Application of swarm intelligence algorithms to the characteristic wavelength selection of soil moisture content

Dongxing Zhang[1,2], Jiang Liu[1,2], Li Yang[1,2*], Tao Cui[1,2], Xiantao He[1,2], Tiancheng Yu[1,2], Abdalla N. O. Kheiry[3]

(1. *College of Engineering, China Agricultural University, Beijing 100083, China*;
2. *The Soil-Machine-Plant key laboratory of the Ministry of Agriculture of China, Beijing 100083, China*;
3. *Department of Agricultural Engineering, College of Agricultural Studies, Sudan University of Science and Technology, Khartoum 999129, Sudan*)

Abstract: Swarm intelligence algorithms own superior performance in solving high-dimensional and multi-objective optimization problems. The application of the swarm intelligence algorithms to visible and near-infrared (VIS-NIR) spectral analysis of soil moisture can contribute to the optimization of the soil moisture prediction model and the development of the real-time soil moisture sensor. In this study, a high-resolution spectrometer was used to obtain spectral data of different levels of soil moisture which were manually configured. Isolation Forest algorithm (iForest) was used to eliminate outliers from the data. Based on the root mean square error of prediction $RMSE_P$ of Back Propagation Neural Network (BPNN) model results, a series of new swarm intelligence algorithms, including Manta Ray Foraging Optimization (MRFO), Slime Mould Algorithm (SMA), etc., were used to select the characteristic wavelengths of soil moisture. The analysis results showed that MRFO owned the best performance if only from the predictive capability perspective and SMA had a better performance when considering the proportion of the selecting wavelengths and the results of the model prediction. By comparing and analyzing the modeling results of traditional intelligence algorithms Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), it was found that the new swarm intelligence had a better performance in selecting the characteristic wavelengths of soil moisture. Integrating the results of all intelligence algorithms used, soil moisture sensitive wavelengths were selected as 490 nm, 513 nm, 543 nm, 900 nm and 926 nm, which provide the basis for the design of real-time soil moisture sensor based on VIS-NIR.

Keywords: soil moisture content, swarm intelligence, characteristic wavelength selection, application, visible and near-infrared spectroscopy

DOI: 10.25165/j.ijabe.20211406.6629

Citation: Zhang D X, Liu J, Yang L, Cui T, He X T, Yu T C, et al. Application of swarm intelligence algorithms to the characteristic wavelength selection of soil moisture content. Int J Agric & Biol Eng, 2021; 14(6): 153–161.

1 Introduction

On-the-go soil moisture content (SMC) measurement can realize the continuous collection of soil moisture information, which plays an important role in agricultural activities such as planting and transplanting[1]. Acquiring soil moisture information in real-time has become one of the most advanced development directions of precision agriculture. Some relevant studies reported that the measurement of SMC based on visible and near-infrared (VIS-NIR) spectrum owns some preferable performance in real-time and continuity[2-4]. One of the most important parts of SMC measurement based on VIS-NIR is to acquire a prediction model with easy interpretation and high operational efficiency. For this reason, it is expected to select characteristic wavelength variables and eliminate redundant wavelengths, and the original spectral information, meanwhile, should be retained to the maximum when analyzing SMC spectrum[5]. Furthermore, finding the sensitive wavelengths is also significant to the design of SMC measurement equipment[6,7]. When using spectrum to predict SMC, the difference between an SMC sensor and a spectrometer lies in that the SMC sensor just selects several sensitive wavelengths as the light source. This is because the real-time measurement requires swift calculation, using only a few wavelengths can greatly reduce the time complexity of the prediction model. Besides, the portable size and economical cost of the sensor should also be considered.

Common methods of characteristic wavelength selection mainly include principal component analysis (PCA), successive projections algorithm (SPA), uninformative variable elimination (UVE), competition adaptive reweighted sampling (CARS), etc.[8] These methods are often used in conjunction with partial least-squares regression (PLSR). Modeling SMC spectrum and the measured data by PLSR, analyzing the model coefficient through the above methods, the characteristic wavelength can be selected. Some researchers have used these methods in SMC characteristic wavelength analysis. Liu[9] has investigated the relationship between the soil reflectance and moisture in 400-

Received date: 2021-03-27 **Accepted date:** 2021-07-07

Biographies: Dongxing Zhang, Professor, research interest: full mechanization of maize production, Email: zhangdx@cau.edu.cn; **Jiang Liu**, MS research interest: precision agriculture, Email: ljiangcau@163.com; **Tao Cui**, PhD, Associate Professor, research interest: full mechanization of maize production, Email: cuitao@cau.edu.cn; **Xiantao He**, PhD, Associate Professor, research interest: precision agriculture, Email: hxt@cau.edu.cn; **Tiancheng Yu**, MS, research interest: precision seeding, Email: 809217014@qq.com; **Abdalla N. O. Kheiry**, PhD, research interest: agricultural machinery management and simulation modeling, Email: abdallakheiry@gmail.com.

***Corresponding author**: **Li Yang**, PhD, Professor, research interest: precision agriculture. College of Engineering, China Agricultural University, Beijing 100083, China. Tel: +86-10-62737765, Email: yangli@cau.edu.cn.

2450 nm wavelength range. Linear stepwise regression was used to find the principal components of soil reflectance spectral data and seven wavelengths of 450, 574, 986, 1400, 1672, 1998, and 2189 nm were extracted according to their residual error rank. The wavelength variables sensitive to SMC were selected from the full spectrum by CARS in Yu's study[10]. The prediction accuracy of PLSR calibration model was an evaluation to locate optimal variables. According to his study, four wavebands were selected in 350-2500 nm which are respectively 443-449 nm, 1408-1456 nm, 1916-1943 nm, and 2209-2225 nm. Various methods such as UVE, SPA, CARS, etc., were used in the determination model optimizing and characteristic wavelength extracting of SMC spectral reflectance in Wu's research[11]. Comparison among the modeling results showed that wavelength variables extracted by β coefficient were optimal. The characteristic wavelengths of the range of 400-1000 nm are 411 nm, 440 nm, 622 nm, 713 nm, and 790 nm.

Some traditional intelligence algorithms like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) have already been applied in characteristic wavelength selection. Combining with PLSR model, algorithms including PSO and GA were used on analyzing the characteristic spectral variables of total nitrogen and nicotine of tobacco leaves in Bin's study[12]. Xue et al.[13] used PSO to determine the characteristic wavelength of dichlorvos residue on the surface of navel orange with VIS-NIR spectroscopy. The application of combining GA and SPA has been studied on the selection of characteristic wavelength to evaluate exudative characteristics in frozen-thawed fish muscle by Cheng et al.[14]. Swarm intelligence (SI) refers to an optimization algorithm that imitates the intelligent behavior of non-intelligent population in nature through group interaction[15]. With the deeper exploration of nature, many new SI algorithms have been proposed. Fewer parameters, better global search ability, and better ability to solve high and multi-objective optimization problems, the merits of these algorithms make them extensively applied in recent years, particularly in path planning, mechanical optimization, or areas like that[16].

