J Arid Land (2023) 15(2): 191–204
https://doi.org/10.1007/s40333-023-0094-4

Science Press Springer-Verlag

Estimation of soil organic matter in the Ogan-Kuqa River Oasis, Northwest China, based on visible and near-infrared spectroscopy and machine learning

ZHOU Qian[1,2,3], DING Jianli[1,2,3*], GE Xiangyu[1,2,3], LI Ke[1,2,3], ZHANG Zipeng[1,2,3], GU Yongsheng[1,2,3]

[1] College of Geography and Remote Sensing Science, Xinjiang University, Urumqi 830046, China;
[2] Xinjiang Key Laboratory of Oasis Ecology, Xinjiang University, Urumqi 830046, China;
[3] Key Laboratory of Smart City and Environment Modelling of Higher Education Institute, Xinjiang University, Urumqi 830046, China

Abstract: Visible and near-infrared (vis-NIR) spectroscopy technique allows for fast and efficient determination of soil organic matter (SOM). However, a prior requirement for the vis-NIR spectroscopy technique to predict SOM is the effective removal of redundant information. Therefore, this study aims to select three wavelength selection strategies for obtaining the spectral response characteristics of SOM. The SOM content and spectral information of 110 soil samples from the Ogan-Kuqa River Oasis were measured under laboratory conditions in July 2017. Pearson correlation analysis was introduced to preselect spectral wavelengths from the preprocessed spectra that passed the 0.01 level significance test. The successive projection algorithm (SPA), competitive adaptive reweighted sampling (CARS), and Boruta algorithm were used to detect the optimal variables from the preselected wavelengths. Finally, partial least squares regression (PLSR) and random forest (RF) models combined with the optimal wavelengths were applied to develop a quantitative estimation model of the SOM content. The results demonstrate that the optimal variables selected were mainly located near the range of spectral absorption features (i.e., 1400.0, 1900.0, and 2200.0 nm), and the CARS and Boruta algorithm also selected a few visible wavelengths located in the range of 480.0–510.0 nm. Both models can achieve a more satisfactory prediction of the SOM content, and the RF model had better accuracy than the PLSR model. The SOM content prediction model established by Boruta algorithm combined with the RF model performed best with 23 variables and the model achieved the coefficient of determination (R^2) of 0.78 and the residual prediction deviation (RPD) of 2.38. The Boruta algorithm effectively removed redundant information and optimized the optimal wavelengths to improve the prediction accuracy of the estimated SOM content. Therefore, combining vis-NIR spectroscopy with machine learning to estimate SOM content is an important method to improve the accuracy of SOM prediction in arid land.

Keywords: soil organic matter content; vis-NIR spectroscopy; random forest; Boruta algorithm; machine learning

Citation: ZHOU Qian, DING Jianli, GE Xiangyu, LI Ke, ZHANG Zipeng, GU Yongsheng. 2023. Estimation of soil organic matter in the Ogan-Kuqa River Oasis, Northwest China, based on visible and near-infrared spectroscopy and machine learning. Journal of Arid Land, 15(2): 191–204. https://doi.org/10.1007/s40333-023-0094-4

1 Introduction

Soil organic matter (SOM) is an essential parameter to evaluate soil fertility and soil quality, plays

*Corresponding author: DING Jianli (E-mail: watarid@xju.edu.cn)
Received 2022-11-12; revised 2023-01-20; accepted 2023-02-01

a critical function in the stability and security of local ecosystems, and is of major significance for regional sustainable development (Ding and Yu, 2014; McBratney et al., 2014). SOM information is traditionally obtained by laboratory chemical analysis; however, this method is relatively complex, inefficient, and uneconomical and cannot meet the needs of smart agriculture. Therefore, the establishment of an efficient, low-cost, and modern method for SOM determination is an urgent task.

In recent years, narrow-band spectra in the visible and near-infrared (vis-NIR) range have attracted much attention in soil property prediction studies due to the maturity of proximity sensing technology, which provides technical support for the accurate estimation of SOM content (Wang et al., 2022). Many scholars have explored the relationship between SOM and soil spectra using vis-NIR spectroscopy technique. Proximity sensing technology with fine spectral resolution is used to obtain continuous spectral information of features at the nanometer level. SOM has a variety of functional groups (including hydroxyl, carboxyl, etc.), which have characteristic absorption in the vis-NIR spectral regions, and the intensity of absorption at different wavelengths corresponds to the molecular structure and concentration of the substance (Zhang et al., 2021; Xie et al., 2022). Therefore, the quantitative estimation of SOM through vis-NIR spectroscopy is of great practical significance. However, since ground object spectra provide hundreds of variables, there is redundancy between variables, and the variables are usually nonlinearly correlated with soil sample properties (Viscarra Rossel et al., 2006). At the same time, there is background noise in the spectra as well as interference from specific physical factors (Tian et al., 2013). In addition, Swierenga et al. (2000) suggested that choosing wavelengths with strong information and less interference from external factors is an effective way to construct stable spectral analysis models. Therefore, the prerequisite for building SOM content prediction model is to determine the appropriate characteristic spectral wavelengths.

In selecting the characteristic spectral wavelengths of vis-NIR spectra, the competitive adaptive reweighted sampling (CARS) (Liu et al., 2021), genetic algorithm (GA) (Chen et al., 2022; Yin et al., 2022), successive projection algorithm (SPA) (Mesquita et al., 2018), and uninformative variable elimination (Song et al., 2020) methods have been more widely used. The CARS algorithm has been shown to be able to select the optimal combination of spectral variables from full wavelength data to reveal the relationship between spectral reflectance and soil properties (Xing et al., 2021). After the spectra were preprocessed, Liu et al. (2021) used the CARS algorithm to screen the response characteristics of SOM and used the random forest (RF) method to build a prediction model to realize an accurate assessment of the organic matter content of agricultural soils. The SPA highly summarizes the information of most of the sample spectra, avoiding overlapping information (Shi et al., 2014). In these studies, most of the traditional feature selection algorithms follow the principle of min-optimality, which makes them overly dependent on the smallest subset of features and leads to errors and uncertainties in the selection of classifications. Compared to other feature selection and importance ranking algorithms, the Boruta algorithm not only provides a simple ranking of variables but also classifies all variables in order and groups them into three categories: strongly correlated variables, moderately correlated variables, and weakly correlated variables (Chen et al., 2022). Additionally, as the Boruta algorithm is based on the RF classification algorithm, the method can be used to detect linear and nonlinear relationships between soil properties and environmental predictors. Therefore, the Boruta algorithm becomes an important approach for feature selection. The partial least squares regression (PLSR) algorithm is a more common modelling method that is better able to solve the problem of multicollinearity between independent variables (Shi et al., 2016). In previous studies, the RF model was used as a hierarchical nonparametric method for estimating complex nonlinear relationships between independent and dependent variables (Zhang et al., 2019). The RF model is not prone to falling into overfitting due to the number of variables being much larger than the number of modelled samples and has good resistance to noise (Ge et al., 2022a). In addition, there have been no uniform standard feature wavelength selection methodologies presented in previous studies, and the results from different feature wavelength selection strategies in combination with various

modelling methods are significantly different. Therefore, it is a challenge to address the adaptability of wavelength selection methods and modelling schemes.

