Geoderma 333 (2019) 23–34

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm

Yao Zhang[a,b], Minzan Li[a,*], Lihua Zheng[a], Qiming Qin[b], Won Suk Lee[c]

[a] Key Laboratory of Modern Precision Agriculture System Integration Research, Ministry of Education, China Agricultural University, Beijing 100083, China
[b] Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing 100871, China
[c] Department of Agricultural & Biological Engineering, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

Editor: Dr. A.B. McBratney

Keywords:
Ant colony optimization
Feature selection
Mutual information
Near infrared
Spectroscopy
Soil total nitrogen

ABSTRACT

Nondestructive and rapid estimation of soil total nitrogen (TN) content by using near-infrared spectroscopy plays a crucial role in agriculture. The obtained original spectrum, however, presents several disadvantages, such as high redundancy, large computation, and complex model, because it generally processes a large amount of data. This study aimed to determine soil TN content-sensitive wavebands with high information quality, considerable predictive ability, and low redundancy. This paper proposes an evaluation criterion in selecting sensitive wavebands based on three factors, namely, degree of relevance with target variables, representative ability of the entire spectral information, and redundancy of the selected wavebands. Based on these three factors, two methods, namely, mutual information (MI) algorithm and the combination of ant colony optimization (ACO) and MI, were innovatively developed to identify soil TN content-sensitive wavebands. After the analysis and comparison, a set of wavelengths, including 943, 1004, 1097, 1351, 1550, 1710, 2123, and 2254 nm, using the ACO–MI combined method was selected as the soil TN content-sensitive wavebands to estimate the TN content of soil samples, under four soil types, collected from different regions. The partial least squares (PLS) models based on full-spectral information, multiple linear regression (MLR) models and support vector machine (SVM) regression models based on the eight selected wavelengths for soil TN content were established separately. After the comparison, the MLR and SVM models achieved higher accuracies than the PLS models based on the full spectral information. In addition, the SVM models got the best results. In the calibration group, the coefficients of determination (R^2) was 0.989, and the root mean square errors (RMSE) of calibration was 0.078 g/kg. In the validation group, the R^2 was 0.96, and the RMSE of prediction was 0.219 g/kg. The residual predictive deviation (RPD) was 5.426. For the soil samples with TN content in the range of 0–1 g/kg, the detection precision also reached a high level. Therefore, the eight sensitive wavebands selected through the ACO–MI method performed good mechanism, universality and predictive ability in soil TN content estimation. The ACO–MI method would be valuable for soil sensing in precision agriculture.

1. Introduction

Soil is the primary support for soil-grown crops. It is an important medium for plant root extension and the main nutrient source for crop growing. The main soil nutrients include TN, OM, available potassium, and available phosphorus (Chacón Iznaga et al., 2014; Sinfield et al., 2010). Among those soil nutrients, soil nitrogen (TN and available nitrogen) plays the most important role in promoting the growth of leaf, root, and stem and is a decisive factor to the crop yield (Bansod and Thakre, 2014). Excessive nitrogenous fertilization however will cause environmental pollution and crop distortion of growth and quality. The amount of nitrogen fertilizer applied therefore needs to be precisely controlled to ensure the crop yield and environmental protection. The fundament of precision fertilizing is effectively acquiring the soil information in the field. Rapid and precise acquisition of soil nitrogenous information in farmlands hence becomes increasingly important. The conventional method of detecting soil nitrogen content usually takes several days and consumes toxic chemicals. The conventional method

Abbreviations: TN, total nitrogen; OM, organic matter; NIR, near-infrared; MI, mutual information; ACO, ant colony optimization; PLS, partial least squares; MLR, multiple linear regression; SVM, support vector machine; R^2, coefficients of determination; SSR, sum of squares for regression; SST, sum of squares for total; SSE, sum of squares for error; RMSE, root mean square errors; RPD, residual predictive deviation

[*] Corresponding author.
E-mail address: limz@cau.edu.cn (M. Li).

https://doi.org/10.1016/j.geoderma.2018.07.004
Received 19 December 2017; Received in revised form 25 June 2018; Accepted 3 July 2018

also has several disadvantages, such as high requirements for detection personnel, expensive testing equipment, low efficiency, and environmental pollution (Debaene et al., 2014; Florinsky et al., 2002; Kuang and Mouazen, 2013; Moore et al., 1993; Nocita et al., 2014). By contrast, spectral analysis techniques are based on the internal relations between radiation energy and the composition and structure of matters. According to the characteristic spectra of matter, the target concentration or properties can be determined rapidly without chemicals (Igne et al., 2010; Vohland et al., 2011). For soil, the spectral information related to most of the organic radical groups containing hydrogen was in the NIR region (Li, 2006). NIR spectroscopy is a rapid, non-destructive, and non-pollutant testing method that plays an growing important role in soil nutrition measurement and exhibits extraordinary development potential in applications of soil TN content detection (Chang et al., 2001; Lucà et al., 2017; Morellos et al., 2016).

In the detection of soil TN content with NIR spectroscopy, the spectra of soil samples are first measured, and the NIR spectral data are then used as the input variables to establish the prediction models. Modern spectrometers possess high spectral resolution, and spectral data measurement generally involves hundreds or thousands of wavelength variables. Three kinds of information variables are involved in measurement of such superlarge-scale data. One is the effective informative variable, which can improve the model predictive ability because it reflects the characteristics of the target substance in the NIR region. The second is redundant or interfering variable, which is related with other targets. The last one is uninformative variable, which is irrelevant to the target material and usually caused by the measurement environment, such as noise. If the prediction model was established by the entire-spectrum information, then the latter two kinds of variables would increase the computation complexity and reduce the target prediction accuracy of the model. The multivariate calibration model is therefore a better choice when the informative variables could be selected appropriately, which can help in simplifying the calibration model and improving the model's predictive ability in terms of accuracy, speed and robustness (Petropoulos et al., 2012; Sorol et al., 2010). Several studies confirmed that the models adopting limited characteristic wavebands only are better than the ones with entire-spectrum information. Cai et al. (2008) employed Monte Carlo uninformative variable elimination method to extract the characteristic bands and then established a PLS model to predict the sugar content of tomato. The results showed that the prediction accuracy and robustness of the proposed model were all better than the model on the basis of the entire band. Gao et al. (2009) proposed a method combining screened contribution and SPA to select soil TN content-sensitive wavebands. The result of the established multivariable model was more precise than the result derived from PLS for the entire spectra. Several scientists took advantage of the specific wavelength of the light source to develop a soil spectral detection device and obtained good detection accuracy (An et al., 2014; Li et al., 2010). The results revealed that the use of limited wavelength of light source in detecting soil nutrients had reached a practical level. The existing characteristic wavelength selection algorithm however revealed that several drawbacks in soil TN prediction remained. The selected wavebands had a weak mechanism interpretation about soil TN, and the accuracies of the established models were relatively low. The methods for selecting the characteristic wavebands therefore need to be further explored for the study of NIR spectral technology on soil TN content detection.

ACO is an evolutionary algorithm used to simulate the natural foraging behavior of ant colonies and was introduced in the early 1990s (Colorni et al., 1991). The technique is based on updating the coordination mechanism of seeking the shortest path to achieve intelligent search and parametric optimization. The ACO algorithm has been extensively applied in feature selection because of its prominent advantages, such as information positive feedback, distributed computation, heuristic search, robustness, and easy to combine with other algorithms (Dorigo et al., 1996). Aghdam and Kabiri (2016) proposed an intrusion detection system with features optimally selected using ACO to improve the performance. Varma et al. (2016) also used fuzzy entropy-based heuristic for ACO to search for the global best smallest set of network traffic features for real-time intrusion detection data set. Selection of the optimal spectral characteristic variable is also a combinatorial optimization issue. The features, such as global, discrete, and probability selection of self-adaptive ACO, are applicable to spectroscopy analysis. Hou et al. (2016) applied ant colony clustering algorithm to detect Grapevine leafroll disease (GLD) spectral anomalies on four GLD-infected vineyards from multi-spectral images for precision disease management. The classification accuracies of Non-, GLD1-, GLD2-, and GLD3-infected grapevines were 94.4%, 75%, 84.6%, and 83.3%, respectively. Guo et al. (2014) selected sensitive wavebands from apple NIR spectroscopy using ACO–PLS optimized algorithm based on the features of heuristic global search and the random selection mechanism of Monte Carlo roulette to predict soluble solid content. The prediction model obtained a good prediction performance for SSC with correlation coefficients of 0.970, and RMSE of prediction of Brix of 0.514. Allegrini and Olivieri (2011) employed the concept of cooperative pheromone accumulation, which is typical of ACO selection methods, and optimized PLS models using a pre-defined number of variables and employing a Monte Carlo approach to discard irrelevant sensors. Ke et al. (2008) proposed an ACO-based algorithm to deal with feature selection in rough set theory and compared its performance with the simulated annealing-, genetic algorithm-, and Tabu search-based algorithms. The results showed the proposed algorithm achieved better performance according to both the classification results and the number of features. Santana et al. (2010) found ACO performed better than genetic algorithm-based feature selection method for ensemble classifiers when the number of individual classifiers was small. Agrawal and Kaur (2018) compared ACO and Hybrid Particle Swarm Optimization in test case selection. The results indicated ACO outperforms Hybrid Particle Swarm Optimization in the calculating efficiency.

