Ciência Rural, Santa Maria, v.52:6, e20210002, 2022

http://doi.org/10.1590/0103-8478cr20210002

ISSNe 1678-4596

CROP PROTECTION

A lettuce moisture detection method based on terahertz time-domain spectroscopy

Xiaodong Zhang[1,2] Zhaohui Duan[1,2] Hanping Mao[1,2] Hongyan Gao[1*,2] Zhiyu Zuo[1,2]

[1]School of Agricultural Engineering, Jiangsu University, 212013, Zhenjiang, China. E-mail: gaohy@ujs.edu.cn. *Corresponding author.
[2] Key Laboratory of Modern Agricultural Equipment and Technology, Ministry of Education, Jiangsu University, Zhenjiang, China.

ABSTRACT: *For non-destructive detection of water stress in lettuce, terahertz time-domain spectroscopy (THz-TDS) was used to quantitatively analyze water content in lettuce. Four gradient lettuce water contents were used . Spectral data of lettuce were collected by a THz-TDS system, and denoised using the S-G derivative, Savitzky-Golay (S-G) smoothing and normalization filtering. The fitting effect of the pretreatment method was better than that of regression fitting, and the S-G derivative fitting effect was obtained. Then a calibration set and a verification set were divided by the Kennan-Stone algorithm, sample set partitioning based on joint X-Y distance (SPXY) algorithm, and the random sampling (RS) algorithm, and the parameters of RS were optimized by regression fitting. The stability competitive adaptive reweighted sampling, iteratively retained information variables and interval combination optimization were used to select characteristic wavelengths, and then continuous projection was used on basis of the three algorithms above. After the successive projection algorithm was re-screened, partial least squares regression was used into modeling. The regression coefficients R_c^2 and RMSEC reach 0.8962 and 412.5% respectively, and R_p^2 and RMSEP of the verification set are 0.8757 and 528.9% respectively.*

Key words: *water stress, successive projection algorithm algorithm, partial least square regression, terahertz time-domain spectroscopy.*

Um método de detecção de umidade de alface baseado em THz-TDS

RESUMO: *Para a detecção não destrutiva de estresse hídrico da alface, espectroscopia no domínio do tempo em terahertz (THz-TDS) foi usada para analisar quantitativamente o conteúdo de água na alface. Quatro gradientes de conteúdo de água de alface foram usados. Os dados espectrais da alface foram coletados por um sistema THz-TDS e denoised usando o derivado S-G, Savitzky-Golay (S-G) suavização e filtragem de normalização. O efeito de ajuste do método de pré-tratamento foi melhor do que o do ajuste de regressão, e o efeito de ajuste da derivada S-G foi obtido. Em seguida, um conjunto de calibração e um conjunto de verificação foram divididos pelo algoritmo Kennan-Stone, particionamento do conjunto de amostra com base no algoritmo de distância X-Y conjunta (SPXY) e o algoritmo de amostragem aleatória (RS), e os parâmetros de RS foram otimizados por ajuste de regressão. A amostragem adaptativa de estabilidade competitiva reponderada, variáveis de informação retidas iterativamente e otimização de combinação de intervalo foram usadas para selecionar comprimentos de onda característicos e, em seguida, a projeção contínua foi usada com base nos três algoritmos acima. Depois que o algoritmo de projeção sucessivo foi reprojetado, a regressão de mínimos quadrados parcial foi usada na modelagem. Os coeficientes de regressão R2 e erro quadrático médio (RMSEP) atingem 0,8962 e 412,50%, respectivamente, e R2 e RMSEP do conjunto de verificação são 0,8757 e 528,93%, respectivamente.*

Palavras-chave: *alface, teor de umidade, THz-TDS, algoritmo SPA, regressão parcial de mínimos quadrados.*

INTRODUCTION

China is the largest producer of lettuce, accounting for approximately 50% of the world's output. Although, lettuce requires abundant water, its demand for water varies with growth stages (TOSIN et al., 2017). Thus, supplying appropriate amounts of water at different growth stages, as well as rapidly and accurately detecting water stress, is essential for maximizing the yield, quality, and taste of lettuce (WANG et al., 2016).

Crop canopy temperature is an indicator of crop water stress. Infrared thermal imaging is a well-studied technique that has been widely used to monitor crop water stress because it permits the water status of crops to be determined rapidly and in a non-destructive manner (O'SHAUGHNESSY et al., 2011). A previous study using thermal imaging to explore the relationship between water stress in papaya and three physiological indexes (stomatal conductance, transpiration rate, and net photosynthesis) under different irrigation conditions demonstrated that thermal imaging is a promising technology for monitoring the physiological state of papaya under drought conditions (LIMA et al., 1999).

Received 01.03.21 Approved 07.29.21 Returned by the author 10.01.21
CR-2021-0002.R2
Editors: Leandro Souza da Silva Alessandro Dal'Col Lucio

S.A. et al. conducted thermal infrared monitoring in a soybean field and reported a negative correlation between leaf water potential and an index of water stress ($R^2 = 0.93$).

Although, the crop canopy temperature obtained by infrared thermal imaging is strongly correlated with water stress, canopy temperature, which is determined by complex energy exchanges of the farmland ecosystem, is affected by several other environmental variables, such as soil evaporation (OSCO et al., 2018). Thus, assessing the water stress status of crops based on the canopy temperature can be misleading (BELEN et al., 2018).

The detection of plant water surplus and deficiency based on hyperspectral technology has been a major focus of research. The use of hyperspectra for detecting plant water status detection has been extensively explored (GALVÃO et al., 2015). A hyperspectral response model of lettuce established using an artificial neural network was able to distinguish water-stressed lettuce from non-stressed lettuce with an accuracy of 93% (AUSTON et al., 2015), which represented a major improvement in the non-contact estimation of water stress.

A previous study showed that spectral information at approximately 1450 nm can best predict water potential of vegetable leaves based on spectral reflectance obtained using a radiation spectrometer for 350–2500 nm and a prediction model of leaf water potential using partial least squares (PLS) (ZHOU et al., 2016). The leaf water potential of grape before dawn was determined using hyperspectral data obtained via a hand-held spectroradiometer (400–1010 nm) (WANG et al., 2017), and the extracted vegetation index and structural variables were used as predictors in a water stress model. The prediction accuracy R^2 of the estimated model obtained from the verification set was 0.73, and that of the severe water stress plants was 0.79. The accuracy and operability of the prediction model indicated that it could be used to monitor the water status of grape and aid irrigation management.