Therefore, in this study, we studied 110 surface soil samples from the Ogan-Kuqa River Oasis in Xinjiang Uygur Autonomous Region, China, and collected and measured vis-NIR spectral data. The objectives of this study were to: (1) analyze the spectral characteristics of the soil in the Ogan-Kuqa River Oasis; (2) obtain preselected significance variables from preprocessed spectra by Pearson correlation analysis and then acquire the spectral response characteristics of SOM using CARS, SPA, and Boruta algorithms; and (3) develop SOM content prediction models for PLSR and RF based on preselected and optimal variables. This result provides methodological guidance for the fast and efficient estimation of SOM content in arid regions using vis-NIR spectroscopy technique.

2 Materials and methods

2.1 Study area and sampling sites

The Ogan-Kuqa River Oasis (41°06′–41°40′N, 82°10′–83°50′E) is located in the northern Tarim Basin of Xinjiang Uygur Autonomous Region, China, with a total area of 9.5×10^3 km^2 (Fig. 1). The temperature difference between day and night is relatively large in the region. Rainfall is low and evaporation is high. The annual average temperature is 10.5°C–14.4°C, the maximum temperature is 41.1°C, the average annual precipitation is only 43.1 mm, and the evaporation is relatively high, which makes this region a typical arid and extremely arid area (Han et al., 2022). The soil texture is mainly clay loam, chalky clay loam, loamy clay, and chalky clay. Land cover and land use types mainly include agricultural land, grassland, bareland, woodland, and saline land.

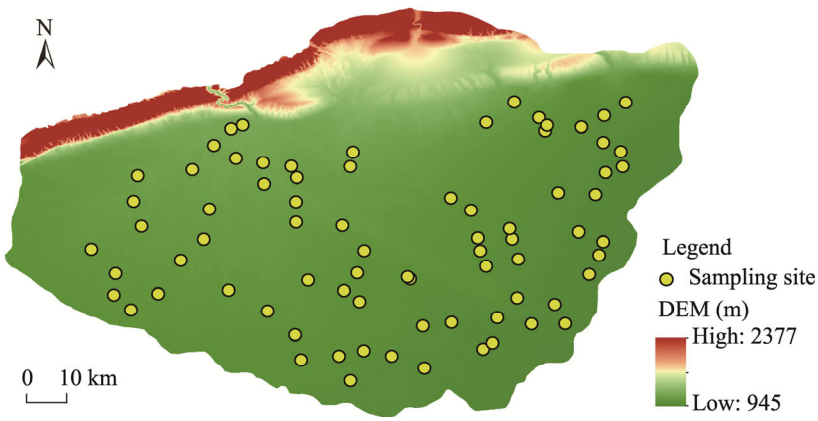

Fig. 1 Overview of the Ogan-Kuqa River Oasis and spatial distribution of sampling sites. DEM, Digital Elevation Model.

2.2 Soil sample collection and chemical analysis

From 7 July to 19 July in 2017, we collected the surface soil (0–20 cm) of the oasis area according to the five-point sampling method, with a collection sample square of 30 m×30 m; five samples were mixed into a single soil sample. A total of 144 soil samples were collected, covering different land cover and land use types in the inner area of the oasis, including agricultural land, wasteland, saline land, and forestland, and the locations of the sampling sites were recorded using GPS (LT500T, CHC Navigation Technology Co. Ltd., Shanghai, China). The accuracy of the GPS measurement is approximately 1 m. The soil samples were retrieved in sealed bags, naturally air-dried, ground, and sieved (≥0.15 mm) in the laboratory after removing debris (stones, plant roots, and humus) from the soil samples. Soil samples were prepared in two parts that were used for spectroscopic measurements and SOM analysis. The SOM content was obtained using the potassium dichromate oxidation method heated using an electric sand bath (Jin et al., 2016).

2.3 Soil spectra collection and preprocessing

Soil reflection spectra were measured by an ASD FieldSpec®3 portable spectrometer (Analytical Spectral Devices, Boulder, Colorado, USA) with a wavelength range of 350.0–2500.0 nm, in which the sampling interval from the range of 350.0–1000.0 nm was 1.4 nm, and the sampling interval from the range of 1000.0–2500.0 nm was 2.0 nm. The number of output wavelengths was 2151. Soil spectra were measured in a dark environment using a 50-W halogen lamp as the light source with a distance of 30 cm between the light source and the soil surface and a zenith angle of 5° for the halogen lamp. Before the measurement, we used a reference white board to obtain the absolute reflectance. Each soil sample was tested five times, and the arithmetic mean was taken as the reflectance of that sample, which was averaged into one spectrum as the final reflectance spectrum.

Spectral data contain both chemical information about the sample itself and irrelevant information and noise, i.e., the linear or nonlinear transformation and signal noise problems caused by absorption and scattering of signal intensity at the soil surface (Jin et al., 2016). Therefore, the edge wavelengths from 350.0 to 399.0 nm and from 2401.0 to 2500.0 nm were removed from the original spectrum. The reflectance spectra of the original band were processed by Savitzky-Golay (SG) smoothing and first derivative (FD) processing. The SG smoothing method reduces the noise to enhance the signal-to-noise ratio (Savitzky and Golay, 1964). The FD processing is used to differentiate overlapping peaks, attenuate feature background interference, repair baseline drift, sharpen spectral features, and capture minute details of the spectral curves (Wang et al., 2018; Ge et al., 2022b). The SG smoothing and FD processing were implemented in R software with the "prospectr package".

In addition, in order to avoid the impact of outlier sample values on the performance of the prediction model, we applied the Monte Carlo outlier detection (MCOD) method to remove sample outliers prior to modelling in this study (Schomberg et al., 2018). The MCOD method was carried out using the toolbox in MATLAB software. The outlier plot of 144 soil samples generated through the MCOD method is shown in Figure 2. The plot was divided into 4 areas; about 34 points were identified as sample outliers and were excluded from subsequent study. The remaining 110 points were used as valid samples for the follow-up study.

