Research on estimating soil organic matter content in Northeast China based on CARS-IRIV and neural network optimization algorithm

Jianyu Fang[1], Chenyi Xu[1], Juchi Bai[1], Shengfan Zhu[1], Honggang Zhang[1], Fenghua Yu[1,2*]

(1. *College of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang 110866, China*;
2. *Key Laboratory of Smart Agriculture Technology in Liaoning Province, Shenyang 110866, China*)

Abstract: Choosing appropriate variable screening methods and models can effectively improve the estimation accuracy of soil organic matter content. This article takes the Haicheng Experimental Field of Shenyang Agricultural University as an example to perform SG (Savitzky-Golay) smoothing on the original soil reflectance. It uses iterative preservation of information variables (IRIV), competitive adaptive reweighted sampling (CARS), and CARS-IRIV hybrid methods to reduce dimensionality and extract relevant features from raw spectral data. A hyperspectral inversion model for total organic matter in soil was established using backpropagation neural network (BPNN), sparse optimization (SSA-BPNN), and chaotic sparse optimization BP neural network (CSSA-BPNN). Determine the coefficient (R^2) and root mean square error (*RMSE*) to evaluate the inversion model. The results show that: (1) the optimized model algorithm is better than the unoptimized algorithm; (2) The combination of CARS-IRIV dimensionality reduction method is better than CARS and IRIV dimensionality reduction algorithms in both results and efficiency; (3) The CSSA-BPNN inversion model based on CARS-IRIV dimensionality reduction has the best prediction performance, with a final prediction of soil total organic matter content R^2=0.839 and *RMSE*=1.705. The inversion accuracy is higher than that of SSA-BPNN and BPNN models, which can provide reference for soil nutrient evaluation in the region.

Keywords: UAV hyperspectral; Soil organic matter element inversion; CSSA optimization algorithm; CARS-IRIV dimension reduction; BPNN modeling

DOI: 10.33440/j.ijpaa.20230601.198

Citation: Fang J Y, Xu C Y, Bai J C, Zhu S F, Zhang H G and Yu F H. Research on estimating soil organic matter content in Northeast China based on CARS-IRIV and neural network optimization algorithm. Int J Precis Agric Aviat, 2023; 6(1): 51–59.

1 Introduction

The quantitative extraction of soil composition information is conducive to provided essential data support for subsequent crop planting. Soil organic matter has always been one of the critical elements in plant nutrition to increase crop yield and improve crop quality[1]. Soil quality affects the growth of crops[2]. In recent years, with the change in agricultural fertilization mode, nutrient matter in the soil is a potential environmental pollution factor[3]. Traditional methods for soil organic matter detection, such as the potassium dichromate-sulfuric acid digestion method, cost a lot of time and human resources, and material resources[4]. Rapid and accurate acquisition of nutrient content in the topsoil layer is of great significance for guiding agricultural production and soil environmental monitoring[5]. The spectral reflection characteristics of soil are closely related to the physical and chemical properties, and a good inversion model can be established faster and more conveniently to obtain the content of soil elements. It simplifies the work of agricultural production personnel and can provide a better guarantee for future agricultural production[6]. Hyperspectral rice fields based on UAV can be planted and produced on a larger scale, and subsequent fertilization and topdressing measures can be determined by measuring organic matter content before cultivation[7].

Soils with different organic matter contents have different spectral characteristics, which opens up new avenues for rapid determination of soil organic matter content. Research has shown that an increase in organic matter content can cause changes in the spectral curve, where the peak position belongs to the N-H expansion vibration peak in amine and amide organic compounds. These organic compounds exist in each particle level aggregate through adsorption, encapsulation, and molecular binding, promoting the accumulation of organic matter in each particle level aggregate. It can be confirmed that soil organic matter has a certain influence on the formation mechanism of spectral curves[8]. In recent years, numerous researchers have conducted extensive research on the prediction of soil organic matter content using hyperspectral techniques. At present, common models for predicting soil organic matter content include linear models such as multiple stepwise regression and partial minimum multiplication regression, as well as nonlinear models such as random forest and support vector machine. In the spectral feature band selection of soil element content, common feature extraction algorithms include Iterative Preserved Information Variable (IRIV) and Competitive Adaptive Reweighting (CARS). In terms of hyperspectral inversion of soil organic matter, many scholars have also conducted related inversion studies[9]: Huang et al.[10] conducted correlation

Received date: 2023-08-03 **Accepted date:** 2023-12-05

Biographies: **Jianyu Fang**, Postgraduate student, research interests: agricultural remote sensing, Email: 2227978388@qq.com; **Chenyi Xu**, Postgraduate student, research interests: agricultural remote sensing, Email: xtyxcy11@163.com; **Juchi Bai**, PhD, research interests: agricultural remote sensing, Email: 773858525@qq.com; **Shengfan Zhu**, Postgraduate student, research interests: agricultural remote sensing, Email: 1984701474@qq.com; **Honggang Zhang**, Postgraduate student, research interests: artificial intelligence, Email: a961972679@163.com.

*Corresponding author: **Fenghua Yu**, Professor, research interests: precision agriculture aviation technology and equipment, Email: adan@syau.edu.cn.

and multiple stepwise regression analysis on the main nutrient content in soil through different spectral reflectance and composite indices of hyperspectral images. The results indicate that there is a moderate correlation between total organic matter in red soil and the near-infrared and green light bands, which is statistically significant. Liu et al.[11] studied the hyperspectral reflectance of typical soil samples in Heilongjiang Province. They established a hyperspectral inversion model for the organic matter content of black soil using multivariate statistical analysis methods. The inversion model indicates a strong correlation between soil organic matter and nitrogen elements. However, these studies are based on indoor spectroscopy and have a certain degree of lag, which is not conducive to the development of agricultural production. Yang et al.[12] sed drone hyperspectral data to obtain soil spectral information. They compared three analysis methods and found that the correlation coefficient between the BP neural network model and total organic matter in soil was as high as 0.76, indicating a strong correlation. These studies demonstrate the feasibility of soil hyperspectral inversion of organic matter. In specific research, due to the low sensitivity of soil spectra to organic matter bands, common feature extraction algorithms cannot guarantee the accuracy and efficiency of variables. Therefore, it is necessary to make relevant improvements to the stability of feature extraction methods when establishing regression models to invert organic matter content. In addition, numerous research results have shown that the content of soil trace elements and spectral reflectance are influenced by multiple factors, and the relationship is extremely complex. It is difficult to provide a reasonable explanation using linear models such as partial least squares. However, the setting of nonlinear model parameters such as neural networks also has a significant impact on the performance of the model. Therefore, a stable feature selection algorithm and a high-precision prediction model are constructed. The spectral inversion of soil organic matter content is particularly important.