Previous research has shown that moisture detection methods based on visible and near-infrared hyperspectra can be used to determine the water status of plants rapidly (MATHANKER et al., 2015) and in a convenient and non-destructive manner. However, hyperspectral technology is greatly affected by environmental factors such as background and light intensity changes; in addition, environmental variables must be strictly controlled during the early cultivation of samples. When moisture is the only variable being monitored, the concentrations of nutrient elements need to remain constant and the moisture gradient needs to be precisely controlled. Furthermore, point source sampling is the most common sampling approach, but this method cannot fully describe the physical changes of leaves and the physiological and biochemical characteristics of internal tissues under water stress; as a result, the measurement accuracy of this approach is suboptimal (PARASOGLOU et al., 2010).

Terahertz time-domain spectroscopy (THz-TDS) radiation comprises electromagnetic waves between far infrared light and microwaves, with frequencies from 0.1 to 10 THz (wavelengths of 0.03–3.00 mm) (YANG et al., 2014). This region is referred to as the “THz Gap” because it is the last spectral region to be explored by humans. Terahertz radiation is penetrating, coherent, and highly sensitive to organisms and polar liquids, such as water molecules. Terahertz technology is effective for biomass detection (GENTE et al., 2013), non-destructive food testing, agricultural product analysis, and quality control and has become one of the most cutting-edge fields of scientific research (PARK et al., 2018).

THz-TDS shows high potential for the rapid, convenient, and non-destructive detection of crop water stress and has been used for moisture detection (LONG et al., 2013). For example, previous studies have established regression models based on the average values of the leaf time-domain amplitude, leaf frequency-domain amplitude, and leaf water content, which suggested that terahertz technology could be used to detect the water content of plant leaves (SONG et al., 2017). Qualitative analysis of THz-TDS changes under different levels of drought stress revealed that the peak time-domain spectrum decreases to the level below the blank reference peak and shows an obvious time delay after water stress decreases (ZHAO et al., 2015). The absorption coefficient and refractive index both decreased gradually as the degree of drought stress increased, indicating that THz spectroscopy provides a feasible approach for characterizing the water content of soybean canopy leaves (ZAHID et al., 2016).

Extraction of the frequency, time, and time-frequency multi-domain THz features, coupled with fused support vector machine, k-nearest neighbor, and decision tree algorithms, for the accurate determination of leaf moisture content provides a rich set of tools for growers to monitor plant health (PAGANO et al., 2016). A previous study examined the feasibility of using an advanced THz-QCL instrument to measure the absolute water content of purple coral leaves. In the graph of the tau L-A as M-w function in this study, the best-fitting regression line R^2 was consistently greater than 0.85, indicating that this method could be combined with plant water stress indicators to improve leaf water management. Infrared and THz spectra of water molecules have been suggested to be useful for monitoring the water content and leaf characteristics of plants, and the calculated dielectric constant indicated that the leaves become increasingly transparent under the action of THz waves with time. The results of this study indicated that the timely monitoring of leaf water stress could improve plant health monitoring; these findings have important implications for precision agriculture.

In a study of the relationship between THz spectra and the leaf water content of winter wheat, the prediction correlation coefficient and root mean square error of the best model established by linear regression were 0.812 and 0.044, respectively, under a 0.3 THz frequency domain amplitude (LI et al., 2013). These results suggested that THz spectra perform well in predicting the leaf water content and could be useful for detecting the water content in winter wheat leaves (NIE et al., 2014). Another study obtained the transmission and absorption spectra of rape leaves by THz-TDS and used the average transmittance and absorption coefficients to analyze changes in water content. The results of this study showed that THz spectra combined with statistical modeling are effective for obtaining physiological information from plants (BALDACCI et al., 2017).

Measurements of the transmittance of six grapes using a THz quantum cascade laser revealed that the leaf moisture content is linearly related to the product of absorbance and the projected area. This method is robust to heterogeneity among varieties and permits the leaf moisture status to be determined quickly, simply, and non-invasively (TORRES et al., 2013). The water status of vines was measured using the THz reflectivity of the trunk, and the results of this method were consistent with measurements taken from tree measuring instruments and humidity probes. Current research suggested that THz-TDS data are strongly correlated with crop water stress (CASTRO et al., 2018). Many molecules have strong fingerprint characteristics in the terahertz band, including rich physical and chemical information. Thus, THz-TDS has been used to detect the moisture in field crops and fruits such as wheat, soybeans, canola, and grapes (KENNARD et al., 2016). However, currently used methods have low detection accuracy and are unable to effectively screen irrelevant variables. There is thus much room for improvement of currently used detection methods.

The water content of lettuce leaves is an important indicator of water stress status. In this study, lettuce samples were exposed to different levels of water stress,

and other irrelevant variables are controlled. THz-TDS was used to predict the water contents of lettuce, establishing a high-precision lettuce leaf moisture content prediction model can understand crop drought conditions, provide a reference for efficient monitoring of lettuce water demand characteristics and scientific irrigation, and is of great significance for exploring precision agriculture models.

MATERIALS AND METHODS

Experiments were conducted in a Venlo-type experimental greenhouse in the Key Laboratory of Modern Agricultural Equipment and Technology of the Ministry of Education, Jiangsu University. Italian year-round tolerant lettuce was cultivated using potted perlite.

Lettuce was planted on August 19, 2018, and seeds with full grains and uniform size were selected. A 30 cm×60 cm rectangular black plastic plug tray was used for cultivation. The cultivation substrate consisted of peat, perlite, and vermiculite. After the seeds were soaked in cold water for 30 min, the water was drained, and the seeds were sown in the nursery during the day. When lettuce plants had five leaves and one heart on September 18, 2018, lettuce plants of the same size were transplanted to 15-cm flowerpots; perlite was used as the substrate, and Yamazaki nutrient solution formula was provided to all plants to ensure that they had access to similar quantities of nutrients. To minimize environmental interference, artificial ventilation and other measures were used to ensure that the temperature and humidity in the greenhouse were appropriate.

The seedling samples were treated with four levels of water stress starting on October 8, 2018, and there were a total of 80 samples (20 plants at each level). The experiment lasted 45 days. The water provided to plants was reduced while ensuring that all plants received normal supplies of nutrients. Group 1 (W1) was provided with a sufficient water supply for the entire 45-day period, and plastic pipes were used to deliver water to the roots of crops for irrigation through a dripper on a capillary with a diameter of approximately 10 mm. Group 2 (W2) plants received 100 ml of water once every two days; group 3 (W3) plants received 100 ml of water once every four days; and group 4 (W4) plants received 100 ml of water once every six days. All plants were watered from 8 am to 9 am. The main purpose of the samples subjected to extreme stress was to explore a greater range of stress conditions and thus improve the accuracy of sample feature recognition.

