Soil organic matter determination based on artificial olfactory system and PLSR-BPNN

OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. **32** (2021) 035801 (14pp)
https://doi.org/10.1088/1361-6501/abc964

Soil organic matter determination based on artificial olfactory system and PLSR-BPNN

Dongyan Huang[1,2,*], He Liu[1,2], Longtu Zhu[3], Mingwei Li[1,2], Xiaomeng Xia[1,2] and Jiangtao Qi[1,2,*]

[1] Key Laboratory of Bionic Engineering, Ministry of Education, Jilin University, Changchun 130022, People's Republic of China
[2] College of Biological and Agricultural Engineering, Jilin University, Changchun 130022, People's Republic of China
[3] College of Engineering, Huazhong Agricultural University, Wuhan 430070, People's Republic of China

E-mail: Huangdy@jlu.edu.cn and qijiangtao@jlu.edu.cn

Received 2 June 2020, revised 26 October 2020
Accepted for publication 11 November 2020
Published 17 December 2020

Abstract

Soil organic matter (SOM) is a key indicator of soil fertility. For accurate measurement of SOM, a novel method based on an artificial olfactory system (AOS) was proposed. The response curves of soil volatile organic compounds (VOCs) were measured using a metal-oxide semiconductor sensor array, and four features (including maximum value, mean differential coefficient, response area, and the transient value at the 20th second) were obtained from the curves and used to build olfactory feature space. Then, prediction models were established using the pattern recognition algorithm. To further enhance the accuracy of AOS measurement, we used Monte Carlo cross-validation (MCCV) to identify and eliminate the abnormal samples of the soil olfactory feature space. Then, the dimension reduction method of the genetic algorithm (GA)back-propagation (BP) was used to find the appropriate feature vectors, and two types of hybrid models were presented. One was the support vector machine (SVM) and group method of data handling (GMDH) combined model—SVM-GMDH. The other was a combination of partial least squares regression (PLSR) and back-propagation neural network (BPNN)—PLSR-BPNN. The forecasting performances of three single models (BPNN, PLSR, support vector regression: SVR) and two combined models (PLSR-BPNN, SVM-GMDH) were comparatively evaluated. The evaluation indices included coefficient of determination (R^2), root mean square error (RMSE), ratio of performance to deviation and relative prediction error (RPE). It was found that the predictive capabilities of all five tested models were improved after elimination of abnormal samples and feature reduction. Moreover, PLSR-BPNN performed the best in predicting SOM concentrations, with $R^2 = 0.952$, RMSE = 1.771, PRD = 4.291, and slight variation of RPE within 0–0.185, and thus can offer a reference for predicting SOM via AOS.

Keywords: artificial olfactory system, soil organic matter, Monte Carlo cross-validation, hybrid model, feature optimization

(Some figures may appear in colour only in the online journal)

* Authors to whom any correspondence should be addressed.

1. Introduction

Soil organic matter (SOM) is a critical property of soils [1] and contributes to soil physical property improvement, plant growth, and crop production [2]. SOM is a key evaluator of soil fertility, and loss of SOM reflects a decline in soil quality [1, 3]. Accurate determination of SOM variation is critical for guidance in crop fertilization and soil quality improvement.

SOM concentrations are usually measured by chemical detection and analysis of soil samples collected in the field, but this method is limited by high time/labor consumption, high costs and destructiveness [4]. Thus, new fast, economical and nondestructive methods for precise prediction of SOM concentrations are increasingly in demand [5]. In recent years, owing to the universal application of proximal soil remote sensing, visible and near-infrared diffuse reflectance spectroscopy (Vis–NIR DRS) has attracted increasing interest among soil scientists and has been considered feasible for soil analysis [6–8]. For instance, Conforti *et al* predicted the spatial variation of SOM using laboratory-based Vis–NIR spectroscopy [9]. Nawar *et al* used different Vis–NIR DRS spectra to detect clay and SOM concentrations [10]. Despite its high accuracy, Vis–NIR spectroscopy is limited by its susceptibility to changes in soil humidity [11], soil particle size [12] and iron oxide [13].

The generation and consumption of soil gases are mainly related to microbic activities in soils [14, 15]. SOM is the major substrate of nutrients and energy needed by the vital activities of soil microbes [16]. The substrate for nutrient and energy supply can generate abundant volatile organic compounds (VOCs) and gases during microbial degradation [15]. This means that the VOCs and gases in soils inevitably correlate with SOM [17]. Such correlation facilitates the fast and low-cost detection of SOM. Gas detection can be achieved at very low costs, especially with methods based on solid-state chemical sensors [18]. However, soil gases are compositionally complex [19] and difficult to identify with a single gas sensor. An artificial olfactory system (AOS, also called electronic nose or e-nose) consisting of non-selective sensor arrays and a pattern recognition model is considered as an efficient means of detection of complex gases [18]. Though the AOS does not present any concrete information or property about volatile gaseous compounds [20], its combination with appropriate pattern recognition algorithms, like artificial neural networks (ANNs) or statistical methods, can identify the gas pattern of specific samples and separate them from other samples [21, 22]. So far, AOS has been extensively applied in foods [23–25], medicine [26, 27], diseases [28], the environment [29], beverages [30] and other fields [31–33]. AOS has also reportedly been used for soil characterization. For instance, Andrzej *et al* used an e-nose to assess soil humidity and research on ten moisture levels in ten types of soils, and indicated that it was a very promising tool [34]. Pobkrut *et al* integrated an e-nose and robots into detection of soil surface VOCs and thereby measured soil fertility [35]. However, there is little research on the use of an AOS in SOM measurement.

Although an AOS has the advantages of low cost, fast detection speed and being lossless [36], it also has some inevitable defects, which mainly reflect the construction of sensor arrays and the selection of pattern recognition algorithm. For different applications, sensor arrays are constructed differently, and selectivity, sensitivity and operating temperature should be considered comprehensively. For gases with known components to be measured, the most popular method is to select specific sensors to construct hybrid sensor arrays. However, the formation mechanism of soil gas is different, and its composition is very complex. It is difficult for a specific sensor to detect an uncertain gas mixture efficiently. Using the same type of non-specific sensor to construct a difference detection array by temperature control seems to be an effective measure to collect uncertain combined gas signals.

The pattern recognition algorithm is a key component of an AOS [37]. Commonly used pattern recognition algorithms include the back-propagation neural network (BPNN), support vector regression (SVR) and partial least squares regression (PLSR). The BPNN is the most widely used form of neural network, and has strong nonlinear mapping ability. SVR is proposed based on support vector machine (SVM) theory, which can effectively simplify the complexity of high-dimensional space. PLSR is particularly useful for predicting a group of dependent variables from a large number of independent variables, especially when there is a linear correlation between variables. These algorithms are widely used in soil property modeling. For example, based on Vis–NIR spectra, Ji *et al* compared the performances of PLSR and SVR in predicting soil information *in situ* in rice fields [38]. Qi *et al* preprocessed the hyperspectral Vis–NIR data of 153 soil samples using different methods and assessed the ability of PLSR, SVR and BPNN in predicting soil available nutrients, including nitrogen (N), phosphorus (P) and potassium (K) [39]. Santana *et al* used the NIR spectral technique and PLSR to minimize the effect of moisture on SOM detection [40]. Generally, all single pattern recognition algorithms are faced with inherent limitations [41]. For instance, BPNNs rely on abundant training data; the parameter selection of SVR is very difficult; and PLSR is not effective for nonlinear data. Guo *et al* also confirmed that no single prediction algorithm can be considered as superior [42]. To solve these problems, researchers have paid much attention to hybrid models in recent years. Compared with a single model, an appropriate combined model can yield more accurate results [43].

The hybrid model refers to the appropriate combination of different pattern recognition algorithms that can comprehensively utilize the useful information of single algorithms and thereby improves the prediction precision to the largest extent. Wen *et al* combined a gray model (GM) (1,1) and BPNN in grain yield prediction, in which the fluctuation of data series was weakened by the gray theory and the nonlinear handling ability of the BPNN was fully utilized [44]. Yin *et al* combined secondary hybrid decomposition, crisscross optimization and an extreme learning machine and thereby predicted wind power [45]. Wang *et al* combined a BPNN and genetic algorithm (GA) for wind speed forecasting [46]. The above studies suggest these combinations can yield better prediction results. The complex components of soil gases may correlate linearly or nonlinearly, or both, with SOM. PLSR is especially

sensors with the same type.

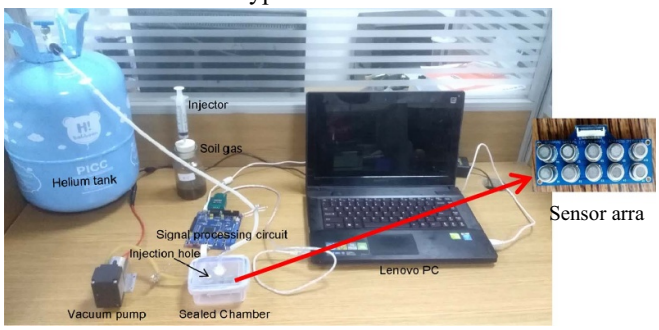


Figure 1. Structure of the AOS.

well suited when the characteristic matrix (or matrix of predictors) has more variables than observations, and when multicollinearity exists among variables in a characteristic matrix [47]. BPNNs, because of their strong nonlinear approximation ability, are widely used in nonlinear modeling [48]. Thus, a hybrid model combining PLSR and a BPNN can effectively utilize the advantages of both and improve the predictive precision. In this study, PLSR and a BPNN were combined for the first time in SOM detection. This combination is called PLSR-BPNN.

