



OPEN

## Retrieval of nicotine content in cigar leaves by remote analysis of aerial hyperspectral combining machine learning methods

Chenyu Tian , Yifei Lu , Hengduo Xie , Yufan Yu  & Liming Lu  

Cigar leaf is a special type of tobacco plant, which is the raw material for producing high-quality cigars. The content and proportion of nicotine and other composite substances of cigar leaves have a crucial impact on their quality and vary greatly with the time of harvest. Hyperspectral remote sensing technology has been widely used in the field of crop monitoring because of its advantages of large area coverage, fast information acquisition, short cycle turnover, strong real-time performance and high efficiency. Therefore, it is important to accurately monitor nicotine content of field crops in a timely manner in the production of high-quality cigar leaf. To this end, this study set out to measure crop reflectance spectra acquired by UAV drones from tobacco field crops by hyperspectral image acquisition. MSC, SG, and SNV were combined and applied to the raw data. The output of these operations was then further processed by CARS, SPA, and UVE algorithms to determine the nicotine sensitive bands. Three machine learning algorithms were then used to analyze the data: PLS, BP, RF, and the SVM. An inversion model of the content of nicotine was established, and the model was evaluated for accuracy. The main research conclusions are as follows: (1) With the increase in the rate of application of nitrogen fertilizer, the nicotine content of cigar leaves increased; (2) Processing data by the CARS, SPA, and UVE methods reduces the degree of data redundancy and information co-linearity in the screening of the content of nicotine sensitive bands; (3) The MSC-SNV-SG-CARS-BP model has the best predictive accuracy on the nicotine content. The prediction accuracy of the testing set was  $R^2 = 0.797$ ,  $RMSE = 0.078$ ,  $RPD = 2.182$ .

Tobacco is an important cash crop. Among, cigar leaf is a special type of tobacco which has the characteristics of unique aroma, strong taste, great strength and strong satisfaction<sup>1</sup>. The intrinsic substances of tobacco leaf directly determine its smoking quality. The quality of cigar leaves includes its appearance quality, chemical composition, and human sensory response to its aroma<sup>2</sup>. Among these, the chemical composition of tobacco, such as sugars, total nitrogen content, nicotine and other substances, in addition, other factors such as the type, proportion, and content level, play a role on the quality of cigar leaves. Fast and accurate monitoring of the quality of cigar leaves is of importance for the production of high-quality cigar leaves<sup>3</sup>. However, in the current production of cigar leaves in China, quality monitoring is still based on traditional methods of manual investigation, and modern information technology is rarely used.

Hyperspectral remote sensing refers to the technology of obtaining multiple, highly dense, and spectrally resolved images from optical targets in the four bands of the electromagnetic spectrum (ultraviolet band, visible band, near infrared band and middle infrared visible band). Hyperspectral remote sensing technology has the advantages of low cost and high efficiency compared with conventional artificial ground survey due to its capacity for large area coverage, fast image acquisition speed, short period, and strong real-time performance. At the same time, hyperspectral remote sensing itself produces a large amount of information collected under a large number of derived variables. In the topic of crop monitoring, especially in rice, wheat, corn and other crops, multiple parameters have been used to characterize crop growth and product yields. By acquiring continuous spectral images of a target object in different wavelength bands, hyperspectral remote sensing acquires, analyzes and compares its reflection and absorption characteristics for the purpose of obtaining detailed information on the composition and structure of the material<sup>4</sup>. This technology utilizes narrow band width, combined with high spectral resolution and density. Its data is parametrized in a variety of ways, which makes it of some consequence in agricultural production. For example, in planting flue-cured tobacco, hyperspectral remote

Agronomy college, Sichuan Agricultural University, Chengdu 611130, Sichuan, China.  email: luliming@sicau.edu.cn

sensing technology has been used to access key indicators such as growth status, quality and stress of tobacco leaves, which is important for achieving efficient and accurate flue-cured tobacco production<sup>5,6</sup>.

Spectral reflectance is a parameter that indicates the ability of an object to reflect light at different wavelength bands. By measuring changes in spectral reflectance factors as a function of wavelength, one can identify crop growth problems in a timely fashion. Rapid scanning rates and high area coverage facilitate timely remediation measures. Daily surveys can enable farmers and producers to establish a scientific basis to identify actions that will improve crop quality and yield, or reduce unnecessary waste of resources. The absorption, reflection and transmission factors of optical radiation being monitored are determined by the assorted biochemical constituents that are active within the crop, or indicative of its state of hydration. Reflectivity parameters that characterize crop growth are perhaps the easiest to monitor from the perspective of surveying an entire field of crops. Measurements of the changes of spectral reflectance obtained at different angles of incident illumination and at different optical wavelengths have been used to characterize crop growth and agricultural production under varying seasonal and environmental conditions. A band selection method called Genetic Programming Spectral Vegetation Index (GP-SVI), was previously reported by<sup>7</sup>. They proposed that a combination of SVI and GP be used to characterize crop growth. The method derives from GA. In addition, maize canopy nitrogen content was correlated with hyperspectral data acquired by the Compact Airborne Spectrographic Imager (CASI) sensor<sup>8</sup>. used binary PSO to select the best feature subset and fed it to SVR to estimate rice N concentration. The features extracted by PCA were incorporated into an artificial neural network and used to estimate the maize N concentration index<sup>9</sup> and rice N concentration<sup>10</sup>. About in tobacco leaf information UAV spectral data modeling inversion, Qin uses PROSAIL to augment hyperspectral images with samples to avoid overfitting. The inversion model combines K-means and XGBoost to form a hybrid model. The results indicate that the hybrid model outperforms the other models on the validation set with  $R^2=0.83$  and  $RMSE=3.9$ <sup>11</sup>. Zhang estimated tobacco LNC using UAV hyperspectral-image data. The results show that all the ensemble learning methods are superior to PLSR ( $R^2=0.680$ ,  $RMSE=5.402$  mg/g, 19.72%). Specifically, the stack-based model achieves the highest accuracy and relatively high stability ( $R^2=0.745$ ,  $RMSE=4.825$  mg/g, 17.98%)<sup>12</sup>. Junying et al.<sup>13</sup> proposed a method to predict the  $K_2O$  content of tobacco based on UAV-borne hyperspectral imaging. The results on the test set show that the  $RMSE$  of the model is 0.40, and the absolute value of the average relative error is 8.04%. Junying et al.<sup>14</sup> proposed a method for predicting total sugar content based on UAV-borne hyperspectral imaging. The model was constructed by combining spectral properties and measured total sugar values. According to the sample test results, the  $RMSE$  of the model is 1.84, and the absolute value of the average relative error is 8.82%. Hayes and Reed<sup>15</sup> conducted a field study using UAV-borne hyperspectral imaging to detect tobacco pests and diseases. The subspace LDA algorithm is utilized to test the recognition ability, and the overall accuracy is 85.7%.

With the continuous development of hyperspectral remote sensing technology over recent years, it is now possible to observe weak spectral differences. This technology is capable of being used directly to obtain quantitative analysis, which gives credibility to the prospects that remote sensing has considerable potential for field applications to characterize plant physiological and biochemical health<sup>16</sup>. This study uses hyperspectral information technology to collect the spectral information of cigar leaves as they grow in the field. It establishes corresponding mathematical models, and inverts the data to reveal and predict the content of nicotine that is related to the quality of cigar leaves at maturity. It aims to broaden the applications of hyperspectral sensing to optimize leaf production and value.

## Experimental design

### Test materials

The test materials in this experiment are leafs from 15 cigar varieties. They are: Cuba 1, Cuba 4, Cuba 5, Cuba 6, Cuba 7, Cuba 8, Cuba 9, Slovenia 1, Norway 1, Dominica 2, Nicaragua 2, Indonesia 1, Desuce 1, Desuce 3, and Shiyan 1.

### Experimental site profile

This work adopted a two-factor split-plot design, with nitrogen fertilizer accretion as the main controlling factor and tobacco variety as the secondary factor. A nitrogen (pure nitrogen) application gradient was established at three levels: level 1, 0 kg/667 m<sup>2</sup>; level 2, 6 kg/667 m<sup>2</sup>; level 3, 12 kg/667 m<sup>2</sup>. During the field growth stage (May to September), the average temperature was 25°C, the average contents of soil organic matter was 29.61 g/kg, the average contents of total nitrogen, available phosphorus and available potassium were 1.88 g /kg, 18.09 mg/kg and 56.90 mg/kg, respectively. The total precipitation in the field during the growth stage was 260 mm. The soil type was yellow brown soil. The tobacco soil has a strong capacity of water and fertilizer, and is convenient for drainage and irrigation. the soil water content is 60–70%, the soil organic matter content is 1.5%, and the pH value is 6.3.