Acquisition of THz-TDS data

First, the fresh weight of lettuce leaves was determined. After 30 days of growth, one leaf was cut in the upper, middle, and lower layers from each pot (total of 240 leaves). Care was taken to ensure that the blade size, thickness, and shape were not affected during the cutting process. After cutting, the surface dust was removed by wiping the surface of the leaves with alcohol and air-drying them in a dry and ventilated place. The leaves were weighed with an electronic balance (Shanghai Jinghai Instrument Co., Ltd., model FA2004N); three measurements were taken, and the average value was used. Leaves were then placed in separate sealed bags and labeled.

A THz-TDS system (TS7400, Advantest Co., Japan) was used to obtain THz spectra. This system is specially designed for the collection of agricultural biological information. Compared with traditional THz-TDS systems, this system has higher precision and can detect samples ranging in size from 3 to 225 cm². The measurement range is 0–4 THz. The sampling interval in the 0–4 THz spectral region was 0.038 THz, and the resolution was 5 GHz. The maximum sample area was 150×150 mm^2.

Figure 1 shows the structure and working principle of the measurement system. The THz measuring unit and the THz detector of the THz transmitter are both fiber-optic docked without adjusting the external optical path. The THz transmitter emits laser pulses, which are divided by a beam splitter into two mutually perpendicular laser beams, including a strong pump light and a weak probe light. The pump light is incident on the emitting crystal, which generates THz pulses that pass through the sample stage via the mirror. These pulses are then transmitted to the terahertz detector through the detection crystal collinear with the probe light and are reflected many times. The detector transmits the

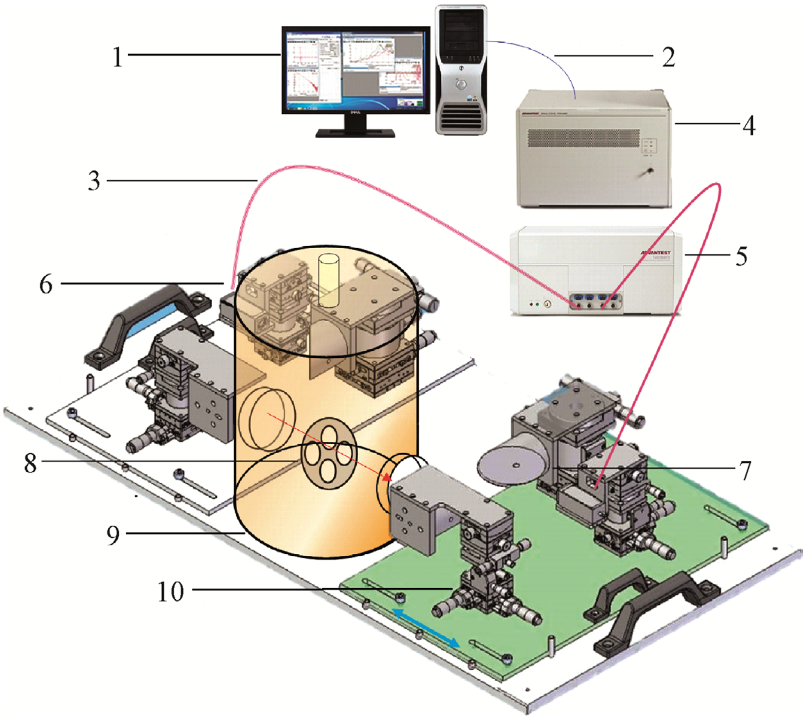

Figure 1 - Structure and working principle of the TS7400 equipment. (1) Control computer, (2) Ethernet, (3) Optical fiber, (4) Analysis unit, (5) Measurement unit, (6) Terahertz transmitter, (7) THz detector, (8) Sample sets, (9) Low temperature and constant temperature transmission module, (10) Movable support.

difference between the two laser beams to the control computer. After receiving the signal, the control computer can directly calculate the refractive index, absorption coefficient, dielectric constant, and other parameters of the sample through the analysis unit, in addition to the THz-TDS spectrum and its distribution information.

Before data acquisition, the measurement area needs to be calibrated. The size of the sample area was calculated, and the acquisition area was located. Before each measurement, the background was calibrated, the blank reference was saved, and THz-TDS spectra of lettuce samples were collected. Clear differences in the THz-TDS spectra corresponding to differences in water content could be observed (Figure 2).

Determination of the dry-basis moisture content

After collecting terahertz information, the collected leaves were placed in an envelope into an oven for 15 min of enzyme deactivation at 108 ºC. After drying at 80 ºC for 8 h, the primary dry weight was taken. After drying was continued for a period of time, the secondary dry weight was taken. The primary dry weight was then compared with the secondary dry weight. If the difference did not exceed 1% of the final measured mass, the secondary dry weight can be used as the final dry weight m$_2$. To ensure optimal sample separation, the dry-basis water content was calculated using the following formula:

$$w=\frac{m_1-m_2}{m_2}\times100\% \qquad (1)$$

where w is the dry-basis moisture content of the sample; m$_1$ is the fresh mass of the sample, g; and m$_2$ is the dry mass of the sample, g.

Figure 3 shows the scatter diagram of the dry-basis moisture content of four lettuce samples with different levels of water applied, arranged from low to high moisture content. The dry-basis moisture content significantly varied among the four treatments.

RESULTS AND DISCUSSION

The acquisition of THz-TDS data was affected by system noise. To prevent noise from affecting subsequent data processing and reducing modeling accuracy, we used Savitzky-Golay (S-G) smoothing combined with the S-G derivative algorithm and normalization to preprocess the data and select the optimal noise reduction method.

