reflectance ratio indices (10,11), and spectral and derivative red-edge indices (12). Yan used the Prospect model to test the applicability of four types of spectral indices, such as visible ratios (eight kinds), vis-NIR ratios (eight kinds), red-edge reflectance ratio indices (six kinds), and derivative red-edge indices (six kinds), to test chlorophyll content extraction. She found that some spectral indices, such as normalized difference vegetation index (NDVI) and the modified chlorophyll absorption in reflectance index (MCARI), are not suitable for retrieving chlorophyll based on the reflectance spectra. Rather, it was necessary to carefully select a spectral index for a specific application (13,14). Using a spectral index to evaluate the chlorophyll content of vegetation is simple and flexible, but there are obvious limitations. One limitation is that the spectral index method does not consider the radiative transfer (RT) mechanism of light in the leaf, which leads to a lack of definite physical meaning (15). The other limitation is that spectral index models are generally adapted to specific databases and are difficult to generalize to other databases.

TABLE I: The spectral simulation test with difference input parameters in Prospect

Dateset name	N	C_{ab} (μg/cm²)	C_{ar} (μg/cm²)	C_{brown}	C_w (cm)	C_m (g/cm²)
Set_C_{ab}	1.5	20–60, step 1	12	1	0.012	0.005
Set_N	1–3, step 0.5	50	12	1	0.012	0.005
Set_C_m	1.5	50	12	1	0.012	0.003–0.015, step 0.002
Set_C_{ar}	1.5	50	5–25, step 5	1	0.012	0.005

TABLE II: The common vegetation indices for predicting leaf chlorophyll content, including the spectral index name, formulas, and sources

Section	Spectral Index	Formula	Source
Spectral index of visible light	native plant conservation initiative (NPCI)	$(R_{680}-R_{430})/(R_{680}+R_{430})$	(34)
	transformed chlorophyll absorption in reflectance index (TCARI)	$3*((R_{700}-R_{670})-0.2*(R_{700}-R_{550})*(R_{700}/R_{670}))$	(35)
	modified chlorophyll absorption in reflectance index (MCARI)	$((R_{700}-R_{670})-0.2*(R_{700}-R_{550}))*(R_{700}/R_{670})$	(2)
Spectral indices of visible and infrared light	MERIS terrestrial chlorophyll index (MTCI)	$(R_{750}-R_{710})/(R_{710}-R_{680})$	(36)
	Triangular vegetation index (TVI)	$(120*(R_{750}-R_{550})-200*(R_{670}-R_{550}))/2$	(37)
	Transformed Chlorophyll Absorption in Reflectance Index/Optimized Soil-Adjusted Vegetation Index (TCARI/OSAVI)	$(3*((R_{700}-R_{670})-0.2*(R_{700}-R_{550})*(R_{700}/R_{670})))/(1.16*(R_{800}-R_{670})/(0.16+R_{800}+R_{670}))$	(35)
Red-edge spectral index	Gitelson & Merzylak (GM)	R_{750}/R_{700}	(38)
	Vogelmann2 (VOG2)	$(R_{734}-R_{747})/(R_{715}-R_{726})$	(39)

Therefore, researchers began to consider optical models with geometric meanings, including the radiative transfer performance of light in vegetation, and the effects of vegetation biochemistry on light. The process of light reflection and absorption in vegetation and the interaction process were mathematically described to retrieve the biochemical components of the vegetation. Optical models, such as the Prospect (16), Liberty (17), and Leafmod (18) models, are not limited to the time, location, and other factors, and they are robust to noise. Yan noted that the reflection and absorption spectra simulated using the Prospect model are nearly the same as those of the measured spectra. Jiang used the Prospect model to simulate spectral data of different spectral scales (different bandwidths (5–65 nm)) and analyzed the validity and sensitivity when applied to the estimation of the chlorophyll content (19). Wang used the Prospect model to simulate the spectra of leaves with different biochemical characteristics, and analyzed the influence of different mathematical combinations of the NDVI spectral index on the elimination of interference factors. The spectral index of the dual NDVI ratio vegetation index was established to estimate the carotenoid content in leaves (20). Zhang used the Prospect model to construct a lookup table from which the chlorophyll content of eucalyptus leaves could be retrieved. The spectral data determined using a statistical method, an estimation model linking the chlorophyll, and carotenoid content with spectral characteristic parameters of eucalyptus leaves, were established (21).

The aim of this study was to construct a new spectral index for estimating chlorophyll content at the leaf level with physical implications, robustness, and universality. First, the relationship between the chlorophyll content and the spectral characteristics of vegetation based on the transmission mechanism of the Prospect optical model was analyzed to construct a new specific band. The "Experimental Materials and Methods" section introduces the experimental materials and methods, and describes the establishment of a new inversion model for chlorophyll content with continuous wavelet transform (CWT). The "Results" section reviews the experimental results, which describe the new inversion model validation of comparison with other vegetation indices. The "Discussion" section focuses on the advantages and disadvantages of new spectral index. Finally, the "Conclusion" section summarizes the findings of the study.

Experimental Materials and Methods

Prospect Leaf Model Simulated Dataset

The Prospect model, among the most mature RT models, has few required parameters and easy inversion. It can simulate the optical properties of vegetation leaves from 400 nm to 2500 nm (22). The Prospect model is a simple "flat model." The leaf is composed of N-plate layers and N-1 air layers. In Prospect-5 (23), the total spectral absorption coefficient (k) of each leaf plate layer is a function of leaf structural parameters (N), chlorophyll content (C_{ab}), water content (C_w), carotenoids (C_{ar}), brown pigment (C_{brown}),

TABLE III: Vegetation indices with different mesophyll structure parameters

Index	CAB1	NPCI	TCARI	MCARI	MTCI	TVI	TCARI/ OSAVI	GM	VOG2
C_{ab} (R^2)	0.9949	0.9321	0.9239	0.9951	0.9805	0.9949	0.7032	0.9900	0.9987
N (SI%)	34.87	1652.53	628.45	1881.76	142.30	125.66	479.88	60.63	69.21
C_m (SI%)	0.56	1.52	6.12	11.06	4.52	10.58	1.40	5.50	8.07
C_{ar} (SI%)	2.28	1.37	5.42	2.17	0.00	0.90	5.42	0.00	0.00
Ava	0.8998	-3.2293	-0.6056	-3.7600	0.6295	0.6476	-0.2249	0.8346	0.8064

and dry matter (C_m). Therefore, the absorption (*A*) of a leaf plate layer can be approximately expressed as a function of the leaf structural parameter (*N*) and a single plate layer absorption coefficient (*k*) (24,25).

$$A(\lambda) = Nk_e(\lambda) + C_{ab}k_{ab}(\lambda) + C_{ar}k_{ar}(\lambda) + C_{brown}k_{brown}(\lambda) + C_m k_m(\lambda) + C_w k_w(\lambda) \quad [1]$$

where K_e is the refractive index of the basic whitening layer; k_{ab} is the corresponding absorption coefficient spectrum of the leaf chlorophyll; k_{car} is the corresponding absorption coefficient spectrum of the leaf carotenoid; k_{brown} is the corresponding absorption coefficient spectrum of the leaf brown pigment; k_w is the corresponding absorption coefficient spectrum of the leaf water; and k_m is the corresponding absorption coefficient spectrum of the leaf dry matter.

K_e, k_{ab}, k_{car}, k_{brown}, k_m, and k_w were measured and fixed (23). Thus, the Prospect model had six input parameters, including the leaf structure (*N*), content of chlorophyll a+b (C_{ab}), carotenoids (C_{ar}), brown pigment (C_{brown}), equivalent water thickness (C_w), and dry matter content (C_m). Its output parameters were the reflectivity and spectral transmittance of the leaf ranging from 400 to 2500 nm, and the sampling interval was 1 nm.

According to different experimental purposes, the input parameters of the Prospect model are set, and the simulated data set with different biochemical components can be simulated. Four simulation data used in this paper, named Set_C_{ab}, Set_*N*, Set_C_m, and Set_C_{ar}, are shown in Table I.

The sensitivity index, SI, is used for accuracy evaluation (equation 2) (26).

$$SI = \frac{|VI|_{max} - |VI|_{min}}{|VI|_{min}} \text{ x}100\% \quad [2]$$

where *VI* is the calculated values of the spectral indices and |*VI*| is the absolute value. The smaller the *SI*, the smaller the influence of the component on the spectral index.

Lopex93 and Angers Measured Data Set

The Lopex93 database was established by the Joint Research Center (JRC) in 1993 (27). Approximately 70 leaf samples representations of more than 50 species were obtained. The database contains 62 samples of spectral data and chlorophyll content data, including 13 monocotyledons and 49 dicotyledons. Spectra were collected over the 400–2500 nm region with a sampling interval of 1 nm. The database not only takes into account the differences of different species, but also the changes of biochemical composition in time. It fully comprises the diversity of biochemical composition content and inner types of leaves. The database cannot only provide a more effective test for spectral index, but also serve as an ideal data source for analyzing the correlation between chlorophyll content and spectral reflectance of mixed species. The Lopex93 data set can be accessed free online (http://opticleaf.ipgp.fr/index.php?page=database).

The Angers database was established at the National Institute for Agricultural Research (INRA) in Angers, France, in June 2003 (23). Approximately 276 leaf samples representations of more than 39 species were obtained. The directional-hemisphere reflectance spectra were measured using an ASD FieldSpec spectrometer with a spectral range of 400–2500 nm and a data interval of 1.4 nm. The biochemical parameters include leaf chlorophyll content, carotenoids, specific leaf weight (SLW), and water content. The Angers data set can also be accessed free online (http://opticleaf.ipgp.fr/index.php?page=database).

Wavelet Analysis Method

Wavelet transform is an effective mathematical tool used to decompose the original spectral signal into multiple scales (28,29). The continuous wavelet transform (CWT) tool can make the wavelet shift smoothly to different positions; thus, each scale component can be compared directly to the input spectral reflectance, and more useful spectral information can be captured (30). In this study, the CWT wavelet analysis method was used in the spectral analysis method. The CWT is defined as follows in equation 3 (31),

$$W_f(a,b) = \int_{-\infty}^{+\infty} f(\lambda)\frac{1}{\sqrt{\alpha}}\Psi\left(\frac{\lambda - b}{\alpha}\right)d\lambda \quad [3]$$

where W_f is the signal to be analyzed; Ψ is a chose wavelet basis function, such as biorthogonal (bior), daubechies (db), coiflets (coif), or symlets (sym) (32); $\Psi a,b\ (\lambda)$ is the prototype to generate child wavelets by adjusting the scale parameter *a* (the corresponding decomposition frequency or band

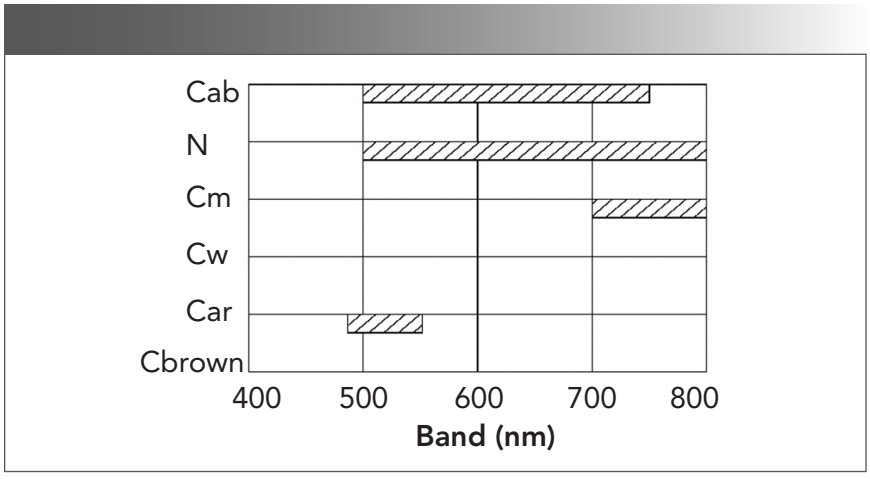


FIGURE 1: The chlorophyll sensitive band areas (in nm) of different dataset contents. Abscissa is spectral band (nm) and ordinate is dataset spectra.

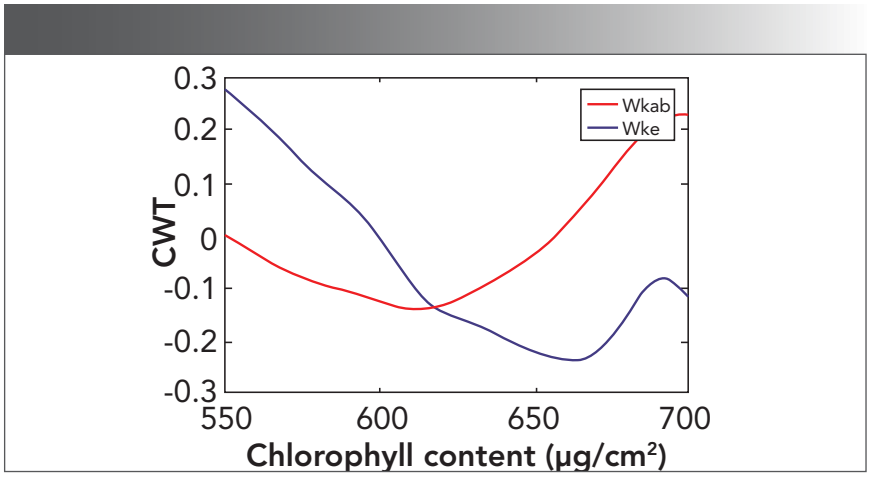


FIGURE 2: The CWT curve W_{kab} (the blue line) and W_{Ke} (the red line) with spectrum scale *a* = 150 nm using wavelet function "bior1.1".

range) and *b* the translation factor (corresponding to the band position); λ is the wavelength; and dλ is the spectral resolution.

In this study, the spectral reflectance signal was processed by continuous wavelet transform, where f(λ) is the spectral reflectance value or absorption coefficient spectrum of the leaf. Commonly used continuous wavelet decompositions are selected as Ψ; *a* is the spectrum scale of 100 nm, 150 nm, and 200 nm; *b* is the band position which should be selected by analysis; λ is the wavelength that can be cut down by analysis; and dλ = 1 nm. After the CWT wavelet analysis, the coefficients are obtained, which can be used for singularity detection or periodic analysis. When the signal is projected into the wavelet transform domain, it is advantageous to extract some useful features.

Reconstruction of a New Leaf Spectral Index

To construct the new spectral index for the chlorophyll content of leaves, the relationship between the reflectance and biochemical components of the vegetation is built first. As is known, the summation of the spectral reflectance of the leaf (*R*), transmittance (*T*), and absorption (*A*), were constantly one. The shape of the leaf reflectance and transmittance were found to be similar in the visible and near-infrared (NIR) radiation, and could be expressed approximately as follows (33)

$$T(\lambda) = \alpha R(\lambda) \quad [4]$$

where α is a constant.