In our previous study, we mainly built a set of AOS for SOM, used a metal-oxide semiconductor (MOS) gas sensor array to detect the response curves of soil gases, extracted four features (V_{max}: maximum value, MDCV: mean differential coefficient value, RAV: response area value, and V_t: transient value at the 20th second), and thereby constructed a soil olfactory feature space (SOFS) [49]. Single prediction models were also assessed, including a BPNN, SVR and PLSR. In this work, we optimized the SOFS prior to the prediction modeling to further improve the SOM forecasting performance. Generally, the optimization processes included the elimination of abnormal samples and the dimension reduction of features. The abnormal samples mainly originated from misoperation, errors of the AOS, or temperature, humidity and other external factors. Abnormal samples largely reduced the precision of the prediction models. Thus, abnormal samples should be identified and eliminated. Monte Carlo cross-validation (MCCV) was confirmed as a very useful method to remove abnormal samples [50]. The dimension reduction of features is a key influencing factor on model performances [51], since the original feature space contains much redundant information unrelated to modeling. The use of an unoptimized feature space in modeling will enlarge the amount of calculation and decrease the precision of prediction. The frequently used dimension reduction methods are based either on statistics (e.g. principal component analysis (PCA)) or optimization (e.g. firefly algorithm, GA) [52]. GA-BP is a combination of a GA and BPNN, which is used as an optimization algorithm for dimension reduction to overcome the shortcoming of BPNN convergence to the local optimum. In this paper, on the basis of the artificial olfactory detection method of SOM, the optimization study of SOFS was carried out, and two new mixed prediction models were proposed to improve the accuracy of

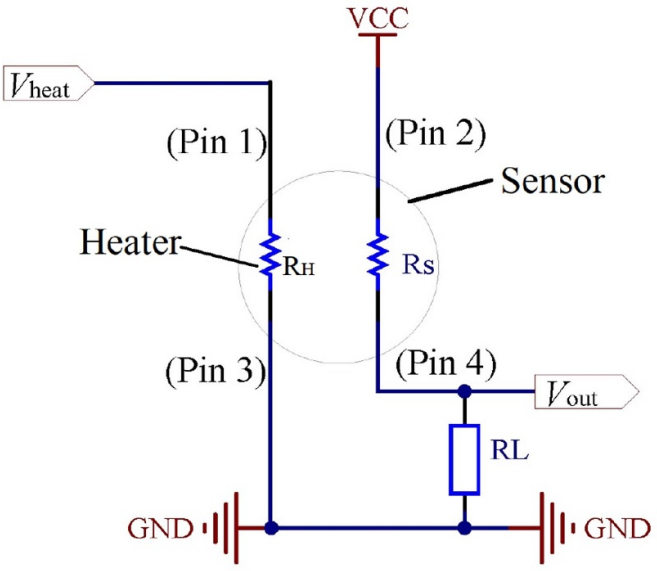


Figure 2. Basic measuring circuit of the sensor.

soil olfactory detection. The aims of this paper are: (a) to discuss the modeling effect of SOFS optimized by MCCV and GA-BP; and (b) to evaluate the performance of three single models (BPNN, SVR, PLSR) and two hybrid models (SVM-GMDH, PLSR-BPNN) for SOM prediction.

2. Materials and methods

2.1. Structure and working principle of AOS

A laboratory-based AOS was used to detect volatile soil gases. The AOS mainly consisted of a sensor array installed in a closed reaction chamber, a signal processing circuit, and a laptop PC (figure 1). The sensor array was a monoclass sensor array, which was composed of multiple sensors of the same type.

The signal processing circuit included multiple temperature modulation circuits and multiple basic measurement circuits. Each sensor in the sensor array corresponded to one temperature modulation circuit and one basic measurement circuit. The temperature modulation circuit was used to set the sensor's working temperature. The basic measuring circuit, as shown in figure 2, was used to collect the gas response signal.

VOCs during SOM degradation included gaseous hydrocarbons (CH_4, C_2H_4, C_2H_6, C_3H_8), H_2S, ammonia, aldehyde, etc [19]. Therefore, the sensor array was selected on the basis of their sensitivity to the VOCs in soil. In this study, ten gas sensors with the same type of IDT SGAS707, purchased from Integrated Device Technology Inc. (San Jose, CA, USA), were used to construct an array for the detection of VOCs in soil gas. The basic metrological parameters of the SGAS707 are shown in table 1. R_{Air}/R_{Gas} in table 1 represents the ratio of air response to VOC gas response.

These sensors were arranged in a 2 × 5 array at a line spacing of 20 mm and column spacing of 10 mm. Each sensor contained a resistance element (referred to as the heater) capable

Table 1. Basic metrological parameters of the sensor.

Sensitive gases	Detection concentration range	Range of sensor response	Range of sensor sensitivity (R_{Air}/R_{Gas})
VOCs, including: ethanols, formaldehyde, toluene, xylenes, acetone, isobutylene, octane, etc	1−1000 ppm	100−100 MΩ	1−100

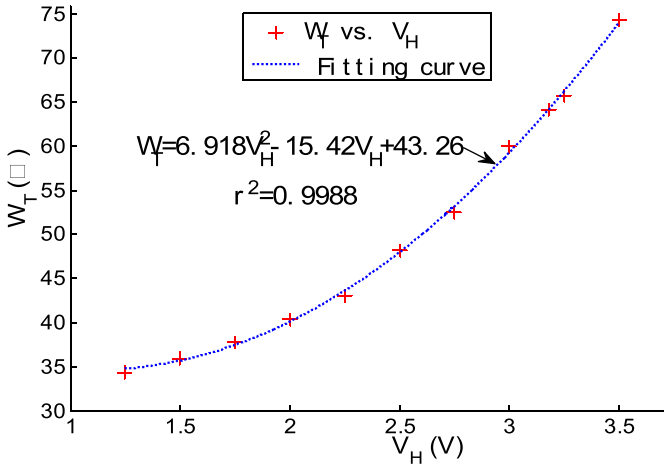


Figure 3. Relationship between heater voltage (V_H) and work temperature (W_T).

of modulating the working temperature, as shown in figure 2. The working temperature is controlled by the heater voltage (V_H). The selectivity of sensors can be enhanced by temperature modulation [53]. There are usually two temperature modulation modes: isothermal modulation and dynamic thermal modulation [54]. In our study, isothermal modulation was adopted. To set the working temperature of sensors, we needed to obtain the relationship between V_H and the sensor working temperature. Since the sensing material of the sensor is located in the metal protective casing, it is difficult to measure without damaging the sensor. Therefore, we used the temperature of the metal casing (W_T) instead of the temperature of the sensing material to carry out temperature modulation work. In this work, the relationship between W_T and V_H was calibrated by a PT1000 platinum resistance thermometer (precision class B) attached to the metal casing of the MOS gas sensor, as shown in figure 3. A polynomial was used to fit the calibration results of W_T and V_H, and the calculation formula (1) with a fitting degree (r^2) greater than 0.99 was obtained:

$$W_T = 6.918V_H^2 - 15.42V_H + 43.26 \quad (1)$$

In the design, the values of V_H for each sensor were set with a step of 0.25 V in a range from 1.25 V to 3.5 V (this range is limited to temperature modulation circuits [49]). The corresponding working temperatures of each sensor were 34.4 °C, 36.0 °C, 37.8 °C, 40.4 °C, 43.0 °C, 48.1 °C, 52.5 °C, 60.0 °C, 65.7 °C and 74.3 °C, respectively, and remained unchanged during the whole measurement process.

Prior to operation, a vacuum pump extracted inertia helium to rinse the test chamber. After the output of the sensor array stabilized, the rinsing was stopped and the pass valve was closed, so the test chamber was closed. After that, measurement was started. During the measurement, soil gases sealed in a 200 ml aluminum foil gas sampling bag were extracted by using a 20 ml injector, and then transferred via the injection hole to the test chamber. In the meantime, the sensor output signals processed by the signal processing circuit were collected at 10 Hz sampling frequency and then stored on hard disks. The sampling continued for 5 min.

It is necessary to verify the correctness (or selectivity and sensitivity) of the selected sensors before making large-scale measurements. Therefore, three typical soil gas samples with the maximum, moderate and minimum organic matter contents were selected from all soil gas samples to be tested for verification. The results show that these sensors have different responses to different soil gas samples and show great differences. This verifies that the selected sensors can realize SOM detection based on an artificial olfactory method. After that, we measured all the gas samples. More information about the verification test and data collection can be found in our previous study [49].