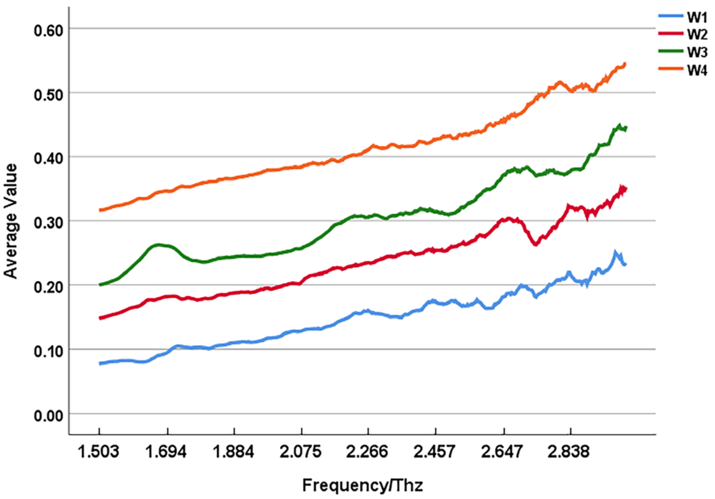

Figure 2 - Terahertz time-domain spectroscopy absorption spectrum mean value of samples with different water content levels. W1: average absorption spectrum of samples irrigated with sufficient water every day, W2: average absorption spectrum of samples irrigated with 100ml water every two days, W3: absorption spectrum of samples irrigated with 100ml water every four days, W4: absorption spectrum of samples irrigated with 100ml water every six days.

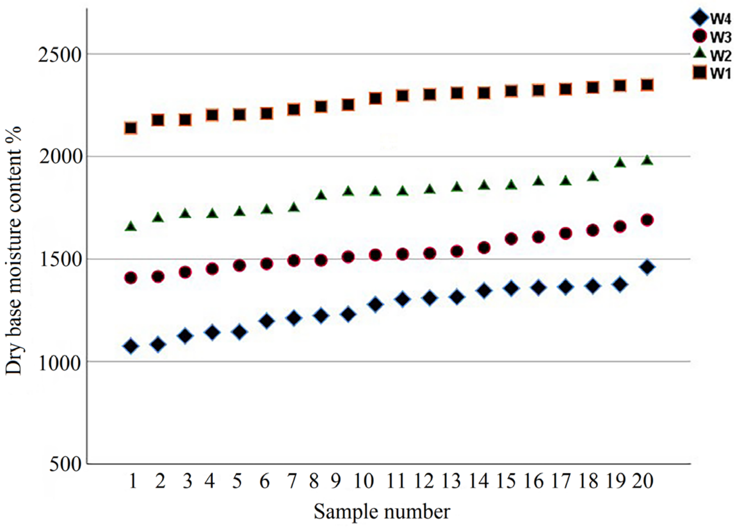

Figure 3 - Scatter diagram of groups with different gradient dry basis moisture contents. W1: dry basis moisture content of samples irrigated with sufficient water every day, W2: dry basis moisture content of samples irrigated with 100ml water every two days, W3: dry basis moisture content of samples irrigated with 100ml water every four days, W4: dry basis moisture content of samples irrigated with 100ml water every six days.

When S-G smoothing is used to preprocess data, the window width and polynomial order are particularly important because different values of these parameters directly affect the filtering performance. In this study, the window widths were either 5, 7, or 9 points, and the optimal window width was selected by comparing the determination coefficient R_c^2 obtained from PLS linear fitting with root mean square errors of calibration (RMSEC). Figure 4 (a) to (d) shows the spectrograms of the original data, as well as 5, 7, and 9 points for smoothing. There was obvious spectral jitter in the original data. When the window width was set to 5 points/time, the spectral shape was significantly smoothed. When it was set to 7 points/time and 9 points/time, a smooth transition was observed, which was accompanied by the loss of information. Calculations of the fitting accuracy are shown in table 1. The fitting effect of 5 points was consistently the highest, and R_c^2 and RMSEC were 0.9476 and 2.523%, respectively.

Division of the samples sets into the calibration set and verification set

To obtain improved modeling results, we used the Kennan-Stone algorithm, the sample set partitioning based on the joint X-Y distance algorithm, and the random sampling (RS) algorithm to divide the sample sets into a sample calibration set and a verification set. The best sample division algorithm was selected based on the correlation and root mean square error of full-band regression fitting. Detailed introductions into the algorithms are provided in previous studies. Results of the three algorithms are shown in table 2. The RS algorithm was the best (correction set: R_c^2 = 0.9307 and RMSEC = 285.1%; verification set: R_p^2 = 0.9253 and RMSEP = 425.7%). Hence, the RS algorithm was used to divide the sample sets.

Extraction of characteristic frequency bands

To accurately characterize the THz spectra of lettuce moisture content, we used three algorithms—stability competitive adaptive reweighted sampling (SCARS), iterative retained information variables (IRIV), and interval combination optimization (ICO)—to extract and compare the characteristic THz frequency bands; reduce the variables, redundancy, and collinearity; and improve the operational efficiency of the model.

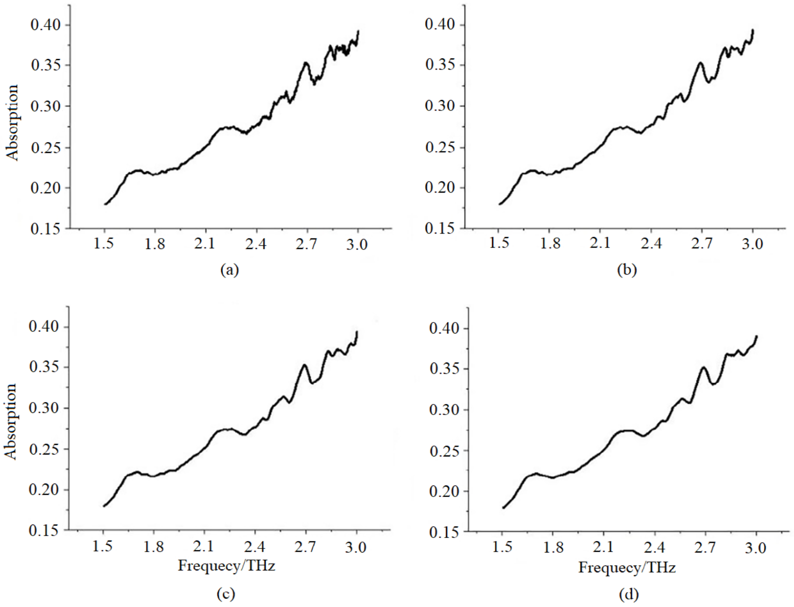

Figure 4 - Terahertz spectrogram smoothed with different window widths. (a) is the original data spectrogram, (b) is the spectrogram when the window width is set to 5, (c) is the spectrogram when the window width is set to 7, and (d) is the spectrogram when the window width is set to 9.

Screening of the characteristic frequency bands based on SCARS

The principle of SCARS is to use the stability of variables as an index, wherein higher stability indicates a greater probability that the selected variables are larger. The frequency band points with high absolute values of the regression coefficient in the PLS model were selected based on adaptive re-weighted sampling and the exponential decay function (EDF). Frequency band points with small weights were removed. The subset with the lowest RMSECV was selected by cross-validation, which can identify the optimal combination of variables.