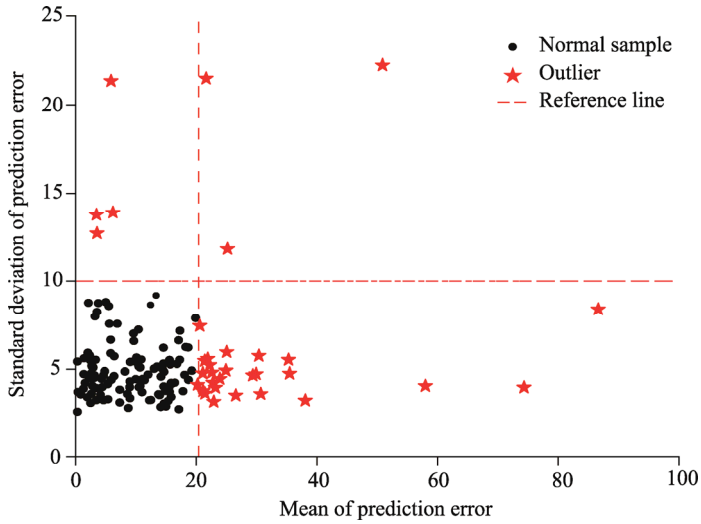

Fig. 2 Plot of outliers detected through the Monte Carlo outlier detection (MCOD) method

2.4 Feature variable selection method

2.4.1 Competitive adaptive reweighted sampling (CARS)

The CARS algorithm is used to select the characteristic wavelengths of soil spectra by imitating the principle of "survival of the fittest" in Darwin's evolutionary theory. In each iteration, the

wavelength variables with large absolute regression coefficient values in the PLSR model are retained, and those with small absolute regression coefficient values are removed by the adaptive reweighted sampling technique to obtain a series of subsets of wavelength variables. Then, the exponential decay function and adaptive reweighted sampling method are used to achieve a competitive selection of variables. The root mean square error of cross-validation (RMSECV) is calculated using the cross-validation method. We selected the best subset of wavelength variables according to the principle of minimization of RMSECV values (Li et al., 2019; Xing et al., 2021).

The CARS algorithm used in this study was run in MATLAB software. The optimal variables were selected by the MCOD method, in which the number of Monte Carlo samples was set to 50, and iterations of the sampling times were performed. By comparing the RMSECV values of each sample, the variables of the corresponding sampling times were selected as the optimal set of variables when their values were minimal.

2.4.2 Successive projections algorithm (SPA)

The SPA is a vector space covariance minimization algorithm for forwarding variable selection. It aims to improve the covariance between variables by quickly filtering multiple feature wavelengths from the full wavelength using simple projection operations so that the covariance between variables is improved and the computational effort is greatly reduced, thus increasing the modelling speed. Details of the SPA operations are given in the literature (Araújo et al., 2001). The SPA was run in MATLAB software.

2.4.3 Boruta algorithm

The Boruta algorithm obtains the importance of all features in the dataset with respect to the target variable, selects the important features, and removes the redundant feature variables (Keskin et al., 2019). This algorithm features a black box prediction model with good prediction accuracy to obtain the importance indices related to the target variables. The essential idea of Boruta algorithm is to evaluate the importance of each feature variable through a circular method. By replicating the original set of features, a random mixture of each feature value is used to construct a shadow feature with randomness; the final sample dataset of the model is a new feature set created by combining the original features and the shadow features. In each iteration of the RF algorithm, we compared the importance scores of the original features and the shadow features to select the optimal set of features for modelling (Kursa et al., 2010). The importance score (Z score) in the Boruta algorithm is based on the out-of-bag error of the RF model. The equation is as follows:

$$\mathrm{MSE_{OOB}} = \left(\mathrm{y}_i - \hat{y}_{i\mathrm{OOB}}\right)^2 / N \,, \tag{1}$$

where $\mathrm{MSE_{OOB}}$ is the out-of-bag error in the RF model; y_i is the observed SOM of sample i (g/kg); $\hat{y}_{i\mathrm{OBB}}$ is the predicted SOM value of the out-of-bag sample of sample y_i (g/kg); and N is the number of samples.

$$Z\ \mathrm{score} = \overline{\mathrm{MSE}}_{\mathrm{OOB}} / \mathrm{SDMSE_{OOB}} \,, \tag{2}$$

where $\overline{\mathrm{MSE}}_{\mathrm{OBB}}$ is the mean of the out-of-bag error; and $\mathrm{SDMSE_{OOB}}$ is the standard deviation of the out-of-bag error.

The final result is based on the maximum Z score of the shadow feature (shadowMax) as the filtering indicator. When the Z score of the feature variable is larger than shadowMax, the feature is considered to be important; otherwise, the variable is considered to be unimportant and is not used for modelling (Ge et al., 2022a).

2.5 Calibration method

2.5.1 Partial least squares regression (PLSR)

The PLSR model combines the advantages of principal component analysis, typical correlation analysis, and multiple linear regression and is used to better address strong covariance and a number of variables exceeding the number of available samples (Chang et al., 2001; Wang et al., 2019). This study used ten-fold cross-validation to determine the root mean square error (RMSE)

to identify the optimal number of latent variables for the PLSR model. The "libPLS package" in R software was used to implement the model.

2.5.2 Random forest (RF) model

The RF model is a decision tree-based classification regression algorithm that uses the bootstrap sampling method to randomly select some samples from the original data and decision tree modelling for each sample data, where each decision tree is not linked to each other, and finally, the predicted value of the model is obtained by combining the voting results of all decision trees (Zhang et al., 2019; Ma et al., 2021). The RF model performs well for many datasets, does not easily overfit, and has some advantages in data modelling. Before applying the model, the parameters in the model need to be optimized, and these parameters have a large impact on the model performance. When running the RF model, there are three parameters to be defined: the number of trees ("ntree"), the minimum node size ("nodeSize"), and the number of input variables randomly selected as candidates at each split ("mtry"). We set the "ntree" to 1000 after repeated tests. Then, we used a grid search technique with ten-fold cross-validation to optimize "mtry" and "nodeSize", and selected the best parameters based on RMSE minimization of the cross-validations. Furthermore, we also set the "mtry" to 2–30 with a step size of 2, and the "nodeSize" to 1–10 with a step size of 1.