2.2. Dataset

The dataset (including a training set and a validation set) was cited from our previous study [49]. A total of 102 soil samples were collected in Jilin Province. The SOM concentration of each sample was measured by the potassium dichromate method and regarded as the observed value. The olfaction response curves of all samples were determined using an AOS device. The SOM concentrations are listed in table 2. The AOS-detected SOM concentrations were presented in a 102 × 40 olfaction feature space, which consisted of 102 samples and 40 features. The 40 features consisted of V_{max}, MDCV, RAV and V_t on the ten curves, and a boxplot of the feature data is illustrated in figure 4. The Si ($i = 1, 2, 3, \ldots, 10$) in figure 4 stood for the ten sensors. As can be seen from the figure, there were some obvious outliers in the attribute parameter, indicating the existence of abnormal samples in the data set. Therefore, it was necessary to remove the abnormal samples before establishing the prediction model.

2.3. Abnormal sample elimination

Abnormal samples can be produced by many factors such as unstable instrument status or imperfect operation. MCCV is proposed based on the basic assumption that the effect of an

Table 2. Organic matter concentrations in soil samples.

Dataset	SOM (g kg^{-1})	Sample number
Training set	20.51; 27.62; 33.50; 20.23; 23.11; 24.43; 28.71; 26.53; 18.88; 26.92; 14.97; 20.48; 17.69; 13.76; 17.38; 19.97; 32.13; 29.87; 28.85; 39.64; 12.37; 17.33; 14.22; 22.85; 15.49; 22.85; 25.27; 22.55; 18.13; 20.52; 25.20; 23.72; 13.44; 16.24; 15.67; 41.10; 22.31; 20.17; 13.29; 19.54; 35.55; 36.28; 43.85; 19.14; 25.42; 19.79; 13.79; 15.90; 30.71; 19.27; 23.16; 30.14; 24.76; 23.80; 27.95; 20.60; 22.88; 24.75; 23.46; 18.67; 35.38; 16.53; 15.32; 16.31; 16.74; 17.78; 22.89; 14.80; 29.65; 38.86; 19.75	1–71
Validation set	33.77; 12.19; 24.15; 25.11; 34.24; 21.32; 25.86; 18.94; 25.85; 25.10; 19.64; 25.94; 18.96; 17.58; 22.71; 21.50; 23.18; 38.92; 28.58; 48.79; 21.13; 28.62; 20.01; 17.78; 13.64; 21.28; 14.72; 19.37; 15.59; 15.71; 27.89	1–31

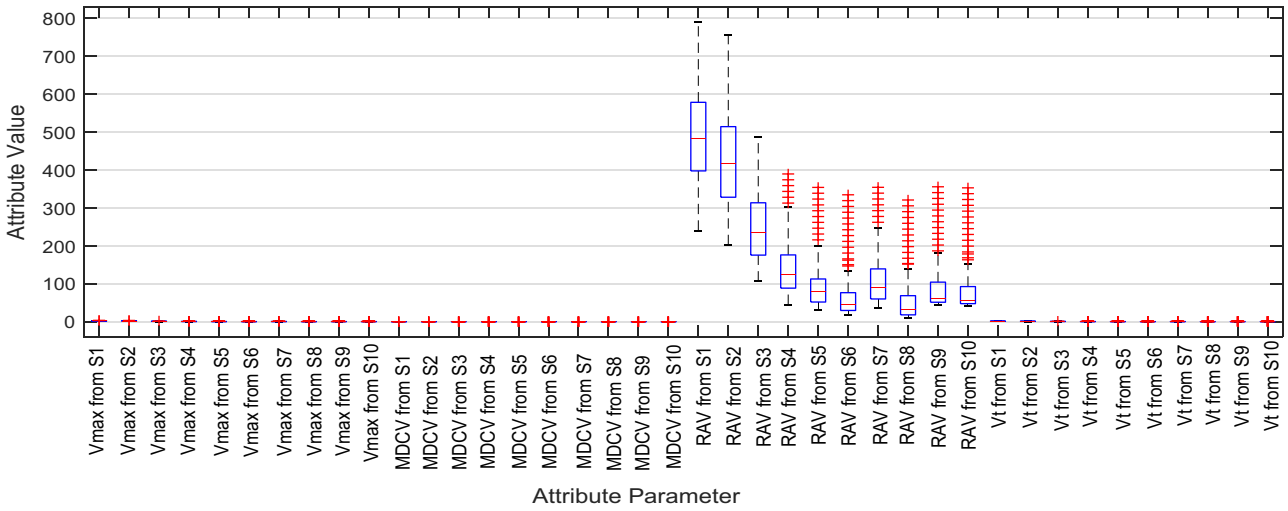


Figure 4. Boxplot of feature data of SOM.

outlier on the model will be different depending on whether it is selected in the calibration set or in the prediction set [50]. Because the outliers are unstable, they are not applicable for the models built based on the rest of the samples [54]. Thus, the outliers can be considered to be abnormal samples. In this study, the MCCV was used to identify the outliers in the SOFS. Firstly, from the training set, 80% of the samples were randomly selected for the establishment of PLSR models, and the remaining 20% were used for prediction. Secondly, the above process was repeated multiple times, so several PLSR models were built. Thirdly, the models were sorted in ascending order according to the predicted residual sum of squares (PRESS). Finally, the abnormal samples were identified according to the accumulative probability (f_{ac}). PRESS and f_{ac} were defined as follows:

$$\mathrm{PRESS} = \sum_{i=1}^{k} (\hat{y} - y_i)^2 \tag{2}$$

where $\hat{y}$ and y_i are the predicted value and observed value of the ith sample respectively, and k is the number of prediction samples.

$$f_{\mathrm{ac}}(m,n) = 100 \times \sum_{n=1}^{N} f_{mn}/N \tag{3}$$

where m is the ordinal number of a sample, and n is the index of ranked PLSR models. N is the number of training set samples, and f_{mn} indicates whether sample m in the randomly selected samples existed in the training set of model n: if yes, $f_{mn} = 1$; otherwise, $f_{mn} = 0$.

According to the definition, f_{ac} represented the probability of each sample appearing in good and bad models, since the models were sorted according to PRESS values. As n increased, the f_{ac} of normal samples was still kept around 80%, but the f_{ac} of abnormal samples deviated from those of normal samples.

2.4. Feature dimension reduction

GA is an evolutionary algorithm that simulates natural selection. Firstly, a population was randomly generated and after crossing, mutation and 'survival of the fittest' selection, the suitable individuals remained in the next generation until certain termination conditions were met [55]. To remove redundant information from the SOFS, we used GA-BP for dimension reduction. The concrete procedures are shown in figure 5.

(a) A population of size 20 was randomly generated, and the chromosome encoding length of each individual in this population was set as the dimension size of the original olfaction feature space (namely 40). Then, the chromosomes were binary-encoded so that each

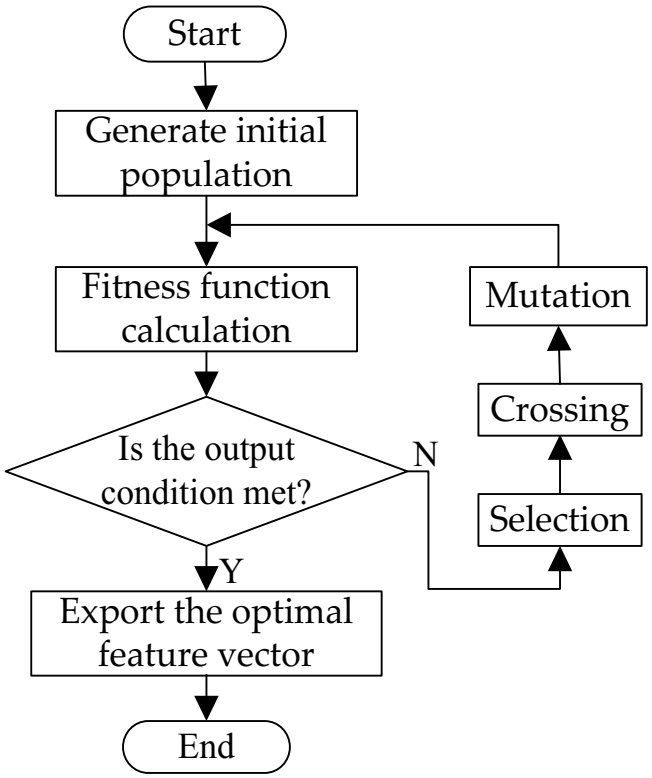


Figure 5. Flowchart of GA-BP.

locus gene on each chromosome corresponded to a feature vector. In each gene, '1' meant the feature vector participated in BPNN modeling, and 0 meant it did not take part in modeling. For instance, if the chromosome of a random genetic individual was encoded '0110010001001100010011001001101101000000', the corresponding feature vectors involved in modeling were 2, 3, 6, 10, 13, 14, 18, 21, 22, 25, 28, 29, 31, 32, 34.

(b) The feature vectors corresponding to the genetic individuals were selected and used in the BPNN model. The model was trained using the training set, and the reciprocal of the sum of squared errors was validated as the fitness function. Let the fitness function be $f(x)$; $\hat{y}_i$ is the predicted value of the ith sample in the validation set, and y_i is the observed value of the ith sample in the validation set. Then, $f(x)$ can be expressed as

$$f(x) = \frac{1}{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{4}$$

where n is the number of samples in the validation set.