The studies discussed above used ACO to carry out feature selection in numerous areas. After the comparison between ACO and other heuristic algorithms, ACO performs more flexible and efficient. It is especially suitable for the relatively small-scale problems (Xue et al., 2016). In order to make further improvement on the performance of ACO, the information theory, as a supplementary, could explore more and deeper information from the variables themselves. When using ACO, few researchers however adopt information theory to improve the performance of ACO in feature selection. Several studies adopted the Monte Carlo roulette principle to select features, which is random and lack connection with the variable information (Allegrini and Olivieri, 2011; Guo et al., 2014). In this way, the selected features were mostly dependent only on the performance of ACO. Merging ACO and information theory could significantly benefit feature selection, especially when the training set is not big enough to represent the whole application space.

In the early 1990s, Battiti (1994) and Lewis (1992)introduced MI theory into variable selection research. The feature selection method based on MI has gained considerable attention after 20 years of development. MI is an excellent tool for quantitatively calculating the common information between two random variables. The MI theory has therefore been extensively applied as an effective indicator for investigating correlation. MI can also be used to measure the arbitrary dependencies between random variables, which makes it suitable for assessing the "information content" of features in complex tasks.

Filter methods are defined by a criterion *J* based on MI, also referred to as a "relevance index" or "scoring" criterion, which is intended to imply the potentially predictive ability of the feature (Duch, 2006). Moreover, there is another widely accepted point that an useful and parsimonious set of features is considered individually relevant and should not be redundant with respect to each other, which means that the selected features should not be highly correlated (Brown et al., 2012). This heuristic has been adopted numerous times. Battiti (1994) presented MI feature selection criterion. Peng et al. (2005) proposed max-relevance min-redundancy criterion. Cheng et al. (2011) proposed a conditional MI feature selection criterion by using an assumption to approximate the terms on the basis of joint MI. Vinh et al. (2016) systematically investigated the issues of employing high-order dependencies for MI-based feature selection based on "relevancy" and "redundancy" criteria.

The variable selection methods above are all based on the principle of maximizing "relevancy" between the independent variables and the dependent variables and minimizing the "redundancy" between the selected independent variables. However, from the information theory perspective, another goal of feature selection is to select a feature subset that can contain most or all information of the original data set. The feature subset has the strongest representative ability of the entire data set.

The studies summarized above thus show that an effective featured waveband selection method is strongly needed for soil TN content prediction. This paper aims to propose a modified ACO algorithm mixed with a new evaluation criterion of waveband selection based on MI theory. This criterion would comprehensively consider three factors that demonstrate the relationships between independent and dependent variables, representative spectral information of the entire spectra, and discrepancies between the selected wavebands. To ascertain the universality and predictive ability of the selected wavebands, the soil samples under different kinds of fertilizers and covering wide TN content range were collected to validate the soil TN content prediction model.

2. Materials and methods

2.1. Experiment

The experiments were conducted from November 2014 to August 2016. Three experimental farms located in two representative areas of China, North China plain, and Northeast China plain were selected for this study, as shown in Fig. 1.

In November 2014, the experiments were conducted in Shangzhuang experimental farm (116.191794°E, 40.144091°N) of the China Agricultural University located in the North China plain. The soil type is sandy loam. A total of 270 sampling spots were selected randomly to collect soil samples. These samples were used to explore the sensitive wavebands of soil TN.

To validate the TN predictive ability and universality of the selected wavebands, two validation experiments were designed as described below.

The first validation experiment was also conducted in the Shangzhuang experimental farm in April 2015. Three different regions were selected with the distance of 50 m to avoid the interference from each other. Three kinds of fertilizers, namely, carbamide ($N \geq 27\%$), calcium nitrate ($N \geq 13\%$), and potassium nitrate ($N \geq 13.5\%$), were used separately to perform variable rate fertilization on the three regions mentioned above. Sixty soil samples (20 in each region) were collected, and the soil TN content of each sample was detected.

The second validation experiment was carried out in Heilongjiang Province in August 2016. Two farms located in the Northeast China plain, namely, Hongwei Farm and Shengli Farm, were selected. Hongwei farm (133.4472°E, 47.3125°N) covers 63,000 hm^2, and the soil type is meadow albic bleached soil. One hundred soil samples were collected in this farm. Shengli farm covers 87,733.33 hm^2, and two sampling areas were selected. The soil type in Shengli-A area (133.7247°E, 47.5277°N) is turfy soil, and the soil type in Shengli-B area (133.7530°E, 47.4044°N) is black soil. One hundred soil samples were collected in each part.

Finally, 630 samples were obtained and used as the target materials for all the subsequent experiments (NIR absorbance spectra measurement and soil TN detection).

2.2. Spectra measurement

All the soil samples were air dried, and the NIR spectroscopy absorbance of each sample was measured in the laboratory by a MATRIX-I type of FT-NIR analyzer with a rotating sample pool (Bruker Optical Company, Germany). Approximately 15 g of each soil sample was placed into a quartz cuvette with 48 mm diameter, and the cuvette was then placed into the rotating sample pool. The NIR spectra were finally measured by using the spectrometer. The spectral measurement range was 12,493–3899 cm^{-1} (800–2564 nm) with resolution of 4 cm^{-1} and scan times of 32. Each sample was scanned 10 times with three replications, and the mean value was considered the relative absorbance.

2.3. Measurement of soil TN content

After air drying, grinding, and sieving, the TN concentrations of soil samples were measured with a Kjeltec TM2300 azotometer (FOSS, Sweden). Air-drying soil powder amounting to 2.0 g was mixed with 6.2 g of $K_2SO_4/CuSO_4{\bullet}5H_2O$ catalyst (30:1), 20 ml of H_2SO_4 was then placed into the mixture, and the digestion was conducted at 420 °C for 1.5 h. The azotometer was finally used to distill cooling liquid and display nitrogen concentrations.

2.4. Data processing method

2.4.1. MI

MI is a key concept of information theory and used as a quantitative index of the correlation between two variables. This concept is also a quantitative index of the amount of information in one variable as that contained by the other variables. The MI theory can effectively eliminate the irrelevant information to the target components (y) from the spectral information (x). If x and y are continuous random variables with the existence of joint density $\mu(x,y)$ and edge density ($\mu_x(x)$, $\mu_y(y)$), then MI can be defined as in Eq. (1).

$$\mathrm{MI}(\mathrm{x},\mathrm{y}) = \iint dxdy\mu(x,y)\log\frac{\mu(x,y)}{\mu_x(x)\mu_y(y)} \tag{1}$$

In Eq. (1), x and y represent two variables. $\mathrm{MI}(\mathrm{x},\mathrm{y})$ is the MI between x and y, which expresses the reductive amount of uncertainty of y after obtaining the information of x. Compared with other correlation analysis methods, MI can consider both linear and nonlinear relationship between variables.

A nonlinear relationship is ubiquitous among spectral information and between spectral response and material concentration (Benoudjit et al., 2004; Hu et al., 2011; Viscarra Rossel et al., 2006). In this research, three MI-related indexes were calculated using Eq. (1). The first index is the MI values ($\mathrm{MI}(f;\mathrm{c})$) between the soil absorbance (f) in the range of 800–2564 nm and TN content (c). The second index is the sum of all MI values ($\mathrm{MI}(f_i;f_j)$) between the absorbance ($f_i(f_i \in F)$) and the other absorbance ($f_j(f_j \in F - f_i)$). F is the set of all wavelength absorbance. The third index is the sum of MI values ($\mathrm{MI}(f;\mathrm{s})$) between the absorbance ($f(f \in F)$) and the selected ones ($\mathrm{s}(\mathrm{s} \in \mathrm{S})$). S is a set of selected wavelengths.

2.4.2. ACO

Combining the characteristics of ACO algorithm and the evaluation criteria for selecting feature variables proposed in this research, the modified algorithm called ACO–MI is shown as Eqs. (2)–(3) for the soil TN content-sensitive wavelength selection.

The probability of the k^{th} ant to move from one wavelength to another can be described as Eq. (2).

$$P_{ij}^{k}(t) = \alpha \times \tau_{ij}(t) + \beta \times \eta_{ij}(t)$$
$$\eta_{ij} = \frac{1}{MI(f_i;s_j)} \tag{2}$$

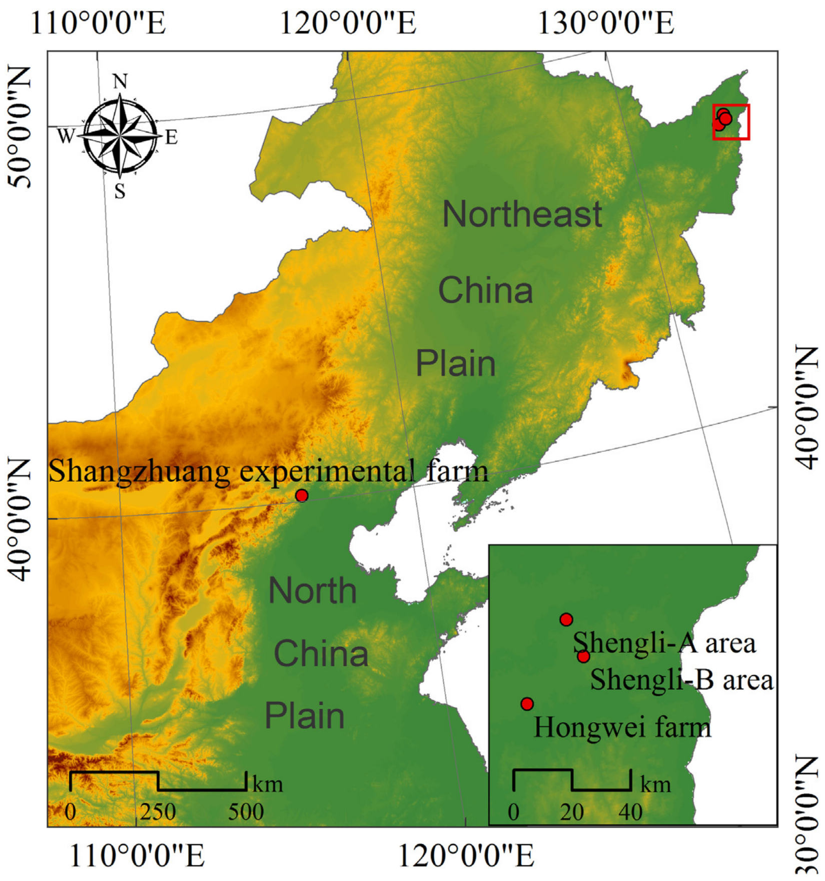

Fig. 1. Location of the study area.