Attempting to apply various kinds of latest SI algorithms like Butterfly Optimization Algorithm (BOA), Crow Search Algorithm (CSA), Grey Wolf Optimization (GWO), etc. on SMC characteristic wavelength selection in the 400-980 nm reflectance spectral band, the purpose of this study mainly lies on the following two perspectives:

1) Discuss the feasibility of the SI algorithms on SMC characteristic wavelength selection and SMC spectral prediction model optimization. Finding the preferable algorithm among them by comparing the modeling results. Making comparisons between the new SI algorithms and traditional intelligence algorithms. Thus, to provide a new thought in soil spectrum analysis.

2) Finding the sensitive wavelengths of SMC by comprehensively analyzing the characteristic wavelength selection results of intelligence algorithms. Trying to use several wavelengths to reach a competitive prediction accuracy. Providing a theoretical basis for the light source selection of the soil moisture measurement sensor using the optical principle.

2 Material and methods

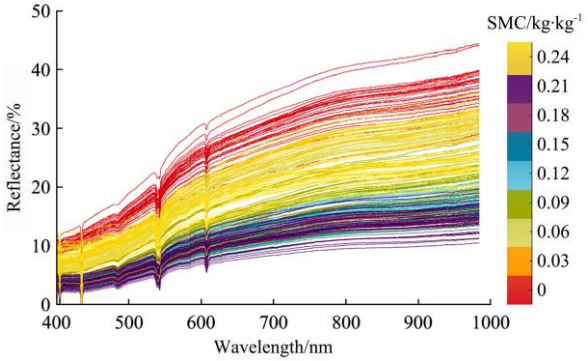
2.1 Acquisition of spectral data

Soil samples were taken from Liuquan Town, Gu'an County, Hebei Province, China (116°24′35″E, 39°22′48″N), belonging to arable soil. And the soil type was mainly sandy loam. In order to obtain different levels of SMC, soil samples were manually configured. Soil taken from field was first put into the oven (DHG-9123A, Shanghai) being dried till the weight of soil no longer changes. Then a 2 mm soil sieve was used to screen. The screened soil was evenly mixed up with a gradient weight of water. Subsequently, soil samples with gradient moisture were filled into aluminum boxes and scrapped and compacted the surface to flat. Spectral data could be collected by putting the samples under optical fiber. The spectrometer that the experiment used was a high-performance QE Pro spectrometer (Ocean Optics, Inc., USA) whose measurable wavelength range was around 185-1100 nm and full width at half maximum (FWHM) is 1.1 nm along with laboratory optical fiber (Ocean Optics, Inc., USA) and HL-2000 halogen light source (Ocean Optics, Inc., USA). In this study, ultraviolet exploration was not included. Thus, the actual wavelength range was 400-980 nm after excluding the noise at the edges of each spectrum. When finishing spectrum data collection, samples were put on an electronic balance to weigh the total weight of aluminum boxes and wet soil which was marked as m_1. Subsequently, put them into the oven to dry at 105℃ for 12 h. When the samples cooled, the total weight of aluminum boxes and dry soil was weighed and marked as m_2. The SMC can be calculated as Equation (1).

$$w = \frac{m_1 - m_2}{m_2 - m_0} \quad (1)$$

where, w is the gravimetric water content of soil, kg/kg; m_0 is the weight of the aluminum box, kg; m_1 is the total weight of the aluminum box and wet soil, kg; m_2 is the total weight of the aluminum box and dry soil, kg.

Three samples were set for each moisture level, and 10 groups of data were collected for each sample. A total of nine levels were set and 270 groups of spectral data were eventually collected. Figure 1 displays the full spectral curve acquired.



Note: SMC is the soil moisture content, kg/kg.

Figure 1 Spectral curve of all SMC levels

2.2 Pretreatment of spectrum

Due to the stability of factors like experiment environment and operation cannot be absolutely guaranteed, there could be some outliers that should be removed from the acquired data. Isolated forests algorithm (iForest)[17] is one of the latest methods in mining anomalies. Its principle is to use a hyperplane to divide the data space into two subspaces, then, divide the subspaces over again, through times of iteration, there will be only one data or the same data in each subspace. Apparently, the anomalies need less time in being divided into a single subspace. So, according to the times that one point needs to be divided into a single subspace in a data space, the outlier can be evaluated. The advantage of this method is that it does not need to calculate the distance between

points or the density of point groups which makes it has low complexity and high accuracy. Using iForest to screen outliers of the collected spectrum, the abnormal data can be found by calculating the anomaly score of each point. Figure 2 shows the data points distribution of spectral data of 0 kg/kg level SMC with the mean and standard deviation as the attributes of iForest model (all dimensions are directly used as attributes in the actual program). It is easy to find that the outliers are separated far away from the group. Eliminating outliers from the spectral data of each gradient SMC by the above method, 243 sets of valid data remained. After the outliers were removed, Savitzky-Golay filtering (SG filtering) was used to smooth the spectrum curve with polynomial order of 3 and frame length of 11.

2.3 Modeling method and fitness function

SI algorithm usually requires a fitness function as the optimization evaluation standard. Fitness function is a standard to evaluate the quality of individuals in a group determined by the objective requirement. In order to find the appropriate fitness function, spectral data needs to be modeled to choose an appropriate modeling evaluation standard as the fitness function. PLSR has commonly been used in spectral modeling analysis, and using a neural network to train spectral data has also been applied to spectral model building in recent years[18]. To choose the method with better modeling results, PLSR and Back Propagation Neural Network (BPNN) were used to model and analyze the full spectrum before and after pretreating. Sample set partitioning based on joint X-Y distance (SPXY) was used to divide the training set and test set, and the dividing results were shown in Table 1. The cross-validation method was 10-fold cross-validation.

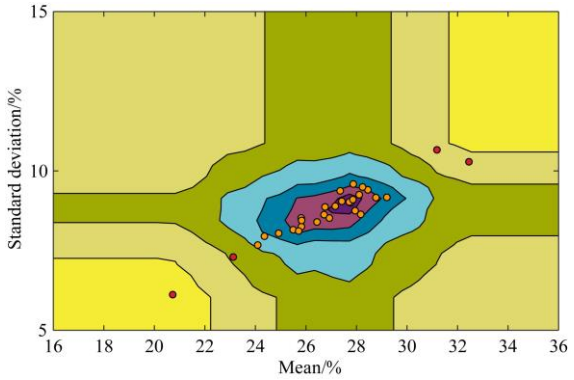


Figure 2 Distribution of outliers of 0 kg/kg level spectrum

Table 1 Data set partitioning of the acquired spectral data before and after pretreatment

Pretreatment	Sample set	Number of samples	Minimum/kg kg^{-1}	Maximum/kg kg^{-1}	Mean/kg kg^{-1}	Standard deviation/kg kg^{-1}
Before	Training set	189	0.0001	0.2458	0.1249	0.0793
	Test set	81	0.0001	0.2458	0.1145	0.0673
After	Training set	170	0.0001	0.2458	0.1047	0.0725
	Test set	73	0.0001	0.2102	0.1170	0.0554

When using PLSR build model, the parameter called Latent Variable (LV) should be determined. For the reason that the cross-validation root mean square error ($RMSE_{CV}$) is often viewed as a standard to determine LV, this study used $RMSE_{CV}$ to seek the proper number of LV. As shown in Figure 3, $RMSE_{CV}$ is small enough and tends to be steady when the number of LV is over 6. Hence the parameter was set as 6 during the modeling process. As for BPNN model, a classical three-layer network was adopted, namely, an input layer, a hidden layer and an output layer. The number of hidden nodes affects the modeling accuracy. The different numbers of hidden nodes were tried in BPNN modeling analysis. As the line graph in Figure 3, a fine modeling accuracy can be achieved by using 3 nodes.