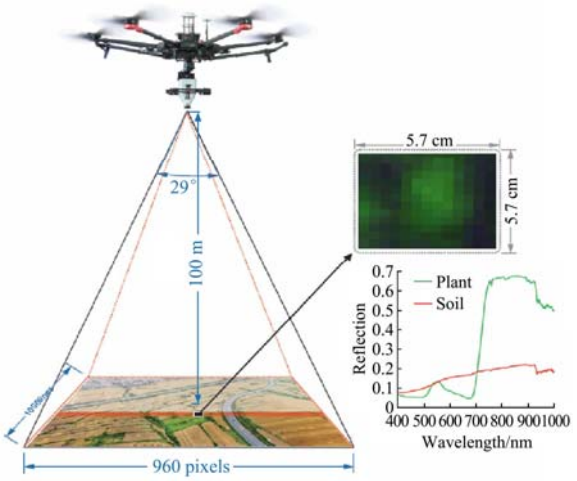
In this paper, we try to perform SG smoothing on soil spectra in the spectral reflection interval of 400-1000 nm for preliminary data processing, and then obtain the spectral characteristics through the combination of iterative retained information variable method (IRIV), competitive adaptive reweighting algorithm (CARS) and CARS-IRIV. Finally, BP neural network, SSA-optimized neural network and CSSA-optimized neural network were used to establish the inversion model of soil organic matter hyperspectral elements, and the model was verified to determine the optimal combination form of spectral transformation and modeling method, which provides a basis for the rapid determination of soil nutrient content in the study area.

2 Research data

2.1 Overview

2.1.1 Overview of the study area

The study area is located in the Haicheng Practice Base of Shenyang Agricultural University, Gengzhuang, Haicheng City, Liaoning Province (122°73'E, 40°97'N), which is one of the national crucial commercial grain bases with rice cultivation as the central part. In this area, the terrain is low and flat, and the land is fertile. The main types of soil are alopecia soil, meadow soil and marsh soil. The total area of the farmland is about 1.73 hm^2, the crop planting in the field is unified and orderly, and the ridges of the field are arranged neatly. The growing season of crops is from May to October each year. Figure 1 shows the farmland image of the collection area. To quickly establish the inversion model of farmland soil organic matter, this study selected three observation sampling points with different regions of each farmland to detect the soil organic matter content in the field. The collection time was before and after planting each year, which can effectively reflect the content of soil elements in different periods.



a. Principles of hyperspectral imaging for drones

b. Hyperspectral drone

Figure 1 Drones and their imaging principles

2.1.2 Rice Soil Properties

Under long-term cultivation conditions, paddy soil undergoes a redox reaction and forms a topsoil layer under the dual effects of artificial water tillage and natural soil formation factors. The texture of the soil plow layer is relatively hard, and it will clump after drying, which has a certain impact on the spectral reflection of the soil. This experiment collected land data after plowing. The plow consists of two 25 centimeter plowshares, each with a width of 25 centimeters and a width of 50 centimeters for the entire plow. After flipping over the ground, the soil shape will be relatively similar, ensuring that the soil properties are unified and not affected by external factors.

2.2 Data Acquisition

2.2.1 UAV image data

In this paper, the UAV hyperspectral imaging remote sensing platform combines Shenzhen DJI Innovation Company's six-rotor flight platform and Sichuan Shuang li He Spectrum Company's miniature built-in push and sweep high airborne spectrometer (Gaiasky-mini). The hyperspectral imager is composed of an ICX285 CCD sensor and optical components. The pixel spacing is 6.45 um, the output is 14bit, and the numerical aperture is F/2.8. Hyperspectral image information was collected simultaneously in the soil collection area. To ensure the quality of hyperspectral data collection, the hyperspectral soil information in this paper is selected between 9:30 and 11:30 on a clear and windless day as far as possible[13]. The hyperspectral imager used in this paper has a band range of 400-1000 nm, a spectral resolution of 2.5±0.5 nm, and a flight altitude of 30 m. Before shooting, it is necessary to

perform whiteboard correction and no light correction on the camera, and collect whiteboard data and no reflectance data of the spectral lens under sunlight exposure. This can eliminate the influence of the environment on the spectrum while ensuring that the spectral reflectance is consistent with the weather conditions of the day. The acquisition method of the Gaiasky-min hyperspectral imager is a built-in push sweep. In a hyperspectral remote sensing image acquisition, the UAV only needs to hover over the measured land mass, and the acquisition time of hyperspectral remote sensing image of a scene is 12 seconds. The collected data was corrected by SpecView and the spectral data of the sampling points is extracted through Envi puzzle.