2.5.3 Assessment of the prediction quality

In this study, we divided 110 samples into three groups using the Kennard–Stone algorithm, with two groups serving as the training set (74 samples) and one serving as the validation set (36 samples). The performance of each model was evaluated by the coefficient of determination (R^2), RMSE, and residual prediction deviation (RPD) (Chang et al., 2001). The smaller the RMSE of the validation set, the larger the R^2; and the greater the RPD, the better the model prediction. According to previous studies (Nocita et al., 2014; Bao et al., 2017; Luo et al., 2022), RPD less than 1.4 denotes that the model is poor and is unable to predict the real sample; when RPD is greater than or equal to 1.4 and less than or equal to 2.0, the prediction results are barely acceptable but need further improvement; and when RPD is greater than 2.0, it demonstrates that the model can achieve better performance. The formulae for the three evaluation indicators are as follows:

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y}_i)^2}, \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}, \tag{4}$$

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSE}}, \tag{5}$$

where R^2 is the coefficient of determination between the predicted SOM and measured SOM; RMSE is the root mean square error of SOM in test set (g/kg); RPD is residual prediction deviation; SD is the standard deviation of the observed SOM (g/kg); and $\overline{y}_i$ is the average of the observed SOM (g/kg).

3 Results

3.1 Descriptive statistics of the soil organic matter (SOM) content

The statistical characteristics of the SOM content are shown in Table 1. The SOM content ranged from 5.49 to 59.86 g/kg with a mean and standard deviation of 29.05 and 11.34 g/kg, respectively. The mean value of the SOM content in the calibration and validation sets was 28.59 g/kg and 29.99 g/kg, respectively. The coefficients of variation for the full sample set, calibration set, and validation set were 39.04%, 39.77%, and 36.57%, respectively, which were moderate variation, implying that the division of samples was reasonable.

Table 1 Statistical characteristics of the soil organic matter (SOM) content

Sample	Number of samples	Minimum (g/kg)	Maximum (g/kg)	Mean (g/kg)	Standard deviation (g/kg)	Coefficient of variation (%)
Full sample set	110	59.86	5.49	29.05	11.34	39.04
Calibration set	74	59.86	5.49	28.59	17.09	39.77
Validation set	36	52.91	9.94	29.99	10.97	36.57

3.2 Soil spectral analysis

The measured soil spectra showed that the reflectance spectral curves of all soil samples had roughly the same trend. In the 400.0–800.0 nm interval, the curves increased with increasing reflectance; after 800.0 nm, the curves were generally smooth except for the moisture absorption valley. Compared with the original spectral curves, the spectra after SG smoothing did not change much, with only the spectral curves becoming smoother. Therefore, FD preprocessing was implemented on the basis of SG smoothing of the spectral curve in this study. As shown in Figure 3, the FD spectral curves showed reduced spacing, increased density, and significantly enhanced spectral feature regions when compared with the original spectral curves.

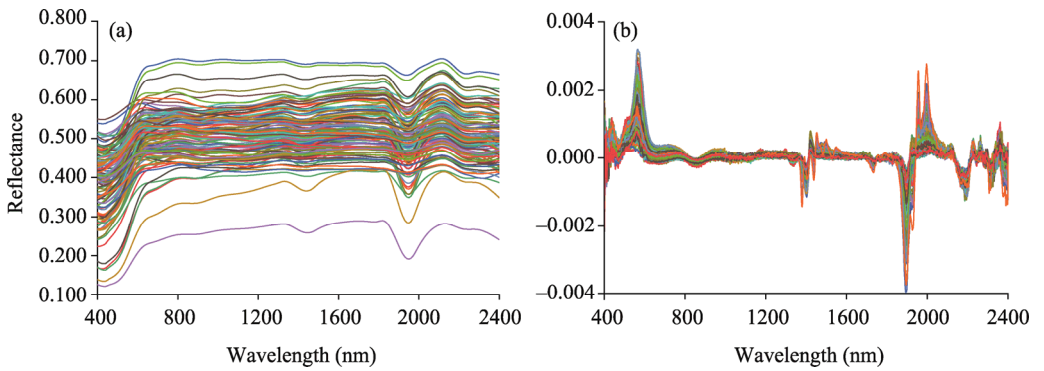

Fig. 3 Reflectance curves of the original and preprocessed soil spectra. (a), original spectra; (b), spectra processed by Savitzky-Golay (SG) smoothing and first derivative (FD) processing. Note that the curves with different color represent the reflectance spectra of different soil samples.

3.3 Correlation analysis of the SOM content with original and preprocessed soil spectra

The correlation coefficient curves are derived by analyzing the correlation between the SOM content and preprocessed soil spectra (Fig. 4). The correlation curve between the original spectra and the SOM content was relatively smooth, and only the 1810.0–1850.0 nm wavelengths passed significance testing at the 0.01 level, indicating that the sensitivity of the original spectra to the SOM content was low. Based on SG smoothing, the overall correlation of the FD-treated spectra was significantly improved, especially at 750.0–950.0 and 1220.0–2350.0 nm, with a maximum absolute correlation coefficient of 0.479 at 843.0 nm. There was a carbon-hydrogen (C-H) bond near this wavelength, which is directly related to the SOM content. Therefore, we selected 442 wavelengths that passed the significance test at the 0.01 level for subsequent comparative analysis and modelling predictions based on the results of the FD processing spectra.

3.4 Characteristic wavelength optimization

3.4.1 CARS algorithm to extract feature variables

Figure 5 shows the variable selection process of the CARS algorithm. It can be seen that the number of retained wavelengths gradually decreased as the number of iterations increased, and the rate of decrease was from fast to slow (Fig. 5a). The RMSECV showed a trend from large to small and then from small to large, and the RMSECV was the smallest (9.44) when the number of iterations was 26 (Fig. 5b). This was because during the variable selection process from 1 to 26, the RMSECV decreased by continuously eliminating wavelengths that were less correlated with

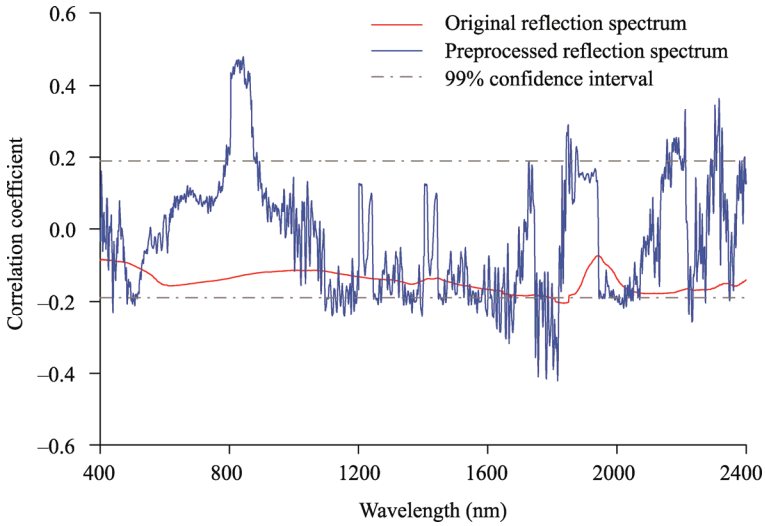

Fig. 4 Correlation coefficient curves between the soil organic matter (SOM) content and preprocessed soil spectra