The spectrum absorption rate, *A*, could be expressed as:

$$A(\lambda) = 1 - (1+\alpha)R(\lambda) \quad [5]$$

By combining equation 1 with equation 5, the formula of the leaf spectrum reflectance rate (*R*) could be obtained.

$$1-(a+1)R(\lambda) = Nk_e(\lambda) + C_{ab}k_{ab}(\lambda) + C_{ar}k_{ar}(\lambda) + C_{brown}k_{brown}(\lambda) + C_m k_m(\lambda) + C_w k_w(\lambda) \quad [6]$$

Different biochemical components have a different influence on the reflectance in the band 400–800 nm (the sensitive spectrum range of the chlorophyll). In the Prospect model, only one biochemical component was set with different ranges and the other input parameters were fixed to obtain many reflectance spectra, which were used to test the influence of the biochemical component on the reflectance. For example, to test the C_{ab} in the band 400–800 nm influence on the reflectance in the Prospect model, C_{ab} was set from 5 to 95 μg/cm^2 (step size was 15 μg/cm^2). The other input parameters were fixed (N = 1.5, C_w = 0.012 cm, C_m = 0.005 g/cm^2, C_{ar} = 12 μg/cm^2, and C_{brown} = 1). Finally, seven reflectance spectra curves were obtained, which were used to find the band region where reflectance is changed with a different C_{ab}. The test result is shown in Figure 1. The range of the influence for C_{ab} was 500–750 nm, but the ranges of 400–500 nm and 750–800 nm were not an influence on the spectra. The range of influence for N was 500–800 nm, for C_m it was 700–800 nm, for C_{ar} it was 480–550 nm, but C_w and C_{brown} did not change, and the content changes did not affect the spectral shape. We found that, in the range of 550–700 nm, the reflectance spectrum was affected only by chlorophyll C_{ab} and N. Thus, to extract the main influence spectrum range of C_{ab} and minimize the effect of the other

components, the final sensitive band range of chlorophyll was determined to be from 550 to 700 nm. Then, equation 6 was further reduced as follows:

$$1-(\alpha+1)R(\lambda) = Nk_e(\lambda) + C_{ab}k_{ab}(\lambda) + b \quad [7]$$

where b is a constant value, meaning C_m, C_{ar}, C_w, and C_{brown} content changes do not affect the reflectance spectra.

In mathematics, the result of CWT computation for constant value is 0. The calculated value of constant coefficients with CWT computation remains unchanged. Therefore, both sides of equation 7 are carried out with the CWT computation. At the same time, we can obtain equation 8

$$-(\alpha+1)*W_R(a,b) = N*W_{k_e}(a,b)+C_{ab}*W_{k_{ab}}(a,b) \quad [8]$$

where the wavelet coefficients $W_R(a,b)$, $W_{ke}(a,b)$ and $W_{kab}(a,b)$ are determined by scaling factor a, the specific band position b, and wavelet base function Ψ.

Finally, two factors were considered to determine the scale basis function and band position of the continuous wavelet decomposition in building the new spectral index. One factor was in the range of 550–700 nm, the CWT curve of the $k_{ab}(a,b)$ should have a wave valley or wave peak at bands w1 and w2, where the chlorophyll absorption was sensitive. The other factor was that the CWT curve of K_e(a,b) should be approximately 0 at bands w1 and w2, eliminating or decreasing the influence of N on the reflectance spectrum. Through a large number of comparative experiments, in the selected band range from 550 to 700 nm, with a spectrum scale of 100 nm, 150 nm, and 200 nm, were conducted to K_{ab}(a,b) and k_e(a,b). We found that "bior1.1" with a spectrum scale of 150 nm can meet the two factors as shown in Figure 2. Thus, the wavelet base function Ψ was chosen as "bior1.1", the wave spectrum scale, s, was 150 nm (a = 150 nm), the wave-peak w1 position was 699 nm, and the wave-

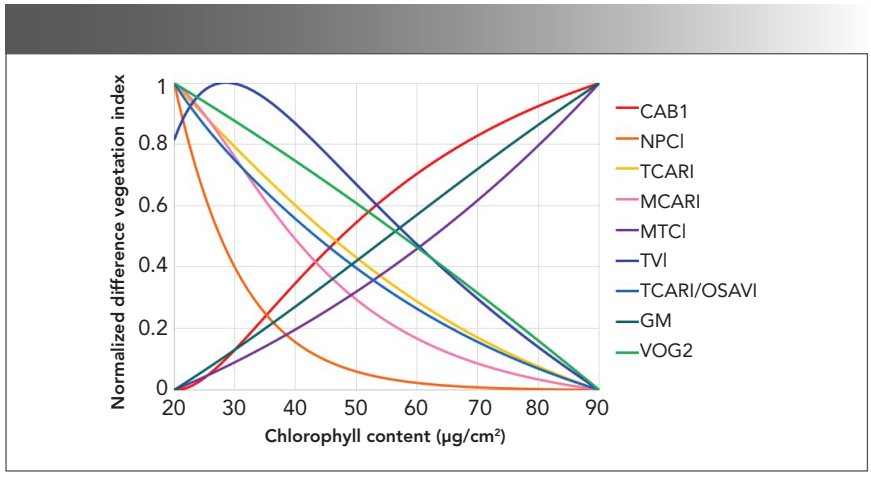


FIGURE 3: Sensitivity analysis of hyperspectral vegetation indices on chlorophyll content.

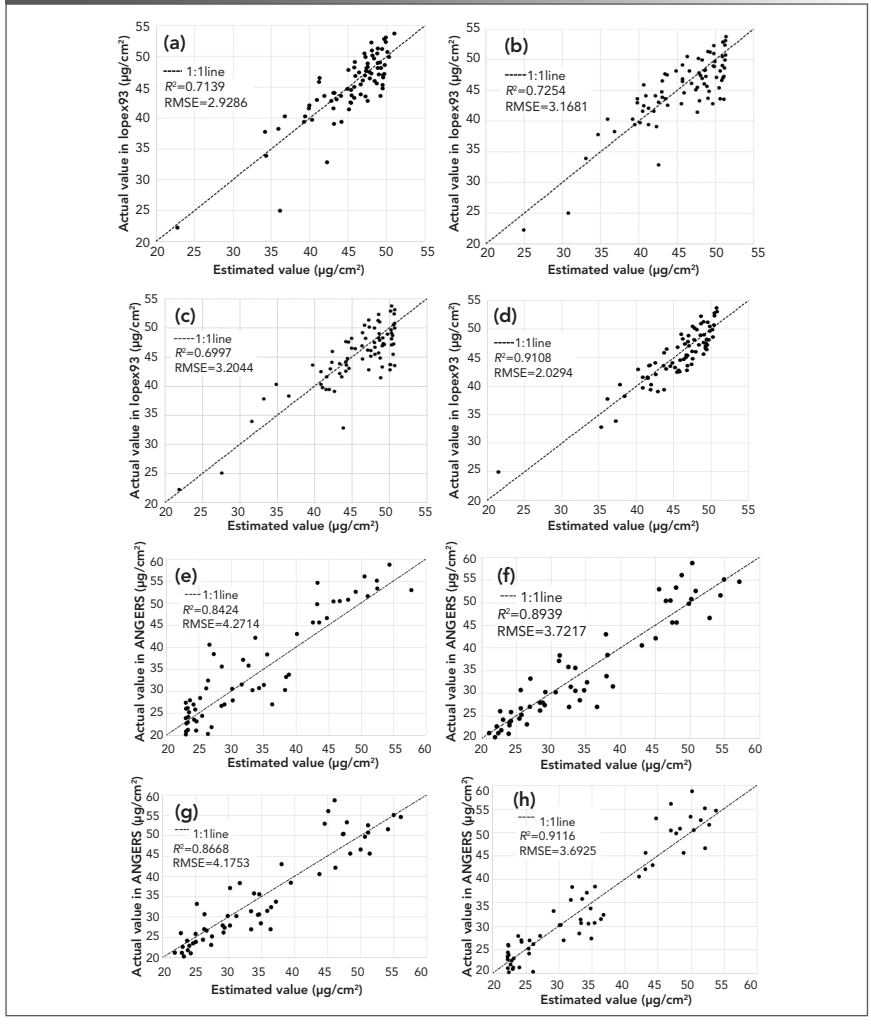


FIGURE 4: Validation results of chlorophyll estimation with Lopex93 data using various approaches: (a) TCARI, (b) MTCI, (c) VOG2, and (d) CAB1; and Angers data: (e) TCARI, (f) MTCI, (g) VOG2, and (h) CAB1.

valley w2 position was 613 nm (b = 699 nm and 613 nm).

At the wave-peak w1 and wave-valley w2, two formulas can be obtained using Formula 8.

$$-(\alpha+1)*W_R(\alpha,w1) = N*W_{k_e}(\alpha,w1) + C_{ab}*W_{k_{ab}}(\alpha,w1) \quad [9]$$

$$-(\alpha+1)*W_R(\alpha,w2) = N*W_{k_e}(\alpha,w2) + C_{ab}*W_{k_{ab}}(\alpha,w2) \quad [10]$$

$$C_{ab} = N* \frac{W_R(a,w1)*W_{k_e}(a,w2) - W_R(a,w2)*W_{k_e}(a,w1)}{W_R(a,w2)*W_{k_{ab}}(a,w1) - W_R(a,w1)*W_{k_{ab}}(a,w2)} \quad [11]$$

where w1 equals 699 nm, w2 equals 613 nm, a equals 150 nm, and the wavelet function is "rbio1.1". According to Figure 2, we can obtain the wavelet coefficients W_{kab} and W_{Ke} as follows:

$$Wk_e(a,\text{w1}) = -0.1113, \quad Wk_e(a,\text{w2}) = -0.1134$$
$$W_{k_{ab}}(a,\text{w1}) = 0.2327, \quad W_{k_{ab}}(a,\text{w2}) = -0.1386 \quad [11a]$$

A new leaf chlorophyll content model, termed CAB1 with the wavelet function "rbio1.1," was obtained as equation 12:

$$CAB1 = C_{ab} = * \frac{-W_R(150,699)*0.1134 + W_R(150,613)*0.1113}{W_R(150,613)*0.2327 + W_R(150,699)*0.1386} \quad [12]$$

CAB1 is only affected by the parameter N. Then, for the same crop type during the same growth period, N was a fixed constant and CAB1 was only affected by the leaf spectrum reflectance rate.

The Spectral Indices Sensitive to Leaf Chlorophyll Contents

To test the established inversion model CAB1, eight commonly used spectral indices (Table II) sensitive to leaf chlorophyll contents were used for comparison, including the spectral index of visible light NPCI, TCARI, and MCARI; spectral indices of visible and infrared light MTCI, TVI, and TCARI/OSAVI; and red-edge spectral index GM and VOG2.

Results

Two aspects should be considered in the spectral index evaluation: The spectral index is sensitive to chlorophyll content, and it can resist the influence of other factors. First, the Prospect optical model was used to simulate the spectral curves with different chlorophyll content, and the advantages and disadvantages of each spectral index were analyzed. According to existing spectral indices (Table II), the band range was selected from 400 to 800 nm. According to the analysis in Figure 1, the influencing factors only included N, C_m, and C_{ar}. Finally, the Lopex93 and Angers real data were used to analyze the spectral index.

Spectral Indices Sensitivity for Leaf Chlorophyll Content C_{ab}

The data of Set_C_{ar} is used for spectral indices sensitivity for leaf chlorophyll content C_{ab}. As shown in Figure 3, the values of CAB1, MTCI, and GM monotonously increase with the increase in chlorophyll content, and NPCI, TCARI, MCARI, TVI, TCARI/OSAVI, and VOG2 decrease. They have a good linear relationship with the chlorophyll content, and their R^2 are greater than 0.9—except NPCI, R^2 = 0.6371) (the first line of Table III)—indicating that these eight spectral indices have good sensitivity to different chlorophyll content. When the chlorophyll content was less than 30 µg/cm^2, TVI exhibited an opposite monotonic feature compared to that when the content was greater than 30 µg/cm^2, which indicates that when the C_{ab} content is low (less than 30 µg/cm^2), TVI is not suitable for retrieving leaf chlorophyll. Similarly, at a chlorophyll content of 40 µg/cm^2, NPCI showed the opposite monotonic characteristics, and the curve gradually changed and usually became smooth, indicating that NPCI was easy to saturate.

Ability to Resist the Influences of Other Leaf Biochemical Components

The test result of Set_N is shown in Table III (the second line). The influence of N on the partial spectral index is large, such as NPCI, TCARI, MTCI, and TCARI/OSAVI. Under the same chlorophyll content, the MTCI value changed approximately 19 times when the N changed from 1 to 3, followed by NPCI, TCARI, and TCARI/OSAVI. The SI of VOG2, GM, and CAB1 are smaller, particularly the SI value of CAB1, which is only 34.87%.

The test result of Set_C_m is shown in Table III (the third line). The influence of C_m on the partial spectral index is small. Only the SI of TVI is over 10%. TCARI/OSAVI, NPCI, and CAB1 had the smallest SI values, particularly for CAB1, which had a SI value of only 0.56%.

The test result of Set_C_{ar} is shown in Table III (the fourth line). The influence of C_{ar} on the partial spectral index is very small. Only the SI of TCARI and TCARI/OSAVI are over 5%. The SI of CAB1, NPCI, MCARI, and TVI are approximately 2%, and the SIs of MTCI, GM, and VOG2 are all approximately 0%.

Spectral Index Verification with Lopex93 and Angers Data

Using the 190 Lopex93 reflectivity spectral data, the quadratic model of CAB1, TCARI, MTCI, and VOG2 were built. The 80 validation spectral data were used to verify the inversion chlorophyll content accuracy. For Angers data, 69 reflectivity spectral data were used to build the quadratic model of CAB1, TCARI, MTCI, and VOG2, and 60 validation spectral data were used as validation data. Because the structural parameter of the leaf N cannot be determined in the Lopex93 and Angers data, N is assumed to be the standard value, which is 2. The R^2 and RMSE are used to evaluate their accuracy (Table IV). Validation results of chlorophyll estimations are done by the quadratic model of CAB1, TCARI, MTCI, and VOG2, which is shown in Figure 4.

As shown in Table IV and Figure 4, the quadratic models of CAB1,

TCARI, MTCI, and VOG2 with Lopex93 data have low RMSE (RMSE < 3.20). The R^2 of MTCI is greater than 0.72, which is better than TCARI and VOG2. The R^2 of CAB1 is 0.9108, and the RMSE values are smaller (RMSE = 2.0294). Compared with TCARI, MTCI, and VOG2, CAB1 is more accurate and robust.

The results of CAB1, TCARI, MTCI, and VOG2 estimating chlorophyll with Angers data are shown in Table IV and Figure 4. The quadratic models of CAB1, TCARI, MTCI, and VOG2 with Lopex93 data have high R^2 (R^2 > 0.84), which indicates that they have high accuracy for chlorophyll estimation. The R^2 of MTCI is greater than 0.89, which is better than TCARI and VOG2. The R^2 of CAB1 is 0.9116, and the RMSE values are smaller (RMSE = 3.6925). Compared with TCARI, MTCI, and VOG2, CAB1 is more accurate and robust.

From the analysis of two sets of data, it was found that the R^2 of CAB1 was the highest, and the RMSE was the smallest among the four spectral indexes, indicating that CAB1 performs the best.

TABLE IV: Results of spectral index TCARI, MTCI, VOG2, and CAB1 estimating chlorophyll with LOPEX93 data

Data	Spectral Index	Fitting Equation	R^2	RMSE (μg/cm²)
Lopex93 data	TCARI	$-91.894x^2 - 44.354x + 53.557$	0.7139	2.9286
	MTCI	$-2.7484x^2 + 20.904x + 11.615$	0.7254	3.1681
	VOG2	$-75.849x^2 + 277.12x - 202.43$	0.6997	3.2044
	CAB1	$-3.6676x^2 - 6.2732x + 52.924$	0.9108	2.0294
Angers data	TCARI	$498.62x^2 - 401.19x + 103.68$	0.8424	4.2714
	MTCI	$-4.0092x^2 + 34.616x + 0.4941$	0.8939	3.7217
	VOG2	$-2304.6x^2 - 775.87x + 1.1046$	0.8668	4.1753
	CAB1	$21.029x^2 + 137.62x + 247.41$	0.9116	3.6925

Discussion

The Performance of Wavelet Analysis Method with Prospect Model in the Reconstruction of Leaf Spectral Index

As shown in equation 8 using the spectral analysis method, the spectral reflectance is limited to 700–800 nm, and the inversion of chlorophyll content has only two important factors: N and C_{ab}. By using continuous wavelet decomposition, the spectral energy can be decomposed to find the band position with the largest C_{ab} effect and the smallest N effect, thus providing the possibility of constructing a new spectral index for chlorophyll inversion (Figure 2).