In Eq. (2), $\tau_{ij}(t)$ represents the pheromone concentration between the wavelength i and the wavelength j at time t. $\eta_{ij}(t)$ represents the intensity of heuristic information, whose value indicates the divergence degree between wavelength i where the ants are and the objective wavelength j. α and β are the parameters that characterize the importance of pheromone concentration and heuristic factors. The greater the probability is, the greater the ants' attractiveness is.

Eq. (2) demonstrates that ants mainly rely on two factors in selecting wavelength, pheromone concentration $\tau_{ij}(t)$, and heuristic information $\eta_{ij}(t)$. The value of $\eta_{ij}(t)$ does not change with time and the iterative updating rules of pheromone concentration because it is only related to the location of ants. Pheromone concentration $\tau_{ij}(t)$ thus also plays a major role in optimizing the global searching ability. After the ant completes a search of the characteristic variable, the pheromone is updated according to Eq. (3).

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t)$$
$$\Delta\tau_{ij}(t) = Q \times L_k$$
$$L_k = Norm_0^1(\mathrm{MI}(f_i;c)) + Norm_0^1\left(\sum_{j=1}^{n}\mathrm{MI}(f_i;f_j)\right) \tag{3}$$

In Eq. (3), $\rho \in (0,1)$ is the factor of evaporation that represents the disappearance speed of the pheromone. The selection of the value will affect the convergence speed of the algorithm. $\Delta\tau_{ij}(t)$ is the new pheromone concentration and is updated on the basis of the objective function and pheromone intensity factor. Q is the new pheromone intensity factor used to adjust the objective function and the convergence rate. L_k is the objective function. In this study, L_k represents the sum of relevance degree with the target material (TN) and the total representative capacity for the entire-spectrum information of the selected wavelength by the current ant. The maximum of $L_k$was then obtained to update the pheromone concentration.

2.5. Model accuracy evaluation methodology

The model accuracy was quantitatively evaluated with three aspects: R^2 calculated from Eq. (4), RMSE based on Eq. (5), and RPD calculated from Eq. (6)(Chakraborty et al., 2017; Mcgladdery et al., 2018; Williams, 1987).

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$
$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2;\ SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$SST = SSR + SSE = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{5}$$

$$RPD = \sqrt{\frac{1/(n-1)\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{1/n\sum_{i=1}^{n_p}(y_i - \hat{y}_i)^2}}_{validation} \tag{6}$$

where, n indicates the number of samples, y_i is the measured value of sample i, $\hat{y}_i$ is the predicted value of sample i, and $\bar{y}_i$ are the mean values of y_i.

3. Results and discussion

3.1. Characteristic analysis of the original VIS/NIR spectra

All spectral curves of the soil samples were obtained by using the

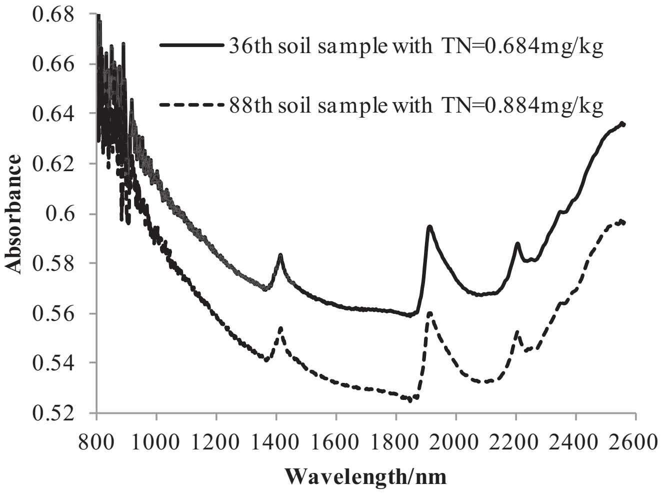

Fig. 2. Absorbance spectra of the soil samples.

spectrometer. Fig. 2 shows the spectral characteristic in the range of 800–2564 nm of the 36th soil sample and the 88th soil sample. The TN contents of the two samples were 0.684 and 0.884 g/kg, respectively. Fig. 2 reveals that the variation tendency of the soil spectral absorbance of the two samples becomes similar with each other with the change of soil TN content. In 800–930 nm, the spectral curve contained a considerable amount of noise because of the vibration of the measurement system. A slow downward trend was observed in the range of 930–1380 nm. An absorbance peak was observed in the range of 1380–1510 nm, which was caused by water absorbance. The soil spectral absorbance was then regained smoothly and steadily until 1850 nm. The soil spectral absorbance then sharply increased and reached its topmost peak at 1940 nm. A curve peak existed at around 2210 nm, which was different with the peaks around 1450 and 1940 nm in peak amplitudes. Viscarra Rossel et al. (2006) revealed that this spectral absorption band at around 2210 nm depended on the O–H group of soil organic compounds.

Although the measured soil samples were air-dried, the absorption peaks were still evident near the wavebands of 1450 and 1940 nm, which were the absorption bands of the O-H functional group in water molecule. Most of the bound water is still in the soil samples and absorbs the light at the wavelengths of 1940 and 1450 nm because the air-dried treatment can only remove free water (Li, 2006).

3.2. Sensitive waveband selection based on MI

As mentioned above, this paper aimed to propose a novel evaluation criteria for selecting feature variables and then develop a method according to MI theory based on the criteria. Three kinds of relationships were considered. The variable selection method focuses on maximizing MI between the candidate spectral variable and the target material (TN), minimizing the redundancy between the candidate and the selected spectral variables, and maximizing the ability of the candidate variables representing for the entire-spectrum variables. Based on the above principles, this study developed a general expression formula shown as Eq. (7).

$$y = Norm_0^1(\mathrm{MI}(f_i;c)) + Norm_0^1\left(\sum_{j=1}^{n}\mathrm{MI}(f_i;f_j)\right) - Norm_0^1\left(\sum_{k=1}^{m}\mathrm{MI}(f_i;s_k)\right) \tag{7}$$

where F is the set of absorbance of candidate wavebands, and $i \in F$. K is the set of absorbance of all wavebands, $j \in K$. S is the set of selected wavebands, $k \in S$. c is the dataset of soil TN contents. n is the number of all wavelengths. m is the number of selected wavelengths.

The process of the feature selection is expressed as follows:

1) Initialization. F and K are sets of absorbance of all wavebands. S is a null set.
2) Computing $\mathrm{MI}(f_i;c)$ and $\sum_{j=1}^{n}\mathrm{MI}(f_i;f_j)$, and then conducting normalization processing in the range of 0–1 between the two indexes.
3) Selecting the first sensitive waveband i_1 to make the sum of $Norm_0{}^1(MI(f_i;c))$ and $Norm_0^1\left(\sum_{j=1}^{n}\mathrm{MI}(f_i;f_j)\right)$ be the maximum, and then $F = F - i_1$, $S = i_1$.
4) Calculating $\mathrm{MI}(f_i;c)$, $\sum_{j=1}^{n}\mathrm{MI}(f_i;f_j)$, and $\Sigma_{s_k}{\in}S\mathrm{MI}(f_i;s_k)$, and then performing normalization processing among the three indexes.
5) The wavelength f_i was used as the next sensitive wavebands, which could make the result of Eq. (7) be maximum, and $F = F - f_i$, $S = S \cup f_i$.
6) Until the number of the selected wavelength reaches the maximum, the sensitive wavelength selection ends, or it will move to Step 4.

As for the maximum number of needed wavelengths, if the number is small, the selected wavelength could not represent the effective spectral information comprehensively. Meanwhile, considering a future possibility of constructing a multispectral device for soil TN content detection, large number of selected wavelength would increase cost and the complexity of structure. According to the prior trials (An et al., 2014; Zhang et al., 2016; Zhang et al., 2015), the maximum of the needed variables in this research is set as eight.