In Table 2, the training correlation coefficient R_T^2, the training prediction root mean square error $RMSE_T$, the cross-validation correlation coefficient R^2_{CV}, the cross-validation root mean square error $RMSE_{CV}$, the prediction correlation coefficient R_P^2, and the prediction root mean square error $RMSE_P$, were respectively obtained by the above two modeling methods. As can be seen from Table 2, after iForest outlier eliminating and SG filtering, $RMSE_{CV}$ of PLSR full-spectrum model decreased by 55.56% and of BPNN full-spectrum model decreased by 55.66%, which indicates that the stability of the full-spectrum model is greatly improved. The correlation coefficients of PLSR and BPNN prediction models are both sharply improved, and $RMSE_P$ is reduced by 48.82% and 73.68%, respectively. The results show that iForest outlier removing and SG filtering has an excellent performance in spectrum pretreatment. Comparing the two modeling methods, BPNN is generally superior to PLSR in training, cross-validation, and prediction results. This study adopted BPNN as the modeling method.

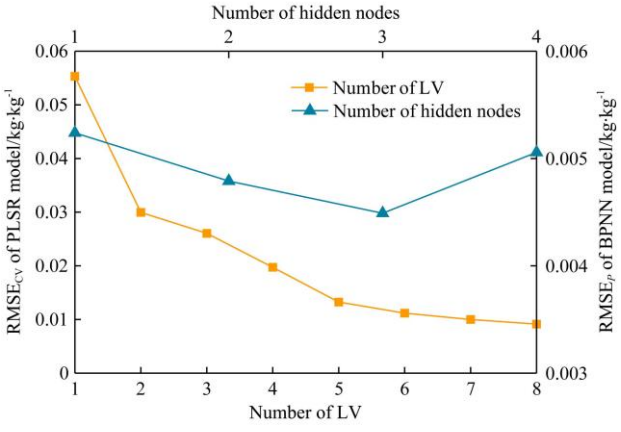


Figure 3 Parameter determination of PLSR and BPNN

Table 2 PLSR and BPNN modeling results before and after pretreatment

Pretreatment	Modeling method	Training set		10-fold cross validation		Test set	
		R_T^2	$RMSE_T$/kg kg^{-1}	R_{CV}^2	$RMSE_{CV}$/kg kg^{-1}	R_P^2	$RMSE_P$/kg kg^{-1}
Before	PLSR	0.9363	0.0200	0.8987	0.0252	0.9008	0.0211
	BPNN	0.9335	0.0204	0.9193	0.0212	0.9346	0.0171
After	PLSR	0.9842	0.0100	0.9804	0.0112	0.9708	0.0108
	BPNN	0.9950	0.0057	0.9839	0.0094	0.9950	0.0045

The purpose of analyzing spectra with different levels of SMC was to predict SMC, so this study adopted $RMSE_P$ as the fitness function f of SI algorithm. And it can be expressed by the following equation:

$$f=\sqrt{\frac{\sum_{i=1}^{n}(y_i-y_{\text{pred}})^2}{n}} \quad (2)$$

where, y_i is the measured value of the test set; y_{pred} is the predicted value obtained by inputting the test set into the BPNN training model; n is the number of labels of the test set.

Steps to program the fitness function are as follows: First, using SPXY to divide the data labels into trainning_label and test_label, respectively. Second, trainning_label and test_label were used to divide the subset selected by SI algorithmin each iteration which can ensure the consistent division of training set and test set of each time. Third, BP neural training using the trainlm training function was carried out on the training set and the training model net was obtained. Finally, substituting the test set into net to calculate y_{pred}, and fitness f could be calculated by Equation (2).

2.4 Swarm intelligence algorithms

The general process of SI algorithm is described in Figure 4. For those algorithms that imitate different natural populations, the main distinction lies in the diverse ways of updating the population position. This study mainly introduces Butterfly Optimization Algorithm[19], Crow Search Algorithm[20], Grey Wolf Optimization[21,22], Harris Hawks Optimization[23] (HHO), Manta Ray Foraging Optimization[24] (MRFO), Slime Mould Algorithm[25] (SMA), Salp Swarm Algorithm[26,27] (SSA), Whale Optimization Algorithm[28,29] (WOA). The brief introduction and key principles of these eight SI algorithms are listed in Table 3. In Table 3, t is the current number of iterations and T is the maximum number of iterations. x_i^t is the current individual position, x_i^{t+1} is the updated individual position, and x_b^t is the optimal individual position (usually described as the prey or food position). ub and lb represent the upper and lower limits of the search range. In this study, ub is 0 and lb is 1. r is a random number in the interval (0, 1). The algorithms were programed and run in MATLAB R2018b software.