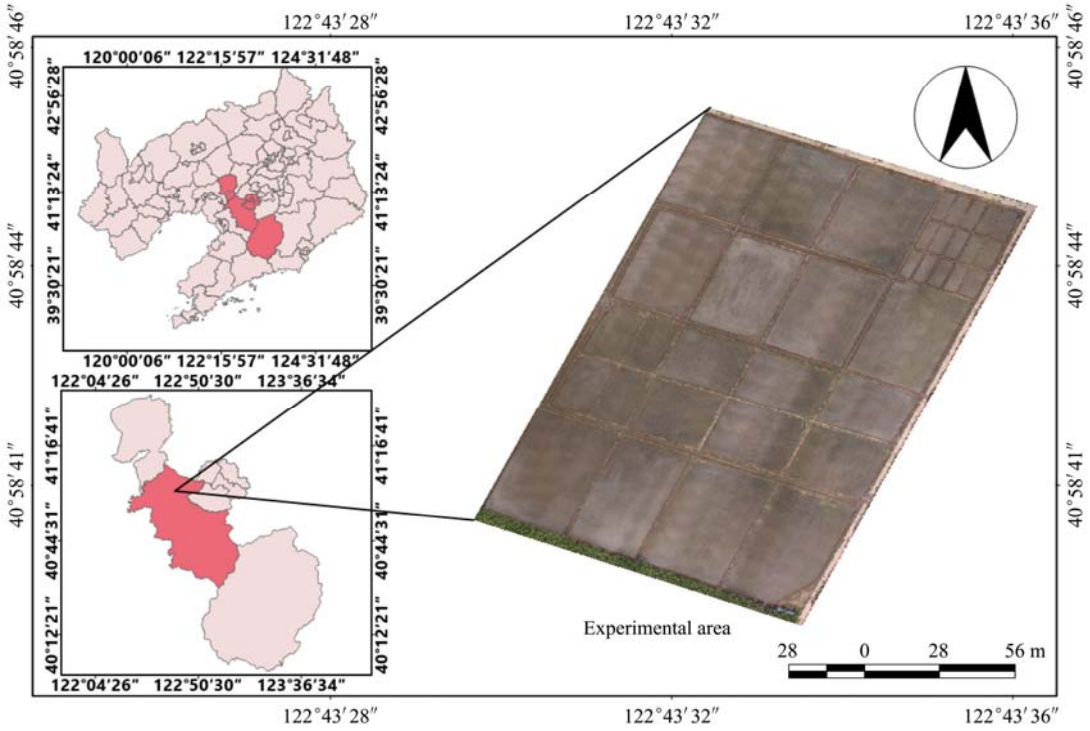


Figure 2 Overview of experimental areas

2.2.2 Soil total organic matter content data

Sampling was carried out according to the room divided by rice planting. The surface soil was collected in squares divided into 20 cm×20 cm in each area. Samples were collected from three different locations in the field area and two different locations in the small field area, and 83 soil samples were collected three times. The soil sample is subjected to natural air drying, grinding, and sieving (0.25 mm) treatment, and then placed in a laboratory environment for 48 hours. The air dried sample is then sieved through a 100 mesh sieve and evenly placed in a container 26[14]. Then weigh 5 g of the sample and use the high-temperature external heat potassium dichromate oxidation volumetric method to determine the organic carbon content. Multiply it by a coefficient of 1.724 and convert it into the organic matter content[15]. The collated data is used as the output end of the neural network for inversion.

3 Research Methods

3.1 Spectral data preprocessing

The spectral information of soil is prone to produce noise bands due to the influence of the environment. How to screens out valuable information related to the soil from mixed data is the key to improving the accuracy of inversion. Usually, due to the device edge (400-450 nm; 950-1000 nm) wavelength is noisy, and relevant bands within the range of 450-950 should be extracted during processing[16]. For the noise bands of burr and sawtooth on the spectral curve S-G convolution smoothing has a good effect on the smoothness of spectral curves and the retention of original spectral information during noise reduction. Therefore, in this study, SG smoothing was used to smooth the original spectral data.

3.2 Spectrum dimension reduction

The feature selection of the spectrum is to select several bands in the original high-dimensional round by some criteria or ways. In this paper, iteratively retains informative variables (IRIV), competitive adaptive reweighted sampling (CARS), and CARS-IRIV algorithm for spectral dimension reduction of original spectral data.

IRIV[17] is a characteristic wavelength selection algorithm based on a binary matrix rearrangement filter. In this study, a possible strategy is proposed to consider the interaction between variables through random combinations, called iterative retained information variable (IRIV). In addition, variables are divided into four categories: substantial information, weak information, non-information, and interference variables. On this basis, IRIV retained both strong and weak information variables in each iteration round until no information and interference variables existed[18], and finally extracted characteristic variables.

Competitive Adaptive Reweighted sampling (CARS) is a feature variable selection method that combines Monte Carlo sampling with regression coefficients of PLS models. The algorithm will take the regression coefficient obtained by each band as the standard to measure its weight. In each sampling process, the exponential attenuation function will be used to remove bands with the relatively small absolute the coefficient value, and the selected band combination will be used to establish the next regression model[19]. This algorithm can overcome the combinatorial explosion problem in variable selection to a certain extent, select the optimal variable subset, improve the prediction ability of the model and reduce the prediction variance, and the effect is good.

Since there are many characteristic variables selected by CARS, and the Monte Carlo sampling process is characterized by solid randomness, the typical variables extracted by the CARS algorithm are also not fixed, and its non-information variables and interference variables will affect the judgment of the results. Therefore, the model results based on the feature variables extracted by this method are unstable[20]. At the same time, the IRIV algorithm can effectively screen the features and obtain a better extraction effect. In this paper, the CARS algorithm is improved by the IRIV algorithm, and the strong and weak information variables are extracted by the IRIV algorithm, which are used as the initial variable set in the CARS algorithm to ensure the rationality and effectiveness of the initial variable set and avoid the randomness of the initial variable set. Eliminate meaningless variables without information and harmful to modeling interference variables, reduce the number of iterations, and improve the reproducibility of results.

3.3 Modeling method

At present, the commonly used modeling methods for soil organic matter content at home and abroad are mainly divided into two categories: the first is the use of the statistical model to establish relations, and the second is the machine learning method. The most commonly used statistical model is the multiple linear regression model, which can carry out a correlation analysis between the extracted critical bands and the soil components to be studied, and reflect the influences of various internal factors on soil organic matter elements[21]. However, the correlation between soil organic matter and spectral reflectance is complex and nonlinear, so the statistical modeling limits the inversion accuracy to some extent. As one of the machine learning methods, the neural network has always been challenging to analyze various indicators, easy to fall into the optimal local solution, and prone to overfitting[22]. However, as the most commonly used inversion method, its advantage is that the neural network usually does not need to know the specific structure, parameters and dynamic characteristics of the object to be modeled. It only needs to take the input and output of the model as the learning object. Through self-learning of neural networks, it is possible to establish an inverse fitting relationship between input parameters and output. Due to the nonlinear relationship between hyper spectrum and soil organic matter content, the neural network has outstanding nonlinear mapping ability and can well process the naturally nonlinear data[23]. In summary, neural network is an effective method suitable for hyperspectral inversion modeling.

3.3.1 Sparrow optimization algorithm

The optimization problem is a hot issue in the field of scientific research and engineering practice. Most intelligent optimization algorithms are inspired by human intelligence, the sociality of biological groups, or the laws of natural phenomena, and carry out global optimization in the solution space. Sparrow Search Algorithm (SSA), first proposed by Wu et al.[24], is a new intelligent optimization algorithm based on the foraging and anti-predation behavior of the sparrow population. It is the same as the particle swarm optimization algorithm and dragonfly optimization algorithm, which belongs to the swarm intelligence algorithm based on social characteristics optimization of groups. The algorithm simulates the foraging and anti-predation behaviors of sparrows by updating individual positions constantly. Compared with the traditional algorithm, the sparrow search algorithm has a simple structure, easy implementation, fewer control parameters and strong local search ability. The performance of this algorithm is better than the traditional algorithms such as particle swarm optimization algorithm and ant colony optimization algorithm on the reference functions of single peak and multi-peak, and it has the characteristics of fast search speed and high stability.