The total water content of the lettuce samples was 80, and the initial number of runs of SCARS was 30, 40 and 50. Figure 5 shows the results

Table 1 - Fitting regression coefficient R^2 and mean square error RMSEC.

Serial number	Pretreatment method	R^2	RMSEC
1	5 points for time smoothing	0.9476	2.523%
2	7 points for time smoothing	0.9378	2.751%
3	9 points for time smoothing	0.9237	3.047%
4	Normalize on a smooth basis	0.9594	2.219%
5	First derivative on a smooth basis	0.9752	1.737%
6	Second derivative on a smooth basis	0.9145	3.224%

On the basis of S-G smoothing, the filtering process was further combined with S-G derivative and normalization. Comparative analysis shows that the determination coefficient R_c^2 and RMSEC of S-G derivative fitting are the best, reaching 0.9752 and 1.737% respectively.

Table 2 - Fitting regression results of the three algorithms.

Methods	Calibration set sample size	Validation set sample size	R_c^2	RMSEC%	R_p^2	RMSEP%
KS	60	20	0.9821	144.5	0.6934	780.1
SPXY	60	20	0.9441	237.1	0.8848	495.6
RS	60	20	0.9307	285.1	0.9253	425.7

when the number of runs was 50. After 30 runs, the number of samples basically stabilized. Therefore, 50 runs achieved the most stable running results, and this setting was used to characterize subsequent patterns. As the number of runs increased, the number of retained variables decreased; under the action of the EDF, the number of retained variables decreased sharply in the first 10 sampling runs, and then the decrease became less sharp, which reflects the primary selection and selection of the algorithm runs. The relationship between the number of reserved characteristic frequency bands and the number of runs indicates that the RMSECV of the interactive verification model was at least 2.520% when the algorithm was run 27 times. The error gradually increased, indicating that the algorithm began to eliminate the characteristic variables affecting the accuracy of the algorithm. Therefore, the subset of characteristic variables obtained in the 27th run was selected as the optimal subset, and the initial 38 terahertz characteristic bands significantly related to the lettuce water content were selected.

Screening of the characteristic frequency bands based on IRIV

IRIV is a feature variable selection algorithm based on a binary matrix rearrangement filter that divides all variables into four categories:

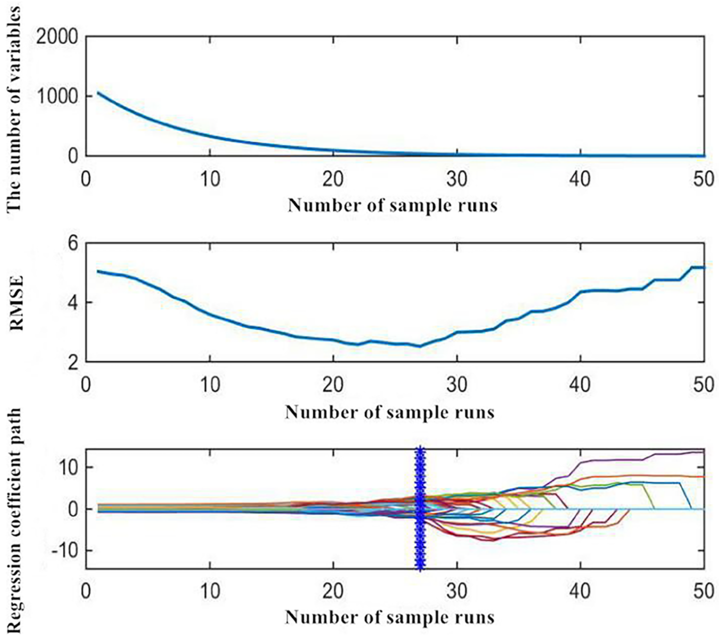

Figure 5 - Running result diagram of stability competitive adaptive reweighted sampling (SCARS). It shows the relationship between the number of running times and the number of characteristic frequency bands when the number of running times is set to 50.

strong, weak, non-interfering, and interfering information variables. IRIV requires many iterations, and in each iteration, the strong and weak information variables are kept, other variables are eliminated, and the optimal variable set is obtained by reverse elimination. IRIV was used to select the characteristic frequency bands of THz spectra filtered by the S-G derivative algorithm, and 36 THz-TDS bands were obtained by primary screening.

Screening of the characteristic frequency bands based on ICO

ICO divides the THz spectrum of each sample into N equal parts with roughly the same width, and each part is a band interval. The RMSECV corresponding to each interval combination subset is calculated by replacing the frequency points in the frequency band interval, which can eliminate accidental errors in the samples. If the inclusion of an adjacent frequency point causes the RMSECV of the model to decrease, the frequency characteristic variable is selected; otherwise, it is eliminated. The optimized interval is searched locally and repeatedly until no new variables affect the RMSECV of the model. The optimized interval is the band interval finally selected by ICO. Four frequency bands were selected by ICO, including 168 characteristic frequency bands.

Optimization of the feature band extraction algorithm

Table 3 shows the numbers of characteristic frequency bands obtained after the THz characteristic screening of lettuce water content by the three algorithms as well as the accuracy of modeling by PLS. When the number of characteristic frequency bands was approximately the same, the calibration set model R_c^2 of IRIV was 0.9648, and RMSEC was 204.9%. The verification set R_p^2 was 0.8990, and RMSEP was 439.1%. The model accuracy was higher compared with that of SCARS. However, ICO returned too many frequency bands, and model accuracy was lower compared with that of IRIV, indicating there is more redundant information. Therefore, we used IRIV for feature screening.

The above findings showed that the correlation coefficient and model accuracy are higher for IRIV, but up to 29 characteristic frequency bands were obtained. This suggested that this model has some redundancy and multicollinearity; an excessive number of variables is not conducive to practical applications of the model. Therefore, the characteristic variables need to be optimized again based on the initial screening of the characteristic frequency bands. Principal component analysis (PCA) and the successive projection algorithm (SPA) were used to optimize the primary features twice. The fitting accuracy was also used as the evaluation standard for algorithm optimization and optimal method selection.