The Prospect model starts with spectral and vegetation components and has a strong theoretical basis. Therefore, the spectral index is constructed step-by-step based on the Prospect formula, which has sufficient theoretical basis. However, the results are still affected by N (equation 12). Although the experimental results show that the effect of N on CAB1 is relatively small, the accuracy of vegetation index would be further improved if the specific value of N can be obtained from the measured data. In addition, the model used in this experiment was Prospect-5, which does not separate chlorophyll from anthocyanin, resulting in a greater impact of C_{ar}, shown in Table III (the fourth line). With the refinement of the Prospect model, the accuracy of building a new vegetation index is expected to be further improved.

The Performance of the CAB1 Method in the Estimation of Leaf Parameters

From the Prospect simulation data, CAB1 is in the upper and middle reaches, shown in Table III. Through the aforementioned analysis, the spectral index on the sensitivity of chlorophyll content and ability to resist the influences of other components such as the structure of mesophyll, dry matter, and carotenoids, were summarized to estimate the effectiveness of assessing chlorophyll content. The comprehensive evaluation index (*Ava*) is defined as follows:

$$Ava = \frac{P \times R^2 + P_i \times (1-SI_i\%) + P_{i+1} \times (1-SI_{i+1}\%) \ldots + P_{i+n} \times (1-SI_{i+n}\%)}{P + P_i + P_{i+1} \ldots P_{i+n}} \quad [13]$$

where R^2 is the spectral sensitivity of the biochemical components, as represented by the index trend line fitting degree value R^2 (Table III); 1 – *SI*% is the ability of spectral indices to resist other biochemical components (*SI*% are selected in Table III); P, P_i … P_{i+n} are the influence weights for each biochemical component, which were set at 1. When the *Ava* of the spectral index is higher, the ability to resist the influence of other components is stronger, indicating that the spectral index comprehensive ability is stronger.

Table III shows that NPCI, TCARI/OSAVI, and MCARI have the low *Ava* value and are not suitable for evaluating chlorophyll content in a vegetation leaf. MTCI and TVI comprehensive evaluations are medium, and they have a certain guiding role for the particular vegetation or a certain period of growth of vegetation but are not used for all vegetation. The *Ava* values of CAB1, GM, and VOG2 are higher (greater than 0.8), and the models are sensitive to chlorophyll content and are relatively stable. The CAB1 integrated capacity evaluations

were equivalent to model sensitivity and stability optimum. CAB1 is based on the Prospect model, so using Prospect simulation data can explain the status quo of the vegetation index. Although CAB1 does perform relatively well based on the Lopex93 and Angers data validation, more measured data are needed for further validation.

Conclusions

This study generated a new spectral index, named CAB1, to estimate the chlorophyll content of a vegetation leaf. Based on the Prospect optical model, the continuous wavelet analysis method was used to establish the CAB1 model, in which the "rbio1.1" wavelet basis function with a scale of 150 nm, as well as reflectance spectra for 613 and 619 nm, were selected. The CAB1 model is derived from the Prospect model, so the CAB1 model has a strict physical meaning. The practical adaptability of some spectral indices, which are based on statistics, was the biggest disadvantage. CAB1 is not based on statistics, but instead on the theory of real light propagation, which is more adaptable.

A good chlorophyll vegetation index should be sensitive to chlorophyll while simultaneously not being sensitive to other factors. To verify the adaptability of CAB1, a comparison was made with eight commonly used spectral indices. According to the Prospect model, the data describing the relationship between the leaf spectrum and the content of each component were simulated. A comparison of the chlorophyll sensitivity and resistance to the mesophyll structure, dry matter, and carotenoid content of the CAB1 model and the common chlorophyll spectral indices was examined. MCARI was found to perform the worst (*Ava* = -3.7600). MCARI is very sensitive to the chlorophyll content but also to other components such that it is only suitable for retrieving the chlorophyll content of the same vegetation leaf within a single period. The VOG2, GM, and CAB1 models performed the best (*Ava* > 0.8) because they can retrieve the chlorophyll content of any vegetation leaf with high accuracy. To verify this conclusion, the actual data of Lopex93 and Angers were selected. CAB1 (proposed method), TCARI (spectral index of visible light), MTCI (spectral indices of visible and infrared light), and GM (red edge spectral index) were selected to retrieve the chlorophyll content of vegetation leaves using the Lopex93 and Angers data. Based on the actual data, MTCI, TCARI, and VOG2 were found to have moderate inversion accuracy (the maximal R^2 is 0.8), which may be because of noise in the actual data. However, CAB1 offers a high level of accuracy ($R^2 > 0.9$), which indicates the CAB1 model is not only more accurate than the other vegetation indexes, but is also very stable. Overall, CAB1 performed well in the leaf layer compared to the commonly used spectral indices NPCI, TCARI, and MCAI in visible regions, and CAB1 is comparable to the red edge spectral index GM and VOG2.

This research was based on the Prospect model simulation data. Because of the influence of the accuracy of the model, and given that the experimental data may be slightly different, the universality of the research results must be further verified. However, the results of this study provide some pointers for combining the physical model and the spectral index method.

Funding Information

This work was supported by the Joint Innovative Center for Safe and Effective Mining Technology and Equipment of Coal Resources, Shandong Province, and Scientific Research Foundation of Shandong University of Science and Technology under Grant [2014TDJH101]; Scientific and Technological Planning Projects of Colleges and Universities in Shandong Province under Grant [J18KA214]; and Funded by Beijing Key Laboratory of Urban Spatial Information Engineering [2019209].

Acknowledgment

We thank the SDUST Laboratory for the use of their equipment. The authors also thank the editor and anonymous referees for their insightful comments, which improved this paper.

References

(1) E. Boegh, H. Soegaard, and A. Thomsen, *Remote Sens. Environ.* **79**(2–3), 329–343 (2002).

(2) C.S.T. Daughtry, C.L. Walthall, M.S. Kim et al., *Remote Sens. Environ.* **74**(2), 229–239 (2000).

(3) Z. Liangpei, *Hyperspectral Remote Sensing* (Wuhan University Press, Wuhan, China, 2011).

(4) S.T. Brantley, J.C. Zinnert, and D.R. Young, *Remote Sens. Environ.* **115**(2), 514–523 (2011).

(5) E.R. Hunt, P.C. Doraiswamy, J.E. Mcmurtrey, et al., *Int. J. Appl. Earth. Obs.* **21**(1), 103–112 (2013).

(6) B. Datt, *Remote Sens. Environ.* **66**(2), 111–121 (1998).

(7) A.A. Gitelson, Y. Gritz, and M.N. Merzlyak, *J. Plant Physiol.* **160**(3), 271–282 (2003).

(8) A.A. Gitelson, G.P. Keydan, and M.N. Merzlyak, *Geophys. Res. Lett.* **33**(11), 431–433 (2006).

(9) D. Haboudane, J.R. Miller, N. Tremblay, et al., *Remote Sens. Environ.* **81**(2–3), 416–426 (2002).

(10) G.A. Carter and B.A. Spiering, *J. Environ. Qual.* **31**(5), 1424–1432 (2002).

(11) G.L. Maire, C. FranOis and E. Dufrêne, *Remote Sens. Environ.* **89**(1), 1–28 (2004).

(12) J.R. Miller, E.W. Hare, and W.U., *Int. J. Remote Sens.* **11**(10), 1755–1773 (1990).

(13) Y. Chun-Yan, N. Zheng, W. Ji-Hua, et al., *J. Remote Sens.* **9**, 742–750 (2005).

(14) I. Baek, M. Kim, and B.K. Cho, et al., *Applied Sciences* **9**(5), 1027–1038 (2019).

(15) J. Sun, Y. Shishou, J. Yang, et al., *Remote Sens. Environ.* **212**, 1–7 (2018).

(16) S. Jacquemoud and F. Baret, *Remote Sens. Environ.* **34**(2), 75–91 (1990).

(17) T.P. Dawson, *Remote Sens. Environ.* **65**(1), 50–60 (1998).
(18) B.D. Ganapol, *Remote Sens. Environ.* **63**(2), 182–193 (1998).
(19) H.L. Jiang, L.F. Zhang, H. Yang, et al., *Spectrosc. Spect. Anal.* **36**(1), 169–176 (2016).
(20) H. Wang, R.H. Shi, P.D. Liu et al., *Spectrosc. Spect. Anal.* **36**(7), 2189–2194 (2016).
(21) Y.H. Zhang, *Acta Ecologica Sinica.* **33**, 876–887 (2013).
(22) W.A. Allen, *J. Opt. Soc. Am.* **63**(6), 664–666 (1973).
(23) J.B. Feret, C. FranOis, G.P. Asner, et al., *Remote Sens. Environ.* **112**(6), 3030–3043 (2011).
(24) L. Liu, B. Song, and S. Zhang, *Remote Sens.* **9**(11), 1113–1137 (2017).
(25) S. Jacquemoud, S.L. Ustin, J. Verdebout, et al., *Remote Sens. Environ.* **56**(3), 194–202 (1996).
(26) R. Shi, D. Zhuang, and Z. Niu, *Chinese Journal of Ecology* **25**, 591–595 (2006).
(27) B. Hosgood and G. Andreoli, *Leaf Optical Properties Experiment 93 (LOPEX93)* (Publications office of the Joint Research Centre: European Commision/Institute for Remote Sensing Applications, 1994).
(28) S. Mallat, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **11**, 674–693 (1989).
(29) D. Chen, B. Hu, X. Shao, et al., *Analyst* **129**(7), 664–669 (2004).
(30) Q. Liao, J. Wang, G. Yang, et al., *J. Appl. Remote Sens.* **7**(1), 3575–3587 (2013).
(31) Y. Guo, Y. Zhou and J. Tan, *J. Thorac Bio.* **370**, 116–120 (2015).
(32) M.S. Bargh, *Wavelet Transform and its Application in the Decomposition and Reconstruction of Signals* (Eindhoven University of Technology, Eindhoven, The Netherlands, 1994).
(33) E.B. Knipling, *Remote Sens. Environ.* **1**(3), 155–159 (1970).
(34) J. Peñuelas, J. A. Gamon, A. L. Fredeen et al., *Remote Sens. Environ.* **48**(2), 135–146 (1994).
(35) F. Meggio, P.J. Zarco-Tejada, L.C. Núuez et al., *Remote Sens. Environ.* **114**(9), 1968–1986 (2010).
(36) J. Dash and P.J. Curran, *Int. J. Remote Sens.* **25**, 5403–5413 (2004).
(37) N. H. Broge and E. Leblanc, *Remote Sens. Environ.* **76**(2), 156–172 (2001).
(38) A. Gitelson and M.N. Merzlyak, *J. Plant Physiol.* **143**(3), 286–292 (1994).
(39) P.J. Zarco-Tejada, J.R. Miller, T.L. Noland, et al., *IEEE Transactions on Geo. & Remote Sens.* **39**(7), 1491–1507 (2001).

Feifei Xie and **Lin Sun** are with the State Key Laboratory of Mining Disaster Prevention and Control at Shandong University of Science and Technology in Qingdao, China. **Xie** and **Fengzhu Liu** are with the Beijing Key Laboratory of Urban Spacial Information Engineering, in Beijing, China. **Jie Wang** is with Inner Mongolia University in Hohhot, China. **Liu** is also with the Beijing Institute of Surveying and Mapping, Beijing, China. Direct correspondence to:xff@sdust.edu.cn ●

● Continued from Page 13

(17) H. Li, J.X. Wang, Z.N. Xing, and G. Shen, *Spectrosc. Spectral Anal.* **31**(2), 362–365 (2011). Doi: 10.3964/j.issn.1000-0593(2011)02-0362-04
(18) W. Liu, Z. Zhao, H.F. Yuan, C.F. Song, and X.Y. Li, *Spectrosc. Spectral Anal.* **34**(4), 947–951 (2014). Doi: 10.3964/j.issn.1000-0593(2014)04-0947-05
(19) K.N. Basri, M.N. Hussain, J. Bakar, Z. Sharif, and M.F.A. Khir, *Spectrochim Acta A* **173,** 335–342 (2017).
(20) M. Asachi, A. Hassanpour, and M. Ghadiri, et al., *Powder Technol.* **320,** 143–154 (2017).
(21) R.J. Barnes, M.S. Dhanoa, and S.J. Lister, *Appl. Spectrosc.* **43**(5), 772–777 (2016).
(22) Q.X. Zhang, Q.B. Li, and G.G. Zhang, *Spectroscopy* **26**(7), 28-39 (2011).
(23) P.Y. Diwu, X.H. Bian, Z.F. Wang, and W. Liu, *Spectrosc. Spectral Anal.* **39**(9), 2800–2806 (2018).
(24) V.K. Keerthi and B. Surendiran, *Perspectives in Science* **8(**C), 510–512 (2016). Doi: 10.1016/j.pisc.2016.05.010
(25) Y.F. Wang, X.D. Ma, and J.J Malcolm, *Renew Energ.* **97,** 444–456 (2016). Doi: 10.1016/j.renene.2016.06.006
(26) K.H. Wong, N.V. Razmovski, K.M. Li, G.Q. Li, and K. Chan, *J. Pharmaceut. Biomed.* **84,** 5–13 (2013). Doi: 10.1016/j.jpba.2013.05.040
(27) L. Sthle and S. Wold, *J. Chemomet.* **1**(3), 185–196 (1987). Doi: 10.1002/cem.1180010306
(28) M. Barker, and W. Rayens, *J. Chemomet.* **17**(3), 166–173 (2003). Doi: 10.1002/cem.785
(29) Q.S. Xu and Y.Z. Liang, *Chemomet. Intell Lab.* **56**(1), 1–11 (2001). Doi: 10.1016/S0169-7439(00)00122-2
(30) C. Cortes and V. Vapnik, *Machine Learning* **20**(3), 273-297 (1995). Doi: 10.1007/BF00994018
(31) J. Kivinen, A.J. Smola, and R.C. Williamson, *IEEE T. Signal Proces.* **52**(8), 2165–2176 (2004). Doi: 10.1109/TSP.2004.830991
(32) J.B. Li, Y.K. Peng, L.P. Chen, and W.Q Huang, *Spectrosc. Spectral Anal.* **34**(5), 1264–1269 (2014).
(33) K.Y. Zheng, T. Feng, W. Zhang, and X.W Huang, et al., *Chemometr Intell Lab.* **191,** 109–117 (2019).
(34) S. Hiroshi, N. Kenji, Y. Hideki, M. YohIchi, and S. Hayaru, *Sci. Technol. Adv. Mat.* **21**(1), 402-419 (2020). Doi: 10.1080/14686996.2020.1773210
(35) F. Guo, and M.A. Clemens, *Spectrochim. Acta A* **211,** 254–259 (2019). Doi:10.1016/j.saa.2018.12.012
(36) C.H. Liu, X.H. Lu, and H.Y. Fan, *Environ. Sci. Manag.* **36**(3), 183–186 (2011)
(37) E. Bonah, X.Y. Huang, R. Yi, J.H. Aheto, and S.H. Yu, *Infrared Phys. Tech.* **105,** (2020). Doi: 10.1016/j.infrared.2020.103220
(38) J. Hui, W.D. Xu, Y.H. Ding, and Q.S. Chen, *Spectrochim. Acta A* **228** (2019). Doi: 10.1016/j.saa.2019.117781

Yong Hao and **Qiming Wang** are with the School of Mechatronics and Vehicle Engineering, at East China JiaoTong University, in Nanchang, China. **Shumin Zhang** is with the Technology Center of Nanchang Customs District, in Nanchang, China. Direct correspondence to: haonm@163.com ●

PEER-REVIEWED RESEARCH
Spectroscopy papers have undergone a double-blind peer review process and are available on our open-access website.