In the sparrow search algorithm, individuals are divided into discoverer, follower and conservative, and each individual position corresponds to a solution. According to the algorithm setting, the proportion of the alert population is 10%~20%, and the finder and follower are dynamic change. That is, one individual becomes the finder inevitably means that another individual will become the follower. According to the division of labor, the discoverer mainly provides foraging direction and area for the whole population, the follower follows the discoverer for foraging, and the alert is responsible for monitoring the foraging site. In the process of foraging, the three positions are constantly updated to complete the acquisition of resources.

3.3.2 Chaotic Sparrow search optimization algorithm

When the sparrow search algorithm (SSA) approaches global optimization, the population diversity decreases, and it is easy to fall into the local optimal solution. Zhang Xin et al.[25] proposed a chaotic sparrow search optimization algorithm (CSSA). First, by improving the initial population of the Tent chaotic sequence, the quality of the initial solution is improved, and the global search capability of the algorithm is enhanced. Secondly, the Gaussian variation method is introduced to strengthen local search ability and improve search accuracy. At the same time, the Tent chaotic series is generated based on the solution of search stagnation, and the chaotic sequence is used to carry out chaotic disturbance to some individuals trapped in the local optimal, which prompts the algorithm to jump out of the limit and continue searching.

In this paper, chaos perturbation is introduced to avoid local optimization, and the global search ability and optimization accuracy are improved. The steps of chaotic perturbation are described as follows:

CSSOA algorithm introduces Tent chaotic search and Gaussian variation, which increases the population diversity, improves the search performance and exploitation performance of the algorithm, and avoids falling into local optimization. The specific implementation steps are as follows:

Procedure Step 1 Initialize the system. It includes population size N, number of discoverers P_{Num}, number of sparrows for reconnaissance and early warning S_{Num}, dimension D of the objective function, upper and lower bound lb and ub of the initial value, the maximum number of iterations T or solving accuracy ε.

Step 2 Apply Tent chaotic sequence to initialize the population, generate N D-dimension vector Z_i, and place each component within the value range of the original problem spatial variable through the carrier. Where, $d_{\max}$ and $d_{\min}$ are respectively the maximum and minimum values of the d-dimension variable X^d_{new}

$$X^d_{new}=d_{\min}+(d_{\max}-d_{\min})Z_d$$

Step 3 Calculate the fitness f_i of each sparrow, and select the current optimal fitness f_g and its corresponding position $x\ b$, and the current worst fitness f_w and its corresponding position xw.

Step 4 Select the first P_{Num} sparrows with good fitness as discoverers and the rest as adders, and update the positions of discoverers and adders according to the corresponding formula of the basic sparrow algorithm.

Step 5 Randomly select a S_{Num} sparrow from the sparrow population for reconnaissance and early warning, and update its position according to the corresponding formula of the basic sparrow algorithm.

Step 6 After completing one iteration, the fitness value f_i of each sparrow and the average fitness value f_{avg} of the sparrow population were recalculated.

1) When $f_i \leqq f_{avg}$, it indicates the phenomenon of "aggregation". Gaussian variation is carried out according to the formula. If the individual is better than the one before the variation, the individual after the variation will be replaced by the one before the variation; otherwise, the original individual will remain unchanged.

$$\text{mutation}(x)=x(1+N(0,1))$$

2) When $f_i \geq f_{avg}$, it indicates a "divergent" trend. At this time, the ith individual is subjected to Tent chaotic disturbance. If the individual performance after disturbance is better, the individual before disturbance is replaced by the individual after disturbance; otherwise, the original individual remains unchanged.

Step 7 updates the optimal position xb and its fitness f_g and the worst position xw and its fitness f_w experienced by the whole population according to the current state of the sparrow population.

Step 8 determines whether the algorithm has reached the maximum number of iterations or the solution accuracy. If so, the loop ends, and the optimization result is output. Otherwise, go back to step 4.

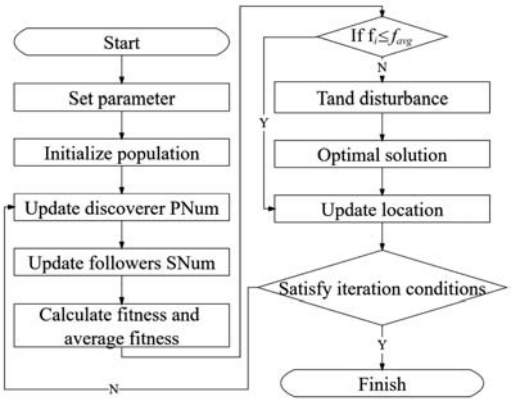


Figure 3 Chaos sparrow optimization algorithm flowchart

3.4 Model accuracy analysis

In this paper, three methods of BPNN, SSA-BPNN, and CSSA-BPNN are used to establish the inversion model and their determination coefficient (R^2) and root mean square error (*RMSE*) are compared. The determination coefficient shows the accuracy of the model. The R^2 is between 0 and 1. The closer it is to 1, the better the fitting effect and the more significant the regression effect. But at the same time, we should also pay attention to the high determination coefficient caused by overfitting. The root means square error can be used to analyze the model's accuracy by amplifying minor differences. It can effectively reflect the degree of dispersion of the model. The smaller the root mean square error is, the lower the dispersion degree of the model is. The model with the best inversion effect is selected after analyzing the two data.

4 Results and analysis

4.1 Soil data collation

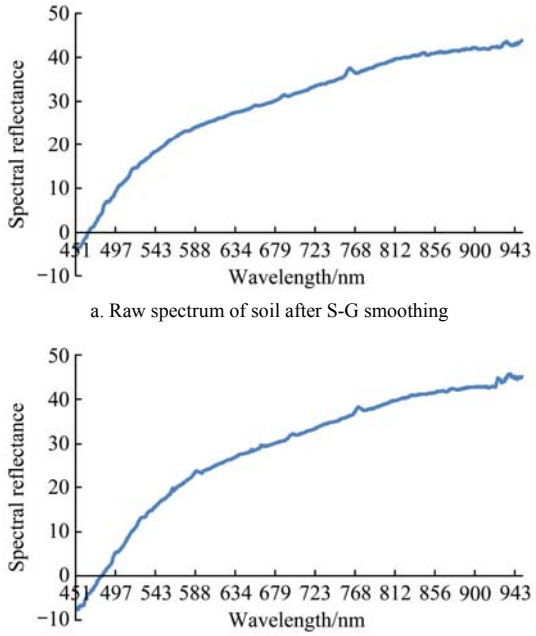
The total organic matter content in the collected data was analyzed by chemical method and statistically sorted out (Table 1). The results showed that the STN content in the two batches of soil samples ranged from 0.328 to 2.156 g/kg, with an average value of 1.539 g/kg, and the data were concentrated in the range of 1.1 to 1.8 g/kg, indicating a high degree of numerical differentiation. It can be used for further inversion modeling analysis.