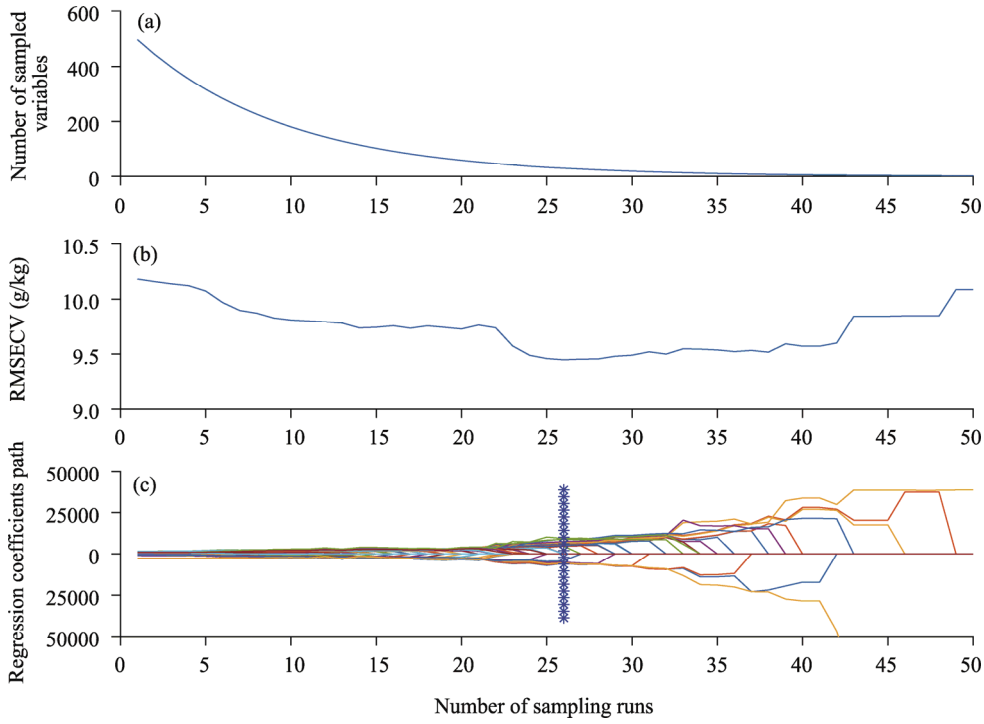

Fig. 5 Process of filtering variables by the competitive adaptive reweighted sampling (CARS) algorithm. (a), changing trend of the number of sampled variables with the increase of sampling runs; (b), changing trend of the root mean square error of cross-validation (RMSECV) with the increase of sampling runs; (c), trend regression coefficient paths with the increase of sampling runs. Note that the curves with different color represent the trend of the stability of each variable with the number of sampling runs, and the positions marked by vertical asterisks correspond to the optimal subset of variables that the RMSECV reached its minimum in the whole variable selection process.

the SOM content and had little impact on the modelling results. After 26 iterations, the wavelengths with strong correlation with the SOM content started to be removed, resulting in an increase in the RMSECV. Figure 5c presents the stability trajectory of the wavelength variables. Each curve in the plot shows the trend of the stability of each variable with the number of iterations, and the optimal

subset of variables with the smallest RMSECV is marked with an asterisk. Thus, the set of variables corresponding to the 26th sampling was the optimal subset of the SOM spectral variables, containing 31 spectral variables: 463.0, 468.0, 476.0, 790.0, 791.0, 792.0, 793.0, 794.0, 795.0, 803.0, 804.0, 805.0, 806.0, 811.0, 812.0, 1338.0, 1347.0, 1348.0, 1349.0, 1350.0, 1816.0, 1817.0, 2177.0, 2178.0, 2211.0, 2274.0, 2303.0, 2316.0, 2325.0, 2385.0, and 2386.0 nm.

3.4.2 SPA to extract feature variables

SPA was used to select the feature variables combined with the spectral data. The range of feature variable variation to be selected was set to from 1 to 10 (Fig. 6), and the settings of the calibration set and prediction set samples were kept constant. Figure 6a shows the RMSE trend with the number of variables included in the model. During the change in the number of feature variables, the horizontal coordinate is the number of variables included in the model, and the vertical coordinate is the RMSE. As the number of variables included in the model increased, the minimum RMSE gradually decreased, reaching a minimum (9.47) when the number of variables included in the model reached 5. When the number of variables included in the model increased to close to 6, further increases introduced wavelength variables that were unrelated to the predicted values or variables with greater noise, and the RMSE then increased. Figure 6b shows the distribution of the feature variables on the first calibration object. The algorithm selected five optimal variables: 835.0, 1347.0, 1769.0, 1874.0, and 2177.0 nm.

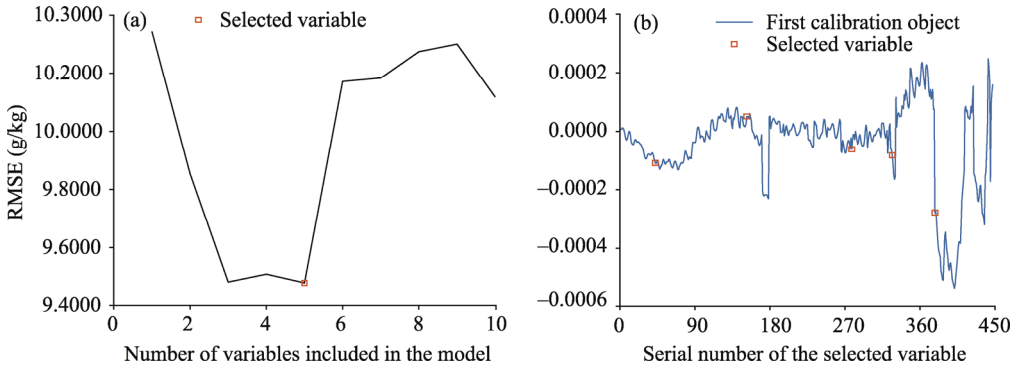

Fig. 6 Process of filtering variables by the successive projections algorithm (SPA). (a), variation in the root mean square error (RMSE) with the number of variables included in the model; (b), distribution of the feature variables on the first calibration object.