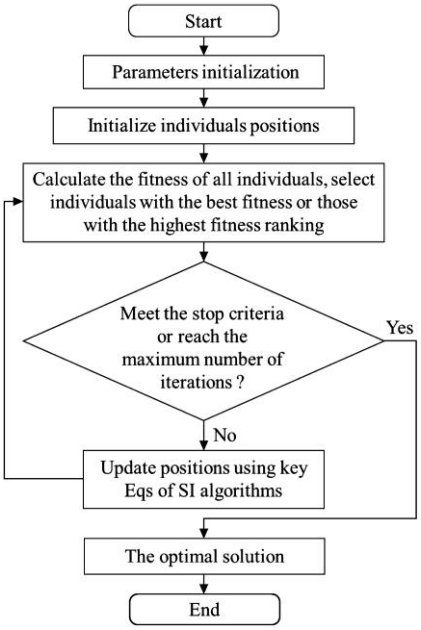


Figure 4 General flow of SI algorithm

Table 3 Introduction of eight SI algorithms

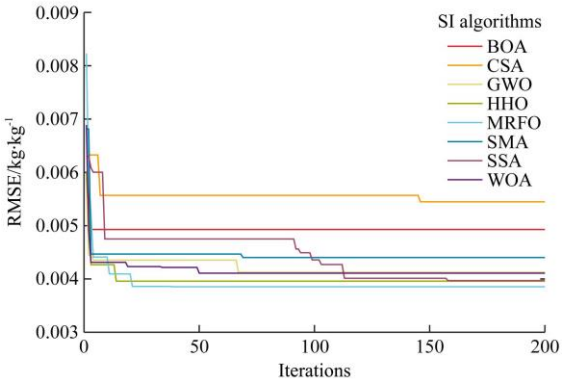
SI	Biological principle	Equations of position updating	Parameter interpretation
BOA	The butterfly that finds food first calls others by producing fragrance, they can update their position through fragrance to find food.	$x_i^{t+1}=\begin{cases}x_i^t+(r^2\cdot x_b^t-x_i^t)\times f_i\\ x_i^t+(r^2\cdot x_j^t-x_k^t)\times f_i\end{cases}$	Where, f_i represents the fragrance, x_j^t and x_k^t represent the jth and kth random butterflies of solution space.
CSA	A crow follows another crow to update its position to steal its food.	$x_i^{t+1}=\begin{cases}x_i^t+r\cdot f\cdot l_i^t\cdot(x_b^t-x_i^t), & r_j\geq \text{AP}_i^t\\ lb+(ub-lb)\cdot\text{rand}, & r_j<\text{AP}_i^t\end{cases}$	Where, $f\,l_i^t$ refers to the flying distance of crow i, AP is the awareness probability of crow j.
GWO	When hunting, the grey wolves will hunt according to the comprehensive information given by wolves in high hierarchy.	$x_i^{t+1}=\sum((x_b^t)_u-A_u\cdot\lvert C_u\cdot(x_b^t)_u-x_i^t\rvert)$	A and C are two coefficient vectors, $u=\alpha, \beta, \delta$ represents the index of leader wolves.
HHO*[a]	In a group predation, Harris eagles will adjust the hunting strategy constantly to confuse the prey and make it exhausted, and then carry out besiege on it.	$x_i^{t+1}=(x_b^t-x_i^t)-E\lvert J\cdot x_b^t-x_i^t\rvert,\ (0.5\leq\lvert E\rvert<1,\ r\geq0.5)$ $x_i^{t+1}=x_b^t-E\lvert x_b^t-x_i^t\rvert,\ (\lvert E\rvert<0.5,\ r\geq0.5)$	Where, J represents the random jump strength of the escaping prey, E is the energy of a prey.
MRFO	There are three foraging strategies of manta rays: chain foraging, cyclone foraging and somersault foraging.	$x_i^{t+1}=\begin{cases}x_i^t+r(x_b^t-x_i^t)+\theta(x_b^t-x_i^t), & i=1\\ x_i^t+r(x_{i-1}^t-x_i^t)+\theta(x_b^t-x_i^t), & i\geq2\end{cases}$ $x_i^{t+1}=\begin{cases}x_s^t+r(x_s^t-x_i^t)+\tau(x_s^t-x_i^t), & i=1\\ x_s^t+r(x_{i-1}^t-x_i^t)+\tau(x_s^t-x_i^t), & i\geq2\end{cases}$ $x_i^{t+1}=x_i^t+K(r_2\cdot x_b^t-r_3\cdot x_i^t)$	θ and τ are weight coefficients, x_s^t has x_r^t and x_b^t two states, K is the somersault factor.
SMA	Slime mold generates veins to approach the food source. Larger food can increase the cytoplasmic flow through the vein, make it thicker. Then the best food position can be located.	$x_i^{t+1}=\begin{cases}\text{rand}\cdot(ub-lb)+lb, & \text{rand}<z\\ x_b^t+vb\cdot(W\cdot x_A^t-x_B^t), & r<p\\ vc\cdot x_i^t, & r\geq p\end{cases}$	Where, vb and vc are parameters imitating biological mechanism. x_A^t and x_B^t are two random individuals. p is a probability related to fitness, z is a controlling parameter, W represents the weight of slime mould.
SSA	Based on the position of the leader, salps form a chain to search food with each individual updating the position according to the previous one's position.	$x_i^{t+1}=\begin{cases}x_b^t+H_1((ub-lb)H_2+lb), & H_3\geq0.5\\ x_b^t-H_1((ub-lb)H_2+lb), & H_3<0.5\end{cases}$ $x_i^{t+1}=\frac{1}{2}(x_i^t+x_{i-1}^t),\ i\geq2$	Where x_1^t is the position of the leader, H_1 is a coefficient related to iterations、H_2 and H_3 are random numbers uniformly generated in [0,1].
WOA	Humpback whales' foraging behavior is called bubble-net feeding method which consists of shrinking encircling mechanism and spiral updating position	$x_i^{t+1}=\begin{cases}x_b^t-A\cdot\lvert C\cdot x_b^t-x_i^t\rvert, & p<0.5\\ \lvert x_b^t-x_i^t\rvert\cdot e^{bl}\cdot\cos(2\pi l)+x_b^t, & p\geq0.5\end{cases}$	Where, A and C are coefficient vectors, b is a constant for defining the shape of the logarithmic spiral, l is a random number in $[-1, 1]$, p is a random number in [0, 1].

Note: More equations of position updating of HHO can be found in Reference [23]. BOA: Butterfly Optimization Algorithm; CSA: Crow Search Algorithm; GWO: Grey Wolf Optimization; HHO: Harris Hawks Optimization; MRFO: Manta Ray Foraging Algorithm; SMA: Slime Mould Algorithm; SSA: Salp Swarm Algorithm; WOA: Whale Optimization Algorithm.

3 Results and analysis

3.1 Characteristic wavelength selection results of SI algorithms

The above eight SI algorithms were used to analyze the pretreated spectral data. $RMSE_P$ was taken as the fitness standard, the initial population was set as 30, the maximum number of iterations was set as 200, and the position threshold was set as 0.95 (it means the variable closer to the best value would be selected). The iterations results of different algorithms are shown in Figure 5.



Note: BOA: Butterfly Optimization Algorithm; CSA: Crow Search Algorithm; GWO: Grey Wolf Optimization; HHO: Harris Hawks Optimization; MRFO: Manta Ray Foraging Algorithm; SMA: Slime Mould Algorithm; SSA: Salp Swarm Algorithm; WOA: Whale Optimization Algorithm.

Figure 5 Optimal individual fitness curve of different algorithms

Figure 5 shows the optimal individual fitness curves of those algorithms. The convergence accuracy and convergence speed of different algorithms are significantly distinguishing. BOA converges the earliest and has reached the optimal fitness of 0.0049 kg/kg in the third iteration, but its convergence accuracy is poor which is a bit better than CSA. The convergence accuracy of CSA is the worst, which is 0.0054 kg/kg. The convergence of SSA is the most prominent which changes greatly in the 114th time and slightly improves around the 160th time, reaching 0.0040 kg/kg. MRFO owns the highest convergence accuracy, and its optimal fitness is 0.0039 kg/kg. HHO and WOA also appear good convergence accuracy.