(c) It was judged whether the relevant parameter (e.g. the value of $f(x)$ or the number of iterations) met the output condition. If so, the optimal feature vector was exported and the operation was stopped. Otherwise, the selection, crossing and mutation of GA were conducted, and then steps 2 and 3 were repeated until the output condition was met.

2.5. Single prediction models

As for prediction of soil properties, commonly used regression prediction algorithms include the BPNN, SVR, PLSR and ELM, etc [38, 39, 56]. Moreover, a BPNN, SVR and PLSR were also used in our previous research [49]. For comparison, these three algorithms were again adopted in this study as prediction models.

The BPNN is a multilayer forward neural network and is the most widely used neural network. Its topological structure consists of an input layer, either one or several hidden layers, and an output layer. The Kolmogorov theory proves that a three-layer network containing one hidden layer can approximate any nonlinear function. Thus, the number of hidden layers was set as 1 in this study. However, the BPNN is largely susceptible to the number of neurons in the hidden layer. In our work, the number of neurons in the hidden layer was optimized by the following empirical formulas:

$$h = \sqrt{n+p} + \alpha \tag{5}$$

where h, n and p are the numbers of hidden-layer neurons, input nodes and output nodes respectively, and α is a positive integer number from 1 to 10. In the BPNN model, the activation function of neurons in the hidden layer was an s-typed transfer function (tansig), and that in the output layer was a linear transfer function purelin. The number of iteratiosn for training was set as 1000, the learning rate was 0.01, and the target error was 0.001.

SVR, with highly similar structures to ANNs, can learn from experimental data [57]. It is one of the most important predictive statistical models. The LIBSVM toolbox offers two types of regression methods, including ε-SVR and ν-SVR [58]. Here ε-SVR was used with the radial basis function (RBF) as the kernel function. The penalty factor c ($c > 0$) and the kernel parameter g are two major influencing factors on the performance of SVR. Here the SVR model was optimized. The parameter combination (c, g) was determined through the mesh search method and five-fold cross-validation, which were also used in our previous studies.

PLSR is very effective in predicting a group of dependent variables from a number of independent variables [59]. This is a multivariable statistical data analytical method that integrates PCA and multivariable linear analysis. When the variables are highly linearly correlated, PLSR can return a very effective prediction. The number of principal component factors (PCFs) is the major cause of over-fitting or under-fitting of PLSR [39]. Here, leave-one-out cross-validation was used to determine the number of PCFs in the PLSR model.

2.6. Hybrid prediction models

2.6.1 SVM-GMDH. SVM is an efficient way to solve nonlinear classification problems [60]. In order to make full use of this advantage of SVM, we put forward a prediction model of pre-classification and later regression. Firstly, the training set was classified into several sub-training sets according to the SOM classification standard, and then an SVM clustering model was established and used to cluster the validation

set. Finally, several regression models were built based on the sub-training sets to predict the values of the clustered validation set. Considering that the number of sub-training sets is small, it can be a challenge to establish a reliable prediction model. For this problem, the group method of data handling (GMDH) provides us with a powerful tool [61]. GMDH is a learning machine on the basis of heuristic self-organization as proposed by Ivakhnenko in 1976 [62], and has been widely used in energy conservation [63], marketing [64], fault recognition [65], and engineering geology [66]. In this study, SVM and GMDH were combined as SVM-GMDH for the first time for SOM detection.

Let the training set be $T = \{(X_j, y_j) \mid j = 1, 2, 3, \ldots, m\}$ and the testing set be $V = \{(X_i, y_i) \mid i = 1, 2, 3, \ldots, n\}$, where X_j and X_i are the feature vectors of the training set and the validation set respectively, and y_j and y_i are the observed SOM concentrations by the training set and the testing set respectively. m and n are the sample numbers of the training set and the testing set, respectively. The algorithm procedures of SVM-GMDH are shown below:

(a) According to the SOM classification standard, T was divided into C_k classes ($k = 1, 2, \ldots,$ m), and there existed:

$$T = \bigcup_{j}^{m} C_k,\ C_k \bigcap C_p = \emptyset \tag{6}$$

where $k \neq p$ ($p = 1,2, \ldots, m$).

(b) The training set T was used to construct an SVM clustering model.

(c) C_k was used as the training set to build GMDH models, and a total of k GMDH models were obtained.

(d) The samples (X_i, y_i) in V were classified into the C_k class by the trained SVM clustering model, and the predicted value of y_i could be obtained by predicting X_i with the kth GMDH model.

2.6.2 PLSR-BPNN. Due to the complex causality between soil gases and SOM, the SOM concentration may be internally related either linearly or nonlinearly to the soil olfaction feature space. Therefore, a hybrid prediction model of PLSR-BPNN was proposed in this study. The PLSR-BPNN is an organic combination of PLSR and BPNN, which effectively utilizes the linear modeling capability of PLSR and the nonlinear mapping capability of BPNN to improve the predictive performance of AOS. The combination of PLSR and BPNN is illustrated in figure 6.

The modeling and forecasting process of PLSR-BPNN is described below:

(a) The training set was used to establish a PLRS model and a BPNN model respectively.

(b) The established PLRS model and BPNN model were used to predict the validation set, respectively. The predicted values y_p (PLRS prediction result) and y_b (BPNN prediction result) could be obtained, where $p = 1, 2, \ldots, n$; $b = 1, 2, \ldots, n$; and n was the sample size of the validation set.

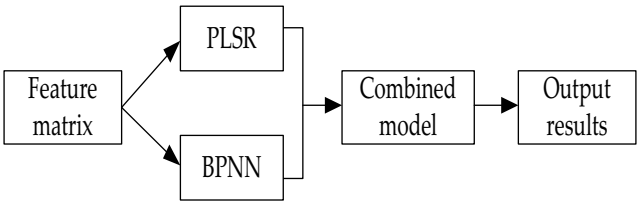


Figure 6. Sketch map of PLSR-BPNN.

(c) The prediction result of $\hat{y}_i$ was output by combining y_p and y_b according to the arithmetic mean, where $i = 1, 2, \ldots, n$. The combination was expressed as follows:

$$\hat{y}_i = k_1 \cdot y_p + k_2 \cdot y_b \tag{7}$$

where k_1 and k_2 are the weighting coefficients of PLRS and BPNN, respectively.

To determine k_1 and k_2, the concept of model validity [67] was introduced here, which is expressed by S. S can be defined as:

$$S = E \cdot (1 - \sigma) \tag{8}$$

$$A_i = 1 - \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{9}$$

$$E = \frac{1}{n} \sum_{i=1}^{n} A_i \tag{10}$$

where A_i ($i = 1, 2, \ldots, n$) represents the prediction precision sequence of the samples, y_i is the observed value of the sample in the validation set, and $\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n} (A_i - E)^2}$ is the predicted value of y_i. In the previous equations, E and σ represent the mean and standard deviation of the A_i sequence respectively.

The above calculation implies that a larger S means a higher prediction precision. We denoted the validity of the PLRS model and the BPNN model as S_p and S_b respectively. In this study, S_p and S_b were normalized as the values of k_1 and k_2 respectively, namely:

$$k_1 = \frac{S_p}{S_p + S_b}, k_2 = \frac{S_b}{S_p + S_b} \tag{11}$$

2.7. Evaluation indices The frequently used performance evaluation indices of soil property prediction models include the coefficient of determination (R^2), ratio of performance to deviation (RPD), root mean square error (RMSE) and relative prediction error (RPE). Compared with previous studies, here we also used R^2, RMSE and RPD to assess the prediction models. Moreover, the RPE was adopted as an evaluation index of parameter optimization such as feature optimization

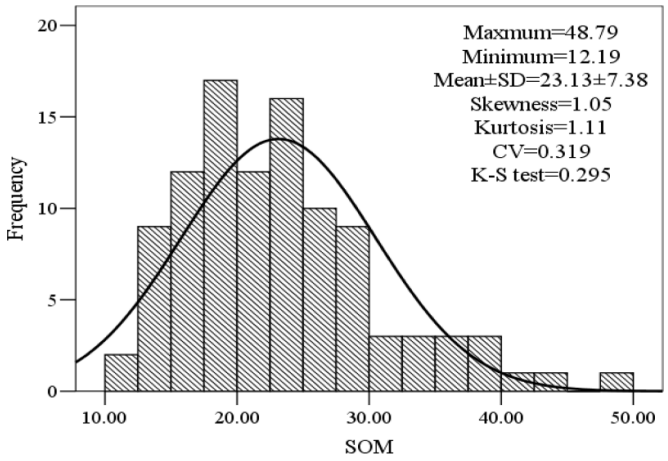


Figure 7. Statistical characteristics of SOM content and significance level of the K–S test.