Inversion of Low-Grade Copper Mining Areas Based on Spectral Information and Remote Sensing Data Using Vis-NIR

Dong Xiao, Hongfei Xie, Yanhua Fu, and Feifei Li

With the continuous exploitation and utilization of mineral resources, the mineral reserves of all countries in the world are decreasing. In this case, the boundary grades and industrial grades of ore are bound to be adjusted downward along with the decrease of mineral resources. Low-grade ore will have mining value and bring economic benefits to enterprises. For low-grade ore, the traditional content determination has the disadvantages of high cost and long time consumption. Therefore, it needs a method that can quickly identify the content of low-grade ore. In addition, mining will destroy the surrounding ecological environment and cause heavy metals in the land to exceed the standard. This paper proposes a method of using spectral information and remote sensing data to determine copper content in mining areas. We trained the calibration model with spectral data as input, and the copper content of the ore as output. Finally, through the remote sensing information of the mining area, the metal content of the entire mining area is inverted. This provides guidance for the later beneficiation technology of ore, and the reclamation of the land after mining.

As one of the earliest nonferrous metals smelted and used by humans, copper is closely related to human activities. More than 7000 years ago, copper was used by humans in waging war (1). Some scholars believe that a Copper Age occurred between the Stone Age and the Bronze Age (2). With the development of modern technology and the progression of society, copper has also been widely used in electronic power, transportation, and construction. In addition to its good performance in traditional industries, copper is also one of the essential elements of human life. Through experimental research, copper has proven to have a beneficial effect on patients with anemia (3). Copper also has shown the ability to prevent cardiovascular diseases (4). Because copper has the effect of eliminating infectious viruses, medical scientists have considered using copper to prevent and treat Covid-19 (5). As the fields of application continue to expand, the copper smelting industry has also been developing, occupying a large proportion of the international economy. However, with the continuous exploitation and utilization of copper resources, the reserves of some mining areas have been exhausted. Because of the different grades of ore, the beneficiation process adopted by copper-related enterprises is also different. For these enterprises, a method that can quickly and accurately determine the heavy metals in ore can save time and bring more economic benefits to enterprises.

There are many methods for measuring copper ore (6–8). The first is the electrochemical method. This method uses the oxidation-reduction reaction of the metal to determine the metal content in the mixture. Although this method is simple, since the electrode material plays an important role in the oxidation-reduction reaction, the choice of electrode material has always been an important issue. The choice of electrode material directly affects the oxidation-reduction reaction in the electrolyte. The flame atomic absorption method is also often used to determine the metal content of the mixture. This method is an elemental measurement and analysis technique

that uses the atomic resonance radiation absorption of the substance to be measured in the vapor state. This method has high sensitivity and strong anti-interference ability, but because of its small working linear range, it has a relatively strict limit on the concentration of the sample. Therefore, for high-concentration samples, the content can be determined only after dilution. As a result, finding a stable and accurate determination of copper content in copper ore is a major problem.

Although mining brings economic benefits to enterprises, it also causes huge damage to the environment as well as ecological imbalance, damaging the health of surrounding residents. Heavy metals can pollute the soil and pose a high risk of carcinogenesis to children near mining areas (9). Because heavy metals cannot be decomposed by microorganisms, they can easily enter the human body through the food chain, causing heavy metal poisoning (10). Sadeq and Beckerman proposed that the excessive intake of copper ions affected the reproduction of Cladocera (11). Khan and associates proposed that human beings accumulate heavy metals mainly by eating heavy metal contaminated food. If the food grown in contaminated land is consumed, both children and adults will take in a large amount of heavy metals that will affect their health (12). Duruibe and associates proposed that mining activities may release a large amount of heavy metals into the environment, causing soil and water pollution (13). Zhuang and colleagues investigated the soil around the Dabaoshan mining area, and found that the heavy metal content in the soil exceeded the maximum value of the agricultural soil heavy metal content standard; if residents eat local rice and vegetables, it will bring great risks to their health (14).

With the increasing scale of urbanization in China, the area of arable land has decreased sharply (15). The government has also been advocating the reclamation of contaminated soil to reuse it. Therefore, this paper proposes the use of spectral data and remote sensing information combined with machine learning to model the content of heavy metals in the soil, analyze the content of heavy metals in the mining area, and provide guidance for the beneficiation and reclamation of enterprises.

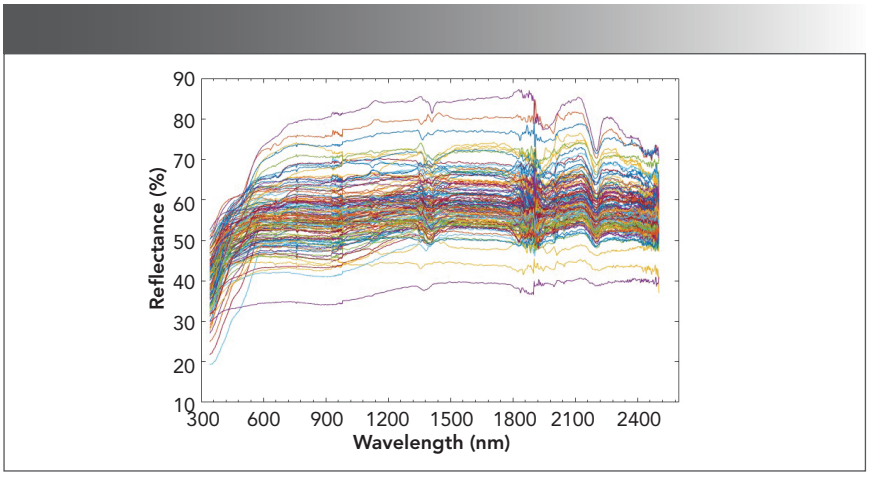


FIGURE 1: Spectral data for copper ore samples.

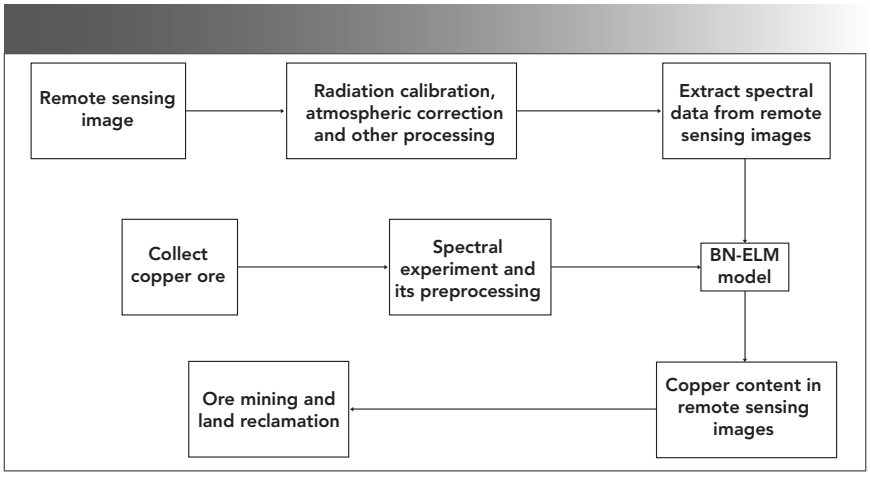


FIGURE 2: Flow chart of heavy metal inversion in the mining area.

TABLE I: Hidden layer activation function

Activation Functions	Mathematical Formula
Sigmoid function	$g(x)=\frac{1}{1+e^{-x}}$
Tanh function	$g(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
Relu function	$g(x)=\max(x,0)$
Sine function	$g(x)=\sin(x)$

The Unugetu Copper Mine in Manchuria, Inner Mongolia, is a porphyry copper deposit with the characteristics of large scale and low grade. The mining area is located in the Hulunbuir grassland, and the local residents live as nomads. Because of continuous mining in the mining area, the soil was polluted by heavy metals. If the soil is not reclaimed in time, it will

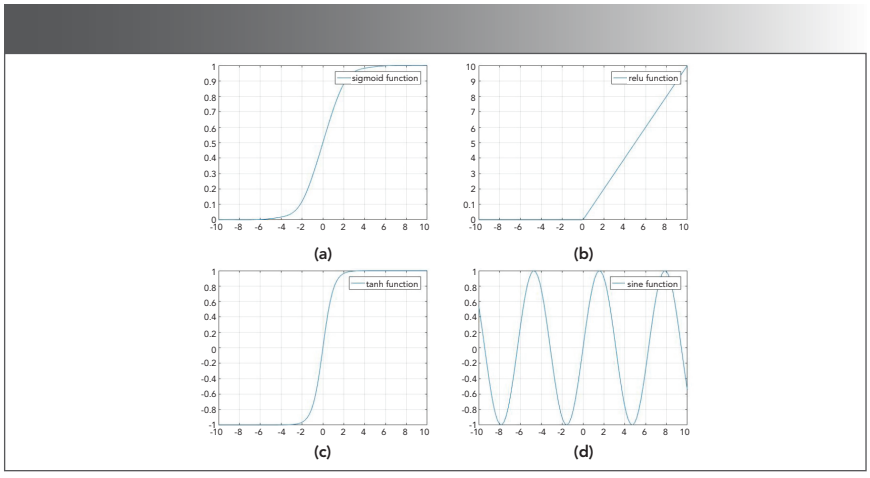


FIGURE 3: Images showing activation functions from Table I.

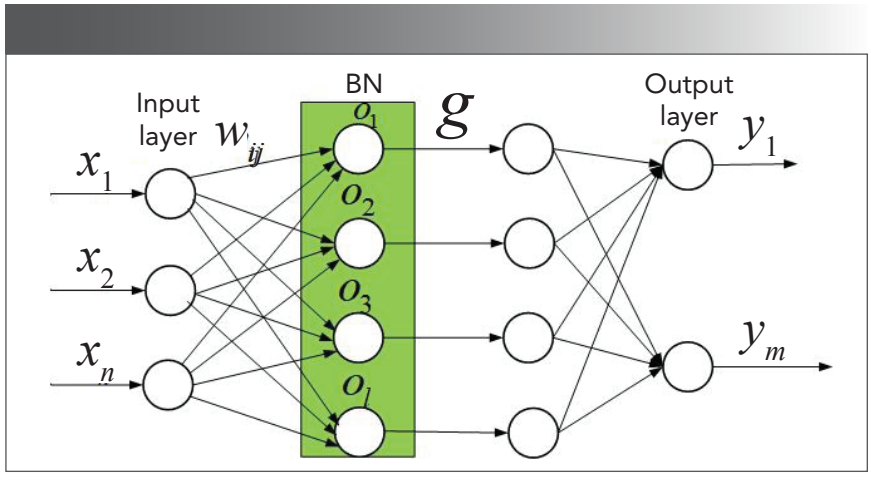


FIGURE 4: Illustration of BN-ELM network structure.

TABLE II: Coefficient of determination (R^2) values for BP, RBF, ELM, and BN-ELM

BP	RBF	ELM	BN-ELM
0.667	0.121	0.664	0.814
0.442	0.003	0.641	0.807
0.276	0.018	0.482	0.776
0.612	0.232	0.609	0.805
0.269	0.156	0.677	0.867

bring health risks to the local inhabitants. Therefore, the soil should be reclaimed while mining in the mining area. Peng and colleagues found that gamma PGA can effectively remove 74.3% copper in soil (16). Through experiments, Jin and associates found that phytoremediation can effectively control soil heavy metal pollution (17). Zhang and Zhou discussed the extraction efficiency of three kinds of chemical extractions for soil heavy metals, and found that different extraction methods should be selected for soils contaminated by different heavy metals (18). Morong and Aggangan discussed the possibility of repairing heavy metal contaminated areas by three tree species of Indian maple, *Acacia mangium* and *Eucalyptus urophylla*, and found that these three tree species can effectively remove copper, lead, and cadmium from the soil (19). This paper proposes the inversion of copper mining areas based on spectral information and remote sensing data, firstly inverting the content of heavy metal copper in the mining area, and then choosing different soil remediation methods according to the mined-out areas with different concentrations of heavy metals.

Spectral analysis is used to determine the reflectance and absorbance at different wavelengths, based on the characteristics of different substances having specific spectral characteristics. Remote sensing can form multi-spectral remote sensing images based on the reflectance and absorption of solar radiation on the earth's surface. Remote sensing technology has the characteristics of fast data acquisition, short measurement period, and ability to measure dynamic sample changes. Wu and colleagues used remote sensing information to monitor the destruction of vegetation and landforms caused by coal mining in the Qinghai-Tibet Plateau (20). Song and associates summarized the progress of remote sensing monitoring in mining area boundary recognition and mining area land cover changes on the basis of previous studies (21). Koruyan and colleagues have shown through research that remote sensing is a valuable tool for management and planning of mining operations (22). Charou and associates used Landsat 5 and Landsat 8 to monitor the surface characteristics and water changes of abandoned land in the mining area (23). The final results show that remote sensing data can be used for long-term environmental management and monitoring of mining area reclamation and restoration.

Machine learning can be divided into three main categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning can also be divided into two categories. The first is classification, and the other is regression. In regression, machine learning can use the training data to adjust parameters and then make predictions on the test set data. This method has very powerful

data fitting capabilities, so more and more researchers have begun to use machine learning to conduct mining area research. Wei and associates used an improved convolutional neural network (CNN) to achieve accurate classification of features of lithofacies (24). Liu and colleagues used the combination of data and machine learning to explore the problem of mineral resource prediction, and used the Zhaojikou lead-zinc deposit in Anhui as an example to establish a CNN model (25). The prediction accuracy of the network model can reach 0.93. Le and associates established a coal classification model using spectral data and multilayer extreme learning machine algorithm, and correctly predicted the distribution of different coals (26).

Extreme learning machine (ELM) (27) is a typical feedforward neural network, and it does not need to update the connection weights by the gradient descent method. In this algorithm, the connection weight of the hidden layer and the input layer and the threshold of the hidden layer are randomly generated, and then the parameter β is solved by the least square method to obtain the unique optimal solution. Moreover, this algorithm only needs to manually set the number of neurons, and does not require parameter adjustment during network operation. Therefore, the ELM algorithm has the advantages of simplicity and fast running speed. Because the ELM algorithm has advantages that traditional neural networks do not have, more and more researchers have begun to study this algorithm. The setting of hidden layer nodes in neural networks can often only be based on empirical formulas, and neural networks are very sensitive to hidden layer nodes. Therefore, Hang and others proposed the incremental extreme learning machine (I-ELM) (28). The I-ELM algorithm adds a neuron to the hidden layer during each learning process until the error generated by the algorithm can reach the ideal value. To further improve the convergence of I-ELM, Huang proposed the Convex incremental extreme learning machine (Convex I-ELM) (29). Contrary to the incremental extreme learning machine algorithm, Rong proposed a pruning extreme learning machine (P-ELM) (30). This algorithm first constructs a neural network with many neurons, then measures the contribution of each neuron according to the set criteria, and finally gradually deletes redundant neurons to achieve the final simplified neural network. Based on P-ELM, Miche proposed an optimal pruning extreme learning machine that can handle classification and regression problems at the same time (31). This paper proposes the batch normalization–extreme learning machine (BN-ELM). In this algorithm, the output matrix of hidden layer is standardized to make the output of hidden layer fall within the sensitive range of activation function as much as possible, and avoid internal covariate shift. This method improves the generalization performance and sample learning ability of ELM.