Table 1 Soil collection-related information

Collection time	No.of samples	Maximum /g·kg^{-1}	Minimum /g·kg^{-1}	Mean /g·kg^{-1}	Standard deviation/g·kg^{-1}
2021-10-13	83	23.8	13.1	18.2	1.49
2022-5-16	83	24.0	13.5	17.9	1.34
2023-5-07	83	24.2	14.0	19.1	1.92

4.2 Spectral pretreatment of soil

The obtained soil hyperspectral data is analyzed, and the soil spectral data is sorted. To retain the characteristic spectral information, reduce the additional influence. In this paper, the spectrum is smoothed by S-G. Comparing the smoothed and pre smoothed spectral curves, we can find that the smoothed curve can effectively eliminate tail noise while smoothing out some prominent parts. By smoothing the spectral band, it can be seen that the original spectral data fluctuated greatly between 634-768 nm and had strong spectral characteristics. Meanwhile, the spectral growth of soil gradually stabilized and leveled off at near-infrared. Figure 4 shows the soil spectral curve in the range of 450-950.



a. Raw spectrum of soil after S-G smoothing

b. Original soil spectral curve

Figure 4 Comparison of spectral curves before and after smoothing

4.3 Selection of soil spectral characteristics

4.3.1 CARS Dimension Reduction

The first derivative data of the spectrum was reduced by the CARS. The figure 5 shows the selection process of soil and spectral characteristic wavelength variables based on CARS algorithm. As can be seen from the upper subfigure, with the increase of running times in the selection process, the number of selected band variables decreases from fast to slow. There is an overall exponential relationship between running times and the number of retained variables. The middle subgraph is the trend graph of residuals obtained by the 50-fold interaction test. In the early stage of operation, the values showed a decreasing trend, indicating that variables unrelated to the properties of sample components were eliminated in the screening process. After a

certain number of runs, the value increases, which may be due to the elimination of crucial variables, resulting in an increase in residual. Each line in the bottom subgraph represents the variation trend of regression coefficients of each variable with the rise of running times. The region corresponding to the dense blue line in the figure is the lowest point, and the model results are optimal now. Figure 5 shows the spectral characteristics selected by isomap.

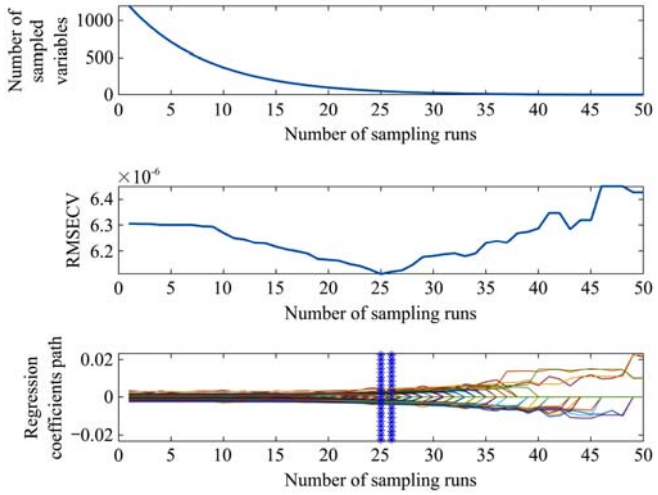


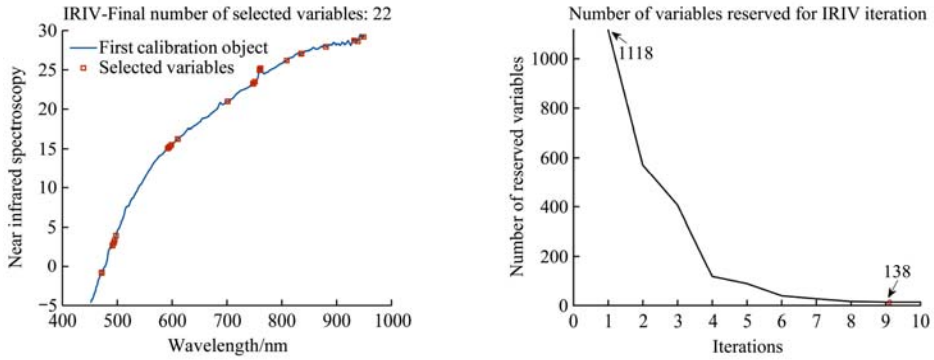
Figure 5 CARS descending correlation curves

4.3.2 Dimension Reduction by IRIV

IRIV dimension reduction was performed on the original spectral data. The figure shows spectral bands selected based on IRIV dimension reduction. The maximum principal factor number in the PLS model is 10. IRIV algorithm carried out a total of 10 rounds. As shown in the figure, the number of iterative variables decreased rapidly in the first four rounds, from 1118 variables to 138 variables, and then the reduction of the number of variables slowed down. After the 7th round of iteration, no information variables and interference variables were completely wholly eliminated, and the reverse elimination operation was carried out. After the 9th round of reverse elimination, 20 characteristic variables related to soil total organic matter were selected. Figure 6 shows the characteristic bands and number of features selected by IRIV

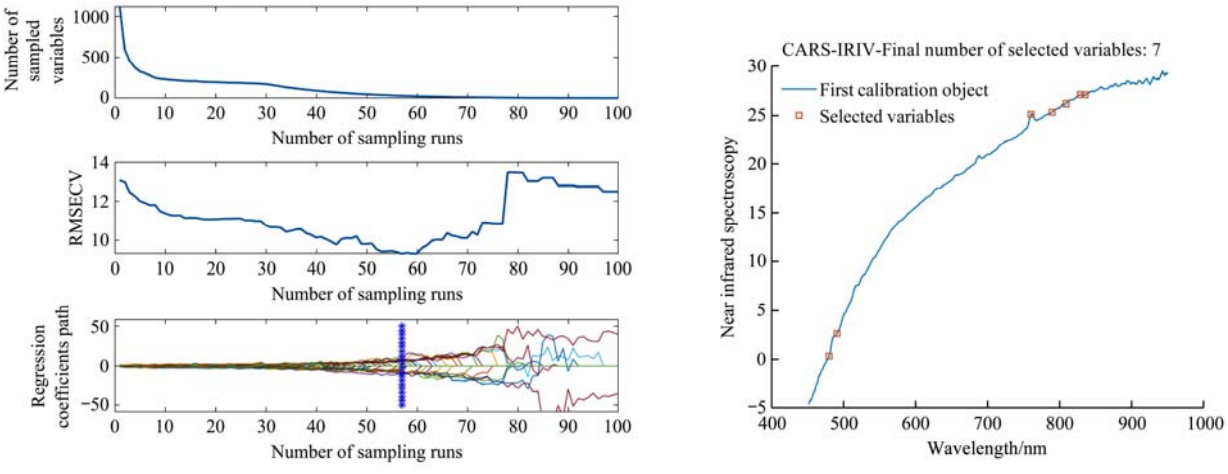
4.3.3 Combined CARS-IRIV dimensionality reduction method