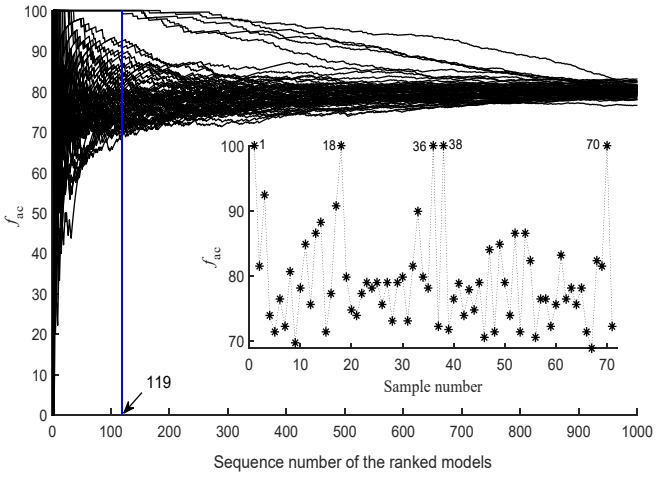


Figure 8. Cumulative frequency (f_{ac}) curves of samples in the training set.

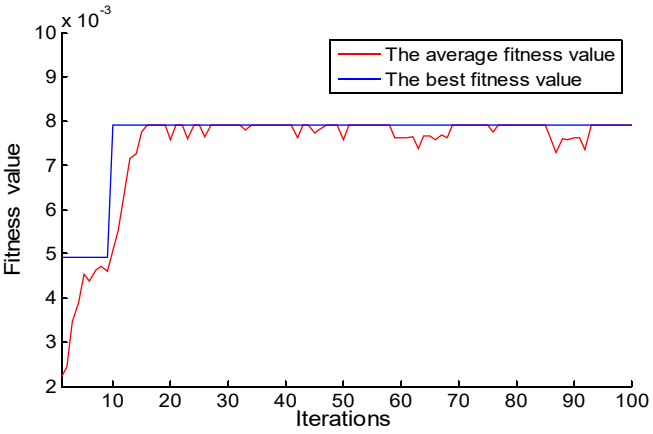


Figure 9. Optimal fitness function evolution curve.

and abnormal sample elimination. The equations of R^2, RMSE and RPD can be found in [49].

R^2 closer to 1 means higher model fitness. PRD can compensate for the limitations of R^2 in predicting nonlinear models, and a larger PRD indicates the forecasting performance is higher. A smaller RMSE implies the prediction precision is higher, and a smaller RPE suggests the predicted value deviates less from the observed value. The RPE is calculated as follows:

$$\mathrm{RPE} = (|\hat{y}_i - y_i|)/y_i (i \leq n) \tag{12}$$

where n is the sample number of the training set or validation set; y_i is the observed value of the ith sample; and $\hat{y}_i$ is the predicted value of y_i.

3. Results

3.1. Descriptive statistics for all samples The observed values of SOM from 102 soil samples were descriptively analyzed on the statistical software SPSS13.0. The normal distribution was tested via the Kolmogorov–Smirnov (K–S) method. The SOM content varied within 12.19–48.79 g kg^{-1}, with a mean of 23.131 g kg^{-1}. The coefficient of variation (CV) was 0.319, indicating SOM content in this study showed a spatial variation trend. The K–S test value was 0.295 ($P > 0.05$), so the null hypothesis of normality cannot be rejected (figure 7).

3.2. Results of abnormal sample elimination The SOFS consisted of a training set (71 samples) and a validation set (31 samples). To eliminate the effects of abnormal samples, we used MCCV to identify the abnormal samples in the training set. In the calculation of MCCV, 57 (71% × 80%) samples in the training set were randomly selected to build 1000 PLSR models, and the remaining 14 (71% × 20%) samples were used for prediction. Figure 8 shows the variation curves of f_{ac} along with the sequence number of ranked models, and the small inset figure shows the f_{ac} of the first 119 models.

Clearly, as the sequence number of ranked models increased, the f_{ac} values of most samples in the training set approached 80%, but the f_{ac} curves of samples 1, 18, 36, 38 and 70 were significantly different from the other curves to some extent because the f_{ac} values of these five samples were still maintained at 100% within a larger model sequence number range. Therefore, the five samples with numbers 1, 18, 36, 38 and 70 were identified as 'abnormal samples', and needed to be removed.

3.3. Results of feature optimization After removing the abnormal samples from the training set, a new training set was obtained. The GA-BP method was used to reduce the dimension of the new training set for optimization. The output condition of the GA-BP was set as 100 iterations. Figure 9 illustrates the fitness function evolution curve.

Obviously, after the number of iterations exceeded 10, the optimal fitness value was unchanged, indicating the optimal effect had been achieved. In this case, a group of optimal feature vectors were screened out: 1, 2, 6, 8, 10, 13, 14, 16, 17, 18, 21, 24, 25, 26, 30, 36, 37, and 40, which meant that the original feature was reduced from 40 dimensions to 18 dimensions. It

can be seen that, after the elimination of abnormal samples and feature dimensionality reduction, the training set was optimized from a matrix of 71 × 40 dimensions to a new matrix of 66 × 18 dimensions, and the validation set was transformed into a new matrix of 31 × 18 dimensions.

3.4. Prediction results by single models

To test the effect of the optimized OFS on modeling performance, BPNN, SVR and PLSR were used to establish three different single prediction models on the new training set (66 samples × 18 features), and the new validation set (31 samples × 18 features) was used to validate these models. In the BPNN modeling, the number of optimized neurons in the hidden layer was eight. In the SVR modeling, the optimal parameters were $c = 1\,048\,576$ and $g = 1.5492 \times 10^{-6}$. In the PLSR modeling, the optimal number of FPCs was four. The prediction results of the models are shown in figure 10.

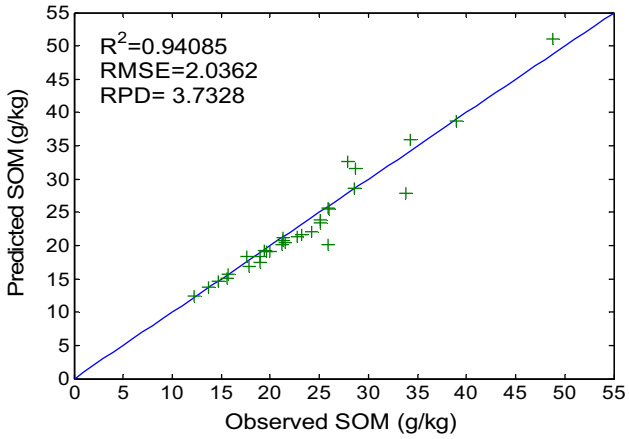
The R^2 of the three single models were 0.941, 0.918 and 0.913, respectively; the RMSEs were 2.036, 2.224 and 2.114, respectively; and the PRDs were 3.733, 3.418 and 3.377, respectively (figure 10). The above results suggest that the three models are all of high predicting ability. In terms of R^2, RMSE or PRD, the SVR model and the PLSR model are not significantly different, but the R^2 and PRD of the BPNN model are both higher than those of the other two models, and the RMSE is lower than those of the other two models. Thus, among the three models, the BPNN model outperforms the other two models and has a higher prediction accuracy.

3.5. Prediction results by hybrid models

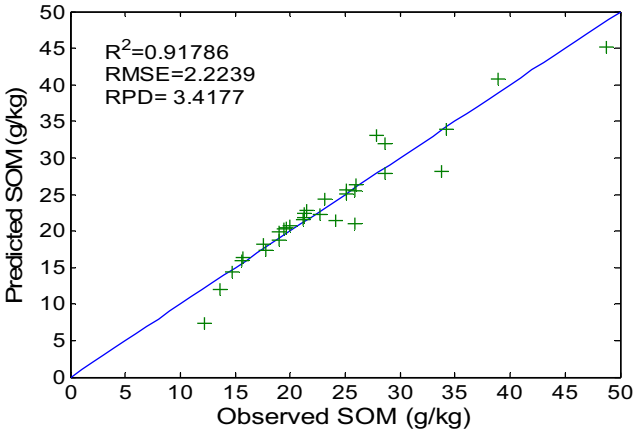
3.5.1 SVM-GMDH. According to the soil nutrient classification standard of the second national soil survey in China, the SOM content can be divided into six levels [68]: level 1—extremely high (>40 g kg^{-1}), level 2—very high (30–40 g kg^{-1}), level 3—high (20–30 g kg^{-1}), level 4—medium (10–20 g kg^{-1}), level 5—low (6–10 g kg^{-1}), and level 6—very low (<6 g kg^{-1}). The classification results of SOM in this study are shown in table 3.

However, since the organic matter content of all soil samples was greater than 10 g kg^{-1}, there were no level 5 or 6 samples in table 2. In addition, when the abnormal samples 1 (20.51 g kg^{-1}), 18 (29.87 g kg^{-1}), 36 (41.10 g kg^{-1}), 38 (20.17 g kg^{-1}) and 70 (38.86 g kg^{-1}) were removed, only one soil sample remained in level 1 of the training set and validation set respectively, which was not conducive to the establishment of a classification model and regression model. Therefore, these two remaining samples (namely, 43.85 g kg^{-1} and 48.79 g kg^{-1}) of level 1 were grouped into level 2. In table 2, C1, C2 and C3 were labeled as three categories respectively: C1 (level 2), C2 (level 3) and C3 (level 4).