### Experimental samples

In this study, the upper, middle and lower leaves of the tobacco plant of each variety were collected at the Modern Agricultural Research Base at Sichuan Agricultural University in Chongzhou, Chengdu, Sichuan Province. Leaves were extracted during the mature picking period from June 20 to July 20, 2022. In total, 15 cigar varieties were picked : Cuba 1, Cuba 4, Cuba 5, Cuba 6, Cuba 7, Cuba 8, Cuba 9, Slovenia 1, Norway 1, Dominica 2, Nicaragua 2, Indonesia 1, Desuce 1, Desuce 3, and Shiyan 1 (Table 1). Next, three nitrogen fertilizer concentrations were established and sampling analysis was carried out at the mature stage. The tobacco leaves are ripe and harvested, dried naturally, and then collected. High-grade cigar tobacco leaves were treated with 2 kg each, and conventional chemical components such as nicotine were analyzed. A total of 270 cigar tobacco samples were measured for nicotine and other chemical components.

Number	Variety
S52	Cuba 1
S53	Cuba 4
S54	Cuba 5
S55	Cuba 6
S56	Cuba 7
S57	Cuba 8
S58	Cuba 9
S59	Slovenia 1
S60	Norway 1
S61	Dominica 2
S62	Nicaragua 2
S63	Indonesia 1
S64	Shiyan No. 1.
S65	Desue 1
S66	Desue 3

**Table 1.** Varieties of cigars.

#### *Partition of sample set*

A total of 270 tobacco spectral data were randomly sampled in sequence, 216 of which were used as the training set and 54 were used as the testing set. The outflow method is divided according to the ratio of the training set to the test set at a ratio of 3:7.

### **Hyperspectral image acquisition and data extraction**

#### *Hyperspectral image acquisition*

After the tobacco growth period, the DJI M600pro six-rotor UAV equipped with Nano-Hyperspec hyperspectral camera was used to take field photos every 7 days until the whole field entered the mature stage. UAV operations should be conducted when good weather conditions can be ensured, when the wind does not exceed level 3, and during the time period around noon. Before take-off, the datum of the survey area should be checked and a spacious, unobstructed hardened road surface selected as the area for take-off and landing. The flight height is 10 m, and the original image obtained by the UAV should be quasi-static. The hyperspectral sensor used has a spectral resolution of 4 nm, The number of axes of the UAV is 6, the wheelbase is 1133 mm, the maximum load is 15.5 kg, the maximum flight speed is 18 m/s, the maximum climb speed is 5 m/s, the maximum descent speed is 3 m/s.

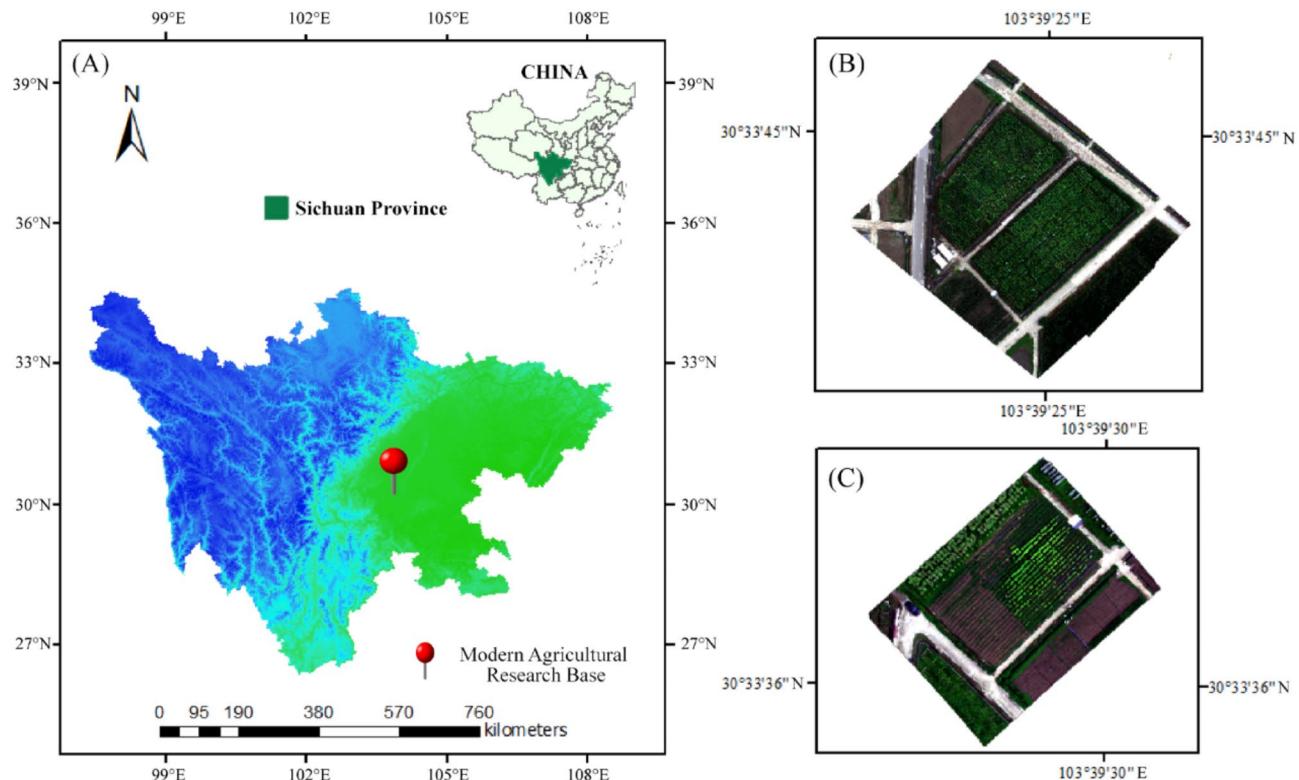
The original images were then stitched into complete images using Pix4DMapper software (Pix4D S.A., Switzerland). The target region is selected in each tobacco field image, and then the spectral information in each region of interest is extracted to obtain the corresponding original spectral data. Part of the hyperspectral image is shown in Fig. 1. Due to the role of hyperspectral sensors, system platforms, atmosphere, and terrain in data acquisition, the generated image pixels are squeezed, stretched, distorted, and shifted relative to the actual locations in the cultivation area. processed using the atmospheric correction and radiometric correction Calibration function of ENVI 5.3 (ENVI/IDL5.3., USA). Radiometric correction and geometric calibration are also essential for reducing noise interference, improving reflection accuracy, and ensuring the precision of radiometric measurements. These operations can convert the raw images into hyperspectral reflectance data for the entire tobacco field.

#### *Spectral enhanced noise reduction and smoothing*

Spectral enhancement mainly includes multiplicative scatter correction (MSC) and standard normal variate (SNV) processing. Matlab 2021 (MathWorks, USA) was used to pre-process the original spectral data for signal enhancement. Spectral smoothing and noise elimination processing incorporate Savitzky-Golay (SG) convolution smoothing and baseline correction. SG convolutional smoothing was proposed by Savitzky and Golay<sup>17</sup> to play the role of noise resolution without changing the original spectrum. Among the various options, the polynomial fitting multiplied by the window size directly affects data smoothing<sup>18</sup>. Matlab 2021 was used to perform spectral smoothing and de-noising on the original spectral data files. These aspects of signal processing are associated with hyperspectral image pre-processing methods. They are useful for reducing and often eliminating the inclusion of irrelevant information in the formed image. Image noise reduction is often critical to providing reliable data source files for calculating subsequent inversion models.

#### *Variable selection of hyperspectral features*

Uninformative variable elimination (UVE) is an effective noise suppression technique that extracts valuable features from a large number of irrelevant variables, and effectively eliminates useless information by simulating the stability of variables. Matlab 2021 algorithms were used to shuffle the original spectral data and add noise through construction of an independent variable matrix. After adding noise, UVE makes a varying judgment call that assigns a regression coefficient of the target matrix based on the independent variable matrix that comprises



**Fig. 1.** Part of the hyperspectral image. (A) Study area; (B) High nitrogen and medium nitrogen treatment areas; (C) Low nitrogen treatment area. ArcGIS (V10.3 Esri).

the spectral variable + noise. The statistical distribution of the regression coefficient is represented by the ratio of the mean and standard deviation. The statistical distribution of regression coefficients was represented by the ratio of mean and standard deviation. The ultimate feature variables are determined by evaluating the upper and lower signal limits and proposing variables that lie within the corresponding range<sup>19</sup>.

Successive Projections Algorithm (SPA) is a widely used, wavelength dependent algorithm<sup>20</sup> that selects spectral features. This algorithm is available in Matlab 2021. It scrambles the original spectral data, then reconstructs the matrix such that it expands the data set to the full band range. A wavelength  $i$  is randomly selected, whereupon a successive projection series is generated, starting from  $i$ , to achieve the best spectral feature selection. This process is repeated until all the bands are covered, and finally the wavelength subset of the whole band is determined.

Competitive adaptive reweighted sampling (CARS) is widely used for uncovering hidden features. It captures the specific characteristic wavelength by the least squares variational method and Monte Carlo sampling and accomplishes effective identification of specific spectral characteristics. Matlab 2021 performs these operations by obfuscating the raw information of the spectrum, then utilizes Monte Carlo sampling and minimizes the variation between (what variables) by partial least squares method in order to pick out features with progressively increasing weights. The extrema of these become the final feature wavelengths<sup>21</sup>.