CARS-IRIV dimensionality reduction was performed on the raw spectral data. The figure shows the spectral bands selected based on IRIV dimensionality reduction, and the maximum number of principal factors in the PLS model is 10. In this dimension reduction, 118 feature bands extracted by CARS were put into IRIV for information extraction, and then the irrelevant variables were eliminated through IRIV. Due to the preliminary screening of CARS, the number of iterations was reduced and the efficiency was significantly improved. The combination finally selected seven feature variables, which not only reduced the number of features but also improved the operation efficiency. At the same time, the characteristic wavelength selection distribution is also similar to the spectral response interval of soil organic matter in Q's other studies, which further confirms the rationality of the CARS-IRIV dimension reduction combination. Figure 7 shows the number of feature selection and sensitive band area after cars-iriv combination.



a. Feature selection point; b. Number of runs when selecting a feature.

Figure 6 IRIV descending correlation curves



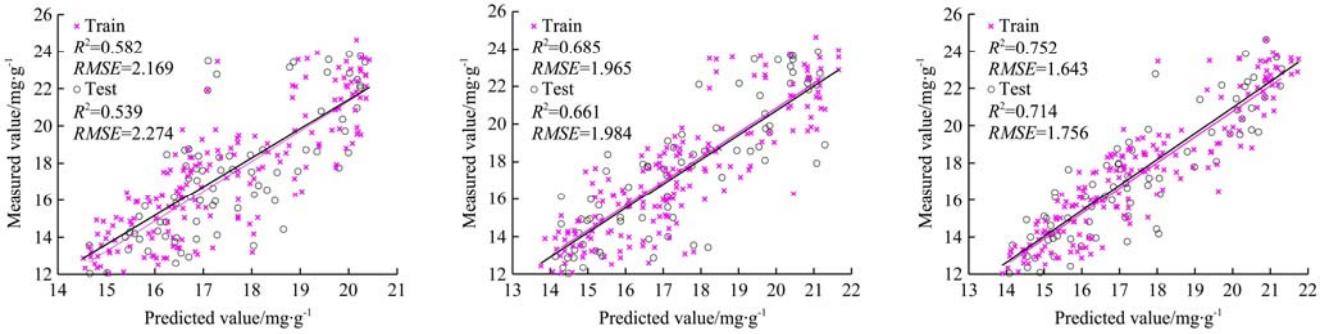
a. CARS descending correlation curves b. Feature selection point.

Figure 7 CARS-IRIV descending correlation curves

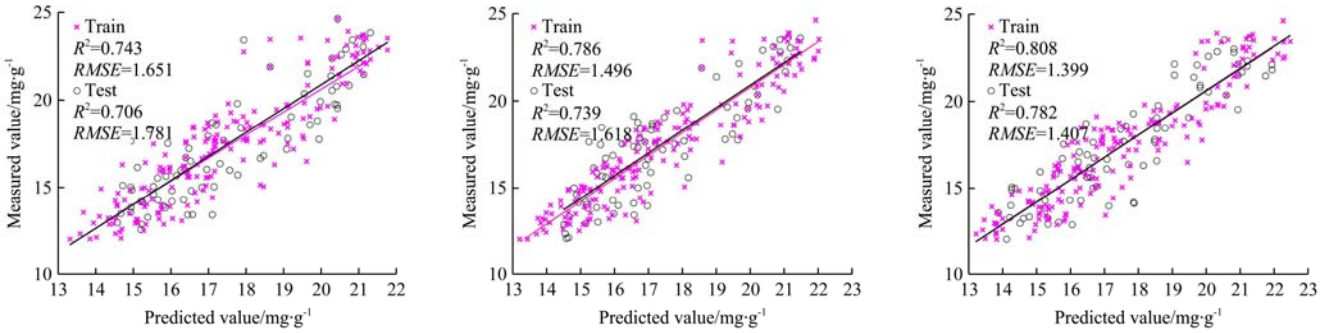
4.4 Inversion results

Based on the spectral features selected by CARS, IRIV, and CARS-IRIV dimensionality reduction methods, BPNN, SSA-BPNN, and CSSA-BPNN methods are used to establish the inversion model. All model parameters are in the best state, and the model accuracy is analyzed respectively.

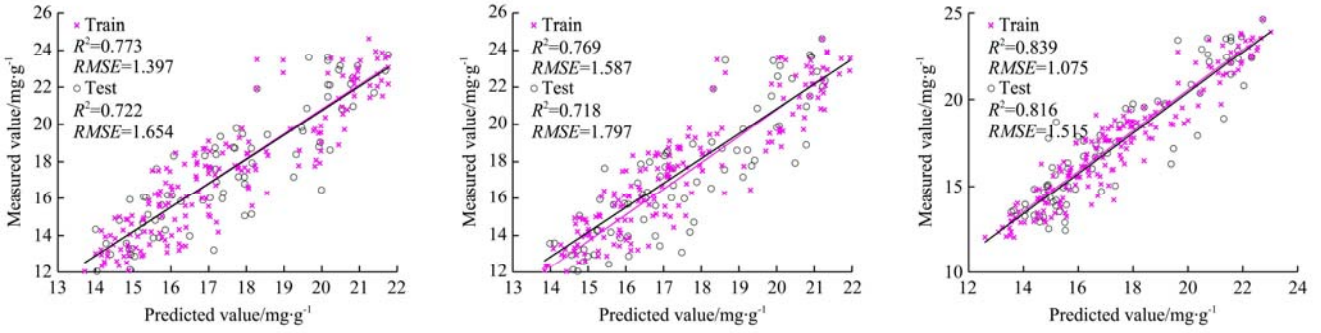
Based on the results obtained by the three-dimensionality reduction methods as the input of inversion modeling, the inversion modeling effect is good, the R^2 of the training set and the validation set are all above 0.539, and the RMSE is less than 2.274 mg/g. Comparing the three dimensionality reduction methods, it can be seen that CARS-IRIV combined with the dimensionality reduction method can effectively improve the model's accuracy, and the R^2 of its inversion model is higher than that of other dimensionality reduction methods. Among the modeling methods, the CSSA-BPNN inversion model has the highest accuracy, the R^2 of the training set and the validation set are 0.839 and 0.816, and the *RMSE* is 1.075 and 1.515 mg/g, respectively. The BPNN model without optimization has the lowest accuracy, with R^2 of 0.773 and 0.722 for the training and validation sets, and *RMSE* of 1.397 and 1.654 mg/g, respectively. The accuracy of the inversion model established by the neural network optimized by SSA-BPNN is between the former two. The R^2 of the training set and the validation set are 0.836 and 0.804, and the *RMSE* is 1.306 and 1.402 mg/g, respectively. It can be seen that the CSSA-BPNN model has better stability and prediction ability. After CARS-IRIV dimensionality reduction, the inversion accuracy of the training set and the test set achieved good results. The neural network inversion effect optimized by the algorithm is also significantly better than that of the non-optimized neural network. It can be confirmed that Tent chaos search and Gaussian mutation can improve the search performance and exploitation performance of the algorithm, and avoid falling into local optimum. After using CSSA to optimize the permutation order, the model accuracy of the neural network has been significantly improved, and its effect is better than that of the non-optimized neural network. Figure 8 shows six inversion result.