Fig. 3 shows two normalizing MI curves. "MI-c" represents the MI between the absorbance and the TN content, which could indicate the correlation degree of the waveband and TN content. Every point on "MI-wav" represents the sum of MI values between the absorbance at this wavelength and each absorbance of all other wavelengths. This curve could represent the ability of different wavebands representing full spectral information. Fig. 3 shows that the MI value between the absorbance and TN fluctuated significantly at the beginning around 800–900 nm, because the spectral curve contained a considerable amount of noise. Numerous local maximum points were observed afterward. MI can express more spectral information of soil TN than the regular linear correlation curve because it contains linear and non-linear relationships between each spectral variable and TN content. The

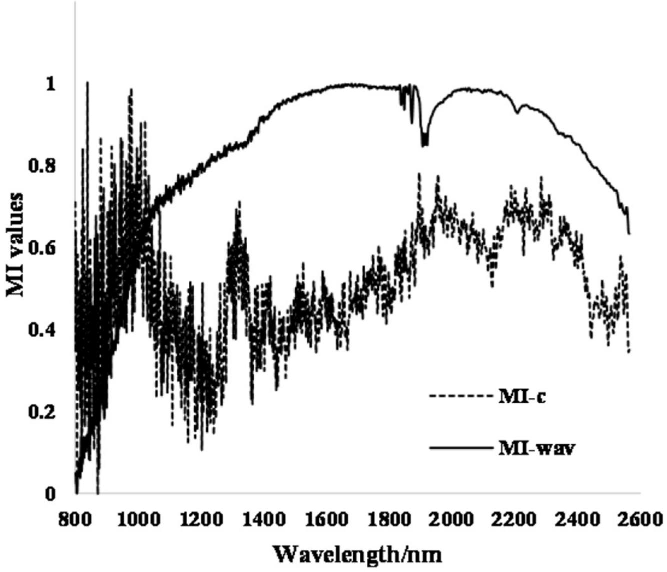

Fig. 3. 0–1 Normalizing MI curves in the range of 800–2564 nm.
(MI-c represents the MI between the absorbance and the TN content of 270 soil samples; MI-wav represents the sum of MI between the absorbance and each absorbance of all other wavelengths.)

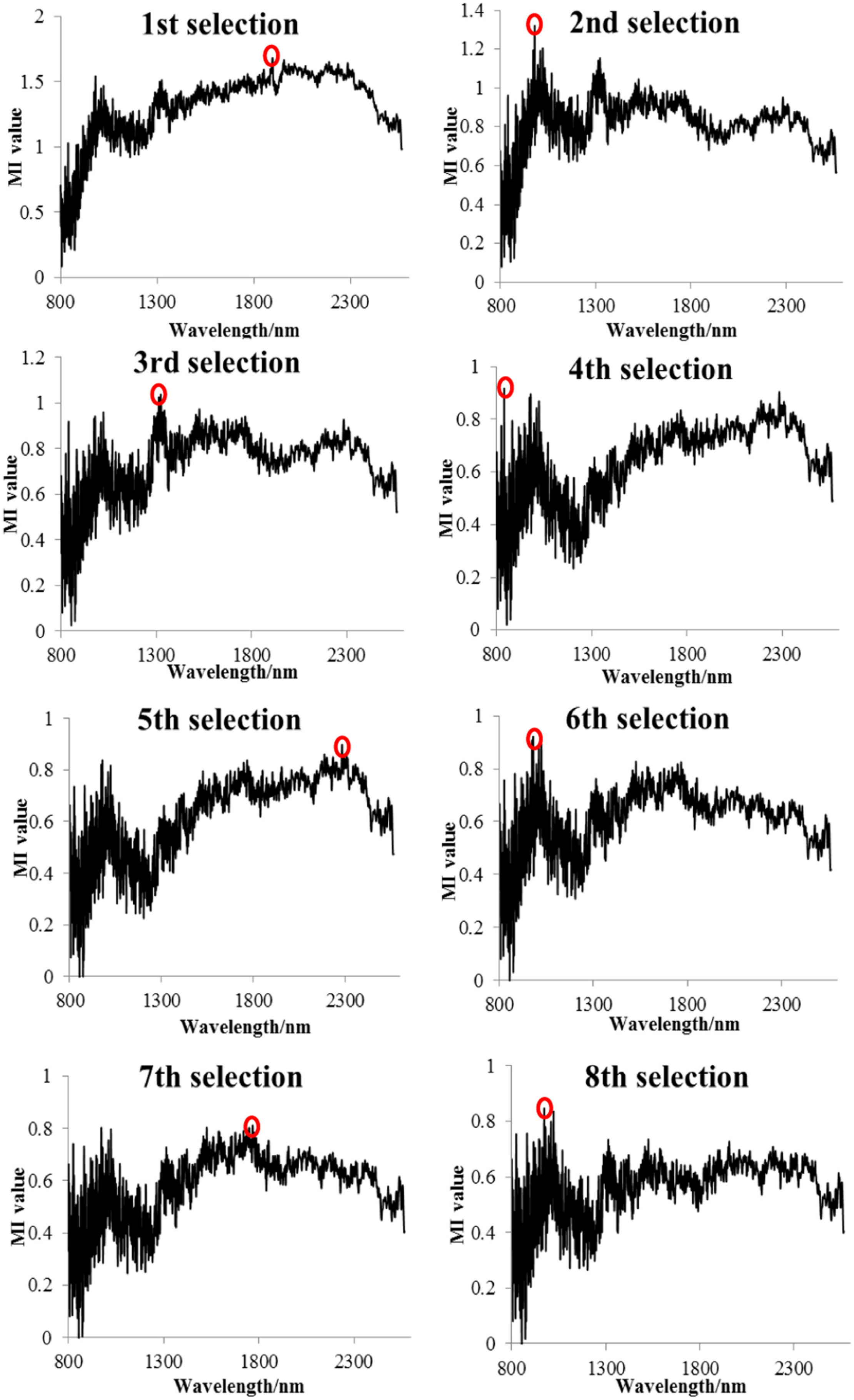

Fig. 4. Featured waveband selection processes.

absorbance at 1895 nm possessed the strongest ability to represent full spectral information.

On the basis of the principles of this selection method, eight sensitive wavebands were selected individually as shown in Fig. 4. In the first selection, 1895 nm was the maximum in the curve, which represents the sum of MI-c and MI-wav. Therefore, 1895 nm was the first selected sensitive waveband. In the second selection, 980 nm was the maximum of the curve, which were calculated as Eq. (7). The same procedure was repeated for the next selections. The first eight selection processes were shown in Fig. 4. Then 1895, 980, 1323, 839, 2284, 978, 1767, and 972 were determined individually as the sensitive wavebands of soil TN content. From the entire featured waveband selection, the MI characteristic curves in the second, sixth, and eighth selection were almost in the similar trend, which lead to the selected wavebands (972, 978, and 980 nm) in these processes to be all approximately 980 nm, which implied that several selected featured bands were in the narrow range with high self-correlation and redundancy. Moreover, 980 nm is related to the O–H groups' stretching vibration from water absorption (Leiva-Valenzuela et al., 2013; Pan et al., 2016). A typical large absorption band is observed in the range of 960–980 nm for the second O–H water overtone (Helgerud et al., 2012). A strong absorption features was centered at 1900 nm which is related to O–H molecules (Demattê et al., 2004). Therefore, using MI method alone for sensitive wavelength selection could not effectively reduce the redundancy among the selected feature wavelengths nor eliminate the interference from soil water. Further research on soil TN content-sensitive waveband selection is needed.

3.3. Sensitive waveband selection based on the ACO–MI method

Considering the outstanding advantages of ACO in feature selection, this research combined ACO and the evaluation criteria based on MI to optimize the results of soil TN content-sensitive waveband selection. The maximum variable number was also set to 8. The flow chart of the entire process of sensitive waveband selection based on ACO is shown in Fig. 5.

1) Initialization. The parameters of ACO algorithm were set after multiple experimental verifications. The maximum iteration time was 3; the ant colony size was 1500; the importance degrees of heuristic factor and pheromone were 0.8 and 0.2, respectively; the evaporation factor was 0.5; and the pheromone intensity factor was 0.0003.
2) Calculate the $MI(f_i;c)$ of all wavelengths, and create an $n \times n$ matrix of $MI(f_i;f_j)$.
3) All the initial pheromone vector values were set to be 1, which implied that each wavelength had the same attractiveness to the ants.
4) Place ants on n wavelengths and activate the ants to select the featured variable based on the probability function shown as Eq. (2).
5) Calculate the objective function using Eq. (3), which indicates the entire-spectrum representative ability and the relevance with the soil TN content of the selected wavelengths, and then update the pheromone. If the number of selected variables does not reach its set maximum, then it will move to Step 4, or it will move to the succeeding step.
6) Evaluate the iteration time to determine whether the set maximum is reached or not. If the set maximum is not reached, then the pheromone would be set as 1, and process is repeated starting with Step 3, or it would go to the succeeding step.
7) Compare the objective function results after all iterations and identify the group of selected wavebands with high objective function result as the sensitive wavebands to soil TN content.