Spectral Experiment and Remote Sensing Data Processing

Location of Unugetu Copper Mine

The Unugetu Copper Mine is located in the Hulunbuir Prairie in Manzhouli, Inner Mongolia. It is currently the first modern large-scale non-ferrous metal mine in the high and cold area of China. The Unugetu Copper Mine is open-pit mining. The latitude and longitude of the mining area is 117° 14′ -117° 22′ east longitude and 49° 22′ -49° 30′ north latitude. Figure 8a is a satellite image of the Unugetu Copper Mine.

Spectral Data

We collected 128 copper ore samples from the Unugetu Copper Mine, on August 20, 2017. We cleaned and dried the surface of the samples, and then grounded them into a powder. After the sample preparation was completed, we used the SVC HR-1024 portable spectrometer to perform spectrum experiments on each sample.

In the spectrum experiment, a whiteboard calibration was performed every ten replicates of the spectrum measurement. We conducted three replicate measurements on each sample to reduce measurement errors. We averaged the three replicates for each sample as its spectral data. After the completion of the spectrum measurements, the final spectral data of the copper ore for calibration was obtained by data preprocessing operations, such as coarsening and band fitting of the measured spectral data. Figure 1 shows the reflectance of copper ore samples at different wavelengths.

Remote Sensing Data

We downloaded Landsat-8 multispectral data from the U.S. Geological Survey website, and then used ENVI to perform radiometric calibration, atmospheric correction, image fusion, and other processing on the downloaded Landsat-8 multispectral data. Finally, as shown in Figure 8b, a multispectral image of the Unugetu copper mining area was obtained. Then used ENVI Class to extract the spectral data corresponding to each pixel in Figure 8b as the final remote sensing data.

Experimental Ideas

We used the spectral data in the training set as input and the copper content as output to train the BN-ELM model. The spectral data in the test set were then used as input to predict the content and evaluate the model. Finally, using the extracted remote sensing data as the input of BN-ELM model, the corresponding copper content of each pixel was obtained. The copper content of the whole mining area can be performed by this method, thus providing guidance for the mining and reclamation of the mining area in the future. The research idea is shown in Figure 2.

Experimental Method

Extreme Learning Machine

ELM is a typical feedforward neural network that can initialize the input weights and biases randomly,

and solve the output weights by the least square method. Suppose there are N arbitrarily different samples, where $x_i = [x_{i1}, x_{i2}, ..., x_{in}]^T \in \boldsymbol{R}^n$, $t_i [t_{i1}, t_{i2}, ..., t_{im}]^T \in \boldsymbol{R}^m$. The ELM calculation formula is:

$$f_L(x) = \sum_{i=1}^{L} \beta_i g_i(x) = \sum_{i=1}^{L} \beta_i g(\omega_i * x_j + b_i), j=1,\ldots ..., N \quad [1]$$

In equation 1, ω is the randomly initialized input weight, b_i is the random initialization bias, L is the number of hidden layer nodes set, g is the activation function, x_j is the number of inputs for each sample, and β_i is the weight of the output layer and the hidden layer. Simplify equation 1 to get:

$$H\beta = T \quad [2]$$

And using equation 2, we expand it to equations 3 and 4,

$$H = \begin{bmatrix} g(\omega_1 * x_1 + b_1) & \cdots & g(\omega_L * x_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(\omega_1 * x_N + b_1) & \cdots & g(\omega_L * x_N + b_L)_{N \times L} \end{bmatrix} \quad [3]$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m} \qquad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \quad [4]$$

We solve equation 2 by the least squares method:

$$\text{Minimize: } \|H\beta - T\| \quad [5]$$

Using the two theorems proposed by Huang, we can get equation 6.

$$\hat{\beta} = H^+ T \quad [6]$$

Where H^+ is the inverse matrix of H obtained by SVD decomposition, and Huang proved that $\hat{\beta}$ exists and is unique.

Batch Normalization-Extreme Learning Machine

In the ELM algorithm, the activation function can introduce nonlinear factors into the model, possibly making the input and output a nonlinear relationship. The activation function can also enable the neural network model to better solve complex problems and improve the generalization performance of the network. Commonly used activation functions are shown in Table I.

Sigmoid function (also known as *logistic function*) is used for the hidden layer neuron output, and its value range is (0,1). The graph of the sigmoid function presents an S-shaped curve. Because the output range of the sigmoid function is (0,1), the sigmoid function can be used to compress data, and is suitable for forward propagation. However, when the absolute value of the input variable is greater than 4, saturation will occur, and the output will become insensitive to small changes in the input. The tanh function is also called the hyperbolic tangent function, and its value range is [-1,1]. As shown in Figure 3; the tanh function also has saturation. By observing Figure 3, we found that when the absolute value of the input variable is greater than 4, the sigmoid and tanh func-

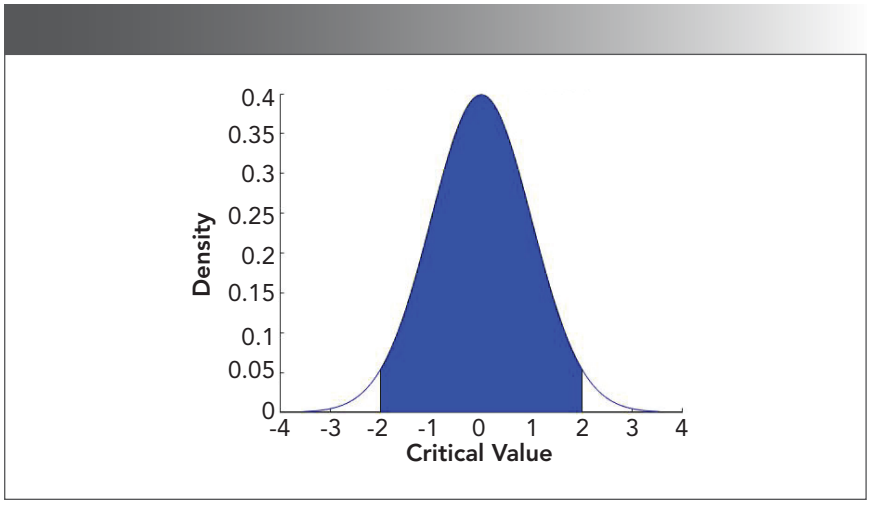


FIGURE 5: Illustration of standard Gaussian distribution probability density; probability between limits is 0.9545.

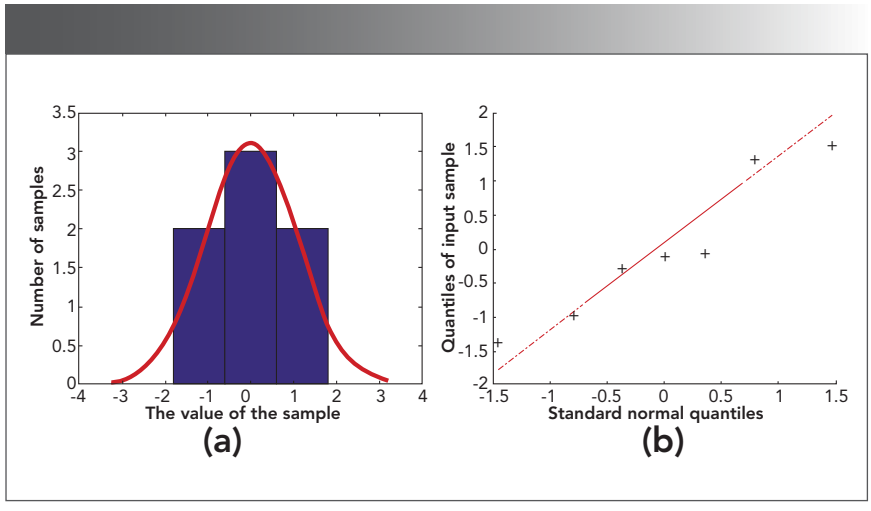


FIGURE 6: (a) Sample distribution showing Gauss distribution test; (b) Q-Q diagram of the sample, showing a plot of sample data versus standard normal data.

tion activation functions are close to output saturation. At this time, the output is not sensitive to subtle changes in the input. Using the rectified linear activation function (ReLU) will cause some neuron necrosis. At this time, the gradient of the neuron is 0, and no longer responds to any data. When the range of the input variable value is too large, the sine function is not a monotonic function.

To solve the saturation problem, we added a BN layer in front of the hidden layer (32). Figure 4 shows the network structure of the BN–ELM. The BN layer can make the output of the hidden layer fall on a normal distribution with a mean of 0 and a variance of 1. It can be seen from Figure 5 that the probability that the value of each data is within [-2,2] is 95.45%. Therefore, the training of the BN layer can not only make the output of the hidden layer fall between [-2, 2] as much as possible, but can also avoid an internal covariate shift.

Algorithm Flow of BN–ELM

In the BN–ELM algorithm, first randomly generate weight *w* and threshold *b* that conform to a Gaussian distribution with a mean value of 0 and a variance of 1. Because the hyperspectral remote sensing data has only seven dimensions, we assume that the 7-dimensional data conforms to the Gaussian distribution, and then normalize the spectral data with a z-score. Figure 6a shows the distribution of spectral data after z-score normalization. Figure 6b is a Q-Q chart of spectral data. It can be seen from Figures 6a and 6b that the spectral data of each group obeys the standard Gaussian distribution.

Then, we assume that the spectral data conforms to a Gaussian distribution with a mean of 0 and a variance of 1, that is: $X \sim N(\mu_1, \sigma_1^2)$. We make the weight *W* and the threshold *B* also conform to the Gaussian distribution, namely: $W \sim N(\mu_2, \sigma_2^2)$., $B \sim N(\mu_3, \sigma_3^2)$. So *TEH* = *WX* + *B* also conforms to the Gaussian distribution, the proof is as follows.

According to the probability density function of Gaussian distribution, we can get:

$$X \sim f(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \quad [7]$$

$$W \sim g(x) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \quad [8]$$

$$B \sim g(x) = \frac{1}{\sqrt{2\pi}\sigma_3} e^{-\frac{(x-\mu_2)^2}{2\sigma_3^2}} \quad [9]$$

First, we prove that *WX* = *g(x) f(x)* conforms to Gaussian distribution. It can be inferred from equation 7 and equation 8:

$$\begin{aligned} g(x)\,f(x) &= \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \\ &= \frac{1}{2\pi\sigma_2\sigma_1} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \\ &= \frac{1}{2\pi\sigma_2\sigma_1} e^{-\left(\frac{(x-\mu_2)^2}{2\sigma_2^2}+\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)} \\ &= \frac{1}{2\pi\sigma_2\sigma_1} e^{-\left(\frac{(x^2+\mu_2^2-2x\mu_2}{2\sigma_2^2}+\frac{x^2+\mu_1^2-2x\mu_1)}{2\sigma_1^2}\right)} \\ &= \frac{1}{2\pi\sigma_2\sigma_1} e^{-\left(\frac{(\sigma_1^2+\sigma_2^2)x^2-(2\mu_2\sigma_1^2+2\mu_1\sigma_2^2)x+\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2}{2\sigma_2^2\sigma_1^2}\right)} \\ &= \frac{1}{2\pi\sigma_2\sigma_1} e^{-\left(\frac{x^2-\frac{(2\mu_2\sigma_1^2+2\mu_1\sigma_2^2)}{\sigma_1^2+\sigma_2^2}x+\frac{\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}{\frac{2\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}\right)} \\ &= \frac{\sqrt{\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}}{2\pi\sigma_2\sigma_1\sqrt{\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} e^{-\left(\frac{(x-\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2+\frac{\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}-(\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2}{\frac{2\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}\right)} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} \frac{\sqrt{\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}}{\sqrt{2\pi\sigma_2^2\sigma_1^2}} e^{-\frac{(x-\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2}{2\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} e^{\frac{\frac{\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}-(\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2}{\frac{2\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} \\ &= \frac{1}{\sqrt{2\pi\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} e^{-\frac{(x-\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2}{2\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{\frac{\frac{\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}-(\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2}{\frac{2\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} \end{aligned} \quad [10]$$

Let:

$$a \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{\frac{\frac{\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}-(\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2}{\frac{2\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} \quad [11]$$

$$b = \frac{1}{\sqrt{2\pi\frac{\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} e^{-\frac{(x-\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2}{2\frac{2\sigma_2^2\sigma_1^2}{\sigma_1^2+\sigma_2^2}}} \quad [12]$$

So, *g(x) f(x)* conforms to the Gaussian distribution with a scaling factor of *a*:

$$WX \sim N\left(\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2}, \frac{\sigma_2^2+\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right) \quad [13]$$

Then:

$$WX + B \sim N\left(\frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2}{\sigma_1^2+\sigma_2^2}+\mu_3, \frac{\sigma_2^2+\sigma_1^2}{\sigma_1^2+\sigma_2^2}+\sigma_3^2\right) \quad [14]$$

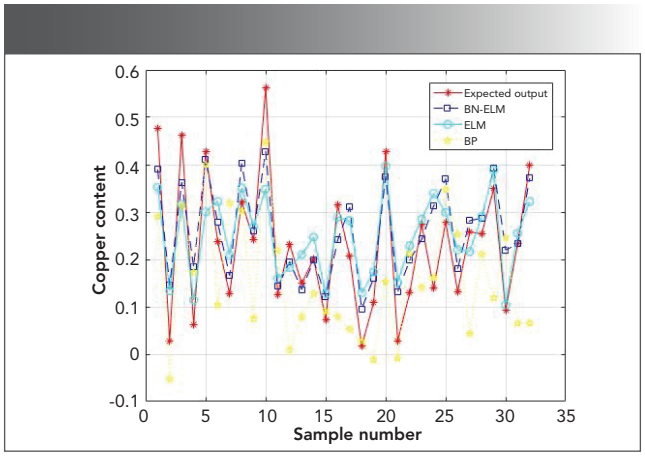


FIGURE 7: Comparison of copper ore content prediction results for various calculations versus actual (expected output in red).

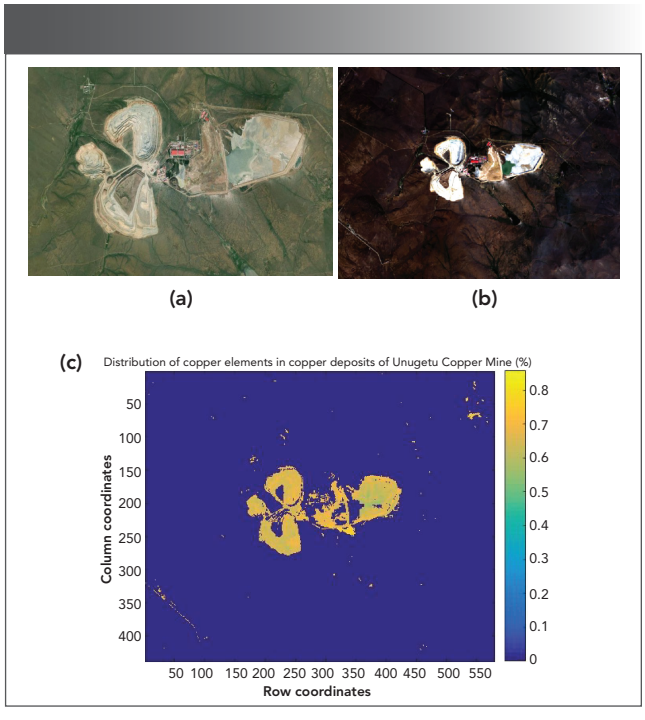


FIGURE 8: (a) Satellite image of copper mining area; (b) remote sensing image of mining area; (c) remote sensing retrieval of copper content in Unugetu Copper Mine.