### Establishment of inversion model

Machine learning regression methods have been successfully applied in many instances, including monitoring crop physiological and biochemical parameters. Inversion models have mainly utilized the partial least squares regression (PLSR)<sup>22</sup>, Back propagation neural network (BP)<sup>23</sup>, support vector regression (SVR)<sup>24</sup>, and random forest regression (RFR)<sup>25</sup>. They all possess the exceptional ability to explore complex non-linear relationships between spectral features and biochemical parameters status indicators in the absence of explicit knowledge of the precise distributions. BP is an artificial neural network ANN. ANN are a common method for developing nonlinear regression and consist of one input layer, multiple hidden layers, and one output layer<sup>26</sup>. The number of hidden layers is determined, in practice, by parameter tuning. Some researchers have suggested that the number of hidden neurons cannot exceed the maximum value obtained by doubling the sum of neurons in the input and output layers<sup>27</sup>. Other ANN parameters, such as initial weights, activation functions, and learning rates and dynamics, also have an impact on the model performance. Two important parameters to tune when dealing with in RFR algorithms are the number of regression trees to construct, and the associated input variables to allow at each node. They are usually determined by tracing out trees and computing errors against the training set of parameters. Optimal use involves estimating trade-offs between computational cost and estimation accuracy.

### The selection of model evaluation index

In this study, the determination of coefficient  $R^2$  and root mean square error (RMSE) were selected as the appropriate indicators to evaluate the estimation model and the validation model. The larger  $R^2$  of the estimation model is, the smaller is the RMSE, indicating the better accuracy of the estimation model. RMSEC is the root mean square error of the training set and RMSEP is the root mean square error of the test set. The larger the  $R^2$  of the verification model, the smaller is the RMSE, indicating the better stability of the estimation model<sup>28</sup>.  $R^2$  and RMSE are calculated using the following formulas (1), (2) and (3).

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}{\sum_{i=1}^n (\bar{Y}_i - \bar{Y})^2} \quad (1)$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}{n}} \quad (2)$$

$$RMSEC = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}{n}} \quad (3)$$

Here,  $n$  is the number of samples,  $Y_i$  and  $\bar{Y}_i$  are the number of samples, the following for nicotine content, respectively.  $\bar{Y}$  represents the average value of nicotine content of the measured samples.

A 5-fold cross-validation (5-fold cross-validation) is used to verify the optimal model selected from different models (the most suitable model). The data set was divided into five sets, and four of them were taken as training data and one as validation data. The corresponding accuracy was obtained for each trial, and the average of the accuracy of 5 times was used as an estimate of the accuracy of the algorithm. At the same time, different models do the same operation as above to get the average ability of each model on a specific data set, which is selected from among.

RPD (Relative Percent Difference) is an indicator used to evaluate the predictive performance of a model, especially in the fields such as spectral analysis or stoichiometry. It represents the ratio between the standard deviation (Standard Deviation, SD) and the predicted root mean square error (Root Mean Square Error of Prediction, RMSEP). The following is the expression form of the RPD calculation formula:

$$RPD = SD/RMSEP$$

$$SD = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

$$RMSECV = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}{n}}$$

SD represents the standard deviation, usually the standard deviation of the actual observations or the reference dataset. A higher RPD value usually indicates a better predictive performance of the model, as it means that the difference between the predicted and actual values is small relative to the overall variability of the data. All image and statistics analyses were evaluated using Matlab 2021a (MathWorks Inc., Natick, MA, USA).

## Results and analysis

### The effects of different nitrogen fertilizer application rates on the chemical composition of cigar leaves

There was strong dependence of nicotine content on the quantity of nitrogen fertilizer applied to tobacco plants, as the greater the weight of nitrogen fertilizer applied, the higher the nicotine content (Table 2). The nicotine content of different varieties of cigars grows with the increased application of nitrogen. However, both high and low nicotine content affected the aroma and taste of cigars. Low nicotine content resulted in reducing the physiological stimulation of aroma and taste of cigars. High nicotine content was found to produce excessive physiological stimulation from its aroma, and adversely affects the taste of cigars.

### Pre-treatment of raw spectra

#### *Spectral data acquisition and pre-processing*

The hyperspectral images of cigar leaves grown under different nitrogen fertilizer applications were collected by a hyperspectral imager mounted on an unmanned aerial vehicle (UAV). Figure 2A shows the original spectra in the visible-near IR band. Figure 3 shows the original spectra in different nitrogen fertilizer applications.

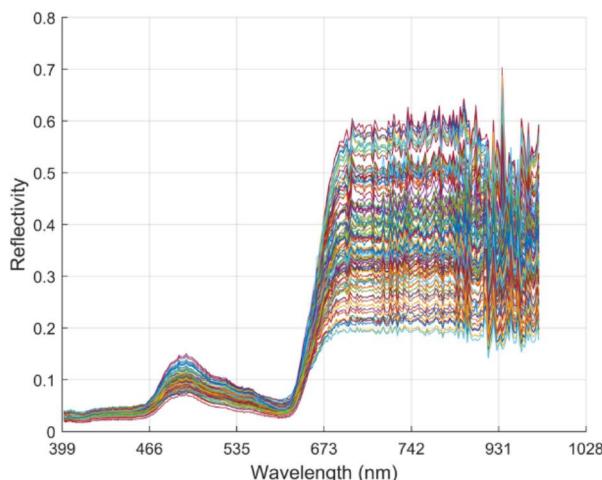
By adopting the joint processing by multivariate scatter correction, standard normal transformation and SG convolution smoothing algorithms, we are able to effectively reduce the interference from variations in the external environment, illumination sources, and instrument dark current on the hyperspectral data. Figure 2B shows the tobacco spectra after pre-processing the data allows us to achieve accurate baseline correction in spectra that originally exhibited a large amount of baseline drift.

According to the reflectance spectral decomposition found in the hyperspectral image, we found a jump in the spectral reflectance of tobacco leaves with increase of fertilizer nitrogen. In particular, the increase of

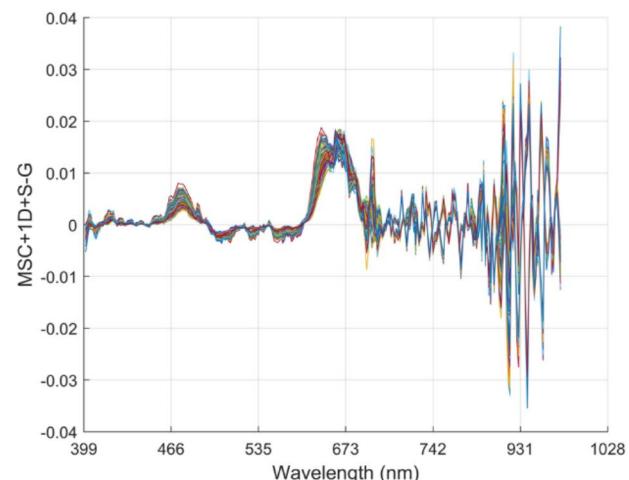
Treatment	0 kg/667 m <sup>2</sup>	6 kg/667 m <sup>2</sup>	12 kg/667 m <sup>2</sup>
S52	1.58 ± 0.23b	1.8 ± 0.18b	2.04 ± 0.12b
S53	1.63 ± 0.23a	1.84 ± 0.18b	2.32 ± 0.12a
S54	1.55 ± 0.23b	1.78 ± 0.19c	2.07 ± 0.11b
S55	1.62 ± 0.22a	1.9 ± 0.19a	2.09 ± 0.12b
S56	1.48 ± 0.22c	1.73 ± 0.19c	2.28 ± 0.12a
S57	1.67 ± 0.22a	1.66 ± 0.19d	2.21 ± 0.11a
S58	1.62 ± 0.22a	1.88 ± 0.18b	2.14 ± 0.11b
S59	1.55 ± 0.22b	1.79 ± 0.18c	1.97 ± 0.11c
S60	1.65 ± 0.21a	1.84 ± 0.18b	2.11 ± 0.12b
S61	1.5 ± 0.21b	1.75 ± 0.18c	2.17 ± 0.13b
S62	1.6 ± 0.2a	1.83 ± 0.18b	1.96 ± 0.13c
S63	1.63 ± 0.2a	1.71 ± 0.17c	2.14 ± 0.14b
S64	1.58 ± 0.2b	1.75 ± 0.16c	1.87 ± 0.15d
S65	1.63 ± 0.19a	1.82 ± 0.14b	2.17 ± 0b
S66	1.63 ± 0.19a	1.82 ± 0.14b	2.17 ± 0b

**Table 2.** Effects of different nitrogen application rates on nicotine in cigar tobacco leaves. LN low nitrogen, MN medium nitrogen, HN high nitrogen.