For the purpose of exploring whether the new SI algorithms have a better performance than the traditional intelligence methods in characteristic wavelength selection, GA and PSO were also programmed in this study to analyze the preprocessed spectral data. Figure 6 depicts the distribution of wavelengths selected by the 10 algorithms in studied spectral range. It can be seen that the results of wavelengths selected by different algorithms can be roughly classified into four situations by the number of selected wavelengths and the extent of dispersion. MRFO, WOA and PSO have large number of selected wavelengths and their distribution is relatively uniform, so it is difficult to find out obvious concentrated bands. The number of wavelengths selected by HHO is relatively large, with obvious piecewise concentrated distribution. The number of wavelengths selected by SMA, SSA and GA is relatively small and dispersed, and some wavelengths are clustered but not quite obvious. The number of selected wavelengths of BOA, CSA and GWO is small and sparsely distributed and there is no significant aggregation band, but it can be clearly seen that the wavelength selection of BOA and CSA is around 590-650 nm (the orange region), CSA in about 910-970 nm (the NIR region), and GWO in 405-460 nm (the purple and blue region), appear a distinct blank.

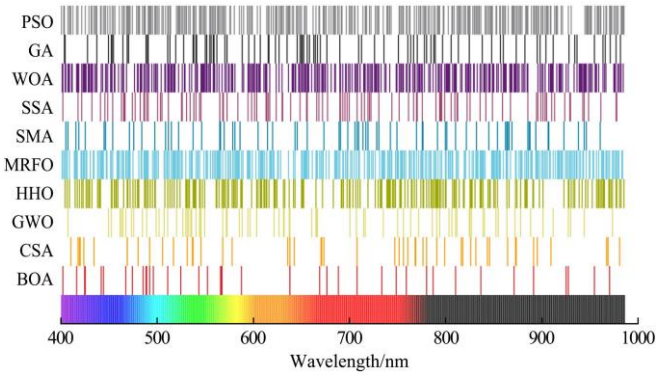


Figure 6 Distribution of wavelengths selected by each algorithm

3.2 Modeling results of the selected wavelengths

In order to compare the performance of different algorithms in predicting SMC, BPNN model using the characteristic wavelengths selected by each algorithm was built and cross-validation was carried out in this study. Other evaluation parameters except $RMSE_P$ could be obtained. Since the initial weights and biases of BPNN are generated randomly, the modeling results varied after each run. The selected results of each algorithm were modeled 20 times and the minimum value among the 20 results was taken as the final modeling result (which is slightly different from the result in Figure 5), as listed in Table 4.

Table 4 Modeling results of wavelengths selected by each algorithm

Algorithms	Training set		10-fold cross validation		Test set	
	R_T^2	$RMSE_T$ /kg kg^{-1}	R_{CV}^2	$RMSE_{CV}$ /kg kg^{-1}	R_P^2	$RMSE_P$ /kg kg^{-1}
BOA	0.9950	0.0056	0.9799	0.0105	0.9914	0.0059
CSA	0.9940	0.0062	0.9812	0.0102	0.9914	0.0059
GWO	0.9951	0.0056	0.9889	0.0081	0.9921	0.0056
HHO	0.9952	0.0055	0.9893	0.0080	0.9937	0.0050
MRFO	0.9963	0.0049	0.9883	0.0079	0.9938	0.0050
SMA	0.9947	0.0058	0.9834	0.0100	0.9935	0.0051
SSA	0.9925	0.0069	0.9865	0.0088	0.9929	0.0053
WOA	0.9956	0.0053	0.9873	0.0087	0.9937	0.0050
GA	0.9943	0.0060	0.9865	0.0089	0.9898	0.0064
PSO	0.9971	0.0043	0.9840	0.0094	0.9893	0.0065

It can be seen from Table 4 that among eight new SI algorithms, MRFO is better than other algorithms in terms of training effect and prediction effect. It has the highest R_T^2 and R_P^2, and the lowest $RMSE_T$ and $RMSE_P$ (0.0049 kg/kg and 0.0050 kg/kg, respectively), which reveals its significant advantage in model prediction ability. And the cross-validation result of MRFO is also at a preeminent level whose $RMSE_{CV}$ is the lowest and R_{CV}^2 is second only to HHO. The R_P^2 and $RMSE_P$ of WOA and HHO are both 0.9937 and 0.0050 kg/kg which is slightly worse than that of MRFO. The R_P^2 of BOA and CSA are evidently lower than that of other algorithms while their $RMSE_P$ are higher than that of other algorithms, which are both 0.9914 and 0.0059 kg/kg, respectively. The cross-validation result of BOA is the worst among all algorithms (R^2_{CV} is 0.9799 and $RMSE_{CV}$ is 0.0105 kg/kg), followed by CSA. The R_T^2 and $RMSE_T$ of CSA are gently better than that of the worst one SSA. In general, BOA and CSA have poor performance in all three types of evaluation standards of the model.

From the perspective of model training results, HHO and GWO are slightly inferior to MRFO and WOA, but their R_T^2 and

$RMSE_T$ are at an excellent level which is followed by BOA and SMA. In terms of the cross-validation effect of the model, GWO is next only to MFRO and HHO while $RMSE_{CV}$ of BOA, CSA and SMA are all at a poor level. R_{CV}^2 and $RMSE_{CV}$ of SSA are relatively good. Evaluating by prediction ability, SMA and SSA are relatively better while GWO is worse in comparison. According to the model training effect, cross-validation effect, and prediction effect, the model results of wavelengths selected by new SI algorithms roughly appear a breakdown as follows. MRFO has the best modeling effect. HHO and WOA have relatively better modeling effects. The modeling effects of GWO, SMA and SSA are relatively poor. BOA and CSA modeling are the worst. In conclusion, judging by the modeling effect, MRFO has a better performance in SMC VIS-NIR wavelength selection than other SI algorithms.

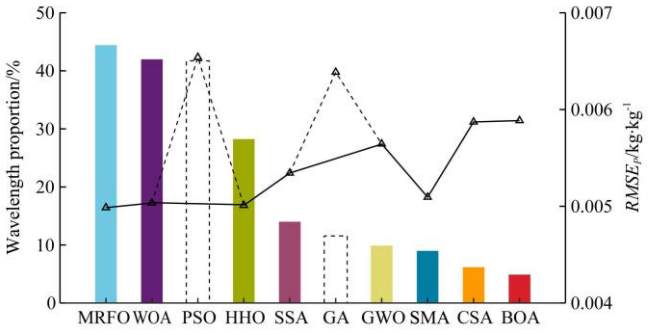
With regard to the traditional intelligent algorithm, the model training effect of GA is only better than that of SSA and CSA (R_T^2 and $RMSE_T$ are 0.9943 and 0.0060 kg/kg, respectively). There is a certain gap with other new algorithms. The R_T^2 and $RMSE_T$ of PSO are 0.9971 and 0.0043 kg/kg, which are the best among all algorithms. However, the R_{CV}^2 and $RMSE_{CV}$ of GA and PSO are merely slightly preferable to BOA, CSA, and SSA, which means they own relatively poor cross-validation ability. The $RMSE_P$ of GA and PSO are 0.0064 kg/kg and 0.0065 kg/kg, respectively, which are remarkably inferior to the results of the new SI algorithms. This demonstrates that, compared with the traditional intelligent algorithms GA and PSO, the new SI algorithms have more advantages in model stability and prediction ability when selecting characteristic wavelength of SMC of VIS-NIR.