TABLE III: Root mean square error (RMSE) values for BP, RBF, ELM, and BN-ELM

BP	RBF	ELM	BN-ELM
6.122×10^{-3}	147.199	6.149×10^{-3}	3.764×10^{-3}
12.672×10^{-3}	2.632	8.327×10^{-3}	4.139×10^{-3}
11.598×10^{-3}	56.254	7.819×10^{-3}	5.392×10^{-3}
11.943×10^{-3}	2000.713	12.313×10^{-3}	3.770×10^{-3}
22.592×10^{-3}	509.659	8.824×10^{-3}	3.655×10^{-3}

From equation 14, we can find that, because of the calculation of the hidden layer fully connected, the output matrix *TEH* no longer conforms to the Gaussian distribution with the mean of 0 and variance of 1. We therefore performed batch normalization processing on *TEH*. First, we calculate the mean and variance of each hidden layer node.

$$\mu_{TEH} = \frac{1}{L}\sum_{i=1}^{L} x_i \quad [15]$$

$$\sigma^2_{TEH} = \frac{1}{L}\sum_{i=1}^{L} (x_i - \mu_{TEH})^2 \quad [16]$$

$$TEH^* = \frac{TEH - \mu_{TEH}}{\sqrt{\sigma^2_{TEH}}} \quad [17]$$

It can be inferred from equation 15 and equation 16 that, if the training sample is large enough, the mean and variance of the hidden layer nodes will be more stable. The mean and variance of the internal offset generated during the training of the training set can be used to replace the mean and variance generated by the internal migration of all data. Therefore, there is no need to recalculate the mean and variance of hidden layer nodes when testing on the test set. After batch normalization processing, the hidden layer matrix obeys a Gaussian distribution with the mean of 0 and the variance of 1.

The output matrix of the hidden layer with the same distribution as the original data is obtained. We then use the activation function to calculate the hidden layer matrix, and finally use the least square method to get the output weight β. After the calculation of the BN layer, not only can the hidden layer value fall within [-2,2] as much as possible, but it will also retain the characteristics of the original input.

Experimental Results and Discussion

Comparison of Neural Network Algorithms

In this paper, four neural network algorithms of BP, RBF, ELM, and BN–ELM are used in the Matlab 2016a environment to compare and verify the feasibility of the BN–ELM algorithm. Multiple cross-validation methods were used for experiments to prove the stability of the BN–ELM algorithm. The experimental results are shown in Tables II and III. From these tables, it can be seen that the BP neural network can easily fall into the local optimal solution during the back propagation, so it presents an unstable phenomenon. The RBF neural network has the worst prediction result. The ELM does not need to calculate the weight *w* and the threshold *b* in reverse, so it will not fall into the local optimal solution. But in the ELM algorithm, because of the saturation of the activation function, the output is not sensitive to small changes in the input. BN–ELM solves the saturation problem of the ELM activation function, and keeps the hidden layer matrix within [-2,2] as much as possible. By comparison, it can be found that BN–ELM has the highest coefficient of deter-

mination and the smallest root mean square error. Figure 7 is a comparison of the best results of the cross-check.

Remote Sensing Inversion of Mining Area

We used NEVI Classic software to obtain all the spectral information in Figure 8b. Because we are using the Landsat 8 multi-spectral image, each pixel can extract the reflectance of seven bands, and each band corresponds to the spectrum of the copper ore collected from the Wunugetushan copper mine data. After BN–ELM simulation, the copper content corresponding to the whole Figure 8b is obtained. Finally, we plotted the copper content distribution of the entire mining area, as shown in Figure 8c.

Conclusion

With the continuous mining of the area studied, high-grade ore has been exhausted. Therefore, attention should be paid to low-grade ores. However, the low-grade ore contains less ore and has low mining value, so it brings less benefit to enterprises. The traditional copper ore grade determination is costly and difficult to be determined in large quantities. This paper puts forward the method of using spectral information and BN–ELM modeling to analyze the grade of copper ore. This method has the advantages of high speed and low cost. Finally, we use Landsat 8 remote sensing data to analyze the content of the whole mining area, providing guidance for future mining and land reclamation.

References

(1) M. Radetzki, *Resour. Policy* **34**(4), 176–184 (2009). DOI: 10.1016/j.resourpol.2009.03.003.

(2) M. Pearce, *J. World Prehistory* **32**(3), 229–250 (2019). DOI: 10.1007/s10963-019-09134-z.

(3) P. Fox, *BioMetals* **16**(1), 9–40 (2003). DOI: 10.1023/A:1020799512190.

(4) X.Y. Ma, S. Jiang, S.M. Yan, M. Li, C.C. Wang, Y.G. Pan, C. Sun, L.N. Jin, Y. Yao, and B. Li, *Biol. Trace Elem. Res.* **197**(1), 43–51 (2020). DOI: 10.1007/s12011-019-01979-x.

(5) S. Raha, R. Mallick, S. Basak, and A. K. Duttaroy, *Med. Hypotheses* **142,** 109814 (2020). DOI: 10.1016/j.mehy.2020.109814.

(6) W.Q. Zhang, S.M. Fan, X.L. Li, S.Q. Liu, D.W. Duan, L.P. Leng, C.X. Cui, Y.P. Zhang, and L.B. Qu, *Microchim. Acta* **187**(1), 69(2020). DOI: 10.1007/s00604-019-4044-y.

(7) K. Supong and P. Usapein, *Water Sci. Technol.* **79**(5), 833–841 (2019). DOI: 10.2166/wst.2019.072.

(8) D. Yildiz and M. Demir, *J. Anal. Chem.* **74**(5), 437–443 (2019). DOI: 10.1134/S1061934819050022.

(9) Z.Y. Li, Z.W. Ma, T.J. van der Kuijp, Z.W. Yuan, and L. Huang, *Sci. Total Environ.* **468,** 843–853 (2014).DOI: 10.1016/j.scitotenv.2013.08.090.

(10) X. Guan and L.N. Sun, *Appl. Mech. Mater.* **675–677,** 612–614 (2014). DOI: 10.4028/www.scientific.net/AMM.675-677.612

(11) S.A. Sadeq and A.P. Beckerman, *Arch. Environ. Contam. Toxicol.* **76**(1), 1–16 (2019). DOI: 10.1007/s00244-018-0555-5.

(12) S. Khan, Q. Cao, Y.M. Zheng, Y.Z. Huang, and Y.G. Zhu, *Environ. Pollut.* **152**(3), 686–692 (2008). DOI: 10.1016/j.envpol.2007.06.056.

(13) J.O. Duruibe, M.O.C. Ogwuegbu, and J.N. Egwurugwu, *Int. J. Phys. Sci.* **2**(5), 112–118 (2007). DOI: 10.1142/S0218127407018087.

(14) P. Zhuang, M.B. McBride, H.P. Xia, N.Y. Li, and Z.A. Lia, *Sci. Total Environ.* **407**(5), 1551–1561 (2009). DOI: 10.1016/j.scitotenv.2008.10.061

(15) M.H. Tan, X.B. Li, and C. Lu, *Land Use Policy* **22**(3), 187–196 (2005). DOI: 10.1016/j.landusepol.2004.03.003.

(16) Y.P. Peng, Y.C. Chang, K.F. Chen, and C.H. Wang, *Environ. Sci. Pollut. Res.* **27**(28), 34760–34769 (2019). DOI: 10.1007/s11356-019-07444-5.

(17) Z.M. Jin, S.Q. Deng, Y.C. Wen, Y.F. Jin, L. Pan, Y.F. Zhang, T. Black, K.C. Jones, H. Zhang, and D.Y. Zhang, *Sci. Total Environ.* **697,** 134148 (2019). DOI: 10.1016/j.scitotenv.2019.134148.

(18) F.Y. Zhang and Y. Zhou, *Soil Sediment Contam.* **29**(2), 246–255 (2020). DOI: 10.1080/15320383.2019.1702921.

(19) L.J.M. Morong and N.S. Aggangan, *Philipp. J. Crop Sci.* **44,** 18–27 (2019).

(20) Q.H. Wu, K. Liu, C.Q. Song, J.D. Wang, L.H. Ke, R.H. Ma, W.S. Zhang, H.Pan, and X.Y. Deng, *Sustainability* **10**(11), 3851 (2018). DOI: 10.3390/su10113851.

(21) W. Song, W. Song, H.H. Gu, and F.P. Li, *Int. J. of Environ. Res. Public Health* **17**(6), 1846. DOI: 10.3390/ijerph17061846.

(22) K. Koruyan, A.H. Deliormanli, Z. Karaca, M. Momayez, H. Lu, and E. Yalcin, *J. South. Afr. Inst. Min. Metall.* **112**(7), 667–672. DOI: 10.1134/S1062739148040235.

(23) E. Charou, M. Stefouli, D. Dimitrakopoulos, E. Vasiliou, and O.D. Mavrantza, *Mine Water Environ.* **29**(1), 45–52 (2010). DOI: 10.1007/s10230-010-0098-0.

(24) Z. Wei, H. Hu, H.W. Zhou, and A.Lau, *Pure Appl. Geophys.* **176**(8), 3593–3605 (2019). DOI: 10.1007/s00024-019-02152-0.

(25) Y.P. Liu, L.X. Zhu, and Y.Z. Zhou, *Acta Petrol. Sin.* **34**(11), 3217–3224 (2018).

(26) B.T. Le, D. Xiao, Y.C. Mao, D.K. He, S.Y. Zhang, X.Y. Sun, and X.B. Liu, *IEEE Access* **6,** 44328–44339 (2018). DOI: 10.1109/ACCESS.2018.2860278.

(27) G.B. Huang, Q.Y. Zhu, and C.K. Siew, *IEEE Intl. Joint Conf. Neural Networks* **2,** 985–990 (2004). DOI:10.1109/IJCNN.2004.1380068.

(28) G.B. Huang, L. Chen, and C.K. Siew, *IEEE Transactions on Neural Networks* **17**(4), 879–892 (2006). DOI: 10.1109/TNN.2006.875977.

(29) G.B. Huang and L. Chen, *Neurocomputing* **70** (16–18), 3056–3062 (2007). DOI: 10.1016/j.neucom.2007.02.009.

(30) H.J. Rong, Y.S. Ong, A.H. Tan, and Z.X. Zhu, *Neurocomputing* **72**(1–3), 359–366 (2008). DOI: 10.1016/j.neucom.2008.01.005.

(31) Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, *IEEE Transactions on Neural Networks* **21**(1), 158–162 (2010). DOI: 10.1109/TNN.2009.2036259.

(32) S. Ioffe and C. Szegedy, *Intl. Conf. on Machine Learning*, *PMLR* **37,** 448–456 (2015). arXiv:1502.03167.

Dong Xiao and **Hongfei Xie** are with the Information Science and Engineering School at Northeastern University, in Shenyang, China. **Yanhua Fu** is with JangHo Architecture College at Northeastern University, in Shenyang, China. **Feifei Li** is with the Liaoning Province Important Technology Innovation and R & D Base Construction Engineering Center in Lialong Province, China. Direct correspondence to: Dong Xiao at xiaodong@ise.neu.edu.cn ●

PEER-REVIEWED RESEARCH
Spectroscopy papers have undergone a double-blind peer review process and are available on our open-access website.

Simultaneous Detection of Nitrate and Nitrite Based on UV Absorption Spectroscopy and Machine Learning

Hang Zhang, Qiong Wu, Yonggang Li, and Sha Xiong

Using spectrophotometry to acquire nitrate and nitrite concentrations is common in water quality monitoring. However, it is challenging to achieve the measurement with high accuracy because of the spectral signal overlapping. In this article, a hybrid machine learning approach is proposed to simultaneously determine nitrate and nitrite based on UV absorption spectroscopy. All spectral data are divided into four subdivisions according to the concentration ratio of nitrate and nitrite. In each subdivision, a regression submodel is established according to the sample characteristics. First, the sample is voted to a category by a joint classifier and then processed by the corresponding submodel to predict the concentrations of nitrate and nitrite. This method has been further optimized by considering the interference of foreign ions. The proposed approach improves the performance of spectral direct detection and is therefore a promising tool for fast determination and continuous monitoring in environmental applications.

Nitrate (NO_3^-) and nitrite (NO_2^-) are the most common forms of nitrogen that are found in environments, physiological systems, and food industries (1,2). Excessive nitrate and nitrite lead to the eutrophication of an ecosystem, which introduces fatal threats to human health, such as methemoglobin syndrome (3). Considering their pollution hazards, regulations have been imposed to set legal limits of nitrate and nitrite in water worldwide. Therefore, it is critical to find convenient and economical methods to monitor these trace analytes. A broad range of techniques have been evaluated to monitor nitrate and nitrite in water, including electrochemical detection, chemiluminescence, colorimetric analysis, and UV spectrophotometry (4–7). Among various approaches, direct UV absorption spectrophotometry has attracted attention in the past decades for its high speed, reagent-free, operational simplicity, and ultralow cost operation (8–10).

The absorption spectra of nitrate and nitrite are similar and nearly overlap in the UV region (11). Hence, it is difficult to separate nitrate and nitrite contributions from the collected spectra. Wetters and Uglum first proposed to use the secondary peak of the absorption spectra to detect nitrate and nitrite at high concentrations (12). After that, Suzuki and Kuroda used the isosbestic absorption points of the second derivative spectra for determining nitrate and nitrite simultaneously (13). Both methods were based on the intrinsic absorption properties of nitrate and nitrite at two special wavelengths, and the models would be greatly affected by the interfering substances. To increase the accuracy, several methods that employed multiple wavelengths were proposed to achieve simultaneous measurements of nitrate and nitrite. Dong and others used a matrix algorithm to select six points in a narrow wavelength interval (14). Rieger and others adopted a multivariate correction algorithm with a total of 256 wavelengths uniformly spaced in the range of 210–400 nm (15). Sandford and others established the reference spectra of nitrate, nitrite, and bromide to achieve the simultaneous measurement based on the deconvolution method (16). However, these approaches for choosing an optimal modeling range are too approximate and inevitably some useful spectral information is lost. In addition, environmental noise, experimental error, and redundant information with low information content is contained in the spectral data. All such

factors reduce the stability and accuracy of the measurement.

Machine learning methods offer versatile and powerful solutions in spectra analysis, such as characteristic extraction, component classification, and concentration prediction (17–19). Recently, the combination of UV spectroscopy and machine learning has been successfully applied in the rapid detection of multiple compounds (20–22). However, most of the machine learning methods are based on a single model, which has prominent limitations in optimizing evaluation accuracy. Hence, some researchers have tried to build a hybrid model based on clustering algorithms to predict certain water quality parameters (23). The hybrid model has demonstrated a higher prediction accuracy than a single model.

In this work, a hybrid machine learning model is developed for direct measurement of nitrate and nitrite based on UV absorption spectroscopy. First, a joint classifier (JC) is utilized to divide the samples into four subregions based on the concentration ratios between nitrate and nitrite. Then, a submodel is selected for regression prediction in each subregion. To the best of our knowledge, this model is the first demonstration of simultaneous detection of nitrate and nitrite using a hybrid machine learning model combining classification and regression algorithms. Compared to other direct spectral methods, the proposed method with the advantages of machine learning can more effectively use the spectral information of the samples, enhancing the sensitivity of detection, especially for extremely low concentrations. For example, the average relative errors in determining nitrate and nitrite are approximately 4–5% by using the second derivative spectroscopy (13) and matrix method (14). The proposed machine learning method reduces the average relative errors to a value below 1%. In addition, the interference effects of wavelength selection and foreign ions on this method has been discussed.