Optimization of the characteristic variables based on PCA

PCA is commonly used for reducing the dimensionality of data sets. A correlation between two variables can be interpreted as overlap or collinearity of the characteristic information reflected by these two variables. The purpose of PCA is to remove the redundancy of the original characteristic variables and establish a new dataset with as few variables as possible that contains most of the original information. The 29 feature variables screened by IRIV were subjected to PCA. The first seven principal components of the water content explained 96.44% of the variation in the data (Figure. 6), indicating that most of the information in the original data set was contained in the transformed data set. When these 7 principal components were used to establish a PLS model, the R_c^2 was 0.9028, and RMSEC was 531.5%. Although, the accuracy was slightly reduced compared with that of the original dataset, the prediction model was significantly simplified.

Characteristic variable optimization based on the SPA

SPA can be used to identify the variable group with the minimum amount of redundant

Table 3 - Model parameters for three algorithms.

Algorithm	Number of features	R_c^2	RMSEC%	R_p^2	RMSEP%
SCARS	38	0.9393	280.1	0.8605	448.3
IRIV	29	0.9648	204.9	0.899	439.1
ICO	168	0.9646	205.0	0.8461	489.4

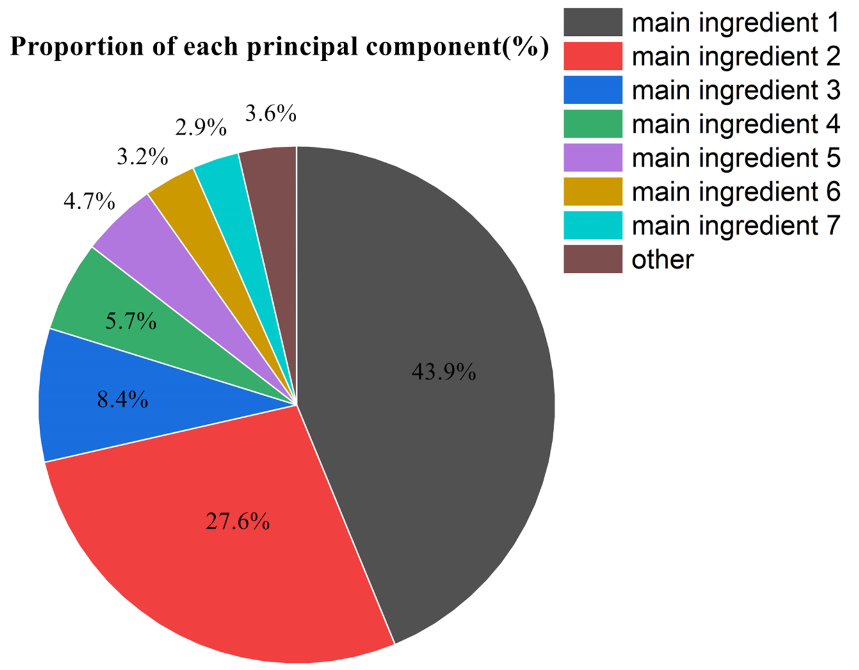

Figure 6 - Results of principal component analysis. The data in the picture shows the contribution ratio of the top 7 principal components.

spectral information, which minimizes the collinearity between variables, greatly reduces the number of variables used in modeling, and improves the speed and efficiency of modeling.

The absorptance of 29 characteristic frequency bands of the water content obtained by IRIV was used as the independent variable X, and the minimum and maximum values of filter results were set to 1 and 29, respectively. The reason why the maximum value was 29 instead of the expected interval was because of the existence of a correlation among these 29 variables. To reduce the collinearity among variables, the variable interval was set to the maximum value, and step-by-step screening was performed. The dimension reduction effect was superior using this approach compared with that of directly setting the parameters to a smaller expected characteristic interval. The six characteristic frequency bands of the lettuce water content obtained by SPA were 1.667, 0.061, 3.105, 0.366, 3.380, and 3.883 THz (Figure. 7). Figure 8 shows the locations of the selected bands. The six THz-TDS water content characteristic frequency bands were mostly distributed in the peaks, valleys, or inflection points.

Based on the band variables selected by SPA, a calibration set and a verification set were established using PLS regression. The R_c^2 and RMSEC of the calibration set fitted by PLS regression were 0.8962 and 412.5% respectively, suggesting that the fitting accuracy of SPA was higher compared with PCA.

The THz detection model of lettuce moisture content was established by PLS regression with SPA-optimized characteristic variables, and the verification set was substituted into the calibration set model. The R_p^2 and RMSEP of the verification set model were 0.8757 and 528.9%, respectively.

Figure 9 shows a correlation model between the predicted and true water content of lettuce. The predicted sample points of the data set were distributed on both sides of the fitting line and were highly correlated, indicating that the prediction model can be used to predict the water content in lettuce leaves.

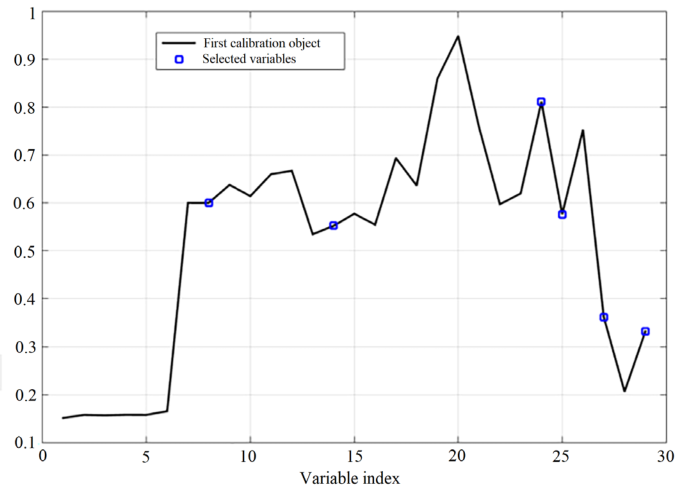

Figure 7 - Selection of successive projection algorithm frequency band for water content of lettuce.

CONCLUSION

Through comparative analysis, when optimizing spectral features and reducing dimensionality, it is concluded that the IRIV algorithm combined with SPA algorithm for feature extraction of lettuce water stress can reduce the number of variables and overcome multicollinearity while ensuring the accuracy of fitting.

A THz-TDS prediction model for lettuce leaf water stress was established. The R_c^2 and RMSEC of the calibration set of the model were 0.8962 and 412.5%, respectively, and the R_p^2 and RMSEP of the validation set were 0.8757 and 528.9%, respectively.