a. The BP neural network inversion model after selecting features based on three dimensionality reduction methods is respectively



b. Based on three dimensionality reduction methods, the features are selected, and the inversion model is established using the BP neural network optimized by SSA



c. The inversion model is established by using the BP neural network optimized by CSSA after selecting features based on three dimensionality reduction methods

Figure 8 Based on three dimensionality reduction methods, nine inversion models after dimensionality reduction

5 Discussion

This article uses three dimensionality reduction methods of CARS, IRIV, and CARS-IRIV, as well as three modeling methods of BPNN, SSA-BPNN, and CSSA-BPNN, to model and validate soil organic matter content. The results indicate that the analysis results based on the three modeling methods all indicate a good response relationship between soil spectra and soil organic matter

content. The CSSA-BPNN method, after dimensionality reduction by CARS-IRIV, has high accuracy and small root mean square error. The fitting effect with soil spectra is good. The accuracy of the established model is much higher than other methods.

At present, neural networks are widely used as the most commonly used method in soil nutrient spectral inversion modeling. Previous research has been based on innovative methods for extracting and modeling complex features. However, these methods are difficult to solve the initial population optimization problem of neural networks. The optimization of the initial population can improve the accuracy of the entire model. This article applies the CARS-IRIV dimensionality reduction method to the dimensionality reduction of hyperspectral soil data and models the inversion of soil organic matter content. Compared with the two prototype dimensionality reduction methods, the prediction accuracy of the soil organic matter content inversion model is more significant in the spectral dataset after CARS-IRIV dimensionality reduction. Combining IRIV with CARS to eliminate non information content and interference variables can effectively fit the physical and chemical properties of soil and improve the accuracy of feature selection. The soil total nitrogen inversion model constructed based on CARS-IRIV extracted spectral features has an R^2 range of 0.714~0.839 and an *RMSE* range of 1.075~1.756/kg; The soil total nitrogen inversion model constructed based on spectral features extracted by the other two methods has an R^2 range of 0.539~0.769 and an *RMSE* range of 1.587~2.274 g/kg. It can be seen that the soil total nitrogen inversion model constructed based on CARS-IRIV extracted spectral features has higher accuracy.

Meanwhile, this article established three inversion models using BPNN, SSA-BPNN, and CSSA-BPNN, and compared the accuracy of the models. When using BPNN for modeling, the R^2=0.773 regression data is affected by local optimal solutions, resulting in lower inversion accuracy. Compared to this, SSA-BPNN and CSSA-BPNN have significantly higher inversion accuracy, with R^2 values of 0.769 and 0.839, respectively, because these two algorithms optimize the model process and improve algorithm efficiency. CSSA-BPNN has the highest prediction accuracy because optimization algorithms can better allocate initial data, improve the accuracy, stability, and generalization ability of the model. However, it should be noted that the weights and thresholds of neural networks are random, requiring multiple training sessions to obtain better fitting results, and have specific uncertainties. Overfitting may also occur during its training [26]. Genetic algorithms cannot completely solve the problem of getting stuck in the optimal local solution, which to some extent affects the accuracy of the results. The idea of confidence testing in future statistical processes can be used as a reference, and the optimized parameters for measurement can be set to the global optimal confidence index[27].

After comprehensive comparison, the CSSA-BPNN model constructed based on CARS-IRIV extraction of sensitive bands has the relatively highest accuracy. Based on the CSSA-BPNN model, the inversion of soil total nitrogen content in the study area found that most of the soil organic matter content in the study area in the range of 12-25 g/kg, this is consistent with the soil total nitrogen content levels of 249 soil samples, as well as the actual situation of planting one season a year, low soil fertility consumption, and annual fertilization in the local area. This indicates that the CSSA-BPNN model can be effectively used for estimating soil total nitrogen content in the field.

6 Conclusions

This article is based on unmanned aerial vehicle hyperspectral data and soil organic matter data collected in the laboratory, analyzing the correlation between soil spectral characteristics and soil organic matter content. CSSA-BPNN is used to construct a soil organic matter inversion model, and the calibration and optimization research of the unmanned aerial vehicle hyperspectral soil organic matter content inversion model is carried out. The following conclusions are drawn:

(1) CARS-IRIV manifold learning method can effectively reduce the dimension of hyperspectral data, and can effectively eliminate and screen the data. The helpful information can be extracted under the premise of maintaining the original basic characteristics of the spectrum, which lays a good foundation for the following modeling.

(2) In terms of model optimization, CSSA genetic algorithm improves the initial population of BP neural network Tent chaotic sequence, improves the quality of the initial solution, and enhances the global search ability of the algorithm. Its R^2=0.839, *RMSE*= 1.075, and the prediction accuracy is better than that of the unoptimized BP neural network and the optimized BP neural network after SSA. It shows that the CSSA optimization algorithm has particular practical value in predicting soil total organic matter content.

(3) The CSSA-BPNN inversion model established by using the spectral characteristics obtained by the CARS-IRIV algorithm has the highest accuracy, the R^2 of the training set and the validation set are more than 0.714, and the *RMSE* is less than 1.756 mg/g. The inversion model established in this paper obtains good prediction results and can provide a new method for accurately detecting soil organic matter conte.