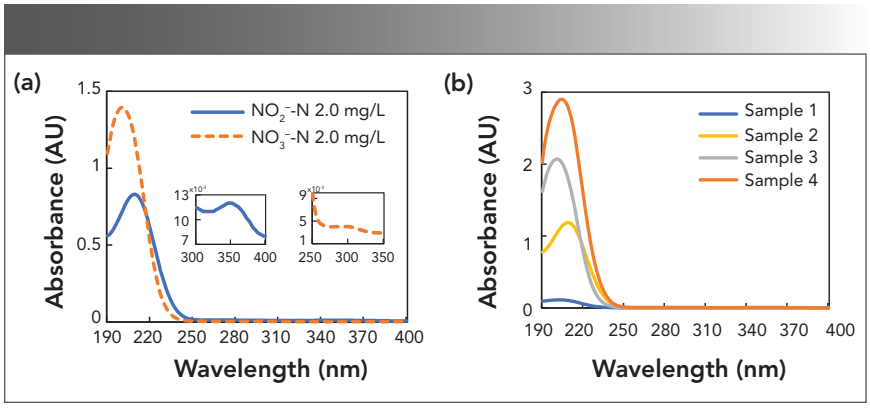


FIGURE 1: The original spectra of the samples. (a) UV absorption spectra of pure nitrite and nitrate solutions with the concentration of 2 mg N/L (Note that nitrate-N (mg/L) = 0.2259 x nitrate-NO_3 (mg/L), and nitrite-N (mg/L) = 0.3044 x nitrite-NO_2 (mg/L)). The insets show the second peaks of nitrite and nitrate, respectively. (b) UV absorption spectra of mixture samples. Sample 1: 0.1 mg N/L nitrate and 0.1 mg N/L nitrite; sample 2: 0.1 mg N/L nitrate and 3 mg N/L nitrite; sample 3: 3 mg N/L nitrate and 0.1 mg N/L nitrite; and sample 4: 3 mg N/L nitrate and 3 mg N/L nitrite.

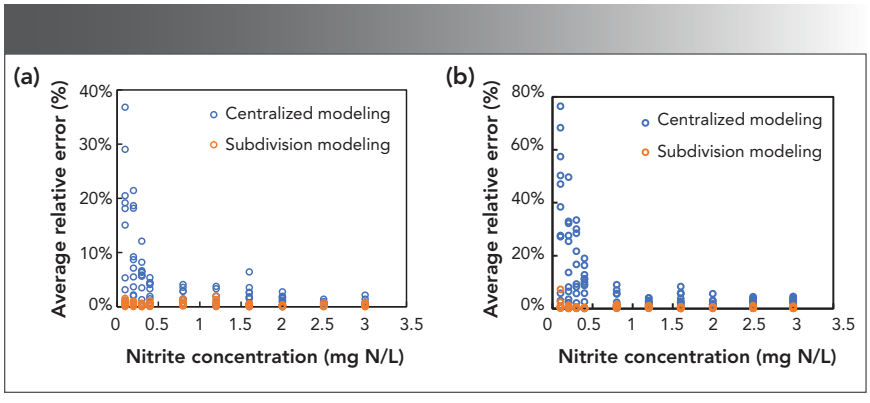


FIGURE 2: Comparison of the average relative errors of centralized modeling and division modeling in determining (a) nitrate and (b) nitrite. The critical concentration used for division modeling was 0.4 mg N/L. Only support vector machines (SVM) was used for classification here.

Theory and Methods

Classification

Support vector machines (SVM) are multiclassifiers based on the statistical learning theory (24). The basic model of SVM is defined as the linear classifier with the largest interval in the feature space. In this work, the spectral data are normalized before being processed by SVM because the convergence of the training network can be accelerated by mapping the data to a range of 0~1. Because abundant data increase the quantity of computation, principal component analysis (PCA) (25), which can eliminate multicollinearity existed among variables, is used to reduce the data dimension of the input layer. The particle swarm algorithm (PSO) (26) optimizes the penalty factor and kernel parameters in the modeling process. The overall program is built on the Libsvm toolbox (27,28).

Logistic regression (LR) is an algorithm that assumes data obey a binomial distribution and applies the maximum likelihood function to achieve binary classification of samples (29). When the samples are going to be divided into multiple categories, LR is employed to build an independent binary classifier for each category.

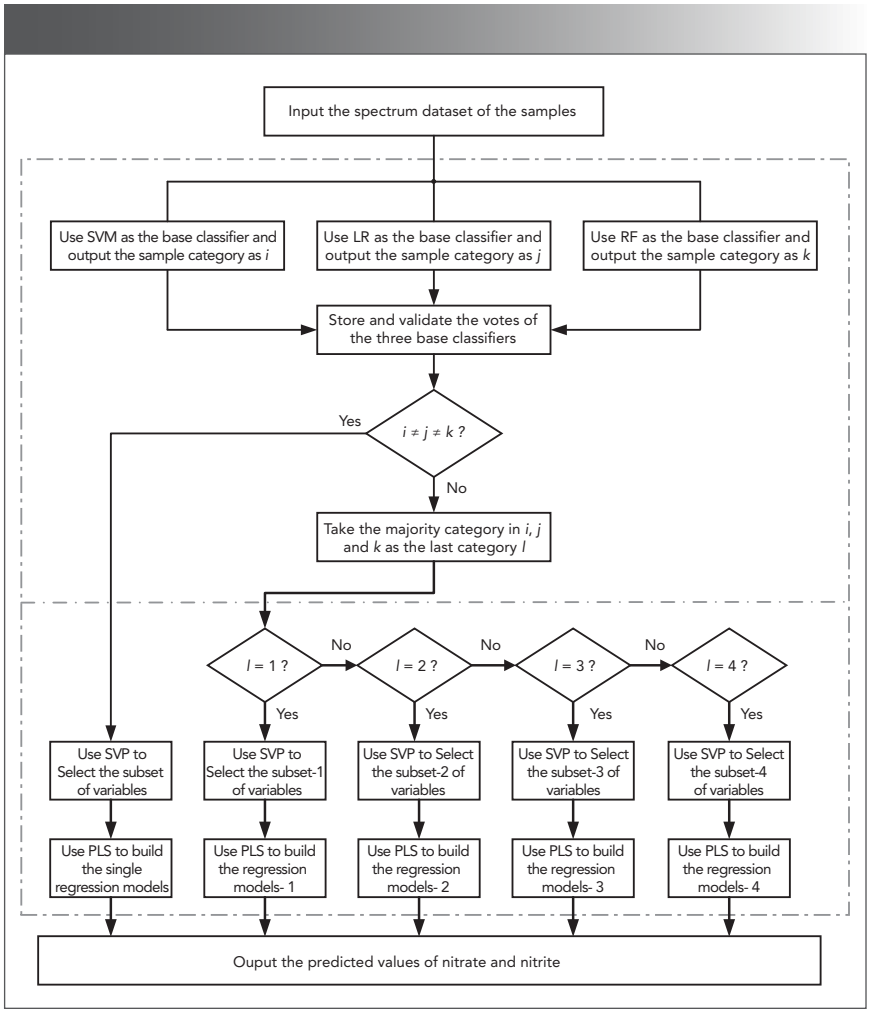


FIGURE 3: Flow chart of the machine learning program framework.

Random forest (RF) is another algorithm for classification (30). It consists of multiple decision trees, in which the training samples are obtained by bootstrap sampling from the original training set. In addition, RF utilizes random feature selection in the growth process of the decision tree, which prevents overfitting. For classification, an unknown sample is sent to each decision tree for prediction, and then voted for classification. The class with the most votes is the final classification result.

The above three classifiers are used as the base classifiers to vote for the categories of the same sample, which can effectively improve the reliability and robustness of the system.

Feature Selection

Modeling with all wavelength points increases model complexity and reduces accuracy because of the huge amount of data and redundant information in the full-range spectrum. Each wavelength point is different in terms of the amount of useful information and the degree of interference by other ions. Therefore, it is necessary to screen out useful variables with high sensitivity and correlation to the target ions, while eliminating redundant variables sensitive to foreign ions. In this work, stability and variable permutation (SVP) is used to choose the characteristic wavelengths. Variables are selected through multiple iterations and competitions in SVP (31). After all iterations are completed, model population analysis is employed to obtain the optimal subset of variables with the minimum mean and relatively low standard deviation value of root mean square error.

Regression Model

A regression model is used to establish the relationship between input variables and output variables. As the most commonly used regression algorithm in spectral multivariate correction analysis, partial least square (PLS) is a perfect combination of multivariate linear regression, canonical correlation analysis, and PCA (32). As an alternative regression algorithm, least squares support vector machine (LSSVM) is an improved version of the SVM algorithm. It uses the least squares linear system as the loss function, and reduces the computational complexity by solving a set of linear equations instead of the more complex quadratic programming method used by the traditional SVM (33).

TABLE I: Comparison of the results of different algorithms for simultaneous determination of nitrate and nitrite

Evaluation parameters	Centralized Modeling (SVP-PLS)		Division Modeling 1 (JC-SVP-PLS)		Division Modeling 2 (JC-SG-SVP-LSSVM)	
	NO_3^-	NO_2^-	NO_3^-	NO_2^-	NO_3^-	NO_2^-
ARE	0.0416	0.1026	0.0044	0.0054	0.0114	0.0158
MRE	0.3684	0.7640	0.0201	0.0723	0.0737	0.0907
R^2	0.9993	0.9966	0.9999	0.9999	0.9998	0.9993
RMSEP	0.0254	0.0575	0.0059	0.0034	0.0090	0.0169

Experiments

All reagents were of analytical grade (Sinopharm Chemical Reagent Co., Ltd) and used without further purification. Nitrate and nitrite stock solutions (100 mg N/L) were prepared by dissolving 0.7221 g of potassium nitrate and 0.4928 g of sodium nitrite in 1 L of deionized water, respectively. A series of measurements were made for nitrate and nitrite solutions with 10 different concentrations, ranging from 0.1 to 3.0 mg N/L. A total of 100 groups of mixture solutions were investigated and used as training data in the machine learning models. Instead of dividing the samples into calibration and validation sets, leave-one-out cross validation (34) was used as an evaluation strategy. These solutions were prepared by serial dilution in deionized water from the nitrate:nitrite stock solution. Sodium chloride, sodium bromide, sodium carbonate, sodium bicarbonate, calcium chloride, magnesium chloride, and humic acid were added in the mixture solutions for the interference studies.

UV spectra were acquired with a dual beam UV-vis spectrophotometer (UV-2600, Shimadzu). Deionized water was used as a reference solution. Samples were scanned between 190 and 400 nm in a quartz cuvette with a 10-mm optical pathlength. Scans were conducted at 120 nm/min with a resolution of 1 nm. Each measurement was repeated three times to ensure reproducibility.

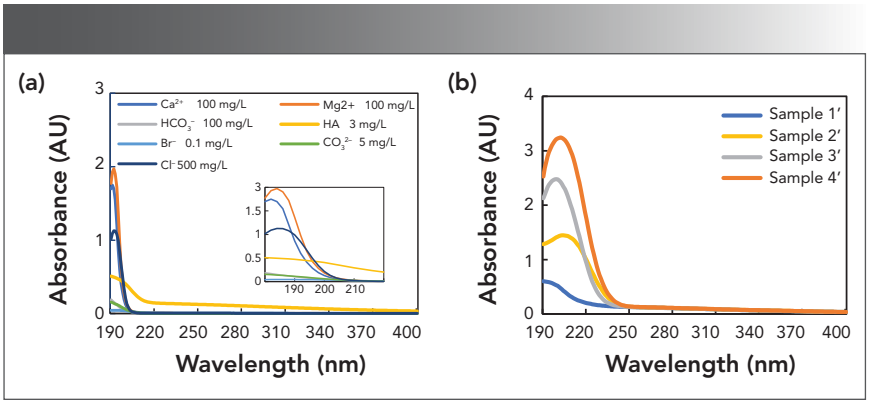


FIGURE 4: (a) UV absorption spectra of foreign ions. (b) The spectra of four samples from different regions affected by HA (3 mg/L). Sample 1': 0.1 mg N/L nitrate and 0.1 mg N/L nitrite; sample 2': 0.1 mg N/L nitrate and 3 mg N/L nitrite; sample 3': 3 mg N/L nitrate and 0.1 mg N/L nitrite; and sample 4': 3 mg N/L nitrate and 3 mg N/L nitrite.

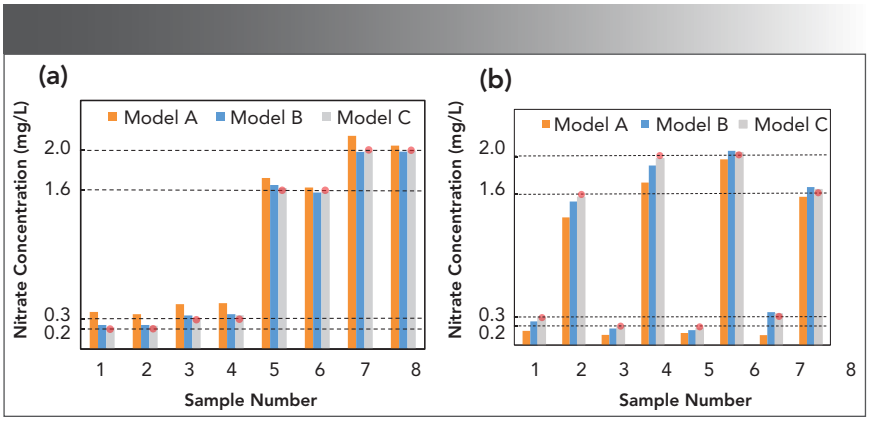


FIGURE 5: Test results of different models when chloride ions are added in the mixture samples. The determination targets are (a) nitrate and (b) nitrite, respectively. A red ball marker passing through the dash line above each sample is used to highlight the true concentration of the analyte.

Results and Discussion

Division Based on Concentration

Figure 1a shows the UV spectra of nitrate and nitrite solutions, which both have a wide absorption peak between 190–250 nm, thus several methods for the measurement of nitrate and nitrite concentrations depend on the absorbance at wavelengths approximately at 200 nm (12–14,16). In fact, there is a second absorption peak for nitrate and nitrite above 250 nm, as shown in the insets of Figure 1a. Because the second absorption peak is relatively weak and has a small value, the spectrum above 250 nm in Figure 1a was flattened by the high absorbance for the region below 250 nm. However, this spectral region may still contain useful information for modeling. Figure 1b shows the spectral curves of four mixture samples with different concentrations. The absorbance difference between the mixtures of nitrate and nitrite with the maximum and minimum concentrations is nearly 40 times. In a preliminary study, we analyzed the spectral data with centralized modeling, in which the mixture solutions with different concentrations were calculated by the same model. It has been found that centralized modeling has insufficient sensitivity to predict components at low concentrations because of the wide modeling range of samples.