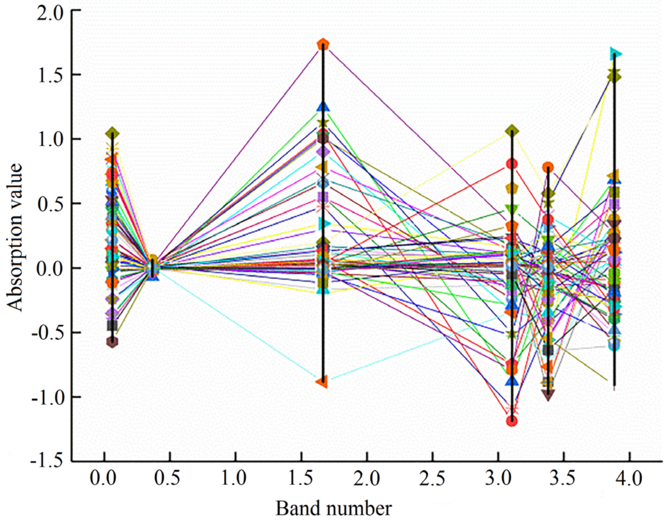

Figure 8 - Characteristic frequency band distribution. At the peak and inflection point, six characteristic bands of lettuce water content were obtained, which were 1.667, 0.061, 3.105, 0.366, 3.380 and 3.883THz, respectively.

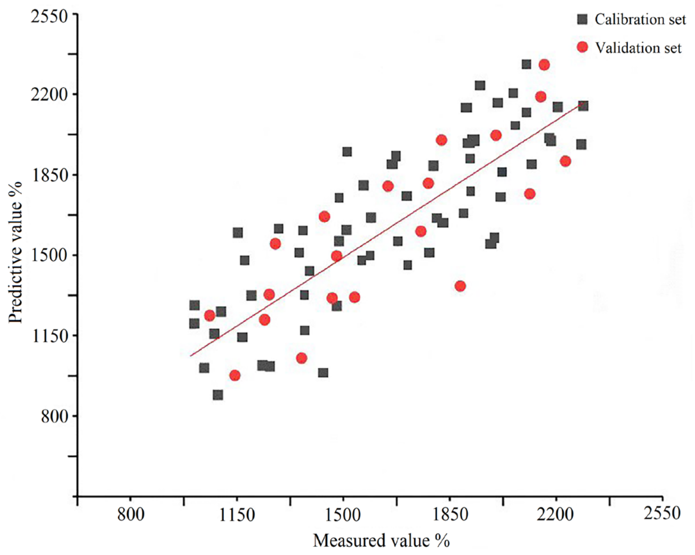

Figure 9 - Water prediction value and measured value of lettuce.

ACKNOWLEDGEMENTS

The study was funded by "the Program of the National Natural Science Foundation of China" (61771224), "Jiangsu Natural Science Foundation" (BK20180864), "Open Fund of Key Laboratory of Modern Agricultural Equipment and Technology" (JNZ201903), "Project of Faculty of Agricultural Equipment of Jiangsu University" (NZXB20200203), "Zhenjiang Science & Technology Program" (NY2019017), "Self made experimental equipment project of Jiangsu University" (ZZYQSB2021009).

DECLARATION OF CONFLICT OF INTEREST

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

AUTHORS' CONTRIBUTIONS

All authors contributed equally for the conception and writing of the manuscript. All authors critically revised the manuscript and approved of the final version.

REFERENCES

LIMA, R. S. N. et al. Linking thermal imaging to physiological indicators in Carica papaya L. under different watering regimes. **Agricultural Water Managementv**, v.164, n.1, p.148-175, 1999. Available from: <https://www.sciencedirect.com/science/article/pii/S0378377415300639>. Accessed: Jul. 10, 2019. doi: 10.1016/j.agwat.2015.07.017.

O'SHAUGHNESSY, S. A. et al. Using radiation thermography and thermometry to evaluate crop water stress in soybean and cotton. **Agricultural Water Management**, v.98, n.1, p.1523-1535, 2011. Available from: <https://www.sciencedirect.com/science/article/pii/S0378377411001119>. Accessed: Jul. 10, 2019. doi: 10.1016/j.agwat.2011.05.005.

OSCO, L. P. et al. Modeling Hyperspectral Response of Water-Stress Induced Lettuce Plants Using Artificial Neural Networks. **Remote Sensing**, v.12, n.21, p.125-145, 2018. Available from: <https://ui.adsabs.harvard.edu/abs/2019RemS...11.2797O/abstract>. Accessed: Jul. 10, 2019. doi: 10.1016/j.agwat.2014.03.015.

TOSIN, R. et al. Estimation of grapevine predawn leaf water potential based onhyperspectral reflectance data in Douro wine region. **Vitis**, v.278, n.585, p.28-77, 2020. Available from: <https://www.sciencedirect.com/science/article/pii/S0304423820306889>. Accessed: Jul. 10, 2019. doi: 10.1016/j.scienta.2020.109860

ZHOU, Z. et al. Terahertz Wave Science and Technology. **Automation Instrumentation**, v.46, n.8, p.1086-1107, 2019. Available from: <https://sci-hub.se/10.1016/j.agwat.2017.02.015>. Accessed: Jul. 10, 2019. doi: 10.1016/j.agwat.2017.02.015.

AUSTON, D H, et al. Cherenkov Radiation from Femtosecond Optical Pulses in Electro-Optic Media. **Physical Review Letters**, v.53, n.16, p.1555-1555, 2019. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.53.1555>. Accessed: Jul. 10, 2018. doi: 10.1016/j.agwat.2008.05.017.

WANG, J. R. et al. THz-TDS combined with a fuzzy rule-building expert system applied to the identification of official rhubarb samples. **Analytical methods**, v.6, n.19, p.7695-7702, 2018. Available from: <https://pubs.rsc.org/en/content/articlelanding/2014/ay/c4ay00555d>. Accessed: Jul. 10, 2019. doi: 10.1039/C4AY00555D.

MATHANKER, S. K. et al. Terahertz(THz) applications in food andagriculture: A review. **Transactions of the ASABE**, v.56, n.13, p.1213-1226, 2015. Available from: <https://elibrary.asabe.org/abstract.asp?aid=42737>. Accessed: Jul. 10, 2019. doi: 10.13031/trans.56.9390.

PARASOGLOU, P. et al.Quantitative water content measurements in food wafers using terahertz radiation. **Terahertz Sci Technol**, v.3, n.4, p.176–82, 2010. Available from: <http://www.tstnetwork.org/10.11906/TST.149-162.2010.12.15>. Accessed: Jul. 10, 2019. doi: 10.11906/TST.149-162.2010.12.15.