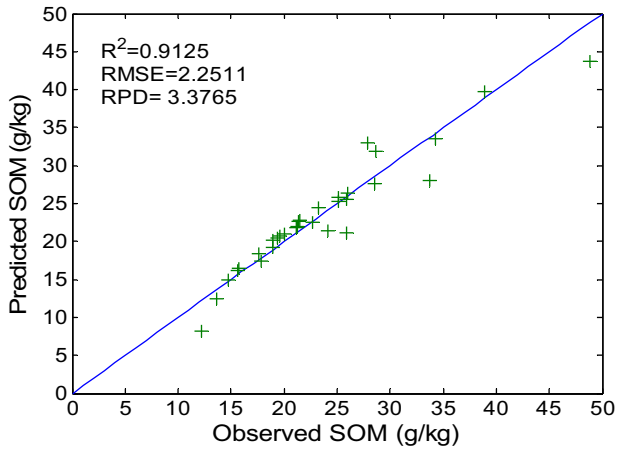
Based on the results in table 3, an SVM-GMDH prediction model was built and the validation set was classified and predicted. The prediction results and performance evaluation indices (R^2, RMSE, PRD, and RPE) are shown in figure 11. The results were $R^2 = 0.79$, RMSE = 3.52 and PRD = 2.16,



(a) BPNN prediction results.



(b) SVR prediction results.



(c) PLSR prediction results.

Figure 10. Prediction results of single models.

indicating the predicting ability of the SVM-GMDH model was not high (figure 11(a)). The reasons were mainly attributed to the low classification accuracy (CA) of SVM, because the CA of SVM-GMDH in the validation set was 83.9% (figure 11(a)). When the CA was set at 100% by manual, the predicting indices of SVM-GMDH were $R^2 = 0.845$,

Table 3. Classification results of SOM content.

Label	SOM content of training set	Number of training set samples	SOM content of validation set	Number of validation set samples
C1	43.85; 33.50; 32.13; 39.64; 35.55; 36.28; 30.71; 30.14; 35.38	9	48.79; 33.77; 34.24; 38.92	4
C2	27.62; 20.23; 23.11; 24.43; 28.71; 26.53; 26.92; 20.48; 28.85; 22.85; 22.85; 25.27; 22.55; 20.52; 25.20; 23.72; 22.31; 25.42; 23.16; 24.76; 23.80; 27.95; 20.60; 22.88; 24.75; 23.46; 22.89; 29.65	28	24.15; 25.11; 21.32; 25.86; 25.85; 25.10; 25.94; 22.71; 21.50; 23.18; 28.58; 21.13; 28.62; 20.01; 21.28; 27.89	16
C3	18.88; 14.97; 17.69; 13.76; 17.38; 19.97; 12.37; 17.33; 14.22; 15.49; 18.13; 13.44; 16.24; 15.67; 13.29; 19.54; 19.14; 19.79; 13.79; 15.90; 19.27; 18.67; 16.53; 15.32; 16.31; 16.74; 17.78; 14.8; 19.75	29	12.19; 18.94; 19.64; 18.96; 17.58; 17.78; 13.64; 14.72; 19.37; 15.59; 15.71	11

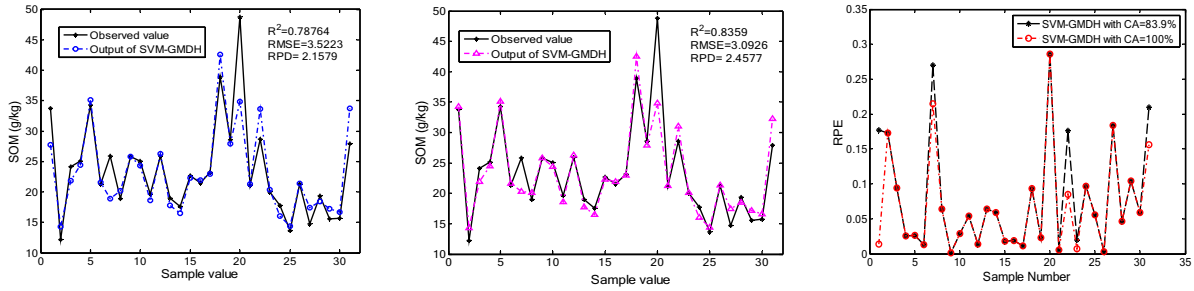


Figure 11. Test results of the SVM-GMDH hybrid model.

Table 4. Preliminary modeling results of three single models.

Models	R^2	RMSE	RPD
BPNN	0.880	2.679	2.837
SVR	0.895	2.531	3.003
PLRS	0.808	3.393	2.240

RMSE = 3.093 and PRD = 2.51 (figure 11(b)), which were all higher than those of SVM-based classification. Furthermore, the RPE was smaller when CA = 100% (figure 11(c)). Admittedly, the low CA decreased the forecasting performance of SVM-GMDH.

3.5.2 PLSR-BPNN. To evaluate the prediction performance of the hybrid model PLSR-BPNN, the optimized SOFS was used in modeling and prediction. In the PLSR-BPNN modeling, the parameters of its two parallel combination branching algorithms PLSR and BPNN were set the same as in the single PLSR model and single BPNN model. Figure 12 shows the predicted results by the PLSR-BPNN.

Clearly, the PLSR-BPNN showed a favorable predicting trend and accuracy, and the R^2, RMSE and PRD were 0.952, 1.771 and 4.291 respectively (figure 12). As per equation (11), the weighting coefficients of the PLSR and BPNN were calculated to be $k_1 = 0.52$ and $k_2 = 0.48$, respectively.

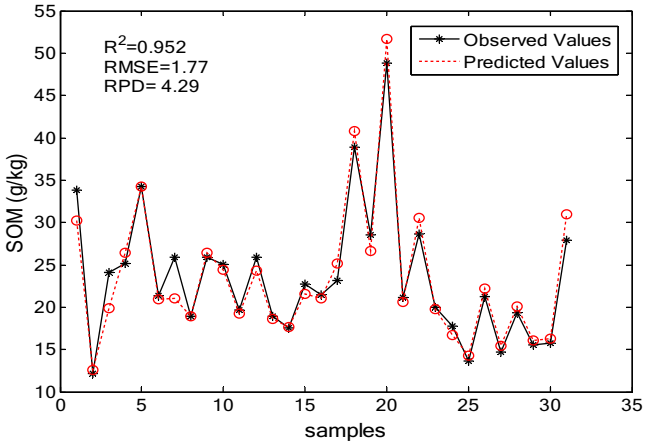
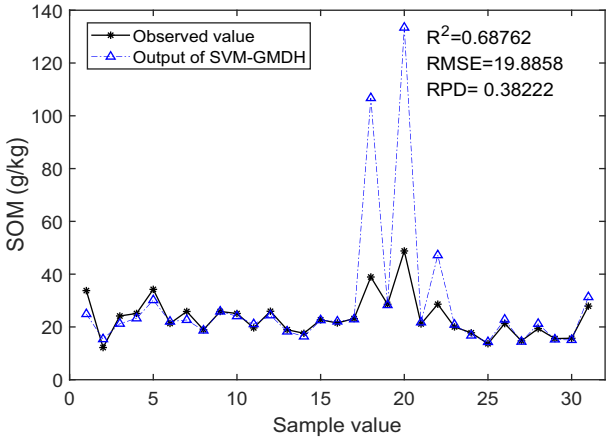


Figure 12. Test results of the PLSR-BPNN.

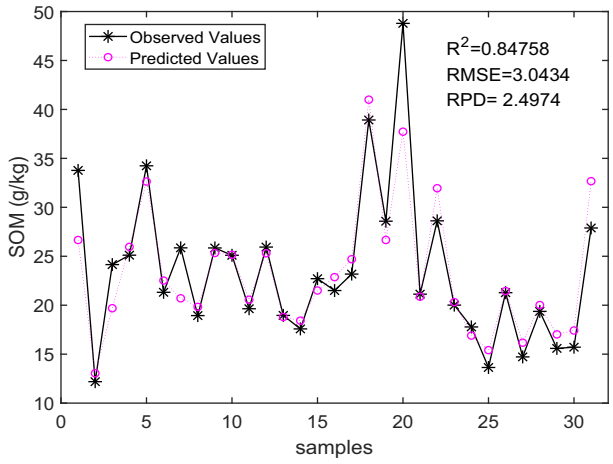
4. Discussion

4.1. Discussion on optimization effect of SOFS

In this paper, we refer to the calibration model based on the unoptimized SOFS the preliminary prediction model. In our previous paper [41], we reported the preliminary prediction results of three single models of a BPNN, SVR and PLSR as listed in table 4.

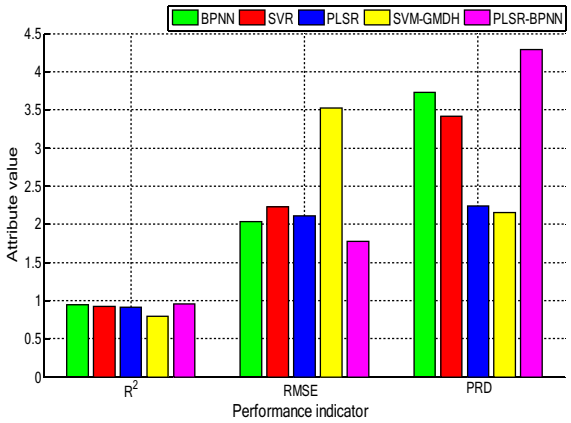


(a) SVM-GMDH prediction results.