3.4.3 Boruta algorithm to select feature variables

When the *Z* score of the feature variable is larger than shadowMax, the feature is considered to be important. As seen in Figure 7, the maximum value of the shadowMax is 3.15, and there were 23 feature wavelengths with a *Z* score larger than the maximum value of the shadowMax, namely, 488.0, 491.0, 806.0, 809.0, 822.0, 823.0, 824.0, 1221.0, 1243.0, 1466.0, 1447.0, 1560.0, 1561.0, 1596.0, 1597.0, 1655.0, 1656.0, 1657.0, 1658.0, 1781.0, 1782.0, 2174.0, and 2175.0 nm. These 23 variables will be selected for modelling later.

3.5 Model construction and comparative analysis

Table 2 shows the results of the PLSR and RF models for the preselected and optimal variables. In the PLSR model, the prediction results based on the optimal wavelength were both better than those based on the preselected wavelength. Among them, the model prediction based on the CARS algorithm was the best, with an R^2 of 0.67 and an RPD of 2.12 in the model validation set, while the prediction accuracy of Boruta algorithm–PLSR (PLSR model based on the Boruta algorithm) on the validation set was second only to CARS–PLSR (PLSR model based on the CARS algorithm). Furthermore, compared to the PLSR model, the RF model based on preselected variables had an R^2 of 0.54 and an RPD of 1.64 for the validation set, showing a slight improvement in modelling results to roughly predict the sample. The best-performing model was Boruta algorithm–RF (RF model based on the Boruta algorithm), which had an R^2 of 0.78 and an

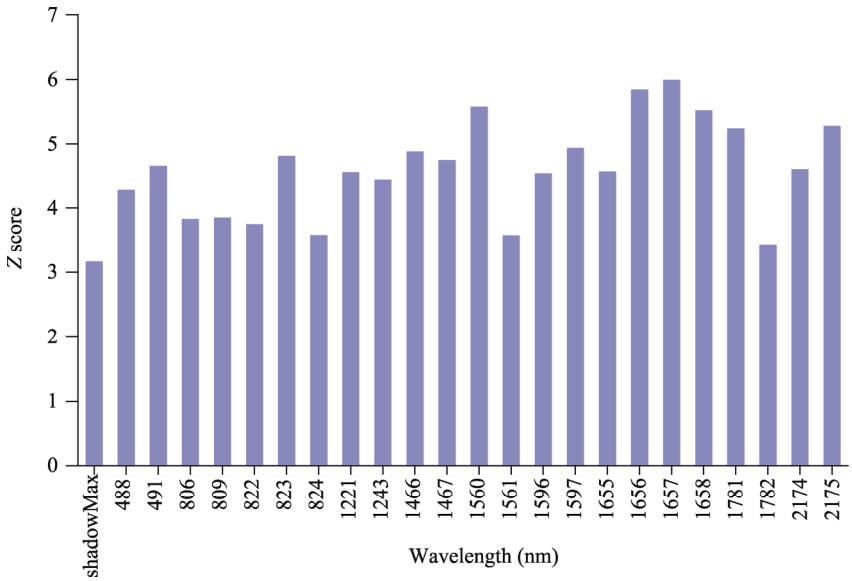

Fig. 7 Importance score (*Z* score) of the different wavelengths identified by the Boruta algorithm

Table 2 Comparison of the coefficient of determination (R^2), root mean square error (RMSE), and residual prediction deviation (RPD) obtained from partial least squares regression (PLSR) and random forest (RF) models based on four wavelength selection methods

Model	Selection method	Variable number	Calibration set (*n*=74)		Validation set (*n*=36)		
			R^2	RMSE (g/kg)	R^2	RMSE (g/kg)	RPD
PLSR	Preselected spectrum	442	0.45	5.23	0.42	5.46	1.24
	CARS	31	0.69	4.38	0.67	4.46	2.12
	SPA	5	0.62	4.56	0.61	4.34	1.82
	Boruta algorithm	23	0.65	4.30	0.63	4.28	2.08
RF	Preselected spectrum	442	0.52	4.96	0.54	4.83	1.64
	CARS	31	0.73	4.24	0.72	4.26	2.36
	SPA	5	0.64	4.37	0.66	4.31	1.86
	Boruta algorithm	23	0.76	4.24	0.78	4.19	2.38

Note: *n*, number of samples; CARS, competitive adaptive reweighted sampling; SPA, successive projections algorithm.

RPD of 2.38 for the validation set. Next, the models built based on the feature wavelengths selected by CARS and SPA showed slightly worse performance results. However, the R^2 of the validation set was higher than that of the preselected variables, and the R^2 of the calibration set was closer to that of the validation set, which indicated that the stability of the built models was better.

4 Discussion

Figure 8 shows the distribution of the feature variables selected by the three variable selection methods. The number of variables selected by the three algorithms was significantly reduced compared with the preselected variables, and the least accounted for only 1.4% of the preselected variables. In addition, the optimal variables obtained by the three variable selection methods had similar distribution ranges. The variables were mainly distributed in the near-infrared spectral regions of 1200.0–1600.0, 1700.0–2000.0, and 2200.0–2400.0 nm. The fundamental and octave vibrational absorption of carbonyl (C=O), carbon-hydrogen (C-H), aluminium-hydroxy (Al-OH), and hydroxide (O-H) bonds is the main manifestations in the near-infrared spectral range (Jin et

al., 2016), which is the main reason why vis-NIR spectra show special absorption peaks at approximately 1400.0, 1900.0, and 2200.0 nm. The absorption feature near 1400.0 nm is associated with hydroxyl (-OH) bonds, while the absorption wavelength near 1900.0 nm is the H_2O spectrum dominated by interlayer water. The absorption wavelength near 2000.0 nm is a combination of -OH stretching vibrations with Al-OH and magnesium hydroxyl (Mg-OH) bending vibrations. However, the CARS and Boruta algorithms also selected a small number of SOM spectral features located in the 400.0–780.0 nm range of the visible spectrum. The result was consistent with the previous studies (Araújo et al., 2001; Nocita et al., 2014; Li et al., 2019). Therefore, this suggests that the preferred wavelength in this study is reasonable.