(A)



(B)

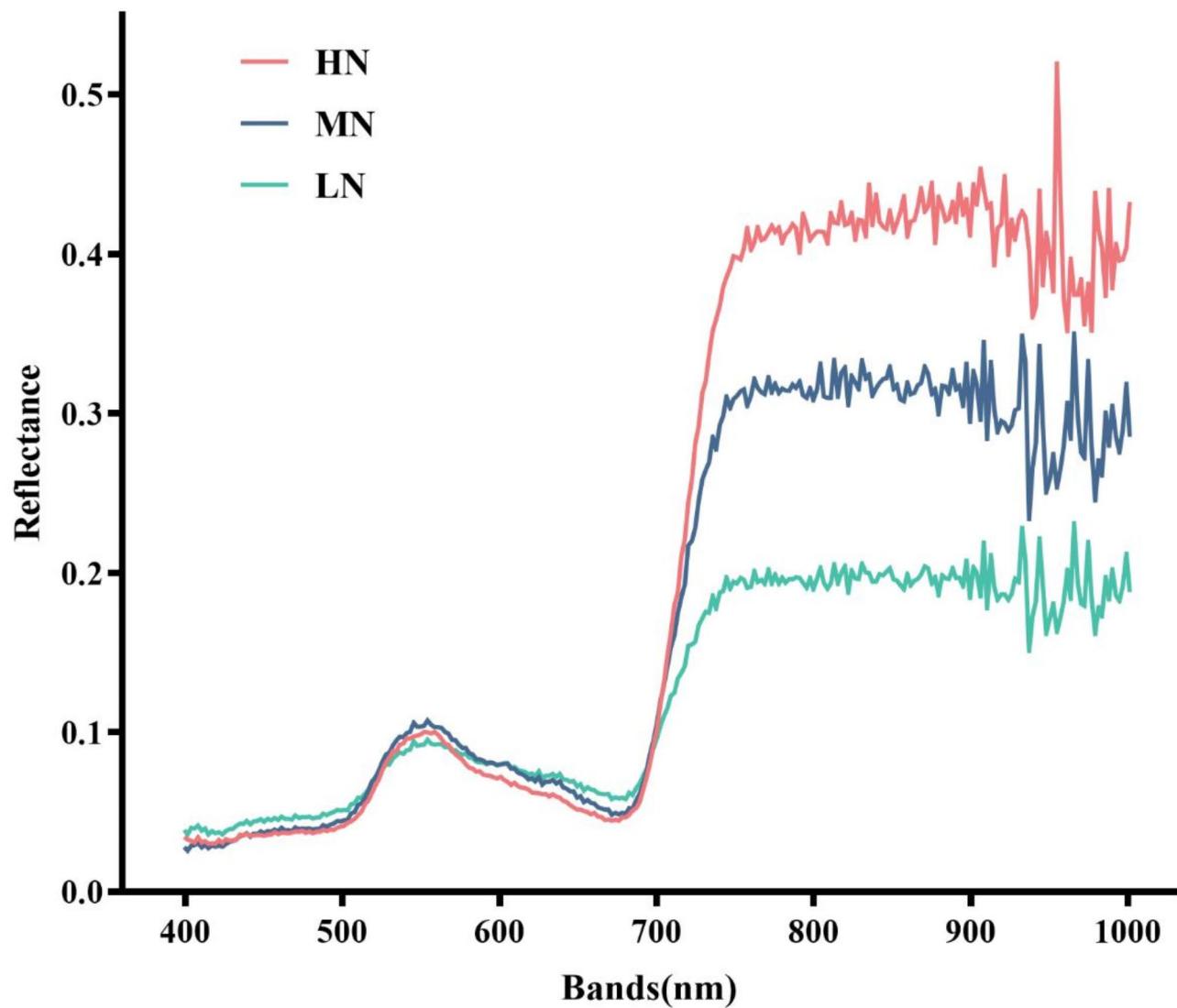


**Fig. 2.** Spectral data. (A) Original spectrograms; (B) Spectrogram of pre-treatment after processing original spectra through MSC + SG + SNV protocols.

spectral reflectance was greatest in the wavelength range between 700 and 1000 nm, and the net change was greatest in the range between 400 and 700 nm. The “green peak” is the locus of strong chlorophyll reflectivity, which peaks at 550 nm, and again within the near-infrared wavelength band between 700 and 1000 nm. These bands distinguish the spectral reflectance of tobacco regardless of the amount of added nitrogen fertilizer, but can be used to identify and quantify nitrogen compounds in samples that have received nitrogen fertilizer. By comparison, under different nitrogen concentration levels, the spectral band reflectances of tobacco plant showed similar characteristics to other green plants, and also showed similar variation patterns. However, there were multiple perceived scattering, baseline deviation, and wave effects in the original spectra that are resolved by multivariate scatter correction, standard normal distribution adjustment, and convolution smoothing. By using the combination MSC + SNV + SG algorithms, the variability errors in the range between 670 and 1000 nm were significantly reduced, and the number of resolved absorption peaks were significantly augmented.

#### Screening of sensitive bands for nicotine content in cigar leaves

With the development of machine learning technology, many traditional modeling methods of full-spectrum spectral data sampling have been unable to meet today's requirements for accuracy and reliability. This is due to their redundancy, overwhelming number of variable adjustments, and large modeling resources. To solve these issues, machine learning algorithms were used to extract data from large-scale spectral files, resulting in improved accuracy of the model. By applying the three competitive de-noising protocols, comprising adaptive re-

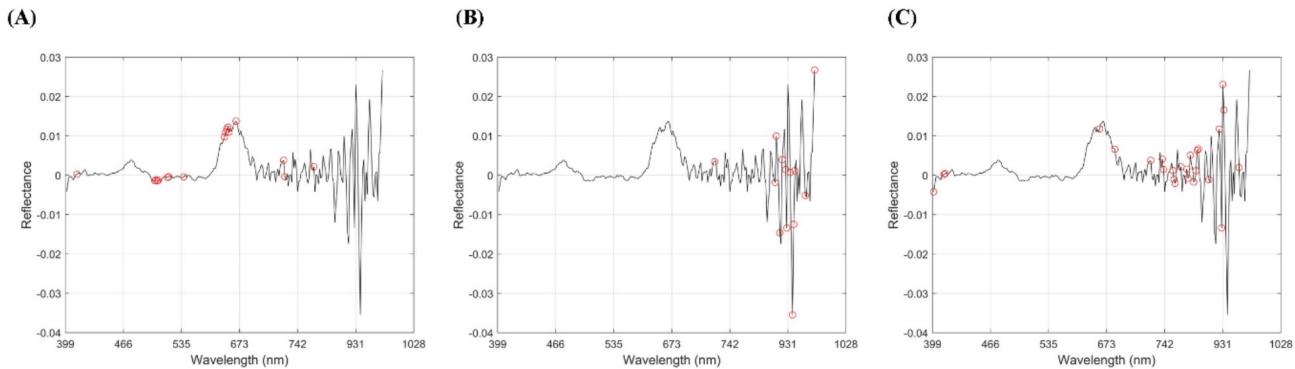


**Fig. 3.** Spectra of cigar leaves treated with different amounts of nitrogen. *LN* low nitrogen, *MN* medium nitrogen, *HN* high nitrogen.

weighting algorithms, continuous projection methods, and data-free factor elimination algorithms, we extracted major feature wavelengths from hyperspectral spectral data sets<sup>29</sup>. The CARS algorithm, SPA algorithm and UVE algorithm are used to screen the cigar nicotine characteristic bands of the preprocessed hyperspectral bands and the results are shown in Fig. 4 and Table 3.

CARS selects the wavelength points with larger absolute values of regression coefficients in the model by adaptive reweighted sampling technique. The wavelength points with small weights are removed, and the subset with the lowest *RMSECV* is selected through interactive verification to effectively identify the optimal variable combination. *RMSECV* shows a trend of first slowly decreasing and then increasing. CARS integrates cross-validation (CV) with PLSR regression model coefficients to screen for important feature variables. It gradually eliminates less important variable information to capture key information variables, and PLSR and CV are used to select the best combination of reflected bands from the full spectrum<sup>30</sup>. The biggest advantage of the spa method is that the collinearity between wavelength variables is eliminated, which ensures the contribution of the screening variable to the estimated variable<sup>31</sup>. However, in the process of screening variables using this method, some valid information in the spectral information may be eliminated, or some wavelength variables with low correlation with the estimated variables may be introduced<sup>32</sup>. UVE builds a series of PLS models by leave-one-out cross validation and calculates the stability of each variable. Uninformative variables are removed by adding random variables with smaller values to the data as thresholds<sup>33</sup>.

According to Fig. 4A, the characteristic bands of cigar nicotine extracted by using UVE algorithm are mainly in the visible wavelength range of green, yellow and orange light from 500 to 673 nm, mainly located in the peak and valley of the wave, and concentrated around the 500 nm and 673 nm bands. According to Fig. 4B, the characteristic bands of cigar nicotine extracted by using SPA algorithm are mainly located in the wavelength range of near-infrared invisible light from 720 to 1000 nm, mainly located at the peak and trough of the wave,



**Fig. 4.** Sensitive band selection of pre-treatment spectrum based on CARS, SPA, and UVE. **(A)** UVE; **(B)** SPA; **(C)** CARS.