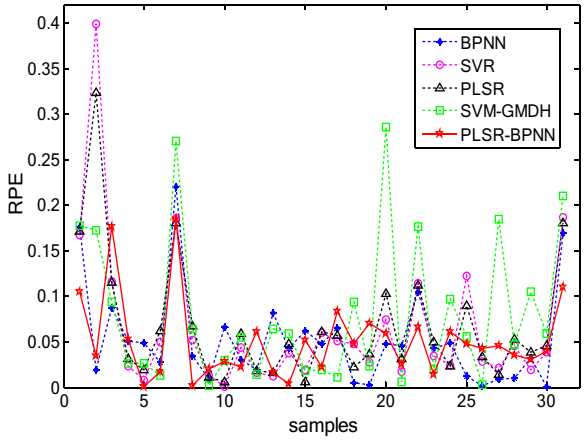


(b) PLSR-BPNN prediction results.

Figure 13. Preliminary prediction results of the hybrid models.



(a) Comparison of R2, RMSE and RPD in five different models



(b) Comparison of relative prediction errors (RPE) in five different models

Figure 14. Comparison of performances on *g* models.

As can be seen from table 3, the R^2 of the BPNN, SVR and PLSR are all greater than 0.8, and all have RPD greater than 2.0. Vohland *et al* reported a detailed evaluation method for predicting goodness [69]: R^2 and RPD values greater than 0.90 and 3.0, respectively, are considered to be 'excellent', whereas values from 0.82–0.9 (R^2) to 2.5–3.00 (RPD) are defined as 'good'. Values between 0.66 and 0.81 (R^2) and 2.0 and 2.5 (RPD) indicate 'approximate quantitative prediction'. When the RPD value is from 1.5 to 2.0 and the R^2 value is from 0.50 to 0.65, the model can only be used to distinguish high and low values. 'Unsuccessful' predictions have RPD and R^2 values lower than 1.5 or 0.50, respectively. Therefore, the preliminary models of the BPNN and SVR have 'good' predictive performance, while the PLSR preliminary model has only approximate quantitative ability.

The preliminary prediction results of hybrid models are shown in figure 13. The results of SVM-GMDH were: $R^2 = 0.69$, RMSE = 19.89 and PRD = 0.382 (figure 13(a)). It is an unsuccessful prediction (RPD < 0.50). On the one hand, as mentioned above, SVM-GMDH is affected by the accuracy of SVM classification. On the other hand, it may be affected by abnormal samples and redundancy features. The preliminary prediction model of PLSR-BPNN has a 'good' prediction, with $R^2 = 0.848$, RMSE = 3.043 and RPD = 2.497 (figure 13(b)).

The optimized SOFS, including the new training set (66 samples × 18 features) and the new validation set (31 samples × 18 features), was used in modeling and prediction.We refer to the calibration model based on the optimized SOFS as the optimized model. The prediction results of the five optimized models (BPNN, PLSR, SVR, SVM-GMDH and PLSR-BPNN) are shown in figures 10(a)–(c), 11(a) and 12, respectively. In comparison with the preliminary prediction model, the results of the optimized mode rise by 6.93%, 2.57%, 13.00%, 14.53% and 12.26%, respectively (R^2); decrease by 24.00%, 12.13%, 37.70%, 82.29% and 41.80% respectively (RMSE); and increase by 31.58%, 13.82%, 50.76%, 464.92% and 95.29% (RPD). These results suggest the predictive capabilities of all five models were largely improved after the optimization of SOFS, indicating that MCCV can effectively identify abnormal samples and GA-BP can effectively eliminate redundant features.

4.2. Comparison of optimized models

To find the optimal prediction model for SOM detection, we visualized the evaluation indices of the five optimized models (figure 14). Except for the SVM-GMDH optimized model ($R^2 = 0.79$, PRD = 2.16), which can only be used for approximate quantitative prediction, the other four optimized models ($R^2 > 0.90$, RPD > 3.0) have 'excellent' prediction.

According to figure 14(a), it can be seen that the prediction performance relationship of the optimized models is PLSR-BPNN > BPNN > SVR > PLSR > SVM-GMDH. Moreover, the RPE curves reflect the fluctuation of prediction errors of the five models (figure 14(b)). Clearly, the RPE of PLSR-BPNN fluctuated within the smallest range (0–0.185) and thus showed the strongest predicting accuracy. Therefore, we believe that the PLSR-BPNN optimized model performs the best in olfactory detection of SOM. This is because there was a linear and non-linear relationship between olfactory feature variables, and the PLSR-BPNN could effectively solve the linear and nonlinear mapping problems by weighting coefficients.

5. Conclusions

The main intellectual merits of this work include the novel approach based on the AOS and PLSR-BPNN and its effectiveness as a method of solving the linear and nonlinear mapping problems by weighting coefficients in detecting SOM. The test results demonstrate that: (a) MCCV and GA-BP were good measures to optimize SOFS, and (b) the PLSR-BPNN model yields higher predictive performance and lower relative predicting errors compared to BPNN, SVR, PLSR and SVM-GMDH models. According to numerical and evaluation index tradeoffs, MCCV + GA-BP + PLSR-BPNN ($R^2 = 0.952$, RMSE = 1.771, and RPD = 4.291) is selected as the most suitable method for predicting SOM in this study.

Compared with the near-infrared spectroscopy in reference [70] (test result: $R^2 = 0.91$) and reference [71] (test result: $R^2 = 0.69$), the method of the current study is superior. However, our samples need to be stored sealed for a period of time, which is time-consuming and not conducive to real-time detection. Therefore, our next work will focus on the influencing factors of artificial olfactory detection of SOM and the rapid processing of soil samples. In addition, the number of soil samples should also be increased to make the determination method more stable and robust.

Acknowledgments

This research was funded by the National Key R&D Plan project, Grant No. 2016YFD070030201, and the Jilin Science and Technology Development Plan (20190302116GX, 20200502007NC). We highly appreciate Yibing Chen, who is a researcher of the Soil and Fertilizer Station of Jilin Province, for providing the soil samples.

ORCID iD

Longtu Zhu https://orcid.org/0000-0001-8699-7902

References

[1] Tziachris P, Aschonitis V, Chatzistathis T and Papadopoulou M 2019 Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters *Catena* **174** 206–16
[2] Karami A, Homaee M and Afzalinia S 2012 Organic resource management: impacts on soil aggregate stability and other soil physico-chemical properties *Agric. Ecosyst. Environ.* **148** 22–8
[3] Salehi M H, Beni O H and Harchegani H B 2011 Refining soil organic matter determination by loss-on-ignition *Pedosphere* **21** 473–82
[4] Kasim N, Sawut R and Qingdong S 2018 Estimation of soil organic matter content based on optimized spectral index *Trans. Chin. Soc. Agric. Mach.* **49** 155–63
[5] Wang J, He T and Lv C 2010 Mapping soil organic matter based on land degradation spectral response units using Hyperion images *Int. J. Appl. Earth Obs. Geoinf.* **12** S171–80
[6] Nawar S and Mouazen A M 2019 On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning *Soil Tillage Res.* **190** 120–7
[7] Shepherd K D and Walsh M G 2002 Development of reflectance spectral libraries for characterization of soil properties *Soil Sci. Soc. Am. J.* **66** 988–98
[8] Conforti M, Matteucci G and Buttafuoco G 2018 Using laboratory vis-NIR spectroscopy for monitoring some forest soil properties *J. Soil Sediment.* **18** 1009–19
[9] Conforti M, Castrignanò A and Robustelli G 2015 Laboratory-based vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content *Catena* **124** 60–7
[10] Nawar S, Buddenbaum H and Hill J 2015 Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy *Soil Tillage Res.* **155** 510–22
[11] Liu W D, Baret F and Gu X F 2002 Relating soil surface moisture to reflectance *Remote Sens. Environ.* **81** 238–46
[12] Bowers S A and Hanks R J 1965 Reflection of radiant energy from soil *Soil Sci.* **100** 130–8
[13] Stoner E R and Baumgardner M F 1982 Characteristic variations in reflectance of surface soils *Soil Sci. Soc. Am. J.* **45** 1161–5
[14] Goutal N, Renault P and Ranger J 2013 Forwarder traffic impacted over at least four years soil air composition of two forest soils in northeast France *Geoderma* **193–4** 29–40
[15] Cesare F D, Mattia E D and Pantalei S 2011 Use of electronic nose technology to measure soil microbial activity through biogenic volatile organic compounds and gases release *Soil Biol. Biochem.* **43** 2094–107
[16] Gougoulias C, Clark J M and Shaw L J 2014 The role of soil microbes in the global carbon cycle: tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems *J. Sci. Food Agric.* **94** 2362–71
[17] Asensio D, Peñuelas J and Ogaya R 2007 Seasonal soil VOC exchange rates in a Mediterranean holm oak forest and their responses to drought conditions *Atmos. Environ.* **41** 2456–66
[18] Llobet E, Jesús B and Vilanova X 1997 Qualitative and quantitative analysis of volatile organic compounds using