YANG, Y. P. et al. Identification of Genistein andBiochanin A by THz (far-infrared) vibrational spectra. **Journal of pharmaceutical and biomedical analysis**, v.62, n.11, p.177-189, 2014. Available from: <https://www.sciencedirect.com/science/article/pii/S0731708511007205>. Accessed: Jul. 10, 2019. doi: 10.1016/j.jpba.2011.12.013.

GENTE, R. et al. Quality control of sugar beet seeds with THz time-domain spectroscopy. **Transactions on terahertz science and technology**, v.6, n.5, p.754-756, 2013. Available from: <https://ieeexplore.ieee.org/document/7536209>. Accessed: Jul. 10, 2019. doi: 10.1109/TTHZ.2016.2593985.

PARK, S. J. et al. Detection of microorganisms using terahertz metamaterials. **Scientific Reports**, v.1, n.7, p.268-298, 2019. Available from: <https://www.nature.com/articles/srep04988/>. Accessed: Jul. 10, 2019. doi: 10.1038/srep04988.

LONG, Y. et al. The preliminary research on isolated leaf moisture detection using terahertz technology. **Spectroscopy and Spectral Analysis**, v.37, n.10, p.3027-3031, 2013. Available from: <http://www.opticsjournal.net/Articles/Abstract/gpxygpfx/37/10/3027.cshtml>. Accessed: Jul. 10, 2019. doi: 10.12677/AAC.2018.81001.

ZHAO, X. et al. The application of terahertz spectroscopyto the detection of soybean canopy water content under drought stress. **Spectroscopy and Spectral Analysis**, v.38, n.8, p.2350-2354, 2015. Available from: <https://www.sciencedirect.com/science/article/pii/S0168169919318204>. Accessed: Jul. 10, 2019. doi: 10.1016/j.saa.2020.118453.

ZAHID, A. et al. Machine learning driven non-invasive approach of water content estimation in living plant leaves using terahertz waves. **Plant Methods**, v.15, n.1, p.1746-4811, 2016. Available from: <https://pubmed.ncbi.nlm.nih.gov/31832080/>. Accessed: Jul. 10, 2019. doi: 10.1186/s13007-019-0522-9.

PAGANO, M. et al. THz water transmittance and leaf surface area: an effective nondestructive method for determining leaf water content. **Sensors**, v.19, n.22, p.27-89, 2016. Available from: <https://pubmed.ncbi.nlm.nih.gov/31698861/>. Accessed: Jul. 10, 2019. doi: 10.3390/s19224838.

ZAHID, A. et al. Characterization and water content estimation method of living plant leaves using terahertz waves. **Applied Sciences-Basel**, v.9, n.14, 2017. Available from: <http://eprints.gla.ac.uk/187606/>. Accessed: Jul. 10, 2019. doi: 10.20944/preprints201907.0125.v1.

LI, B. et al. Measurements and analysis of water content in winter wheat leafbased on terahertz spectroscopy. **International Journal of Agricultural and Biological Engineering**, v.11, n.3, p.178-182, 2013. Available from: <https://www.researchgate.net/>. Accessed: Jul. 10, 2019. doi: 10.25165/j.ijabe.20181103.3520.

NIE, P. et al. Detection of water content in rapeseed leaves using terahertz spectroscopy. **Sensors**, v.17, n.12, 2014. Available from: <https://www.researchgate.net/>. Accessed: Jul. 10, 2019. doi: 10.25165/j.ijabe.20181103.3520.

BALDACCI, L. et al. Non-invasive absolute measurement of leaf water content using terahertz quantum cascade lasers. **Plant Methods**, v.16, n.51, 2017. Available from: <https://plantmethods.biomedcentral.com/>. Accessed: Jul. 10, 2019. doi: 10.1186/s13007-017-0197-z.

TORRES, V. et al. Monitoring water status of grapevine by means of THz waves. **Journal of Infrared Millimeter and Terahertz Waves**, v.37, n.5, p.507-513, 2013. Available from: <https://link.springer.com/journal/10762/volumes-and-issues/37-5>. Accessed: Jul. 10, 2019. doi: 10.1007/s10762-016-0269-6.

ZHANG, X. et al. Water stress diagnosis of rape based on multispectral vision technology. **Transactions ofthe Chinese Society of Agricultural Engineering**, v.27, n.3, p.152-157, 2018. Available from: <https://scialert.net/abstract/?doi=ajps.2017.1.8>. Accessed: Jul. 10, 2019. doi: 10.3923/ajps.2017.1.8.

JIANG, Y. et al. Research on the quantitative detectionof wheat maltose based on terahertz imaging technology. **Spectroscopy and Spectral Analysis**, v.38, n.10, p.3017-3022, 2015. Available from:<https://pubmed.ncbi.nlm.nih.gov/>. Accessed: Jul. 10, 2019. doi: 10.1016/j.foodchem.2019.125533.

KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. **Technometrics**, v.11, n.1, p.137-148, 2016. Available from: <https://www.jstor.org/stable/1266770>. Accessed: Jul. 10, 2019. doi: 10.3760/cma.j.issn.1002-0098.2019.10.012.

GALVÃO, R. K. H. et al. A method for calibration and validation subset partitioning. **Talanta**, v.67, n.4, p.736-740, 2015. Available from: <https://www.sciencedirect.com/science/article/pii/S003991400500192X>. Accessed: Jul. 10, 2019. doi: 10.1016/j.talanta.2005.03.025.

CASTRO, A. I. D. et al. An automatic random Forest-OBIA algorithm for early weed mapping between and within crop rows using UAV imagery. **Remote Sensing**, v.10, n.2, 2018. Available from: <https://www.researchgate.net/>. Accessed: Jul. 10, 2019. doi: 10.3390/rs10020285.

SONG, X. Z. et al. A novel algorithm for spectral interval combination optimization. **Analytica Chimica Acta**, v.948, n.15, p.19-29, 2017. Available from: <https://pubmed.ncbi.nlm.nih.gov/27871606/>. Accessed: Jul. 10, 2019. doi: 10.1016/j.aca.2016.10.041.

WANG, S. et al. A robust principal component analysis (PCA) algorithm. **System Engineering Theory and Practice**, v.1998, n.1, p.10-14, 2016. Available from: <https://pubmed.ncbi.nlm.nih.gov/>. Accessed: Jul. 10, 2019. doi: 10.1109/EMBC.2013.6611074.

ARAÚJO, M. C. U. et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. **Chemometrics and Intelligent Laboratory Systems**, v.57, n.2, 2014. Available from: <https://www.sciencedirect.com/science/article/pii/S0169743901001198>. Accessed: Jul. 10, 2019. doi: 10.1016/S0169-7439(01)00119-8.