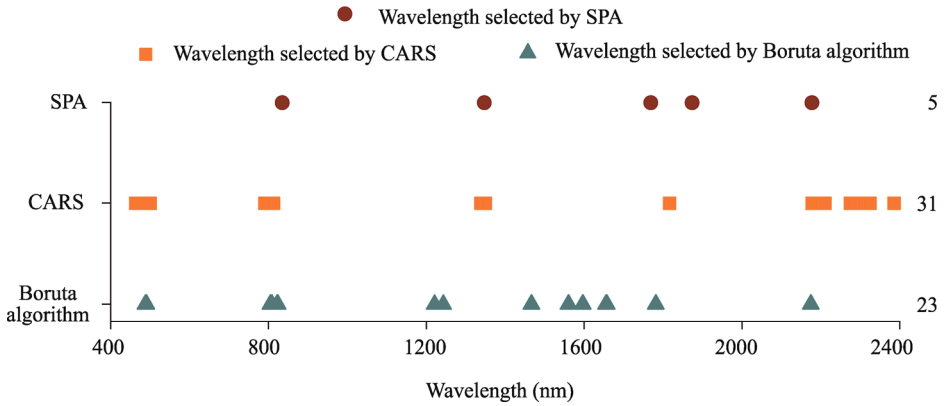

Fig. 8 Distribution of feature variables selected by SPA, CARS, and Boruta algorithms. Note that the numbers on the right side of the figure represent the number of optimal variables selected by SPA, CARS, and Boruta algorithms.

Conventional selection methods for soil spectral variables are performed by Pearson correlation analysis. Correlation analysis only considers the simple linear pattern between the independent variable itself and the dependent variable, while the exploration of deeper nonlinear implied relationships and the elimination of the information redundancy phenomenon appear to be weak (Wang et al., 2019; Ge et al., 2021). Therefore, we suggest correlation analysis as a way to preselect variables. As shown in Table 2, the RPD of the two models was 1.24 and 1.64 for modelling by the significance wavelengths obtained from Pearson correlation analysis, indicating that the models could only achieve a relatively coarse estimation of soil information. This may be due to the presence of more redundant or irrelevant information among the selected variables, resulting in lower model accuracy (Nocita et al., 2014). However, the accuracy of the PLSR and RF models based on the three feature variable selection algorithms was further improved compared to the accuracy of the preselected wavelength model, and the R^2 of the validation set was improved by 25% on average, indicating the importance of optimal variable selection for the preselected wavelength. Compared to traditional linear regression models, machine learning algorithms have significant advantages (Araújo et al., 2014; Li et al., 2021). The poor performance of the PLSR model based on vis-NIR spectroscopy may be due to the indirect spectral response of SOM (Dharumarajan et al., 2022). The same variable selection methods used in the RF model in this study showed an increase in the R^2 and RPD of the test set, while the RMSR decreased. The results exhibited by the variable selection methods were not consistent for different modelling schemes. In the PLSR model, the CARS algorithm showed greater competitiveness, while the Boruta algorithm was second only to the CARS algorithm. Actually, the CARS algorithm is a linear method, while the PLSR model can better handle the linear information between spectra and SOM. The combination of PLSR and the CARS variable selection method can effectively improve the model accuracy, which is consistent with previous research results (Vohland et al., 2014). Among the nonlinear models, Boruta algorithm combined with the RF model had the best prediction accuracy among all the combined models, with R^2

improving by 0.10 and RMSE decreasing by 0.33 on average compared with other algorithms. This is because both Boruta algorithm and RF are nonlinear algorithms, and in addition, the Boruta algorithm is based on the RF classifier so that better prediction accuracy can be achieved (Hong et al., 2021). The poor performance of the SPA in both models may be because the SPA aims to eliminate the covariance between variables, while the projection is performed without including soil property information, and some of the spectral wavelengths with rich information are not selected, thus leading to a lower model performance (Araújo et al., 2001). In addition, as mentioned by Chen et al. (2001), for small datasets (fewer than 200 samples), cross-validation or repeated random splitting leads to more robust model evolution.

Although the spectral ranges selected by the three methods were approximately the same, the application of the different models showed very different results. Therefore, we suggest that when building the SOM content prediction model, a suitable modelling scheme should be implemented according to different variable selection strategies. This method was effective and fast in estimating the SOM content but lacked spatial expressiveness. Furthermore, the soil type was not taken into account in this study due to the different effects of different types of soil texture and composition on the spectral characteristics. Further research is needed on how to improve the spatial expressivity of the SOM content and on how to combine the SOM content prediction of different soil types to improve model accuracy.

5 Conclusions

The original spectra were preprocessed and preselected by Pearson correlation analysis, and then the CARS, SPA, and Boruta algorithms were used to select spectral feature wavelengths, and the PLSR and RF models were combined to construct SOM content prediction models for the selected feature variables. Among the three variable selection algorithms, the RF model based on the Boruta algorithm had the best accuracy in the prediction of the SOM content. The RF model based on the Boruta algorithm improved the R^2 to 0.78 and the RPD to 2.38, achieving accurate SOM content prediction. The regression model coupled with the variable selection algorithm greatly reduced the complexity of the model while ensuring the accuracy of the model and provided technical support for the rapid and nondestructive estimation of the SOM content of arid land using spectral analysis technology, with promising applications.

Acknowledgements

This study was supported by the Key Project of Natural Science Foundation of Xinjiang Uygur Autonomous Region, China (2021D01D06) and the National Natural Science Foundation of China (41961059). We thank anonymous reviewers for their insightful comments, which help improve the quality of this manuscript.

References

Araújo M C U, Saldanha T C B, Galvão R K H, et al. 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems, 57(2): 65–73.

Araújo S R, Wetterlind J, Demattê J A M, et al. 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. European Journal of Soil Science, 65(5): 718–729.

Bao N S, Wu L X, Ye B Y, et al. 2017. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. Geoderma, 288: 47–55.

Chang W C, Laird D A, Mausbach M J, et al. 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Science Society of America Journal, 65(2): 480–490.

Chen Y, Ma L X, Yu D S, et al. 2022. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. Ecological Indicators, 135: 108545, doi: 10.1016/j.ecolind.2022.108545.

Chen S C, Xu H Y, Xu D Y, et al. 2021. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. Geoderma, 400: 115159, doi: 10.1016/j.geoderma.2021.115159.

Ding J L, Yu D L. 2014. Monitoring and evaluating spatial variability of soil salinity in dry and wet seasons in the Werigan-Kuqa Oasis, China, using remote sensing and electromagnetic induction instruments. Geoderma, 235–236: 316–322.

Dharumarajan S, Lalitha M, Gomez C, et al. 2022. Prediction of soil hydraulic properties using VIS-NIR spectral data in semi-arid region of Northern Karnataka Plateau. Geoderma Regional, 28: e00475, doi: 10.1016/j.geodrs.2021.e00475.

Ge X Y, Ding J L, Jin X L, et al. 2021. Estimating agricultural soil moisture content through UAV-based hyperspectral images in the arid region. Remote Sensing, 13(8): 1562, doi: 10.3390/rs13081562.