After the selection shown in the flow chart (Fig. 5), eight wavelengths of 943, 1004, 1097, 1351, 1550, 1710, 2123, and 2254 nm were

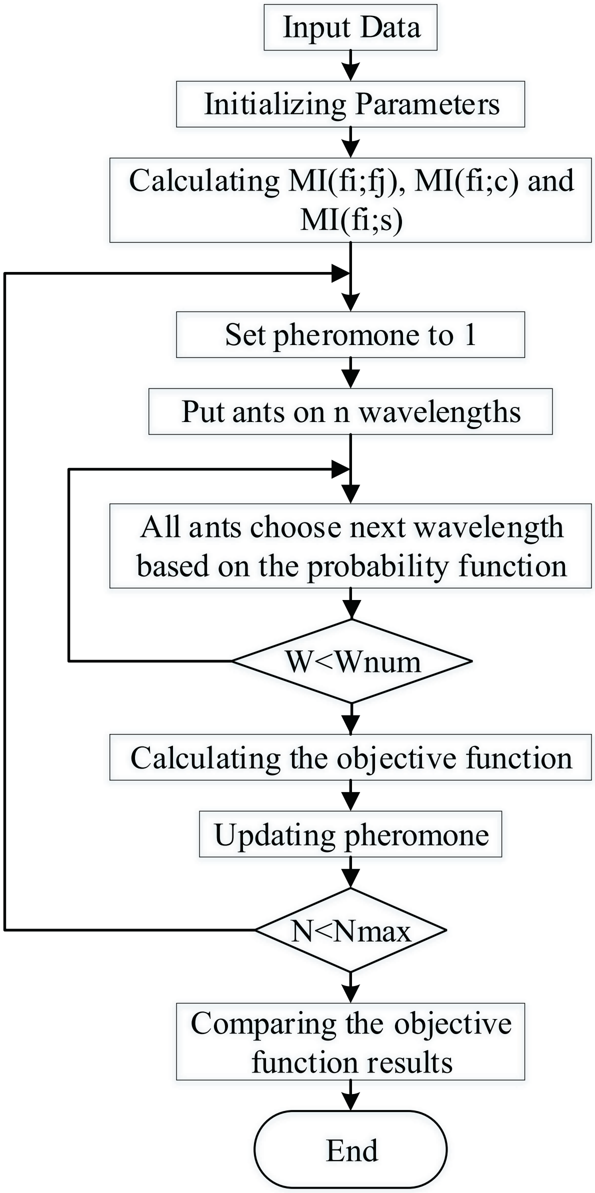
Input Data
Initializing Parameters
Calculating MI(fi;fj), MI(fi;c) and MI(fi;s)
Set pheromone to 1
Put ants on n wavelengths
All ants choose next wavelength based on the probability function
W<Wnum
Calculating the objective function
Updating pheromone
N<Nmax
Comparing the objective function results
End

Fig. 5. Flow chart of the sensitive waveband selection based on ACO–MI approach.

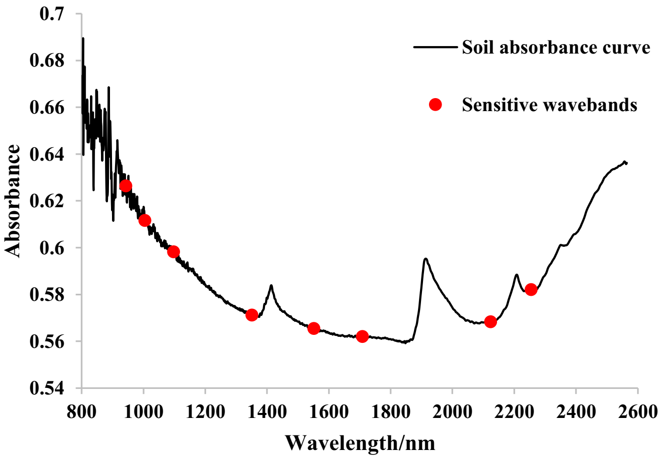

Fig. 6. Sensitive wavebands of soil TN content.

determined as the sensitive wavebands. As shown in Fig. 6, the selected wavelengths were distributed in the entire spectral curve. The integration of ACO and the selection criteria of sensitive wavebands based on MI could therefore effectively avoid repeated selection in a narrow waveband and eliminate the interference from soil moisture compared with the method of adopting MI technology only.

The previous study revealed that the wavelength of approximately 940 nm was sensitive to soil OM, which was identified by stepwise MLR and PLS (Kweon and Maxton, 2013). Palacios-Orueta and Ustin (1998) found that the area of approximately 1000 nm was related to Fe and OM content in soil. The reflectance at wavelength 1100 nm was also the OM estimators (Daniel et al., 2004). Sun et al. (2009) selected 1350 nm as one of the sensitive wavelengths of soil OM according to the correlation analysis. TN is widely known to have a strong relationship with OM content in soil, because N is a major component of OM. Among the eight selected wavelengths, An et al. (2014) and Bansod and Thakare (2014) adopted three wavelengths, namely, 940, 1100, and 1550 nm, to develop portable soil nitrogen detectors. Both detectors are reliable in soil nitrogen content measurement. Aside from 1000 nm, 1710 nm was also related to Fe. A curvilinear relationship between Fe_2O_3 and the reflectance at 1710 nm produces a large squared correlation coefficient ($r = -0.91$ or $r^2 = 0.85$) (Galvão et al., 2001). Fe could help fertilizer N transfer into available N in soil (Ali et al., 1998). In addition, 2120 and 2250 nm were crucial for TN estimation based on the correlation analysis of differential spectral information (Shi et al., 2013; Zhang et al., 2014).

All the eight wavelengths selected through the modified ACO algorithm therefore had direct and close relationship with soil TN content, which verified the effectiveness of the ACO–MI method in wavelength selection of soil TN content. The ACO–MI method also successfully eliminated the interference from the soil moisture, which was critical to TN content-sensitive wavelength selection.

3.4. Soil TN content modeling and validation

To further verify the TN predictive ability and universality of the selected sensitive wavelengths, this study created a mixed dataset containing the soil samples collected from different regions (i.e., North China plain and Northeast China plain), in different soil types, including sandy loam, meadow albic bleached soil, turfy soil, and black soil. All the spectral curves were checked prior to the modeling. Three samples were lost in the field experiment, and the spectral information of one sample collected from Northeast China plain was abnormal, which may have been caused by misoperation in the spectra measurement. A total of 626 soil samples were thus used to establish the prediction model. The TN contents of 626 soil samples in the mixed dataset were in the range of 0.061–2.743 g/kg. The PLS models based on full-spectral information, MLR and SVM regression models based on the eight selected wavelengths for soil TN content were established separately. Samples were divided into two groups, that is, 450 samples were under the calibration group and the remaining 176 samples were under the validation group. The input data of PLS models were the absorbance in the range of 800–2564 nm. The MLR model was established as Eq. (8). SVM regression models adopted parameter optimization method to establish the model. The optimal penalty C was 955.43, and the optimal kernel function g was 0.87.

$$y = -1.9 + 9.23x_{943} - 1.44x_{1004} - 19.15x_{1097} - 38.93x_{1351} + 183.37x_{1550} - 119.59x_{1710} + 1.44x_{2123} - 10.18x_{2254} \quad (8)$$

Where y is the TN content of the soil samples, and x_i is the absorbance value of wavelength i.

According to the established model, the 1:1 relationship diagrams were drawn between the prediction and observation to demonstrate the reliability and consistency of the selected model. The results are shown in Figs. 7–9.

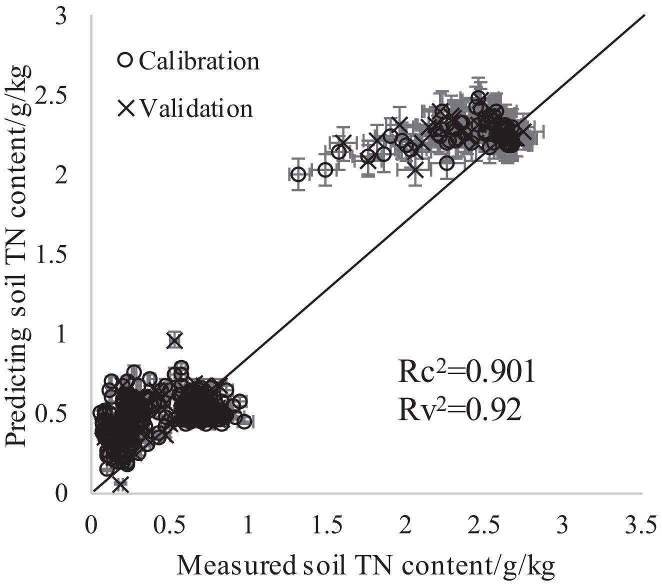

Fig. 7. Calibration and validation of soil TN prediction of the PLS regression models.

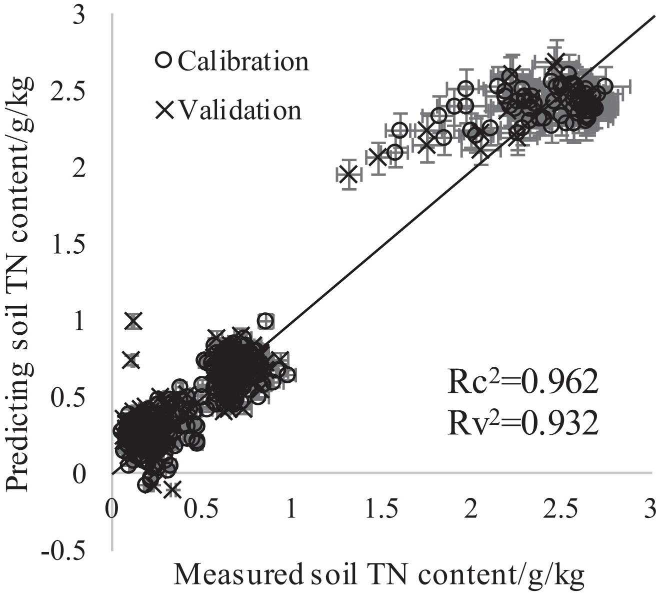

Fig. 8. Calibration and validation of soil TN prediction of the MLR models.