(4) The overall accuracy of the model is high, which can confirm the influence of soil spectral sensitivity band on soil spectral curve.

[References]

[1] Qiong, H. O. U., et al. Effects of manure substitution for chemical fertilizers on rice yield and soil labile nitrogen in paddy fields of China: A meta-analysis, 2022. Doi: 10.1016/j.pedsph.2022.09.003

[2] Li X N, et al. Soil pollution and site remediation policies in China: A review. Environmental Reviews, 2015, 23.3: 263–274. Doi: 10.1139/er-2014-0073

[3] Chang D, Zhang Y X. Farmland nutrient pollution and its evolutionary relationship with plantation economic development in China. Journal of Environmental Management, 2023, 325: 116589. Doi 10.1016/j.jenvman.2022.116589

[4] RAYA-MORENO, Irene, et al. Comparing current chemical methods to assess biochar organic carbon in a Mediterranean agricultural soil amended with two different biochars. Science of the Total Environment, 2017, 598: 604–618. Doi: 10.1016/j.scitotenv.2017.03.168

[5] Niu G X, et al. Effects of decadal nitrogen addition on carbon and nitrogen stocks in different organic matter fractions of typical steppe soils. Ecological Indicators, 2022, 144: 109471. Doi: 10.1016/j.ecolind.2022.109471

[6] LEVI, Nathan; KARNIELI, Arnon; PAZ-KAGAN, Tarin. Airborne imaging spectroscopy for assessing land-use effect on soil quality in drylands. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 186: 34–54. Doi: 10.1016/j.isprsjprs.2022.01.018

[7] ODEBIRI, Omosalewa, et al. Modelling soil organic carbon stock distribution across different land-uses in South Africa: A remote sensing and deep learning approach. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 188: 351–362. Doi: 10.1016/j.isprsjprs.2022.04.026

[8] Hasegawa T, Hasegawa T. Infrared Spectroscopy as a Vibrational Spectroscopy. Quantitative Infrared Spectroscopy for Understanding of a Condensed Matter, 2017: 1–36. Doi: 10.1007/978-4-431-56493-5_1

[9] Qi H J, et al. Evaluating calibration methods for predicting soil available nutrients using hyperspectral VNIR data. Soil and Tillage Research, 2018, 175: 267–275. Doi: 10.1016/j.still.2017.09.006

[10] Huang Y B, et al. Agricultural remote sensing big data: Management and applications. Journal of Integrative Agriculture, 2018, 17(9): 1915–1931. Doi: 10.1016/S2095-3119(17)61859-8

[11] Liu H J, Zhang Y Z, Zhang B. Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China. Environmental monitoring and assessment, 2009, 154: 147–154. Doi: 10.1007/s10661-008-0385-4

[12] Yang C B, et al. Study on hyperspectral estimation model of soil organic carbon content in the wheat field under different water treatments. Scientific Reports, 2021, 11.1: 18582. Doi: 10.1038/s41598-021-98143-0

[13] BAUMGARDNER, Marion F., et al. Reflectance properties of soils. Advances in agronomy, 1986, 38: 1–44. Doi: 10.1016/S0065-2113(08)60672-0

[14] Yang C, et al. Study on hyperspectral monitoring model of soil total nitrogen content based on fractional-order derivative. Computers and Electronics in Agriculture, 2022, 201: 107307. Doi: 10.1016/j.compag.2022.107307

[15] TAVARES TR, et al. Spectral data of tropical soils using dry-chemistry techniques (VNIR, XRF, and LIBS): A dataset for soil fertility prediction. Data in Brief, 2022, 41: 108004. Doi: 10.1016/j.dib.2022.108004

[16] DE ALMEIDA MINHONI RT, et al. Multitemporal satellite imagery analysis for soil organic carbon assessment in an agricultural farm in southeastern Brazil. Science of The Total Environment, 2021, 784: 147216. Doi: 10.1016/j.scitotenv.2021.147216

[17] Yun Y, Wang W, Liang Yi. A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. Analytica Chimica Acta, 2014, 807: 36–43. Doi: 10.1016/j.aca.2013.11.032

[18] Zhang L, He S, Cheng J, et al. Research on neural network wind speed prediction model based on improved sparrow algorithm optimization. Energy Reports, 2022, 8: 739–747. Doi: 10.1016/j.egyr.2022.09.202

[19] Gao D S, L, Y Z, Xu Q S, et al. A new strategy of outlier detection for QSAR/QSPR. Journal of Computational Chemistry, 2010, 31.3: 592–602. Doi: 10.1002/jcc.21351

[20] Fan S X, Guo Z M, Zhang B H, et al. Using Vis/NIR diffuse transmittance spectroscopy and multivariate analysis to predicate soluble solids content of apple. Food Analytical Methods, 2016, 9.5: 1333–1343. Doi: 10.1007/s12161-015-0313-5

[21] Zhang T T, Zeng S L, Gao Y, et al. Using hyperspectral vegetation indices as a proxy to monitor soil salinity. Ecological Indicators, 2011, 11.6: 1552–1562. Using hyperspectral vegetation indices as a proxy to monitor soil salinity. Ecological Indicators, 2011, 11(6): 1552–1562. Doi: 10.1016/j.ecolind.2011.03.025

[22] He C, Ma M, Wang P. Extract interpretability-accuracy balanced rules from artificial neural networks: A review. Neurocomputing, 2020, 387: 346–358. Doi: 10.1016/j.neucom.2020.01.036

[23] Lu B, Liu N, Li H, et al. Quantitative determination and characteristic wavelength selection of available nitrogen in coco-peat by NIR spectroscopy. Soil and Tillage Research, 2019, 191: 266–274. Doi: 10.1016/j.still.2019.04.015

[24] Wu R, Huang H, Wei J, et al. An improved sparrow search algorithm based on quantum computations and multi-strategy enhancement. Expert Systems with Applications, 2023, 215: 119421. Doi: 10.1016/j.eswa.2022.119421

[25] Zhang C, Ding S. A stochastic configuration network based on chaotic sparrow search algorithm. Knowledge-Based Systems, 2021, 220: 106924. Doi: 10.1016/j.knosys.2021.106924

[26] Gholizadeh A, Saberioon M, Rossel R A V, et al. Spectroscopic measurements and imaging of soil colour for field scale estimation of soil organic carbon. Geoderma, 2020, 357: 113972. Doi: 10.1016/j.geoderma.2019.113972

[27] Deiss L, Margenot A J, Culman S W, et al. Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. Geoderma, 2020, 365: 114227. Doi: 10.1016/j.geoderma.2020.114227