4 Discussion

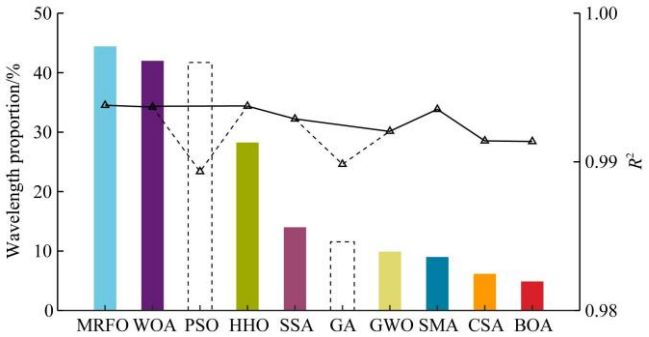
4.1 Modeling effect and selected wavelength proportion

The selected wavelength proportion refers to the percentage of the number of wavelengths extracted in the number of full-spectrum wavelengths. As it is known that the process of selecting characteristic wavelength is also a process of minimizing spectral redundancy information. The higher the proportion of wavelength is, the more variable information it covers, which means the redundant information is not be effectively reduced. For the more eligible characteristic wavelength selection method, redundant information should be minimized but the original spectral information should be retained to the maximum extent simultaneously. Generally speaking, better modeling accuracy is expected to be achieved with less wavelength, that is, redundant variables should be removed to the maximum extent. Hence it is necessary to comprehensively consider both the number of selected wavelengths and the modeling effect to choose the more appropriate characteristic wavelength selection method. Figure 7 shows the relationship between the prediction results of the model and the proportion of selected wavelength of the total of 10 algorithms.

It is clear from Figure 7 that both the R_P^2 and the $RMSE_P$ of MRFO model are at the optimal level while its selected wavelength also accounts for the largest proportion which reaches nearly 44.54%. In other words, half of the wavelengths of the whole is adopted to achieve the optimal model. And in combination with Figure 6, it can be seen that these wavelengths are evenly distributed in the whole spectrum range without obvious aggregation. On the one hand, excessive selected characteristic wavelengths may indicate that the data dimensionality reduction is not well realized, which may be not conducive to improving the portability of calculation. On the other hand, the selection of multiple and scattered wavelengths makes no difference to the determination of sensitive wavelengths or wavebands, which is not quite useful in practical application. It can be said that the excellent modeling effect of MRFO is actual the sacrifice of the limitation on the number of characteristic wavelengths, indicating that it doesn't achieve considerable redundant data reduction. The wavelength proportion of BOA and CSA are 5.01% and 6.29%, respectively. According to the foregoing analysis, both their prediction correlation coefficient and root mean square error are at the worst level. In fact, it can be seen from Figure 7 that, for the new SI algorithms, except for SMA, with the decrease of wavelength proportion, the model R_P^2 generally appears a downward trend, and the model $RMSE_P$ generally shows an upward trend. However, it is obvious that the variation of R_P^2 and $RMSE_P$ is much smaller than that of the wavelength proportion. For instance, $RMSE_P$ of BOA is 18.05% higher compared with MRFO, but its percentage of selected wavelength is 88.75% lower than the optimal one MRFO. In addition, what stands out in Figure 7 is that SMA seems like an anomaly along with the tendency. Its wavelength proportion is 9.11%, which is only approximately 1/5 of the optimal MRFO, but its R_P^2 is only 0.03% lower than MRFO and $RMSE_P$ is only 2.18% higher than MRFO.



a. Comparison between the wavelength proportion and model $RMSE_P$ of each algorithm



b. Comparison between the wavelength proportion and model R_P^2 of each algorithm

Note: The dotted line represents the results of traditional intelligence algorithms GA and PSO.

Figure 7 Comparative analysis of modeling results of selected wavelength and their proportion of all wavelength by using different algorithms

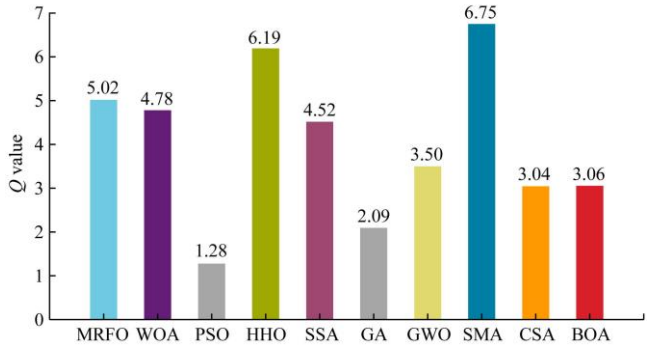
In order to better discuss the relationship between model prediction effect and wavelength proportion, this study defines Equation (3) as follows to illustrate.

$$Q = \frac{1 - \mathrm{wp}}{(\mathrm{RMSE_{PB}} - \mathrm{RMSE_{PA}}) / \mathrm{RMSE_{PA}}} \quad (3)$$

where, $RMSE_{PA}$ is the predicted root mean square error of the full-spectrum modeling after pretreatment in Table 2, kg/kg; $RMSE_{PB}$ is the predicted root mean square error of the optimized

wavelength modeling by each algorithm, kg/kg; wp is the percentage of selected wavelengths.

The numerator of Q represents the loss of number between the full spectrum and the selected characteristic wavelengths. The denominator of Q infers the reduction of prediction effect of selected wavelengths model and of full wavelengths model. The larger Q is, the better the ability to bear the wavelength proportion and the model optimization effect is. The smaller Q is, the worse the ability to consider selected wavelengths number and model prediction effect is. Figure 8 is the bar chart of the Q values of the total 10 algorithms. It is noticeable that the Q value of SMA is the highest, which is 6.75, followed by HHO. Apparently, when considering the proportion of wavelength, the Q value of MRFO, the one with the best modeling effect, is worse than the first two. So, it can be concluded that by comprehensively considering the percentage of selected wavelength and ability to improve model prediction effect, SMA shows a preferable performance in SMC VIS-NIR characteristic wavelength selection among all the new SI algorithms. Although MRFO may have a better modeling effect, it does not remove the redundant information very well.



Note: The larger the Q is, the better the ability to bear the wavelength proportion and the model optimization effect is. The smaller Q is, the worse the ability to consider selected wavelengths number and model prediction effect is.

Figure 8 Q values of all algorithms

To compare the new SI algorithms with the traditional intelligence algorithm GA and PSO which are plotted by the dotted line in Figure 7. It is easy to see that R_P^2 and RMSE_P of GA and PSO far deviate from the overall trend of new types of SI algorithms. And their Q values are 2.09 and 1.28, respectively, which are remarkably lower than all the new SI algorithms. Not only with relatively poor model prediction ability, but they also failed to well minimize redundant information. It further illustrates the relatively poor feasibility of GA and PSO in SMC VIS-NIR characteristics wavelength extraction.