The accuracies of the three models are shown in Table 1. Both of the two models established by the selected wavelengths obtained higher overall predictive abilities than the one based on the full spectral information. While for the MLR model, five soil samples in the whole dataset (626 points) were predicted to be negative values. The TN contents of the five points were in the range of 0.1–0.2 g/kg. It indicated that the MLR model has a limitation in predicting soil TN with a low content. For the SVM model, the accuracies of the calibration and validation were increased significantly. And the drawbacks of predicted negative values of TN contents of MLR model were mended. While for the three models, when the soil samples with TN content around 2 g/kg, some prediction TN values were higher than the measured values. These soil samples mainly belonged to turfy soil and contained a high amount of plant debris, humus, and several minerals, which were not thoroughly broken down. The prediction TN values using spectral information, including the TN in soil and a part of TN in the

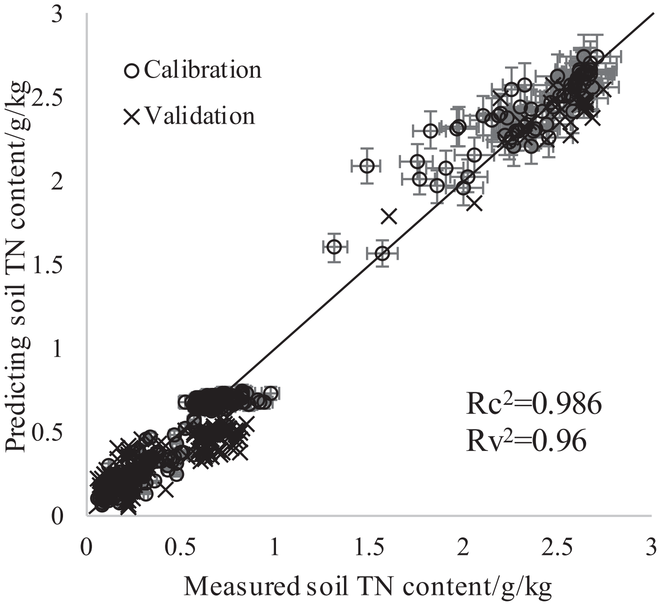

Fig. 9. Calibration and validation of soil TN prediction of the SVM regression models.

Table 1
Accuracies of PLS, MLR and SVM models.

	PLS	MLR	SVM
Calibration R^2	0.901	0.96	0.986
RMSE of calibration (g/kg)	0.238	0.153	0.078
Validation R^2	0.92	0.94	0.96
RMSE of prediction (g/kg)	0.224	0.168	0.219
RPD	5.198	7.253	5.426

undecomposed materials in soil, were thus higher than the measured TN values. The values would be closer if the soil samples' spectra were collected after grinding and sieving.

The TN contents of 528 soil samples were in the range of 0–1 g/kg, which were the major part of all soil samples. Moreover, the majority of soil in cultivated land were in this nutritional level. The PLS, MLR and SVM models were therefore established among this part. The 378 samples were used for calibration and the remaining 150 samples were for validation. The MLR regression model was established as Eq. (9). The optimal penalty C and the optimal kernel function g of the SVM regression models were 512 and 0.25 after the optimization.

$$y = -0.53 - 2.42x_{943} - 2.47x_{1004} + 12.49x_{1097} - 12.99x_{1351} + 28.9x_{1550} - 20.39x_{1710} + 5.54x_{2123} - 6.64x_{2254} \tag{9}$$

Where y is the TN content of the soil samples, and x_i is the absorbance value of wavelength i.

The results of the three models were highlighted in Figs. 10–12.

The accuracies of the three models are shown in Table 2. No matter for the linear regression model (MLR) or the non-linear regression model (SVM), the predicting accuracies of soil TN contents in this range all got better results than the PLS model, which demonstrated that the selected eight sensitive wavebands could represent the full spectral information effectively, and predict the TN, especially in farmland soil, stability, universality, and veracity.

The verification of the samples under four soil types collected from different farms revealed that although the eight sensitive wavelengths were selected from 270 original soil samples collected from a specific region with a single type, the wavelengths still performed with higher universality and predictive ability than the full spectrum in TN content detection of different soil types. The eight selected sensitive wavebands had close mechanism correlation with the material of TN.

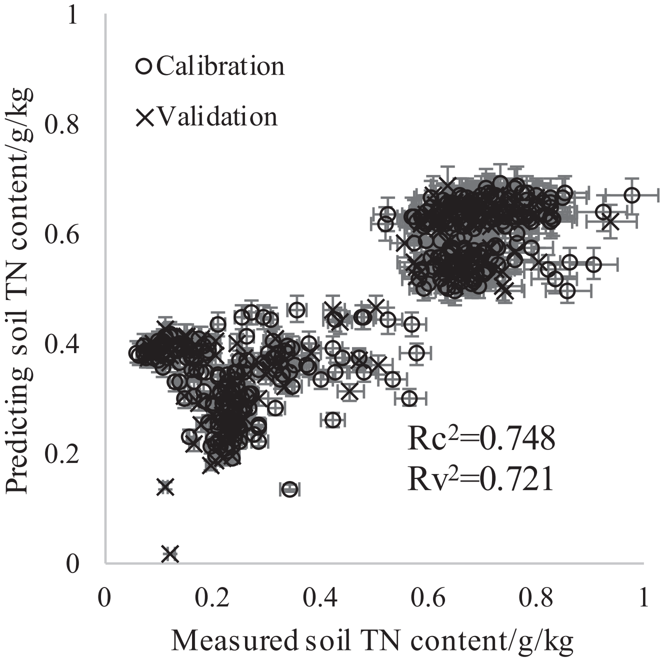

Fig. 10. Calibration and validation of soil TN prediction of the PLS regression models.

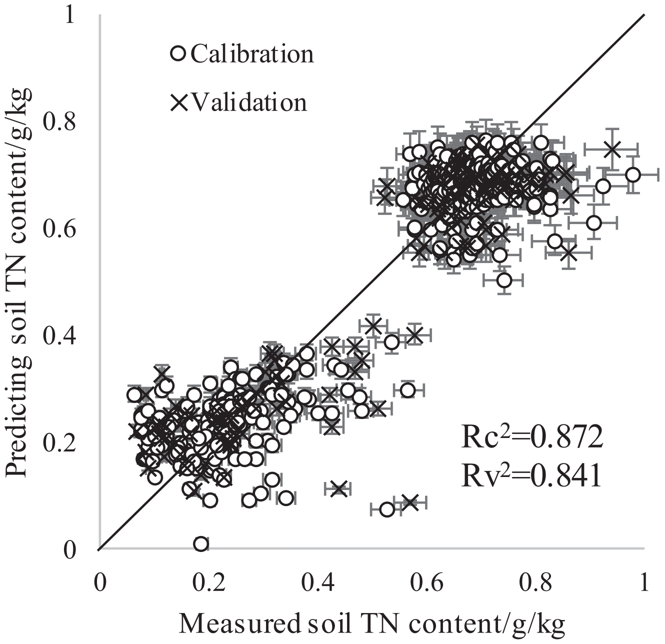

Fig. 11. Calibration and validation of soil TN prediction of the MLR models.

3.5. Discussion

Prior work has documented the advantages of soil nutrients estimation using limited wavelengths. However, the selected wavelengths had a weak mechanism interpretation about soil TN, which might lead to the lack of universality and predictive ability of the established models. In this study, we put forward to a novel method based on ACO and MI to select the sensitive wavebands of soil TN content and verified universality and estimation ability of the selected wavelengths from the aspects of mechanism, model types and sample's composition.

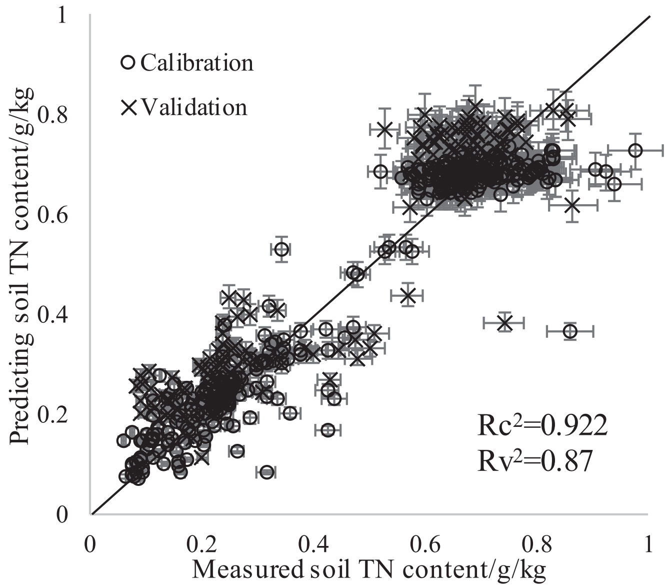

Fig. 12. Calibration and validation of soil TN prediction of the SVM regression models.