Methods	Numbers	Sensitive bands
UVE	17	419 nm, 567 nm, 569 nm, 571 nm, 574 nm, 591 nm, 594 nm, 622 nm, 700 nm, 702 nm, 704 nm, 706 nm, 709 nm, 722 nm, 813 nm, 815 nm, 870 nm
SPA	13	722 nm, 811 nm, 926 nm, 928 nm, 935 nm, 939 nm, 946 nm, 948 nm, 955 nm, 959 nm, 961 nm, 964 nm, 985 nm
CARS	25	399 nm, 419 nm, 421 nm, 716 nm, 745 nm, 813 nm, 835 nm, 838 nm, 855 nm, 857 nm, 860 nm, 871 nm, 882 nm, 884 nm, 888 nm, 895 nm, 899 nm, 901 nm, 903 nm, 923 nm, 943 nm, 948 nm, 950 nm, 952 nm, 979 nm

**Table 3.** Number and distribution of sensitive bands identified by different sensitive band filtering methods.

and concentrated near the band from 900nm to 1000 nm. According to Fig. 4C, The characteristic bands of cigar nicotine extracted by CARS algorithm are mainly located in the wavelength range of 720–1000 nm near infrared invisible light, a small amount is located in the visible light range, and most of them are located in the peak and trough of the wave, and are concentrated near the 900–1000 nm band.

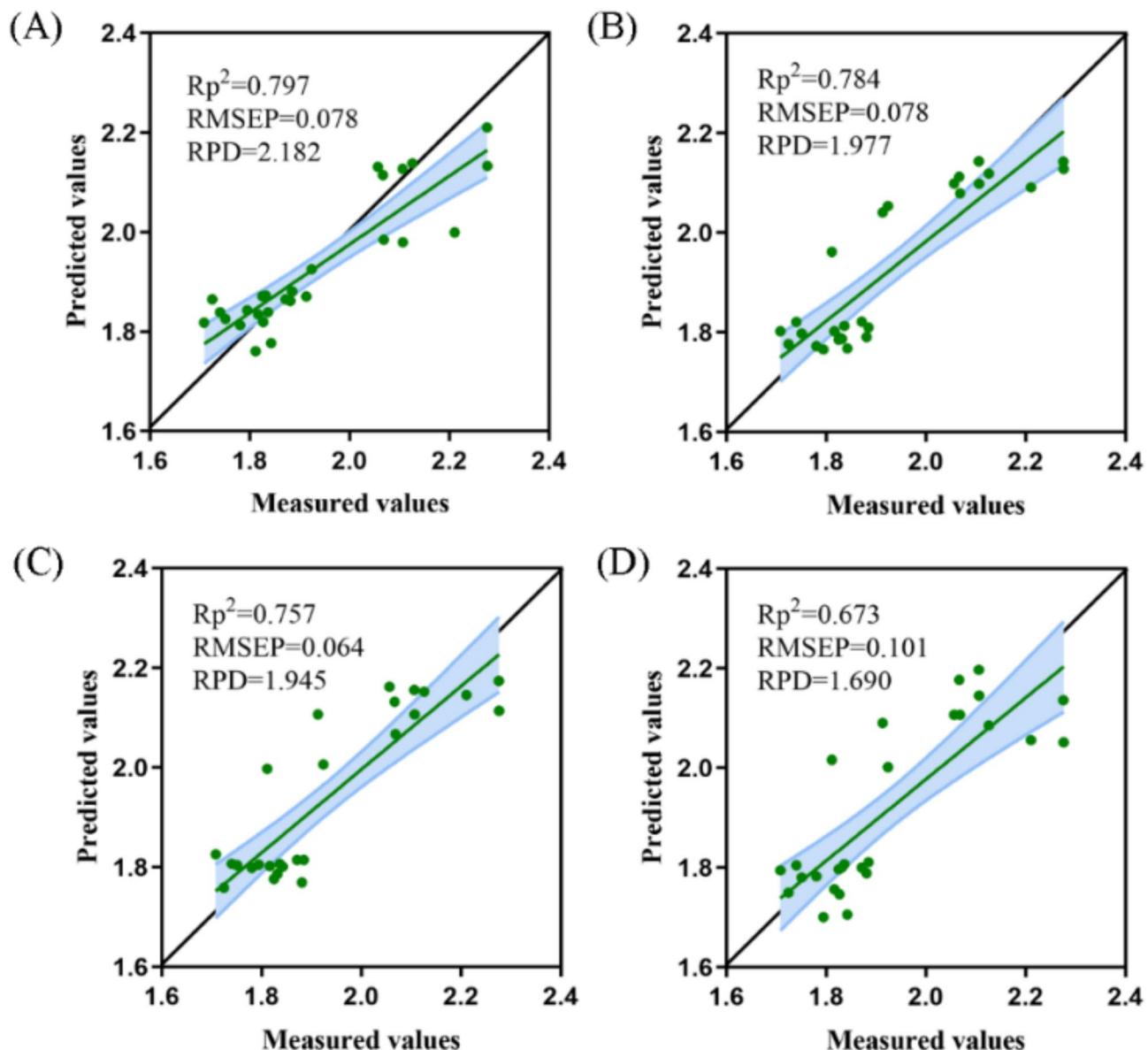
Compared with the original spectral image, in the non-visible band, the band change trend of the pre-processed spectral image is larger than that of the visible band, and the spectral reflectance is increased. This indicates that the band carries a large amount of physiological and biochemical parameter information of tobacco plants. It may be that the alkaloid characteristic absorption peak of tobacco plants is mainly concentrated in the near-infrared band due to the influence of physiological characteristics such as leaf structure and canopy structure. According to the comparison of the results, the cigar tobacco nicotinic sensitive bands screened by the SPA and CARS algorithm are mostly located in the near infrared band, and the cigar tobacco nicotinic sensitive bands screened by the UVE algorithm are mostly located in the visible band. The nicotine sensitive characteristic bands of cigar leaves screened by CARS and SPA are more in the invisible range than UVE spectra. All required variables are searched over a large range of possible correlation values and interference from multiple wavelengths is suppressed. The overlap of data is reduced and the complexity of the model is simplified. On the premise of ensuring data quality, the redundant information is reduced as much as possible. The resonance of multi-wavelength variables is minimized, reducing the complexity of the model.

#### Establishment and verification of nicotine retrieval model

Four algorithms including BP, PLSR, RFR and SVR were used to carry out the inversion modeling of nicotine content of cigar leaves. The results are as follows.

The optimized modeling results in Fig. 5 demonstrate the performance of three different band screening methods based on the MSC-SNV-SG spectrum preprocessing technique. It is evident from the figure that the BP neural network modeling yields the best results for cigar-tobacco nicotine sensitive bands screened by the CARS algorithm, with a test set  $R^2$  of 0.797 and a test set RMSE of 0.078. The relative prediction deviation was 2.182. Additionally, the random forest regression modeling for cigar tobacco nicotine sensitive bands screened by CARS algorithm also performed well, achieving a test set  $R^2$  of 0.784, a test set RMSE of 0.078, and a relative prediction deviation of 1.977. Furthermore, support vector machine regression modeling produced excellent results for cigar-nicotine sensitive bands screened by CARS algorithm, with a test set  $R^2$  of 0.757, a test set RMSE of 0.064, and a relative prediction deviation of 1.945.

It is evident from Table 4 that the cigar-tobacco sensitive band screening method, based on the CARS algorithm, demonstrates superior modeling effectiveness. The nonlinear models' test sets have achieved  $R^2$  values of 0.7–0.8, and through five-fold cross validation, they have attained the minimum RMSECV and maximum relative prediction deviation. In comparison with SPA algorithm and UVE algorithm, the linear model partial least squares method based on UVE algorithm selected cigar tobacco sensitive band screening exhibits the best modeling effect. The test set  $R^2$  is 0.673, and according to the results of five-fold cross validation, RMSECV is 0.08 with a relative prediction deviation of 1.69. As indicated in Table IV, it can be inferred that the nonlinear model's



**Fig. 5.** Results of training set and testing set of cigar nicotine modeling. (A) CARS-BP; (B) CARS-SVM; (C) CARS-RF; (D) UVE-PLSR.

modeling effect surpasses that of the linear model due to potential nonlinearity in the correspondence between cigar tobacco nicotine and its corresponding sensitive spectral bands.