4.2 Determination of SMC sensitive wavelengths

The design of soil moisture sensor based on optics needs to select the light source according to the wavelength sensitive to soil moisture in VIS-NIR. However, the above optimal SI algorithm SMA selected 71 characteristic wavelengths, and even the least number of characteristic wavelengths selected by BOA reaches 39. If these characteristic wavelengths are used to design the soil moisture sensor, it is certainly not conducive to the determination of the light source. To select the SMC sensitive wavelengths and make full use of the results of tall algorithms mentioned before, this paper counted the times of each wavelength selected by different algorithms, as shown in Figure 9. One was selected as the characteristic wavelength by various algorithms, indicating that its sensitivity to SMC is quite subtle. Therefore, it can be considered that the more times it is selected, the more sensitive the wavelength is to SMC.

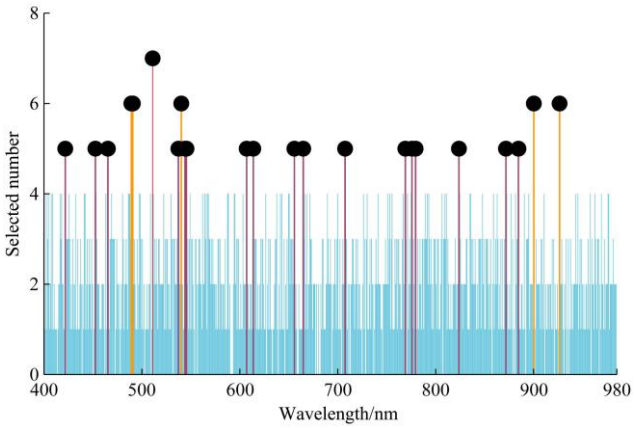


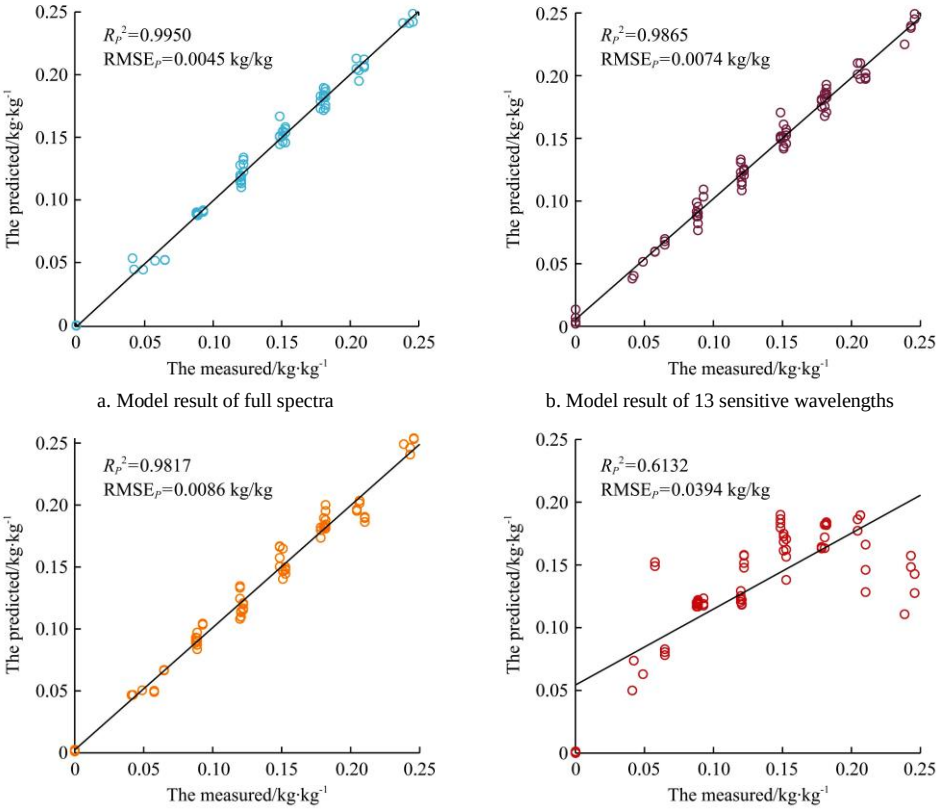
Figure 9 The counted number of selected wavelengths

As depicted in Figure 9, 13 wavelengths (wavelengths closely distributed are counted as one wavelength) have been selected more than or equal 5 times, respectively around 422 nm, 453 nm, 490 nm, 513 nm, 543 nm, 611 nm, 669 nm, 711 nm, 778 nm, 826 nm, 880 nm, 900 nm and 926 nm. These wavelengths roughly include the sensitive wavelengths selected within 400-980 nm in Reference [9] (450 nm, 574 nm), Reference [10] (443-449 nm), and Reference [11] (411 nm, 440 nm, 622 nm, 713 nm, 790 nm), which demonstrates the feasibility of comprehensively using a variety of SI algorithms to conduct SMC sensitive wavelength selection. The slight distinction among them may be caused by the differences in soil type of different regions.

Among the above 13 selected wavelengths, 513 nm was selected the most and was selected for 7 times. 490 nm, 543 nm, 900 nm, and 926 nm were selected 6 times. In this study, the wavelengths with more selected times by the algorithms are used as the sensitive wavelengths for modeling analysis. The modeling results are shown in Figure 10. R_P^2 can reach 0.9865 and RMSE_P 0.0074 kg/kg when 13 sensitive wavelengths are used to build models. The R_P^2 and RMSE_P of the 5 wavelengths (selected 6 times or more) models are 0.9817 and 0.0086 kg/kg, respectively. The R_P^2 and RMSE_P of model of 1 wavelength that was selected 7 times are only 0.6132 and 0.0394 kg/kg, respectively. Obviously, even though 513 nm was selected the most times, the prediction result of SMC is not good when just using a single wavelength to model. In Figure 10, it also can be seen that compared with the model result of full spectra, the modeling results of using 13 wavelengths and 5 wavelengths are not bad. Although their RMSE_P may loss to some extent, there is still a considerable linear correlation between the measured and the predicted. For the development of soil moisture sensor, certainly, the wavelengths should be selected according to the requirement of accuracy. In fact, the model built with 5 wavelengths can be able to achieve a satisfying prediction effect, which can adequately meet the need of the SMC measurement. Therefore, it can be considered that the sensitive wavelengths suited for SMC determination are 490 nm, 513 nm, 543 nm, 900 nm and 926 nm within the range of 400-980 nm.