Table 2
Accuracies of PLS, MLR and SVM models.

	PLS	MLR	SVM
Calibration R^2	0.748	0.872	0.922
RMSE of calibration (g/kg)	0.136	0.087	0.069
Validation R^2	0.721	0.841	0.871
RMSE of prediction (g/kg)	0.145	0.1	0.099
RPD	2.626	3.896	3.864

(1) According to the final criterion of the sensitive wavelength selection proposed in this research, an innovative ACO–MI method was proposed and used to conduct the spectral features extraction of the absorbance values of 270 soil samples in the range of 800–2564 nm. Some selected wavelengths are in general agreement with previous researches (Jia et al., 2017; Kawamura et al., 2017). After the comprehensive comparison, the ACO–MI method outperformed others in reducing the autocorrelation and redundancy of the selected wavelength and eliminating the interference from soil moisture.

(2) The comparison among full spectral PLS models, Linear (MLR) and non-linear (SVM) regression models demonstrated that the SVM regression models got the best results in TN estimation of whole 626 soil samples and 528 samples with TN content in the range of 0–1 g/kg. According to the evaluation parameters (validation R^2 and RMSE of prediction), our study obtains slightly high estimation accuracies. It could be observed from Table 3, the validation R^2 of 528 samples was lower than the one of 626 samples, which is because of a lower SSR with the narrow range of TN (0–1 g/kg). Even though, the obtained validation R^2 was still comparable to the similar studies (as shown in Table 3). Besides, the RMSE of prediction of the SVM model with 528 soil samples was much lower than the others, which indicated this SVM model established by the eight selected wavelengths was sensitive to the tiny change of soil TN content. Moreover, it should be noted that the mixed sample set in our study contains the soil samples in different soil types collected from different regions and the number of samples is dramatically larger than the rest, which would all lead to relatively low accuracies. This research further verify the universality and predictive ability of the sensitive wavelength in predicting soil TN content.

Table 3
Validation results for soil TN content.

N_{sample}	Range (g/kg)	Method	R_v^2	RMSE of prediction	Authors
96	0.2/2.6	PLSR	0.58	0.35	Shi et al., 2013
130	0.3/4.7	PLSR	0.87	0.3	(Ji et al., 2014)
335	0.3/4.7	PLSR	0.77	0.32	Ji et al., 2016
140	0.59/1.42	LS-SVM	0.732	0.076	Morellos et al., 2016
62	0.6/4.4	ISE-PLS	0.949	0.19	Kawamura et al., 2017
626	0.06/2.74	SVM	0.96	0.219	This study
528	0.06/0.98	SVM	0.87	0.09	This study

Note: N_{sample} is the number of samples; R_v^2 is validation R^2.

In light of the fact that for different TN ranges, there may exists optimal estimation models with best accuracies, a further accuracy improvement can be achieved by training at multiple range steps and develop an automatic range-model selection instrument.

Given the recent advances in deep learning, another effort can be invested by employing a separate part in the model for adaptation of computing SVM model based on the “interrelations merit model” of the eight bands. That is the SVM computing (or visible layer) may be affected by another hidden layer that was trained to estimate the “relevance” of each band to the other seven and adapt the SVM parameters accordingly. In that way, some small or even large “noisy” influence on one or several bands (i.e. by water content or OM or other influential constituents) could be eliminated or reduced, or even auto band reject, and thus increase the accuracy of the estimate.

The research provide an innovative and flexible spectral feature extraction method, which could be also applied into other areas. The results of this study are significant for the processing of hyperspectral information and the development of multi-spectral sensors in the future.

4. Conclusions

To effectively extract the sensitive wavebands of soil TN content, a selection criterion based on MI and ACO methods was innovatively proposed to screen the sensitive wavebands of soil TN content. The obtained bands were then used to predict the TN content in the soil samples, which were collected from different farms and under four soil types, also including the samples under different fertilization conditions, to verify the universality and predictive ability. The main conclusions are as follows:

(1) After wavelength selection using ACO-MI method, 943, 1004, 1097, 1351, 1550, 1710, 2123, and 2254 nm were determined as soil TN content-sensitive wavebands. According to the mechanism analysis, all the eight wavelengths had direct and close relationship with TN content of soil, which verified the effectiveness of the ACO–MI method in wavelength selection of soil TN content.

(2) The overall accuracies of the MLR and SVM models based on the selected wavebands achieved higher precision than the full spectral PLS models. In addition, the SVM model reached a highest accuracy in soil TN prediction. All the results of the models indicated that the sensitive wavebands selected using ACO-MI method in this research performed well with high universality and predictive ability in predicting the soil TN content.

Acknowledgments

This paper was supported by the National Key Research and Development Program of China (2017YFD0201500-2017YFD0201501), the National Natural Science Fund of China (31471409), and the China Post-doctoral Fund (No. 2017M620500).

References

Aghdam, M.H., Kabiri, P., 2016. Feature selection for intrusion detection system using ant colony optimization. Int. J. Netw. Secur. 18, 420–432. https://doi.org/10.1007/978-3-319-32213-1_27.

Agrawal, A.P., Kaur, A., 2018. A comprehensive comparison of ant colony and hybrid particle swarm optimization algorithms through test case selection. Data Eng. Intell. Comput. 542, 397–405. https://doi.org/10.1007/978-981-10-3223-3.

Ali, Z.I., Abdel Malik, E.M., Babiker, H.M., Ramraj, V.M., Sultana, A., Johansen, C., 1998. Communications in Soil Science and Plant Analysis Iron and nitrogen interactions in groundnut nutrition Iron and Nitrogen Interactions in Groundnut Nutrition 1. Commun. Soil Sci. Plant Anal. 29, 2619–2630. https://doi.org/10.1080/00103629809370138.

Allegrini, F., Olivieri, A.C., 2011. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis. Anal. Chim. Acta 699, 18–25. https://doi.org/10.1016/j.aca.2011.04.061.

An, X.F., Li, M.Z., Zheng, L.H., Liu, Y.M., Sun, H., 2014. A portable soil nitrogen detector based on NIRS. Precis. Agric. 15, 3–16. https://doi.org/10.1007/s11119-012-9302-5.

Bansod, S.J., Thakare, S.S., 2014. Near infrared spectroscopy based a portable soil nitrogen detector design. Int. J. Comput. Sci. Inf. Technol. 5, 3953–3956.

Bansod, S.J., Thakre, S., 2014. Near infrared spectroscopy based soil nitrogen measurement -a review. Int. J. Curr. Eng. Technol. 26844.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. 5. https://doi.org/10.1109/72.298224.

Benoudjit, N., François, D., Meurens, M., Verleysen, M., 2004. Spectrophotometric variable selection by mutual information. Chemom. Intell. Lab. Syst. 74, 243–251. https://doi.org/10.1016/j.chemolab.2004.04.015.

Brown, G., Pocock, A., Zhao, M.J., Luján, M., 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J. Mach. Learn. Res. 13, 27–66.

Cai, W., Li, Y., Shao, X., 2008. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. Chemom. Intell. Lab. Syst. 90, 188–194. https://doi.org/10.1016/j.chemolab.2007.10.001.

Chacón Iznaga, A., Rodríguez Orozco, M., Aguila Alcantara, E., Carral Pairol, M., Díaz Sicilia, Y.E., de Baerdemaeker, J., Saeys, W., 2014. Vis/NIR spectroscopic measurement of selected soil fertility parameters of Cuban agricultural Cambisols. Biosyst. Eng. 125, 105–121. https://doi.org/10.1016/j.biosystemseng.2014.06.018.

Chakraborty, S., Li, B., Deb, S., Paul, S., Weindorf, D.C., Das, B.S., 2017. Geoderma Predicting soil arsenic pools by visible near infrared diffuse reflectance spectroscopy. Geoderma 296, 30–37. https://doi.org/10.1016/j.geoderma.2017.02.015.

Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65, 480–490.

Cheng, H., Qin, Z., Feng, C., Wang, Y., Li, F., 2011. Conditional mutual information-based feature selection analyzing for synergy and redundancy. ETRI J. 33, 210–218. https://doi.org/10.4218/etrij.11.0110.0237.

Colorni, A., Dorigo, M., Maniezzo, V., 1991. Distributed optimization by ants colonies. In: Proc. ECAL - Eur. Conf. Artif. Life, Paris, Fr vol. 12.

Daniel, K.W., Tripathi, N.K., Honda, K., Apisit, E., 2004. Analysis of VNIR 400 1100 nm spectral signatures for estimation of soil organic matter in tropical soils of Thailand. Int. J. Remote Sens. 25, 643–652.

Debaene, G., Niedźwiecki, J., Pecio, A., Żurek, A., 2014. Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. Geoderma 214, 114–125. https://doi.org/10.1016/j.geoderma.2013.09.022.

Demattê, J.A., Campos, R.C., Alves, M.C., Fiorio, P.R., Nanni, M.R., 2004. Visible–NIR reflectance: a new approach on soil evaluation. Geoderma 121, 95–112. https://doi.org/10.1016/j.geoderma.2003.09.012.

Dorigo, M., Maniezzo, V., Colorni, A., 1996. Ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybern. B 26, 29–41.