According to the results of the picture and table, comprehensive comparison of test set  $R^2$  and  $RMSE$  and other indicators of each model shows that the modeling effect of most nonlinear regression selected by the same characteristic band algorithm is better than that of linear regression. The linear model maintains a certain stability of modeling results and is not prone to poor modeling results. Compared with different feature band screening algorithms, the nonlinear regression model established by the feature band screened by the CARS algorithm has the best effect, and the prediction accuracy of the three established nonlinear regression models is the best. (Comprehensive modeling accuracy: CARS-BP > UVE-BP > SPA-BP, CARS-RF > SPA-RF > UVE-RF, CARS-SVM > SPA-SVM > UVE-SVM, UVE-PLSR > CARS-PLSR > SPA-PLSR). The partial least squares regression model established by the feature bands screened by UVE algorithm is better than other algorithms.

Among them, the  $RPD$  of the nonlinear model constructed by the features screened by the CARS algorithm is the highest among all the models, indicating that its actual value is the best consistent with the predicted value. The  $RPD$  value of the CARS-BP model is greater than 2, which has the best prediction ability. The characteristic bands extracted by CARS algorithm and SPA algorithm are used as the independent variables of nonlinear model, and the modeling results are better than the characteristic bands extracted by uve algorithm as the independent variables of nonlinear model. In this experiment, the prediction accuracy of the nonlinear model established by the feature bands extracted by CARS algorithm and SPA algorithm is better than that of the nonlinear model

Target	Characteristic band selection	Model	Training sets			Test sets		
			Rc <sup>2</sup>	RMSEC	RMSECV	Rp <sup>2</sup>	RMSEP	RPD
Nicotine	UVE	PLSR	0.749	0.095	0.080	0.673	0.101	1.690
		BP	0.762	0.081	0.079	0.744	0.088	1.930
		RF	0.882	0.086	0.067	0.681	0.097	1.648
		SVM	0.766	0.093	0.102	0.664	0.090	1.705
	CARS	PLSR	0.876	0.068	0.067	0.643	0.127	1.673
		BP	0.721	0.044	0.063	0.797	0.078	2.182
		RF	0.909	0.059	0.059	0.757	0.064	1.945
		SVM	0.797	0.087	0.087	0.784	0.078	1.977
	SPA	PLSR	0.723	0.090	0.100	0.671	0.097	1.636
		BP	0.767	0.094	0.088	0.546	0.117	1.459
		RF	0.888	0.065	0.057	0.749	0.085	1.895
		SVM	0.758	0.095	0.095	0.733	0.098	1.562

**Table 4.** Prediction modeling results of nicotine content in cigars.

established by the feature bands extracted by UVE algorithm. However, the prediction accuracy of the linear model established by the feature bands extracted by UVE algorithm is better than that of the linear model established by the feature bands extracted by CARS algorithm and SPA algorithm. It may be related to the band position of the feature band screening.

## Discussion

### Effect of combined spectral pretreatment method based on MSC-SNV-SG

Compared with the original spectra, the original spectral curves of cigar leaves receiving different nitrogen fertilizer applications exhibited similar green vegetation reflectances and showed similar trends in wavelength variability. They also showed noticeable ambient scattering noise, baseline drift, and fluctuation noise. After introducing multivariate scattering corrections, standard normalization corrections, and convolution smoothing, the coincidence uniformity of the spectra improved, and the influence of scattering noise diminished. The fluctuation error of the curves across the wavelength range of 670–1000 nm was also moderated. The combined pre-screening protocol of MSC-SNV-SG reduces baseline drift and increases the number of absorption peaks in the curves.

### Analysis of characteristic band distribution

According to this research results, most of the feature bands are concentrated in the near infrared band for the following reasons. the mean reflectance spectra between tobacco varieties were distinct, which makes it possible to estimate tobacco attributes by hyperspectral remote sensing. The difference is more pronounced in the Vis and NIR regions. In the former (Vis), chlorophyll and carotenoids, such as chlorophyll and carotenoids, the higher the level of these pigments<sup>34</sup>, the more intense the change. In the NIR region, higher reflectance values and significant differences between varieties are related to radiation scattering within the leaf mesophyll<sup>35</sup>, because plants with more developed mesophyll have more developed porous spaces and higher reflectance in the NIR region. In addition, naturally, tobacco varieties, which are quite distinct regarding the structure and thickness of their mesophyll, as well as biochemical properties of their leaves, have different reflectance spectra, mainly in the NIR region, which is highly influenced by such characteristics<sup>36</sup>.

### Inversion effect analysis of cigar nicotine modeling

To successfully apply machine learning to reliably predict cigar tobacco nicotine content, it is crucial to mine data to gather effective spectral variable characteristics into sets of parameters that uniquely represent the presence of nicotine. Machine learning approaches set, such as the number of hidden layers, the number of neurons in the hidden layer, the type of activation function, the training algorithm, the initial weight range, and the learning rate. In this paper, the BP neural network provided the best modeling outcome, which provide insight into the methods for assessing cigar nicotine content using hyperspectral data, but they also promote the translation of research results for practical applications of assessing other crop metabolites. Feature mining was achieved by feature (band) selection and feature extraction. Feature extraction uses subspace transformation to map the original high-dimensional data into a lower-dimensional feature space<sup>37</sup>. The extracted features often do not accord with the physical meaning in the original data. Therefore little progress is made by combining feature extraction methods and machine learning algorithms to evaluate nicotine status in cigar leaves. In this study, the CARS algorithm, SPA algorithm and UVE algorithm were used to extract spectral feature parameters. Among these algorithms, the CARS algorithm produces the most effective extraction of related hyperspectral feature variables to evaluate cigar tobacco nicotine. It also is the basis for identifying the parameters that provide the least noise for inverse modeling compared with the other two algorithms.

The MSC-SNV-SG-CARS-BP model has the best prediction accuracy in terms of the content of cigar nicotine. The prediction effects of the models (in terms of evincing least error) were ranked BP > SVM > RF > PLSR. ANNs have more parameters to reflects the capacity for nonlinear mapping and encompasses broad generalization

capability. The BP neural network essentially produces a strong mapping function from input to output, so it is able to approximate any nonlinear, continuous function with any accuracy. This makes it especially suitable for solving problems with complex internal mechanisms, that is, BP neural network has a strong non-linear mapping ability<sup>38</sup>.

The RF algorithm has high operating efficiency and good anti-noise stability, and the power to deal with high-dimensional data without need of selecting features. RF has a strong generalization ability, fast training speed, and superior parallelization of calculations that overcome the overfitting problem<sup>39</sup>. In the present study, the prediction of nicotine content relied on the collection of spectral feature variables that were affected by multiple uncertain factors, themselves highly random and non-linear. For the spectral feature variables falling under different bands, the attributes of these bands have a greater impact on the random forest than other inversion methods, and thus affect the modeling accuracy.

The SVM model effectively overcomes the shortcomings of slow convergence speed and poor prediction ability of the neural network, and has unique advantages for prediction models with small sample size. The SVM model transforms the linear problem into a high-dimensional space through nonlinear changes and solves it linearly in the space, realizing the regression prediction model<sup>40</sup>. In this paper, the modeling accuracy of the SVM model is lower than that of the BP neural network, which realizes the efficient “transductive inference” by transitioning between training samples to prediction samples. By avoiding the traditional process of transitioning from induction to deduction methodology, it greatly simplifies the regression prediction process.

Dou et al.<sup>41</sup> evaluated the relationship between 11 spectral parameters and nicotine content in leaves. Statistical analysis shows that the combination of power function and  $(SDR-SDy)/(SDr+SDy)$  has the best effect when  $R^2=0.8112$ ,  $RMSE=0.2272$ , and relative error ( $RE$ ) = 14.42%. Divyant et al.<sup>42</sup> applied hyperspectroscopy and four machine learning algorithms to predict nicotine content in tobacco. The results showed that PLSR ( $R^2=0.93$ ,  $RMSE=0.21\%$ ) was superior to SVR ( $R^2=0.89$ ,  $RMSE=0.36\%$ ), RF ( $R^2=0.90$ ,  $RMSE=0.35\%$ ) and PLSR-VIP ( $R^2=0.91$ ,  $RMSE=0.30\%$ ). Coefficient of determination and other model evaluation indexes of this study is lower than that of previous studies. Because this experiment is a field experiment, the spectral data obtained by using a UAV equipped with hyperspectral sensors have more noise dimensions and other disturbances than that obtained by indoor spectrometers, and the resolution of hyperspectral images obtained is also different. Typically the upper leaves of the cigar plant are smaller and thicker and contain more nicotine and other alkaloids, while the middle and lower leaves are larger and thinner and contain less alkaloids<sup>43</sup>. As a result, choosing one leaf is hardly representative of the overall variability of sugars, alkaloids, and yield. In addition, the reflectance spectra of plant leaves not only reflect the physiological and biochemical parameters inside the leaves, but also are affected by leaf morphology (such as leaf inclination and leaf surface characteristics). Some researchers believe that the mirror reflectance of multi-tilt blades is a factor that reduces the performance of the LCC estimation model. To solve this problem, the researchers used indicators such as the reflectance difference ratio (MDATT)<sup>44</sup> and the improved MDATT index (IMDATT)<sup>45</sup>. The LCC linear regression model based on MDATT and IMDATT is not affected by the observation Angle and leaf surface properties. In the next step, modeling methods such as reflectance difference ratio index will be introduced to further reduce error interference and enhance the accuracy of modeling results.