5 Conclusions

Based on the laboratory spectral experiment of SMC and SI algorithms was carried out on the characteristics wavelength selection to explore the feasibility of some latest SI algorithms for SMC spectral analysis. At the same time, through the integration of a variety of algorithms to select the sensitive wavelengths of SMC, this study provides a theoretical basis for SMC measurement based on optical principle. The specific conclusions are as follows:



a. Model result of full spectra

b. Model result of 13 sensitive wavelengths

c. Model result of 5 sensitive wavelengths

d. Model result of 1 sensitive wavelength

Figure 10 Modeling results of different numbers of sensitive wavelengths

1) Using the latest anomalies mining method iForest to remove the outliers of the spectrum and SG filter to smooth the spectrum, the pretreatment has greatly improved the results of both PLSR and BPNN prediction model of SMC.

2) Modeling and analysis of the characteristic wavelengths selected by eight new SI algorithms, BOA, CSA, GWO, HHO, MRFO, SMA, SSA, and WOA, the results indicate that if only judging from model effect, MRFO is the best SI algorithm to select characteristic wavelength of SMC; if considering both the model effect and the ability of decrease redundant variables, SMA shows superior performance in SMC VIS-NIR characteristic wavelength selecting. It is also found that compared with the traditional intelligence algorithms GA and PSO, the new SI algorithms have more advantages in the selection of characteristic wavelengths of SMC.

3) Integrating the selected results of the total 10 algorithms, it is found that modeling with 5 sensitive wavelengths can achieve considerable prediction accuracy. The sensitive wavelengths are 490 nm, 513 nm, 543 nm, 900 nm, and 926 nm, respectively, which provides theoretical support to the development of a real-time SMC sensor based on VIS-NIR.

Acknowledgements

This study was financially supported by the National Natural Science Foundation of China (Grant No. 32071915), and China Agriculture Research System of MOF and MARA-Food Legumes (CARS-08).

[References]

[1] Mouazen A M, Baerdemaeker J D, Ramon H. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. Soil & Tillage Research, 2004; 80(1): 171–183.

[2] Li M, Zheng L, An X, Sun H. Fast measurement and advanced sensors of soil parameters with NIR spectroscopy. Transactions of the Chinese Society for Agricultural Machinery, 2013; 44(3): 73–87. (in Chinese)

[3] Soriano-Disla J M, Janik L J, Viscarra Rossel R A, Macdonald L M, McLaughlin M J. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Applied Spectroscopy Reviews, 2014; 49(2): 139–186.

[4] Stenberg B, Viscarra Rossel R A, Mouazen A M, Wetterlind J. Visible and near infrared spectroscopy in soil science. Advances in Agronomy, 2010; 107: 163–215.

[5] Zhang B. Advancement of hyperspectral image processing and information extraction. Journal of Remote Sensing, 2016; 20(5): 1062–1090. (in Chinese)

[6] Yang L, Xu R, Lei T, Li J, Ouyang T. Design of near-infrared soil moisture measuring instrument. Transactions of the CSAE, 2015; 31(20): 1–9. (in Chinese)

[7] Peng Z, Yao Z, Wei Y, Li M, Zhen L, Liu X. Development and performance test of an in-situ soil total nitrogen-soil moisture detector based on near-infrared spectroscopy. Computers and Electronics in Agriculture, 2019; 160: 51–58.

[8] Balabin R M, Smirnov S V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. Analytica Chimica Acta, 2011; 692(1): 63–72.

[9] Liu W, Bareta F, Gu X, Tong Q, Zheng L, Zhang B. Relating soil surface moisture to reflectance. Remote Sensing of Environment, 2002; 81: 238–246

[10] Yu L, Zhu Y, Hong Y, Xia T, Liu M, Zhou Y. Determination of soil moisture content by hyperspectral technology with CARS algorithm. Transactions of the CSAE, 2016; 32(22): 138–145. (in Chinese)

[11] Wu L, Wang S, He J. Study on soil moisture mechanism and establishment of model based on hyperspectral imaging technique. Spectroscopy and Spectral Analysis, 2018; 38(8): 2563–2570. (in Chinese)

[12] Bin J, Fan W, Zhou J, Li X, Liang Y. Application of intelligent optimization algorithms to wavelength selection of near-infrared spectroscopy. Spectroscopy and Spectral Analysis, 2017; 37(1): 95–102. (in Chinese)

[13] Xue L, Cai J, Li J, Liu M H. Application of particle swarm optimization

(PSO) algorithm to determine dichlorvos residue on the surface of navel orange with Vis-NIR spectroscopy. Procedia Engineering, 2012; 29: 4124–4128.

[14] Cheng J H, Sun D W, Pu H B. Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen-thawed fish muscle. Food Chemistry, 2016; 197: 855–863.

[15] Mavrovouniotis M, Li C H, Yang S X. A survey of swarm intelligence for dynamic optimization: Algorithms and applications. Swarm and Evolutionary Computation, 2017; 33: 1–17.

[16] Lin S, Dong C, Chen M, Zhang F, Chen J. Summary of new group intelligent optimization algorithms. Computer Engineering and Applications, 2018; 54(12): 1–9. (in Chinese)

[17] Fei T L, Kai M T, Zhi-Hua Z. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012; 6(1): 1–39.

[18] Mouazen A M, Kuang B, De Baerdemaeker J, Ramon H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma, 2010; 158(1): 23–31.

[19] Arora S, Singh S. Butterfly optimization algorithm: a novel approach for global optimization. Springer Berlin Heidelberg, 2019; 23(3): 715–734.

[20] Askarzadeh A. A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. Computers and Structures, 2016; 169: 1–12.

[21] Seyedali M, Seyed M M, Andrew L. Grey wolf optimizer. Advances in Engineering Software, 2014; 69: 46–61.

[22] Seyedali M, Shahrzad S, Seyed M M, Leandro dos S.C. Multi-objective grey wolf optimizer: A novel algorithm for multi-criterion optimization. Expert Systems with Applications, 2016; 47: 106–119.

[23] Heidari A A, Mirjalili S, Faris H, Aljarah I, Mafarja M, Chen H L. Harris hawks optimization: Algorithm and applications. Future Generation Computer Systems, 2019; 97: 849–872.

[24] Zhao W G, Zhang Z X, Wang L Y. Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications. Engineering Applications of Artificial Intelligence, 2020; 87: 103300. doi: 10.1016/j.engappai.2019.103300.

[25] Li S M, Chen H L, Wang M J, Heidari A A, Mirjalili S. Slime mould algorithm: A new method for stochastic optimization. Future Generation Computer Systems, 2020; 111: 300–323.

[26] Seyedali M, Amir H G, Seyedeh Z M, Shahrzad S, Hossam F, Seyed M M. Salp swarm algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software, 2017; 114: 163–191.

[27] Chen T, Wang M, Huang X. Time difference of arrival passive location based on salp swarm algorithm. Journal of Electronics & Information Technology, 2018; 40(7): 1591–1597. (in Chinese)

[28] Seyedali M, Andrew L. The whale optimization algorithm. Advances in Engineering Software, 2016; 95: 51–67.

[29] Long W, Cai S, Jiao J, Tang M, Wu T. Improved whale optimization algorithm for larger scale optimization problems. Systems Engineering-Theory & Practice, 2017; 37(11): 2983–2994. (in Chinese)