Duch, W., 2006. Feature extraction: foundations and applications. Stud. Fuzziness Soft Comput. 3, 89–117.

Florinsky, I., Eilers, R., Manning, G., Fuller, L., 2002. Prediction of soil properties by digital terrain modelling. Environ. Model. Softw. 17, 295–311. https://doi.org/10.1016/S1364-8152(01)00067-6.

Galvão, L.S., Pizarro, M.A., Epiphanio, J.C.N., 2001. Variations in reflectance of tropical soils: spectral-chemical composition relationships from AVIRIS data. Remote Sens. Environ. 75, 245–255. https://doi.org/10.1016/S0034-4257(00)00170-X.

Gao, H.Z., Lu, Q.P., Ding, H.Q., Peng, Z.Q., 2009. Choice of characteristic near-infrared wavelengths for soil total nitrogen based on successive projection algorithm. Spectrosc. Spectr. Anal. 29, 2951–2954.

Guo, Z.M., Huang, W.Q., Peng, Y.K., Wang, X., Tang, X.Y., 2014. Adaptive Ant Colony Optimization Approach to Characteristic Wavelength Selection of NIR Spectroscopy. Chin. J. Anal. Chem. 42, 513–518. https://doi.org/10.3724/SP.J.1096.2014.30340.

Helgerud, T., Segtnan, V.H., Wold, J.P., Ballance, S., Knutsen, S.H., Rukke, E.O., Afseth, N.K., 2012. Near-infrared spectroscopy for rapid estimation of dry matter content in whole unpeeled potato tubers. J. Food Res. 1, 55–65.

Hou, J., Li, L., He, J., 2016. Detection of grapevine leafroll disease based on 11-index imagery and ant colony clustering algorithm. Precis. Agric. 17, 488–505. https://doi.org/10.1007/s11119-016-9432-2.

Hu, Q., Zhang, L., Zhang, D., Pan, W., An, S., Pedrycz, W., 2011. Measuring relevance between discrete and continuous features based on neighborhood mutual information. Expert Syst. Appl. 38, 10737–10750. https://doi.org/10.1016/j.eswa.2011.01.023.

Igne, B., Reeves, J.B., McCarty, G., Hively, W.D., Lundc, E., Hurburgh, C.R., 2010. Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils. J. Near Infrared Spectrosc. 18, 167–176. https://doi.org/10.1255/jnirs.883.

Ji, W., Shi, Z., Huang, J., Li, S., 2014. In situ measurement of some soil properties in paddy soil using visible and near-infrared spectroscopy. PLoS One 9 (159785). https://doi.org/10.1371/journal.pone.0105708.

Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R.A., Mouazen, A.M., 2016. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. Soil Tillage Res. 155, 492–500. https://doi.org/10.1016/j.still.2015.06.004.

Jia, S., Li, H., Wang, Y., Tong, R., Li, Q., 2017. Hyperspectral imaging analysis for the classification of soil types and the determination of soil total nitrogen. Sensors 17, 2252. https://doi.org/10.3390/s17102252.

Kawamura, K., Tsujimoto, Y., Rabenarivo, M., Asai, H., Andriamananjara, A., Rakotoson, T., 2017. Vis-NIR spectroscopy and PLS regression with waveband selection for estimating the total C and N of paddy soils in Madagascar. Remote Sens. 9. https://doi.org/10.3390/rs9101081.

Ke, L., Feng, Z., Ren, Z., 2008. An efficient ant colony optimization approach to attribute reduction in rough set theory. Pattern Recogn. Lett. 29, 1351–1357. https://doi.org/10.1016/j.patrec.2008.02.006.

Kuang, B., Mouazen, A.M., 2013. Non-biased prediction of soil organic carbon and total nitrogen with vis–NIR spectroscopy, as affected by soil moisture content and texture. Biosyst. Eng. 114, 249–258. https://doi.org/10.1016/j.biosystemseng.2013.01.005.

Kweon, G., Maxton, C., 2013. Soil organic matter sensing with an on-the-go optical sensor. Biosyst. Eng. 115, 66–81. https://doi.org/10.1016/j.biosystemseng.2013.02.004.

Leiva-Valenzuela, G.A., Lu, R., Aguilera, J.M., 2013. Prediction of firmness and soluble solids content of blueberries using hyperspectral reflectance imaging. J. Food Eng. 115, 91–98. https://doi.org/10.1016/j.jfoodeng.2012.10.001.

Lewis, D.D., 1992. Feature selection and feature extract ion for text categorization. In: In Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics, Chicago, pp. 212–217.

Li, M.Z., 2006. Spectral Analysis Technique and Its Application. Science Press, Beijing.

Li, M.Z., Pan, L., Zheng, L.H., An, X.F., 2010. Development of a portable SOM detector based on NIR diffuse reflection. Spectrosc. Spectr. Anal. 30, 1146–1150.

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., Buttafuoco, G., 2017. Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma 288, 175–183. https://doi.org/10.1016/j.geoderma.2016.11.015.

Mcgladdery, C., Weindorf, D.C., Chakraborty, S., Li, B., Paulette, L., Podar, D., Pearson, D., Yaw, N., Kusi, O., Duda, B., 2018. Elemental assessment of vegetation via portable X-ray fluorescence (PXRF) spectrometry. J. Environ. Manag. 210, 210–225. https://doi.org/10.1016/j.jenvman.2018.01.003.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57, 443–452. https://doi.org/10.2136/sssaj1993.03615995005700020026x.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., Mouazen, A.M., 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst. Eng. 152, 104–116. https://doi.org/10.1016/j.biosystemseng.2016.04.018.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022.

Palacios-Orueta, A., Ustin, S.L., 1998. Remote sensing of soil properties in the Santa Monica Mountains I. Spectral analysis. Remote Sens. Environ. 65, 170–183. https://doi.org/10.1016/S0034-4257(98)00024-8.

Pan, L., Lu, R., Zhu, Q., Tu, K., Haiyan, Cen, 2016. Predict compositions and mechanical properties of sugar beet using hyperspectral scattering. Food Bioprocess Technol. 1–10.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.

Petropoulos, G.P., Arvanitis, K., Sigrimis, N., 2012. Hyperion hyperspectral imagery analysis combined with machine learning classifiers for land use/cover mapping. Expert Syst. Appl. 39, 3800–3809. https://doi.org/10.1016/j.eswa.2011.09.083.

Santana, L.E.A., Silva, L., Canuto, A.M.P., Pintro, F., Vale, K.O., 2010. A comparative analysis of genetic algorithm and ant colony optimization to select attributes for an heterogeneous ensemble of classifiers. IEEE Congr. Evol. Comput. 1–8. https://doi.org/10.1109/CEC.2010.5586080.

Shi, T., Cui, L., Wang, J., Fei, T., Chen, Y., Wu, G., 2013. Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. Plant Soil 366, 363–375. https://doi.org/10.1007/s11104-012-1436-8.

Sinfield, J.V., Fagerman, D., Colic, O., 2010. Evaluation of sensing technologies for on-the-go detection of macro-nutrients in cultivated soils. Comput. Electron. Agric. 70, 1–18. https://doi.org/10.1016/j.compag.2009.09.017.

Sorol, N., Arancibia, E., Bortolato, S.A., Olivieri, A.C., 2010. Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice: a test field for variable selection methods. Chemom. Intell. Lab. Syst. 102, 100–109. https://doi.org/10.1016/j.chemolab.2010.04.009.

Sun, H., Li, M.Z., Zhao, Y., Li, X.H., 2009. Estimation of soil parameters in Fuxin opencast coal mine. Intell. Autom. Soft Comput. 15, 1–7.

Varma, P.R.K., Kumari, V.V., Kumar, S.S., 2016. Feature selection using relative fuzzy entropy and ant colony optimization applied to real-time intrusion detection system. Procedia Comput. Sci. 85, 503–510. https://doi.org/10.1016/j.procs.2016.05.203.

Vinh, N.X., Zhou, S., Chan, J., Bailey, J., 2016. Can high-order dependencies improve mutual information based feature selection? Pattern Recogn. 53, 46–58. https://doi.org/10.1016/j.patcog.2015.11.007.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.

Vohland, M., Besold, J., Hill, J., Fründ, H.-C., 2011. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma 166, 198–205. https://doi.org/10.1016/j.geoderma.2011.08.001.

Williams, P.C., 1987. Variables affecting near-infrared reflectance spectroscopic analysis. Near-Infrared Technol. Agric. Food Ind. 143–147.

Xue, B., Zhang, M., Browne, W.N., Yao, X., 2016. A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. 20, 606–626. https://doi.org/10.1109/TEVC.2015.2504420.

Zhang, Y., Zheng, L., Li, M., An, X., Sun, H., 2014. Soil nitrogen content modeling based on spectral analysis. In: 2014 ASABE Annu. Int. Meet, https://doi.org/10.13031/aim.20141912501.

Zhang, Y., Li, M., Zheng, L., Yang, W., 2015. Prediction of soil total nitrogen content in different layers based on near infrared spectral analysis. Trans. Chinese Soc. Agric. Eng. 31, 121–126.

Zhang, Y., Li, M., Zheng, L., Zhao, Y., Pei, X., 2016. Soil nitrogen content forecasting based on real-time NIR spectroscopy. Comput. Electron. Agric. 124, 29–36. https://doi.org/10.1016/j.compag.2016.03.016.