## Future research prospects

The ripening time of tobacco leaves in different parts is different (generally 20 days apart). The order of harvesting was bottom, middle and top. Some observation tools, such as drones, can only obtain canopy reflectance data. Therefore, it is easy to see that the canopy spectrum does not fully represent the bottom and middle. In practical studies, this problem may cause the results predicted by the model to differ greatly from the actual values. In addition, the growth conditions of tobacco seedlings from transplanting to harvesting may also affect the final quality. However, many studies collect plant samples from one stage (e.g., regreening, root elongation, flourishing, or maturation). Therefore, whether the canopy spectral data of one stage can predict the final tobacco leaf quality is also a problem that needs to be verified. Therefore, multiple tobacco growth period spectral images should be selected to improve the input independent variables of the model and obtain more tobacco spectral information, so as to determine the best period of nicotine content inversion modeling in cigar tobacco leaves.

External factors will affect image quality during data acquisition, such as measurement time, light intensity, sun altitude Angle, etc. Due to the limitation of endurance, the drone must complete the task in a limited time (about 30 min). To ensure image quality, the drone should be kept at the right height (about 50–100 m). The obtained hyperspectral image contains the spectral information of all the ground objects in the lens, such as crops, soil, roads, weeds, etc., which may cause the noise of target resolution. How to balance image quality with flight altitude, time, spatial resolution and coverage area still needs further research. In order to estimate crop canopy variables, confounding factors affecting the estimation must be considered, such as leaf or canopy structure and understory vegetation during multiple scattering, soil parameters and some external parameters<sup>46</sup>.

Studies on high-throughput phenotypic inversion modeling of tobacco plants are scattered, mainly focusing on one agronomic parameter, and establishing inversion models based on the corresponding hyperspectral reflectance data. As for the deeper mechanism of action, few studies have explored. Therefore, the relationship between various biophysical indexes is an important basis for spectral inversion. How to use these relationships to monitor tobacco agronomic parameters that are not sensitive to spectral response is also worth investigating.

Future multi-sensor fusion technologies need to continue to feature fusion and decision fusion, and continue to explore the fusion of multiple types of sensors (such as spectral and lidar fusion, and chlorophyll fluorescence and depth information fusion), so as to provide technical and information support for determining more accurate plant breeding strategies. Image quality and spectral data quality are greatly affected by light conditions. Although the current research focus is on deep learning, machine learning retains its advantages in this application. Deep learning, or deep neural networks, is a data-driven approach that requires large amounts of

training data to adequately estimate even millions of parameters. This is due to the lack of large publicly available nicotine content datasets for comprehensive parametric training of deep networks. Therefore, for small-scale (field) crop parameter estimation, mainstream machine learning methods and shallow neural networks provide the required performance. In addition, machine learning methods require less hardware and are more suitable for practical applications.

## Conclusion

In this study, a hyperspectral camera was delivered to tobacco crop locations by a UAV and used to perform remote hyperspectral sensing. It transmits remote reflectance data files from field tobacco crops to a central computer facility where various signal processing methods were used to refine the original data. The computer executes artificial intelligence algorithms that run various inversion models to evaluate crop nicotine content. Fifteen types of cigar leaves were monitored with the aim to establish an inversion model with wider universality and greater ability to interpret nicotine content. The results show that the CARS-BP model based on MSC-SNV-SG pre-processing is the most representative model among the nonlinear models, and has better modeling accuracy, root mean square error (RMSE) and RPD than any other tested models. The MSC-SNV-SG-CARS-BP model has the best prediction accuracy in terms of the content of cigar nicotine.

The process from data collection to model development requires careful consideration and analysis. It is important to improve the quality of data obtained from UAV-based imagery, ground-based observations, and modeling approaches, which may be subject to errors. Therefore, a standardization procedure for the above process is needed. Future work can be investigated using different image collection methods, data processing algorithms, and modeling approaches to estimate biomass; In addition, the biophysical properties of plants and error sources can be further explored. In addition, the uncertainty and transferability of development models and application tests need to be evaluated. While the study achieved remarkable success in modeling cigar nicotine inversion using four different machine learning models, it must be acknowledged that the performance of machine learning-based methods such as BP neural networks can be affected by the specific characteristics of the target plant species and the environmental conditions in which they are grown. Different plant species may have different shapes, sizes, and growth patterns, which may affect the accuracy of object detection. The accuracy and robustness of cigar tobacco nicotine value estimation can be affected by factors such as plant structure, leaf thickness, plant cover, and soil environment, which vary by growth stage. In addition, variations in environmental factors such as light, soil type, and plant density may pose a challenge to the general applicability of the model<sup>47</sup>. Therefore, further research is needed to assess the adaptability of the method to a wider range of cigar species and diverse environmental Settings.

## Data availability

Raw spectral data, cigar nicotine content data are available upon request. Please contact the corresponding author at 13258326593@163.com.

Received: 5 April 2024; Accepted: 23 January 2025

Published online: 31 January 2025

## References

1. Tan, C., Wang, J. & Wu, T. Determination of nicotine in tobacco samples by near-infrared spectroscopy and boosting partial least squares. *Vib. Spectrosc.* **54**, 35–41. <https://doi.org/10.1016/j.vibspec.2010.06.002> (2010).
2. Wu, C., Chen, P., Chen, J., Vembu, S. & Ho, K. Detection of nicotine based on molecularly imprinted TiO<sub>2</sub>-modified electrodes. *Anal. Chim. Acta* **633**, 119–126. <https://doi.org/10.1016/j.aca.2008.11.038> (2009).
3. Frederico, L. F. S., Marcelo, C. A., Liliane, M. F., Oscar, F. S. & Samuel, K. Inline simultaneous quantitation of tobacco chemical composition by infrared hyperspectral image associated with chemometrics. *Microchem. J.* **151**, 104225. <https://doi.org/10.1016/j.microc.2019.104225> (2019).
4. Brooke, B., Bettina, B. & Megan, L. Approaches, applications, and future directions for hyperspectral vegetation studies: an emphasis on yield-limiting factors in wheat [J]. *Plant. Phenome J.* **3** (1), 2598–2620. <https://doi.org/10.1016/j.rse.2020.112121> (2020).
5. Zhang, M. et al. UAV-borne hyperspectral estimation of nitrogen content in tobacco leaves based on ensemble learning methods. *Comput. Electron. Agric.* **211**, 108008. <https://doi.org/10.1016/j.compag.2023.108008> (2023).
6. Zhang, M. et al. Hyperspectral remote sensing for tobacco quality estimation, yield prediction, and stress detection: a review of applications and methods. *Front. Plant Sci.* **14**, 1073346. <https://doi.org/10.3389/fpls.2023.1073346> (2023).
7. Chion, C., Landry, J. A. & DaCosta, L. A genetic-programming-based method for hyperspectral data information extraction: agricultural applications. *IEEE Trans. Geosci. Remote Sens.* **46**, 2446–2457. <https://doi.org/10.1109/TGRS.2008.922061> (2008).
8. Tan, K., Wang, S., Song, Y., Liu, Y. & Gong, Z. Estimating nitrogen status of rice canopy using hyperspectral reflectance combined with BPSO-SVR in cold region. *Chemometr. Intell. Lab. Syst.* **172**, 68–79. <https://doi.org/10.1016/j.chemolab.2017.11.014> (2018).
9. Chen, P. F. et al. Critical nitrogen curve and remote detection of nitrogen nutrition index for corn in the northwestern plain of Shandong Province, China. *IEEE J. Sel. Top. Appl. Earth Obs Remote Sens.* **6**, 682–689. <https://doi.org/10.1109/JSTARS.2012.2236302> (2013).
10. Yi, Q., Huang, J., Wang, F. & Wang, X. Evaluating the performance of PC-ANN for the estimation of rice nitrogen concentration from canopy hyperspectral reflectance. *Int. J. Remote Sens.* **31**, 931–940. <https://doi.org/10.1080/01431160902912061> (2010).
11. Qin et al. Remote Sensing Inversion of Tobacco SPAD Based on UAV Hyperspectral Imagery. In *IGARSS 2023–2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA*, pp. 3478–3481. <https://doi.org/10.1109/IGARSS52108.2023.10282002> (2023).
12. Zhang, M. et al. UAV-borne hyperspectral estimation of nitrogen content in tobacco leaves based on ensemble learning methods, computers and electronics in Agriculture. **211**, 108008, ISSN 0168–1699 (2023).
13. Jun Ying, L. et al. Method for predicting potassium oxide content in tobacco leaves based on unmanned aerial vehicle (UAV) hyperspectral field, involves generating prediction data of potassium oxide content corresponding to field tobacco leaves. Patent No CN112697724-A (2020).