Ge X Y, Ding J L, Teng D X, et al. 2022a. Exploring the capability of Gaofen-5 hyperspectral data for assessing soil salinity risks. International Journal of Applied Earth Observation and Geoinformation, 112: 102969, doi: 10.1016/j.jag.2022.102969.

Ge X Y, Ding J L, Teng D X, et al. 2022b. Updated soil salinity with fine spatial resolution and high accuracy: The synergy of Sentinel-2 MSI, environmental covariates and hybrid machine learning approaches. CATENA, 212: 106054, doi: 10.1016/j.catena.2022.106054.

Han L J, Ding J L, Wang J J, et al. 2022. Monitoring oasis cotton fields expansion in arid zones using the Google Earth Engine: A case study in the Ogan-Kucha River oasis, Xinjiang, China. Remote Sensing, 14(1): 225, doi: 10.3390/rs14010225.

Hong Y S, Chen Y Y, Shen R L, et al. 2021. Diagnosis of cadmium contamination in urban and suburban soils using visible-to-near-infrared spectroscopy. Environmental Pollution, 291: 118128, doi: 10.1016/j.envpol.2021.118128.

Jin X L, Du J, Liu H J, et al. 2016. Remote estimation of soil organic matter content in the Sanjiang Plain, Northest China: The optimal band algorithm versus the GRA-ANN model. Agricultural and Forest Meteorology, 218–219: 250–260.

Keskin H, Grunwald S, Harris W G. 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma, 339: 40–58.

Kursa M B, Jankowski A, Rudnicki W. 2010. Boruta–a system for feature selection. Fundamenta Informaticae, 101(4): 271–285.

Li X H, Ding J L, Liu J, et al. 2021. Digital mapping of soil organic carbon using sentinel series data: A case study of the Ebinur Lake Watershed in Xinjiang. Remote Sensing, 13(4): 769, doi: 10.3390/rs13040769.

Li Q Q, Huang Y, Song X Z, et al. 2019. Moving window smoothing on the ensemble of competitive adaptive reweighted sampling algorithm. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 214: 129–138.

Liu J B, Dong Z Y, Xia J S, et al. 2021. Estimation of soil organic matter content based on CARS algorithm coupled with random forest. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 258: 119823, doi: 10.1016/j.saa.2021.119823.

Luo C, Wang Y A, Zhang X L, et al. 2022. Spatial prediction of soil organic matter content using multiyear synthetic images and partitioning algorithms. CATENA, 211: 106023, doi: 10.1016/j.catena.2022.106023.

Ma G L, Ding J L, Han L J, et al. 2021. Digital mapping of soil salinization based on Sentinel-1 and Sentinel-2 data combined with machine learning algorithms. Regional Sustainability, 2(2): 177–188.

Mcbratney A, Field D J, Koch A. 2014. The dimensions of soil security. Geoderma, 213: 203–213.

Mesquita D P P, Gomes J P P, Rodrigues L R, et al. 2018. Building selective ensembles of Randomization Based Neural Networks with the successive projections algorithm. Applied Soft Computing, 70: 1135–1145.

Nocita M, Stevens A, Toth G, et al. 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biology and Biochemistry, 68: 337–347.

Savitzky A, Golay M J E. 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8): 1627–1639.

Schomberg J, Ziogas A, Anton-Culver H, et al. 2018. Identification of a gene expression signature predicting survival in oral cavity squamous cell carcinoma using Monte Carlo cross validation. Oral Oncology, 78: 72–79.

Shi T Z, Chen Y Y, Liu H Z, et al. 2014. Soil organic carbon content estimation with laboratory-based visible–near-infrared reflectance spectroscopy: Feature selection. Applied Spectroscopy, 68(8): 831–837.

Shi T Z, Wang J J, Chen Y Y, et al. 2016. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. International Journal of Applied Earth Observation and Geoinformation, 52: 95–103.

Song X Z, Huang Y, Tian K D, et al. 2020. Near infrared spectral variable optimization by final complexity adapted models combined with uninformative variables elimination-a validation study. Optik, 203: 164019, doi: 10.1016/j.ijleo.2019.164019.

Swierenga H, Wülfert F, De Noord O E, et al. 2000. Development of robust calibration models in near infra-red spectrometric applications. Analytica Chimica Acta, 411(1–2): 121–135.

Tian Y C, Zhang J J, Yao X, et al. 2013. Laboratory assessment of three quantitative methods for estimating the organic matter content of soils in China based on visible/near-infrared reflectance spectra. Geoderma, 202–203: 161–170.

Viscarra Rossel R A, Walvoort D J J, Mcbratney A B, et al. 2006. Visible, near infrared, mid infrared or combined diffuse

reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 131(1–2): 59–75.

Vohland M, Ludwig M, Thiele-Bruhn S, et al. 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma, 223–225(1): 88–96.

Wang J Z, Ding J L, Ma X, et al. 2019. Capability of Sentinel-2 MSI data for monitoring and mapping of soil salinity in dry and wet seasons in the Ebinur Lake region, Xinjiang, China. Geoderma, 353: 172–187.

Wang X P, Zhang F, Ding J L, et al. 2018. Estimation of soil salt content (SSC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR), Northwest China, based on a Bootstrap-BP neural network model and optimal spectral indices. Science of the Total Environment, 615: 918–930.

Wang Z, Ding J L, Zhang Z P. 2022. Estimation of soil organic matter in arid zones with coupled environmental variables and spectral features. Sensors, 22(3): 1194, doi: 10.3390/s22031194.

Xie S G, Ding F J, Chen S G, et al. 2022. Prediction of soil organic matter content based on characteristic band selection method. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 273: 120949, doi: 10.1016/j.saa.2022.120949.

Xing Z, Du C W, Shen Y Z, et al. 2021. A method combining FTIR-ATR and Raman spectroscopy to determine soil organic matter: Improvement of prediction accuracy using competitive adaptive reweighted sampling (CARS). Computers and Electronics in Agriculture, 191: 106549, doi: 10.1016/j.compag.2021.106549.

Yin G C, Chen X L, Zhu H H, et al. 2022. A novel interpolation method to predict soil heavy metals based on a genetic algorithm and neural network model. Science of the Total Environment, 825: 153948, doi: 10.1016/j.scitotenv.2022.153948.

Zhang Y, Sui B, Shen H O, et al. 2019. Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors. Computers and Electronics in Agriculture, 160: 23–30.

Zhang Z P, Ding J L, Zhu C M, et al. 2021. Bivariate empirical mode decomposition of the spatial variation in the soil organic matter content: A case study from NW China. CATENA, 206: 105572, doi: 10.1016/j.catena.2021.105572.