To achieve a more accurate prediction, the concentrations of nitrate and nitrite are divided into four subregions for separate modeling. Each subregion has its own distinct characteristics related to the concentration ratio between nitrate and nitrite. In region 1, the concentrations of nitrate and nitrite are both low; in region 2, the nitrite concentration is much higher than that of nitrate; in region 3, the nitrite concentration is much lower than that of nitrate; and in region 4, the concentrations of nitrate and nitrite are both high. In this way, each submodel has higher prediction accuracy than centralized modeling

TABLE II: The influence of foreign ions on the simultaneous determination of nitrate and nitrite by the proposed method

Parameter		Foreign Ion						
		Ca^{2+} (100)[a]	Mg^{2+} (100)[a]	CO_3^{2-} (5)[a]	HCO^{3-} (100)[a]	Br^- (0.1)[a]	Cl^- (500)[a]	HA (3)[a]
Number of misclassified samples		1	1	1	1	0	0	8
ARE	Nitrate	0.0272	0.0226	0.0849	0.0712	0.0550	0.0395	>0.15
	Nitrate	0.0584	0.0483	0.0594	0.1058	0.0722	0.0559	>0.15

a: The unit of the number in parentheses is mg/L.

as it adapts to the sample characteristics of each region.

We performed modeling analysis on 100 experimental samples, and compared the performance of centralized modeling and division modeling in predicting the concentrations of nitrate and nitrite. The results are shown in Figure 2. Although the average relative errors in centralized modeling are small (<10%) with relative high concentration of analytes, it greatly increases when the concentration is lower than 0.4 mg N/L. By contrast, the division modeling always gives a stable and satisfied performance with an average relative error less than 5%. In view of the insufficient predictive sensitivity at low concentrations, the critical concentration used to divide subregions is chosen to be placed at a lower position. We compared the modeling results, which set the critical concentrations as 0.3, 0.4, and 0.8 mg N/L, respectively. The division modeling gives a highest accuracy (~98%) and lowest average relative errors (~0.44%), when the critical concentration is 0.4 mg N/L. Therefore, the critical concentration is set to 0.4 mg N/L for the following modeling procedures.

Classification

The overall program framework is shown in Figure 3. First, the collected spectral data are classified by three classifiers (SVM, LR, and RF) independently. Based on voting results, the data are sent to the submodels or analyzed by the centralized modeling. In each model, SVP is used to select characteristic wavelength, and PLS is utilized to build the regression model. Finally, the predicted concentrations of nitrate and nitrite are given out simultaneously.

Because the program is executed sequentially, and to avoid the classification error from affecting the subsequent regression accuracy later, a joint classifier composed of SVM, LR, and RF is used to vote on the sample category. The sample is classified into the category with the majority of votes (≥2). If the categories selected by the three classifiers are all different, the classification result is determined to be unreliable and a corresponding reminder is given. In this case, a single non-classified regression model is used for prediction, so that the accuracy can be at least consistent with that of centralized modeling. It is worth mentioning that each submodel is built with samples distributed on the classification boundary to avoid larger prediction errors caused by classification errors. Because the probability of classifiers making errors at these points are greater. In the experiments, the classification accuracy using RF, LR, and SVM reached 94%, 97%, and 98%, respectively. There were 11 samples that were misclassified by the three base classifiers among the 100 experimental samples. Three samples were actually located on the classification boundary, so their influence on the prediction results can be ignored. In addition, other samples that were misclassified in a single classifier finally got the right classification results because of the voting mechanism. The fault tolerance and robustness of the system are greatly improved because of the joint classifier.

Wavelength Selection

The optimal wavelength set is supposed to combine the specific characteristics of nitrate or nitrite. For each subregion, SVP selects only one subset of variables (that is wavelength) for nitrate and nitrite, respectively. Parameters in SVP are optimized using grid search. The model using the specialized subset of variables can achieve better performance because it adapts to the characteristics of the target ions in a narrow concentration range. The numbers of optimal variables are different in the submodels, which are changing from eight to 34. The number is related to the spectral similarity of the samples in each subregion. The regression models are built based on these variables.

Performance Evaluation

Four classical parameters were employed to evaluate the performances of the established models, including average relative error (ARE), maximum relative error (MRE), root mean square error of prediction (RMSEP), and determination coefficient (R^2). The RMSEP is expressed as:

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

where n is the total number of samples, $\hat{y}_i$ is the actual value of sample i, and y_i is the predicted value of sample i. The determination coefficient (R^2) is expressed as:

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

where $\bar{y}_i$ is the average value of all the testing samples. According to equations 1 and 2, the model performs better with lower RMSEP and higher R^2.

The performances of different algorithms are compared in Table I. The centralized modeling only builds a single model, in which SVP and PLS are used for wavelength selection and regression modeling, respectively. Division modeling 1 uses the joint classifier for classification, followed by submodeling with SVP and PLS. Division modeling 2 also employs joint classifier. Then, the Savitzky-Golay method (35) is used to obtain the second derivative of the original spectral data, which are subsequently processed by SVP. Finally, the regression model is established using LSSVM. To enhance the generalization ability of the model in the entire concentration range, the centralized modeling inevitably sacrifices the local accuracy, which has much greater error values as shown in Table I. Compared with it, division modeling 1 uses the same feature selection and modeling methods, but only employs the joint classifier to classify the samples in advance. This method improves the accuracy of prediction significantly, especially for low concentration samples. When the centralized modeling is replaced by division modeling 1, the average relative error is reduced from 4.16 to 0.44% for nitrate, and from 10.26 to 0.54% for nitrite. The maximum relative error of nitrate has even decreased 18 times. It proves that division modeling has the potential to expand the modeling range without reducing the prediction accuracy. Division modeling 2 employs the differential preprocessing to expand the distance between the spectral peaks of nitrate and nitrite that are almost overlapped. However, the results show that the effect is not significant, because the spectra of the two analytes in the 200–250 nm range are approximately parallel, which yields the same information trends using the same pretreatment **.** The maximum relative error of division modeling 2 is close to four times that of division modeling 1 in the prediction of nitrate. It is because most of the spectral signals still have linear additivity in a two-component experimental system with relatively simple spectra. It is consistent with the modeling scenario of PLS, which is a multiple correction method based on linear regression. In contrast, LSSVM is a nonlinear correction method. In addition, PLS can also overcome the interference of nonlinear factors to a certain extent, bringing advantages in spectral multivariate correction analysis.

Influence of Foreign Ions

In actual water samples, there are many other ions that also absorb UV light, which probably affect the measurement of nitrate and nitrite (14,36,37). The influence of several common ions in water has been investigated. The individual spectra of these substances are shown in Figure 4a. It can be seen that most of the foreign ions only have absorption at 190–205 nm, whereas humic acid (HA) representing organic matters has an absorption band after 205 nm that is not negligible.

Wavelength selection plays an important role in reducing the influence of foreign ions. For instance, 500 mg/L chloride ion was added into eight groups of nitrate and nitrite mixtures with four concentration levels (0.2, 0.3, 1.6, and 2 mg N/L). The concentrations of the eight samples are symmetrically distributed in four subregions. We compared the performance of three models with different modeling methods. Model A uses 200 nm as the starting wavelength for centralized modeling. Model B and Model C use 200 nm and 205 nm as the starting wavelengths for division modeling, respectively. The final prediction results are shown in Figure 5. It can be seen that the predicted concentrations using model A are far from the true values. The relative errors are even more than 50% for low concentration samples (<0.4 mg N/L). On the contrary, Models B and C are much closer to the true values, which verifies the improvement by using the division modeling method. The average relative errors are 18% and 6% in Models B and C, respectively. It means that the chloride ions have little influence on the prediction of nitrate and nitrite, when the starting wavelength of modeling is delayed to 205 nm.

The concentrations of nitrate and nitrite mixtures were measured with various interfering ions as shown in Table II. Concentration of a foreign ion was chosen when its absorbance equaled 0.1 approximately at 205 nm. Each type of foreign ions was added to the eight groups of nitrate and nitrite mixtures. Table II shows the influence of these foreign ions on the prediction results of Model C. The joint classifier still worked well in the presence of foreign ions, except for HA. Misclassification occurs when samples are assigned to a different category than the one they should be in. For example, a sample in region 1 is assigned to region 2. The experimental results showed that one sample was misclassified after adding Ca^{2+}, Mg^{2+}, CO_3^{2-}, and HCO_3^-. It is worth noting that the misclassified samples are all in region 3 with the same concentrations of nitrate (1.6 mg N/L) and nitrite (0.2 mg N/L). A possible explanation comes from the concentration ratio of nitrate to nitrite in this sample, which is the highest among the eight samples. The absorption spectra of nitrate and nitrite are almost overlapped. When the concentration ratio of nitrate to nitrite is increased, it is more difficult to identify the contribution from nitrite in the spectra of mixture samples. Because the proportion of nitrate and nitrite in the total absorbance is a potential internal consideration in the model, the classifier may be affected by the concentration ratio. For the same reason, the relative error in the determination of nitrite is often higher than that of nitrate.

The average relative error caused by the interference of foreign ions

mostly can be controlled within 10%. Because of the wide absorption band of HA in the range of 205–400 nm, after adding HA to the mixture, the absorbance of the sample in this spectral range will increase as shown in Figure 4b. In addition, the increase at each wavelength point may not be the same. This change causes the classifier to make wrong judgments on the concentration of nitrate and nitrite, resulting in misclassification and further large prediction errors. Thus, all eight representative samples have a high possibility to be misclassified. Therefore, incorporating the organic ions into the modeling components when there are more organic interferences in the water samples is recommended.

Conclusion

To summarize, we proposed a hybrid machine learning method to tackle the challenge of predicting nitrate and nitrite simultaneously with UV absorption spectroscopy. Compared with centralized modeling in other spectroscopic methods, the proposed model provides higher accuracy by employing a joint classifier before the regression modeling. The influence of classification, wavelength selection, and foreign ions are discussed to optimize the model. This method is fast, reagent-free, and potentially useful for developing in situ sensors for monitoring trace species in marine and aquatic environments. In addition, the proposed methodology can also be applied to determine multiple composite substances. The proposed methodology holds broad application prospects in water quality monitoring, food safety, and soil properties qualification.

Funding

This work was supported by the State Key Program of the National Natural Science Foundation of China (Grant No. 61890932).

References

(1) M.J. Moorcroft, J. Davis, and R.G. Compton, *Talanta* **54**, 785–803 (2001).
(2) Q.H. Wang, L.J. Yu, Y. Liu, L. Lin, R.G. Lu, J.P. Zhu, L. He, and Z.L. Lua, *Talanta* **165**, 709–720 (2017).
(3) A.M. Fan, C.C. Willhite, and S.A. Book, *Regul. Toxicol. Pharm.* **7**, 138–148 (1987).
(4) P. Singh, M.K. Singh, Y.R. Beg, and G.R. Nishad, *Talanta* **191**, 364–381 (2019).
(5) M. J. Moorcroft, L. Nei, and J. Davis, *Anal. Lett.* **33**, 3127–3137 (2000).
(6) T. Aokia, S. Fukuda, Y. Hosoi, and H. Mukai, *Anal. Chim. Acta* **349**, 11–16 (1997).
(7) R.B.R. Mesquita, M.T.S.O.B. Ferreira, R.L.A. Segundo, C.F.C.P. Teixeira, A.A. Bordalo, and A.O.S.S. Rangel, *Anal. Methods* **1**, 195–202 (2009).
(8) H.E. Khorassani, P. Trebuchon, H. Bitar, and O. Thomas, *Water Sci. Technol.* **39**, 77–82 (1999).
(9) R.S. Brito, H.M. Pinheiro, F. Ferreira, J.S. Matos, and N.D. Lourenco, *Urban Water J.* **11**, 261–273 (2014).
(10) M.L.C. Passos and M.L.M.F.S. Saraiva, *Measurement* **135**, 896–904 (2018).
(11) A. Mašić, A.T.L. Santos, B. Etter, K.M. Udert, and K. Villez, *Water Res.* **85**, 244–254 (2015).
(12) J.H. Wetters and K.L. Uglum, *Anal. Chem.* **42**, 335–340 (1970).
(13) N. Suzuki and R. Kuroda, *Analyst* **112**, 1077–1079 (1987).
(14) H.R. Dong, M.Y. Jiang, and Q. Zhang, *Anal. Lett.* **24**, 305–315 (1991).
(15) L. Rieger, G. Langergraber, D. Kaelin, H. Siegrist, and P.A. Vanrolleghem, *Water Sci. Technol.* **57**, 1563–1569 (2008).
(16) R. Sandford, A. Exenberger, and P. Worsfold, *Water Sci. Technol.* **41**, 8420–8425 (2007).
(17) H.Q. Zhu, S.J. Wu, Y.G. Li, F. Cheng, and X.L. Wang, *Optik* **194**, 163065 (2019).
(18) P.J. Huang, Y.H. Li, Q.J. Yu, K. Wang, H. Yin, D.B. Hou, and G.X. Zhang, *Spectrosc. Spec. Anal.* **40**, 2267–2272 (2020).
(19) L. Guan, Y. Tong, J. Li, S. Wu, and D. Li, *RSC Adv.* **9**, 11296–11304 (2019).
(20) S. Fogelman, M. Blumenstein, and H. Zhao, *Neural Comput. Appl.* **15**, 197–203 (2005).
(21) X. Qin, F. Gao, and G. Chen, *Water Res.* **46**, 1133–1144 (2012).
(22) J. Chen, C. Yang, H. Zhu, Y. Li, and J. Gong, *Optik* **181**, 703–713 (2019).
(23) M. Ay and O. Kisi, *J. Hydrol.* **511**, 279–289 (2014).
(24) C. Cortes and V.N. Vapnik, *Mach. Learn.* **20**, 273–297 (1995).
(25) H. Abdi and L.J. Williams, *Wiley Inter. Rev. Comput. Stat.* **2**, 433–459 (2010).
(26) J. Kennedy and R. Eberhart, *Proceedings of the IEEE International Conference on Neural Networks (ICNN '95),* 1942–1948, 1995.
(27) Faruto, LIBSVM-Faruto Ultimate Version: a toolbox with implements for support vector machines based on libsvm, 2011. Software available at http://www.matlabsky.com.
(28) C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
(29) J.S. Cramer, Tinbergen Institute Discussion Papers, No 02-119/4 (2002).
(30) L. Breiman, *Mach. Learn.* **45**, 5–32 (2001).
(31) J.M. Chen, C.H. Yang, H.Q. Zhu, Y.G. Li, and W.H. Gui, *Chemometr. Intell. Lab. Syst.* **182**, 188–201 (2018).
(32) S. Wold, M. Sjöström, and L. Eriksson, *Chemometr. Intell. Lab. Syst.* **58**, 109–130 (2001).
(33) J.A.K. Suykens and J. Vandewalle, *Neural Process. Lett.* **9**, 293–300 (1999).
(34) S. Arlot and A. Celisse, *Stats. Surv.* **4**, 40–79 (2010).
(35) P. A. Gorry, *Anal. Chem.* **62**, 570–573 (1990).
(36) O. Thomas, S. Gallot, and N. Mazas, Fresen. *J. Anal. Chem.* **338**, 238–240 (1990).
(37) A. Tudorache, D.E. Ionita, N.M. Marin, C. Marin, and I.A. Badea, *Accredit. Qual. Assur.* **22**, 29–35 (2017).

Hang Zhang, Qiong Wu, Yonggang Li, and **Sha Xiong** are with the School of Automation at Central South University, in Hunan, China. Direct correspondence to: xiongsha@csu.edu.cn ●

Spectroscopy®

SOLUTIONS FOR MATERIALS ANALYSIS

More learning tools for **spectroscopists** exclusively online

Webcasts

View our educational webcasts **live or on demand** in topics such as ICP methods, raman fundamentals, atomic spec, microplastics, and many more!

eBooks

Choose from topics such as food safety, sample prep, raman imaging, atomic absorption and more!

Visit spectroscopyonline.com today!

AN MJH life sciences™ BRAND