14. Jun Ying, L. et al. Method for predicting total sugar content of tobacco leaves in hyperspectral field based on unmanned aerial vehicle, involves loading coordinate of hyperspectral image data and extracting and processing corresponding original spectral curve. (2020). Patent No CN112697725-A.
15. Hayes, A. & Reed, T. D. Hyperspectral reflectance for non-invasive early detection of black shank disease in flue-cured tobacco. *J. Spectr. Imaging.* **10**, 1–10. <https://doi.org/10.1255/jsi.2021.a4> (2021).
16. Wu, X., Xiao, Q., Wen, J., You, D. & Andreas H. advances in quantitative remote sensing product validation: overview and current status. *Earth Sci. Rev.* **196**, 102875. <https://doi.org/10.1016/j.earscirev.2019.102875> (2019).
17. Savitzky, A. & Golay, M. Smoothing and differentiation of data by simplified least squares procedures [J]. *Anal. Chem.*, **36** (8), 1627–1639. <https://doi.org/10.1021/ac60214a047> (1964).
18. Xu, Z. et al. Comparative study of different wavelength selection methods in the transfer of crop kernel qualitative near-infrared models. *Infrared Phys. Technol.* **123**, 104120. <https://doi.org/10.1016/j.infrared.2022.104120> (2022).
19. Skiera, B., Reiner, J. & Albers, S. Regression analysis. In: (eds Homburg, C., Klarmann, M. & Vomberg, A.) *Handbook of Market Research*. Springer, Cham. [https://doi.org/10.1007/978-3-319-57413-4\\_17](https://doi.org/10.1007/978-3-319-57413-4_17) (2022).
20. Mário, C. U. A. et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis [J]. *57*(2), 65–73. [https://doi.org/10.1016/S0169-7439\(01\)00119-8](https://doi.org/10.1016/S0169-7439(01)00119-8) (2001).
21. Bartholomeus, H. M. et al. Spectral reflectance based indices for soil organic carbon quantification[J]. *Geoderma* **145** (1–2), 28–36. <https://doi.org/10.1016/j.geoderma.2008.01.010> (2008).
22. Guebel, D. V. & Torres, N. V. Partial least-squares regression (PLSR). In: (eds Dubitzky, W., Wolkenhauer, O., Cho, K. H. & Yokota, H.) *Encyclopedia of Systems Biology*. Springer, New York, NY. [https://doi.org/10.1007/978-1-4419-9863-7\\_1274](https://doi.org/10.1007/978-1-4419-9863-7_1274) (2013).
23. Mas, J. F. & Flores, J. J. The application of artificial neural networks to the analysis of remotely sensed data. *Int. J. Remote Sens.* **29**, 617–663. <https://doi.org/10.1080/0143160701352154> (2008).
24. Giorgos, M., Im, J. & Ogole, C. Support vector machines in remote sensing: a review. *ISPRS J. Photogramm Remote Sens.* **66**, 247–259. <https://doi.org/10.1016/j.isprsjprs.2010.11.001> (2011).
25. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
26. Katja, B. et al. Retrieval of aboveground crop nitrogen content with a hybrid machine learning method,International. *J. Appl. Earth Observation Geoinf.* **92**, 102174. <https://doi.org/10.1016/j.jag.2020.102174> (2020).
27. Huang, J. et al. Estimation of Overstory and Understory Leaf Area Index by Combining Hyperion and Panchromatic QuickBird Data Using Neural Network Method. *Sens. Lett.* **9**, 964–973. <https://doi.org/10.1166/sl.2011.1380> (2011).
28. Jozsef, P., Csaba, M., Miklos, I., Norbert, K. & Amir, M. Modeling Organizational Performance with Machine Learning. *J. Open. Innovation: Technol. Market Complex.* **8**, 177. <https://doi.org/10.3390/joitmc8040177> (2022).
29. Wan, L. et al. Grain yield prediction of rice using multi-temporal UAV-based RGB and multispectral images and model transfer-a case study of small farmlands in the South of China. *Agric. For. Meteorol.* **291**, 108096. <https://doi.org/10.1016/j.agrformet.2020.108096> (2020).
30. Adão, T. et al. Hyperspectral imaging: a review on UAV-Based sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sens.* **9**, 1110. <https://doi.org/10.3390/rs9111110> (2017).
31. Rogers, M. et al. Wavelength and texture feature selection for hyperspectral imaging: a systematic literature review. *Food Measure.* **17**, 6039–6064. <https://doi.org/10.1007/s11694-023-02044-x> (2023).
32. Zhang, R., Benoit & Rogge, D. The successive projection algorithm (SPA), an algorithm with a spatial constraint for the automatic search of endmembers in hyperspectral data. *Sensors* **8** (2), 1321–1342 (2008).
33. Li Y. K. et al. Variable selection methods in spectral data analysis [J]. *Spectrosc. Spectral Anal.* **41**(11), 3331–3338 (2021)
34. Liu, J. H. X. I. Zhang,nondestructive detection of rape leaf chlorophyll level based on Vis-NIR spectroscopy,[J].*Spectrochim Acta Part A Mol. Biomol. Spectrosc.* **222**, 117202. <https://doi.org/10.1016/j.saa.2019.117202> (2019).
35. de Rodrigues, M. R. R.B.de Oliveira,Vis-NIR spectroscopy:from leafdry mass production estimate to the prediction of macro-and micronutrients in soybean crops.[J]. *Appl. Remote Sens.* **14**, 44505 (2020).
36. Tianjun, Z. J. Y. P. L. W. Y. Y. H. Ying,L.Changyu, Rapid and Automatic Classification of Tobacco leaves using a hand-held DLP based NIR Spectroscopy device. *J. Braz Chem. Soc.* <https://doi.org/10.21577/0103-5053.20190105> (2019).
37. Jia, X., Kuo, B. C. & Crawford, M. M. Feature mining for hyperspectral image classification. *Proc. IEEE.* **101**, 676–697. <https://doi.org/10.1109/JPROC.2012.2229082> (2013).
38. Ravi, A., Ritesh, K., Mishra & Rijwan, K. Chapter 9 - plant diseases detection using artificial intelligence, editor(s): Mohammad Ayoub Khan, Rijwan Khan, Mohammad Aslam Ansari, Application of Machine Learning in Agriculture, Academic, 173–190, <https://doi.org/10.1016/B978-0-323-90550-3.00007-2> (2022).
39. Bappa, D. et al. Quantitative monitoring of sucrose, reducing sugar and total sugar dynamics for phenotyping of water-deficit stress tolerance in rice through spectroscopy and chemometrics. *Spectrochim Acta A.* **192**, 41–51. <https://doi.org/10.1016/j.saa.2017.10.076> (2018).
40. Cardoso, V. G. K. & Poppi, R. J. Non-invasive identification of commercial green tea blends using NIR spectroscopy and support vector machine. *Microchem J.* **164**, 106052. <https://doi.org/10.1016/j.microc.2021.106052> (2021).
41. Dou, Y. Q., Cheng, S., Li, X. J., Liu, Y. & Yuan, X. L. Estimation of nicotine content in tobacco leaves based on hyperspectral imaging. *Appl. Ecol. Environ. Res.* **15**, 1419–1426. [https://doi.org/10.15666/aeer/1504\\_14191426](https://doi.org/10.15666/aeer/1504_14191426) (2016).
42. Divyanth, L. et al. Non-destructive prediction of nicotine content in tobacco using hyperspectral image-derived spectra and machine learning. *J. Biosyst Eng.* **47**, 106–117. <https://doi.org/10.1007/s42853-022-00134-0> (2022).
43. Zhang, L. et al. Metabolic profiling of tobacco leaves at different growth stages or different stalk positions by gas chromatography–mass spectrometry[J]. *Ind. Crops Prod.* **116**, 46–55. <https://doi.org/10.1016/j.indcrop.2018.02.041> (2018).
44. Li, W., Sun, Z., Lu, S. & Omasa, K. Estimation of the leaf chlorophyll content using multiangular spectral reflectance factor. *Plant. Cell. Environ.* **42** (11), 3152–3165. <https://doi.org/10.1111/pce.13605> (2019).
45. Zhao, Y. et al. Estimation of chlorophyll content in intertidal mangrove leaves with different thicknesses using hyperspectral data[J]. *Ecol. Ind.* **106**, 105511. <https://doi.org/10.1016/j.ecolind.2019.105511> (2019).
46. Zhang, H. et al. High-throughput phenotyping of plant leaf morphological, physiological, and biochemical traits on multiple scales using optical sensing[J].*Crop journal*, **11**(5):1303–1318. (2023). <https://doi.org/10.1016/j.cj.2023.04.014>
47. Lin, H. et al. Automated counting of Tobacco plants using multispectral UAV Data. *Agronomy* **13**, 2861. <https://doi.org/10.3390/agronomy13122861> (2023).

## Author contributions

T wrote the main manuscript text and prepared all the pictures and tables. X and Y assisted in processing the materials, L was responsible for guiding the experimental plan, and all the authors reviewed the manuscript.

## Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

Correspondence and requests for materials should be addressed to L.L.

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025