transient and steady-state responses of a thick-film tin oxide gas sensor array *Sensors Actuators* B **41** 13–21
[19] Vermoesen A, Ramon H and Cleemput O V 1991 Composition of the soil gas phase: Permanent gases and hydrocarbons *Pedologie* **41** 119–32
[20] Balasubramanian S, Panigrahi S and Kottapalli B 2007 Evaluation of an artificial olfactory system for grain quality discrimination *LWT—Food Sci. Technol.* **40** 1815–25
[21] Srivastava A K 2003 Detection of volatile organic compounds (VOCs) using SnO_2 gas-sensor array and artificial neural network *Sensors Actuators* B **96** 24–37
[22] Eklov T, Johansson G, Winquist F and Lundstrom I 1998 Monitoring sausage fermentation using an electronic nose *J. Sci. Food Agric.* **76** 525–32
[23] Chatterjee D, Bhattacharjee P and Bhattacharyya N 2014 Development of methodology for assessment of shelf-life of fried potato wedges using electronic noses: sensor screening by fuzzy logic analysis *J. Food Eng.* **133** 23–9
[24] Amy L, Silvia C and Ganesh K M 2015 Electronic noses for food quality: a review *J. Food Eng.* **144** 103–11
[25] Antti R, Erik V and Antti T 2014 Detection of prostate cancer by an electronic nose: a proof of principle study *J. Urol.* **192** 230–4
[26] Shih C-H, Lin Y-J and Lee K-F 2010 Real-time electronic nose based pathogen detection for respiratory intensive care patients *Sensor Actuators* B **148** 153–7
[27] Wojciech W, Tomasz D, Jacek G and Jacek N 2019 Electronic noses in medical diagnostics *Curr. Med. Chem.* **26** 197–215
[28] D'Amico A, Pennazza G, Santonico M, Martinelli E, Roscioni C, Galluccio G, Paolesse R and Di Natale C 2010 An investigation on electronic nose diagnosis of lung cancer *Lung Cancer* **68** 170–6
[29] Capelli L, Sironi S and Del Rosso R 2014 Electronic noses for environmental monitoring applications *Sensors* **14** 19979–20007
[30] Huo D, Wu Y and Yang M 2014 Discrimination of Chinese green tea according to varieties and grade levels using artificial nose and tongue based on colorimetric sensor arrays *Food Chem.* **145** 639–45
[31] Macías M M, Agudo J E, Manso A G, Orellana C, Velasco H and Caballero R 2013 A compact and low cost electronic nose for aroma detection *Sensors* **13** 5528–41
[32] Jacek G and Bartosz S 2018 Discrimination of selected fungi species based on their odour profile using prototypes of electronic nose instruments *Measurement* **116** 307–13
[33] Jacek G, Tomasz D and Jacek N 2014 Monitoring of odour nuisance from landfill using electronic nose *Chem. Eng. Trans.* **40** 85–90
[34] Andrzej B, Katarzyna J G and Łukasz G 2016 Evaluating soil moisture status using an e-nose *Sensors* **16** 886
[35] Pobkrut T and Kerdcharoen T 2014 Soil sensing survey robots based on electronic nose *Proc. of the 2014 14th Int. Conf. on Control, Automation and Systems (ICCAS)* (*Piscataway, NJ*) pp 1604–9
[36] Boeker P 2014 On 'Electronic Nose' methodology *Sensors Actuators* B **204** 2–17
[37] Lotfivand N, Abdolzadeh V and Hamidon M N 2016 Artificial olfactory system with fault-tolerant sensor array *ISA Trans.* **63** 425–35
[38] Ji W, Shi Z and Huang J 2014 *In situ* measurement of some soil properties in paddy soil using visible and near-infrared spectroscopy *Plos One* **9** e105708
[39] Qi H, Paz-Kagan T and Karnieli A 2018 Evaluating calibration methods for predicting soil available nutrients using hyperspectral VNIR data *Soil Tillage Res.* **175** 267–75
[40] De Santana F B, De Giuseppe L O and De Souza A M 2019 Removing the moisture effect in soil organic matter determination using NIR spectroscopy and PLSR with external parameter orthogonalization *Microchem. J.* **145** 1094–101
[41] Du P, Wang J and Yang W 2019 A novel hybrid model for short-term wind power forecasting *Appl. Soft Comput.* **80** 93–106
[42] Guo Z, Wu J and Lu H 2011 A case study on a hybrid wind speed forecasting method using BP neural network *Knowl-Based Syst.* **24** 1048–56
[43] Ning W and Xu Z 2013 Application of combination forecast model to the railway freight volume forecasting *Technol. Econ. Areas Commun.* **15** 78–81
[44] Wen J and Lei L 2010 A combined forecasting method of grain yield in China based on GM(1,1) and BP network *Third Int. Conf. on Information and Computing* (IEEE) pp 75–78
[45] Yin H, Dong Z and Chen Y 2017 An effective secondary decomposition approach for wind power forecasting using extreme learning machine trained by crisscross optimization *Energy Convers. Manage.* **150** 108–21
[46] Wang S, Zhang N and Wu L 2016 Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method *Renew. Energy* **94** 629–36
[47] Zhang Y and Fearn T 2015 A linearization method for partial least squares regression prediction uncertainty *Chemometrics and Intelligent Laboratory Systems* **140** 133–40
[48] Xiao W and Ye J 2009 Improved PSO-BPNN algorithm for SRG modeling *2009 Int. Conf. on Industrial Mechatronics and Automation* (IEEE)pp 245–8
[49] Zhu L, Jia H and Chen Y 2019 A novel method for soil organic matter determination by using an artificial olfactory system *Sensors* **19** 3417
[50] Liu Z, Cai W and Shao X 2008 Outlier detection in near-infrared spectroscopic analysis by using Monte Carlo cross-validation *Sci. China* **51** 751–9
[51] Zhou T, Lu H and Wang W 2019 GA-SVM based feature selection and parameter optimization in hospitalization expense modeling *Appl. Soft. Comput.* **75** 323–32
[52] Dasa S R, Mishra D and Rout M 2019 Stock market prediction using Firefly algorithm with evolutionary framework optimized feature reduction for OSELM method *Expert Syst. Appl.* **4** 100016
[53] Hossein-Babaei F and Amini A 2014 Recognition of complex odors with a single generic tin oxide gas sensor *Sensors Actuators* B **194** 156–63
[54] Herrero-Carrón F, Yáñez D J and Rodríguez F 2014 An active, inverse temperature modulation strategy for single sensor odorant classification *Sensors Actuators* B **206** 555–63
[55] Li X, Wei Y and Xu J 2018 Quantitative visualization of lignocellulose components in transverse sections of moso bamboo based on FTIR macro- and micro-spectroscopy coupled with chemometrics *Biotechnol. Biofuels* **11** 263
[56] Prasad R, Deo R C and Li Y 2018 Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition *Geoderma* **330** 136–61
[57] Wu C L, Chau K W and Li Y S 2008 River stage prediction based on a distributed support vector regression *J. Hydrol.* **358** 96–111
[58] Chang C C and Lin C J 2011 LIBSVM: a library for support vector machines *ACM Trans. Intell. Syst. Technol.* **2** 1–27
[59] Rossel R A V, Walvoort D J J and Mcbratney A B 2003 Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties *Geoderma* **131** 59–75
[60] Gove R and Faytong J 2012 Machine learning and event-based software testing: classifiers for identifying infeasible GUI event sequences *Adv. Comput.* **86** 109–35

[61] Xiao J, Sun H and Hu Y 2015 GMDH based auto-regressive model for China's energy consumption prediction *2015 Int. Conf. on Logistics, Informatics and Service Sciences (LISS)* (IEEE) pp 1–6
[62] Najafzadeh M, Barani G-A and Hessami Kermani M R 2013 GMDH based back propagation algorithm to predict abutment scour in cohesive soils *Ocean Eng.* **59** 100–6
[63] Srinivasan D 2008 Energy demand prediction using GMDH networks *Neurocomputing* **72** 625–9
[64] Mehrara M, Moeini A and Ahrari M 2009 Retracted: investigating the efficiency in oil futures market based on GMDH approach *Expert Syst. Appl.* **36** 7479–83
[65] Witczak M, Korbicz J and Mrugalski M 2006 A GMDH neural network-based approach to robust fault diagnosis: application to the DAMADICS benchmark problem *Control Eng. Pract.* **14** 671–83
[66] Kalantary F, Ardalan H and Nariman-Zadeh N 2009 An investigation on the Su-NSPT correlation using GMDH type neural networks and genetic algorithms *Eng. Geol.* **104** 144–55
[67] Chen S and Wang W 2004 Grey neural network forecasting for traffic flow *J. Southeast Univ.* **45** 34–49
[68] Li L, Yang T and Redden R 2016 Soil fertility map for food legumes production areas in China *Sci. Rep.* **6** 26102
[69] Vohland M, Besold J, Hill J and Fründ H 2011 Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy *Geoderma* **166** 198–205
[70] Zhang X and Tang N 2018 Research on soil's organic matter content prediction based on wavelength optimization of near infrared spectrum *Mod. Electron. Tech.* **22** 126–9
[71] Nowkandeh S M, Noroozi A A and Homaee M 2018 Estimating soil organic matter content from Hyperion reflectance images using PLSR, PCR, MinR and SWR models in semi-arid regions of Iran *Environ. Dev.* **25** 23–32