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

Authors: ZHOU Qian, DING Jianli, GE Xiangyu, LI Ke, ZHANG Zipeng, GU Yongsheng, DING Jianli

Date: 2023-02-15T00:00:00+00:00

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 (R2) 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.

Full Text

Preamble

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 Sciences, 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 enables fast and efficient determination of soil organic matter (SOM). However, effective removal of redundant information is a prerequisite for accurate SOM prediction using this technique. This study evaluated three wavelength selection strategies to identify the spectral response characteristics of SOM.

SOM content and spectral information were measured for 110 soil samples from the Ogan-Kuqa River Oasis under laboratory conditions in July 2017. Pearson correlation analysis was applied to preselect spectral wavelengths from preprocessed spectra that passed the 0.01 significance level test. The successive projection algorithm (SPA), competitive adaptive reweighted sampling (CARS), and Boruta algorithm were then used to detect optimal variables from the preselected wavelengths. Finally, partial least squares regression (PLSR) and random forest (RF) models combined with the optimal wavelengths were developed for quantitative estimation of SOM content. The results demonstrate that the selected optimal variables were mainly located near spectral absorption features (i.e., 1400.0, 1900.0, and 2200.0 nm), while the CARS and Boruta algorithms also selected a few visible wavelengths in the 480.0–510.0 nm range. Both models achieved satisfactory SOM predictions, with the RF model demonstrating better accuracy than the PLSR model. The SOM prediction model established by combining the Boruta algorithm with the RF model performed best, using 23 variables and achieving a coefficient of determination (R^2) of 0.78 and a residual prediction deviation (RPD) of 2.38. The Boruta algorithm effectively removed redundant information and optimized the wavelength selection, thereby improving prediction accuracy. Therefore, combining vis-NIR spectroscopy with machine learning represents an important method for improving SOM prediction

accuracy in arid lands.

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 for evaluating soil fertility and quality, plays a critical role 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). Traditional laboratory chemical analysis methods for obtaining SOM information are relatively complex, inefficient, and uneconomical, and cannot meet the needs of smart agriculture. Therefore, establishing an efficient, low-cost, 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 considerable attention in soil property prediction studies due to advances in proximal sensing technology, which provides technical support for accurate SOM content estimation (Wang et al., 2022). Many scholars have explored the relationship between SOM and soil spectra using vis-NIR spectroscopy. Proximal sensing technology with fine spectral resolution obtains continuous spectral information of features at the nanometer level. SOM contains various functional groups (including hydroxyl, carboxyl, etc.) that exhibit characteristic absorption in the vis-NIR spectral regions, with absorption intensity at different wavelengths corresponding to the molecular structure and concentration of the substance (Zhang et al., 2021; Xie et al., 2022). Therefore, quantitative estimation of SOM through vis-NIR spectroscopy is of great practical significance. However, since ground object spectra provide hundreds of variables, redundancy exists between variables, and the variables are usually nonlinearly correlated with soil sample properties (Viscarra Rossel et al., 2006). Additionally, background noise and interference from specific physical factors are present in the spectra (Tian et al., 2013). Swierenga et al. (2000) suggested that selecting wavelengths with strong information and less interference from external factors is an effective way to construct stable spectral analysis models. Therefore, determining appropriate characteristic spectral wavelengths is a prerequisite for building SOM content prediction models.

In selecting characteristic spectral wavelengths from vis-NIR spectra, methods such as 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) have been widely used. The CARS algorithm can 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). Liu et al. (2021) used the CARS algorithm to screen SOM response characteristics after spectral preprocessing and applied the random forest (RF) method to build a prediction model for accurate assessment of agricultural soil organic matter content. The SPA effectively summarizes information from most sample spectra, avoiding overlapping information (Shi et al., 2014). However, most traditional feature selection algorithms follow the principle of min-optimality, making them overly dependent on the smallest subset of features and leading to errors and uncertainties in classification selection.

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, grouping them into three categories: strongly correlated, moderately correlated, and weakly correlated variables (Chen et al., 2022). Additionally, since the Boruta algorithm is based on the RF classification algorithm, it can detect both linear and nonlinear relationships between soil properties and environmental predictors, making it an important approach for feature selection. The partial least squares regression (PLSR) algorithm is a common modeling method that better addresses multicollinearity between independent variables (Shi et al., 2016). In previous studies, the RF model has been 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 overfitting even when the number of variables far exceeds the number of modeled samples and exhibits good resistance to noise (Ge et al., 2022a). However, no uniform standard feature wavelength selection methodologies have been presented in previous studies, and results from different feature wavelength selection strategies combined with various modeling methods differ significantly. Therefore, addressing the adaptability of wavelength selection methods and modeling schemes remains a challenge.

This study examined 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 were to: (1) analyze the spectral characteristics of soils in the Ogan-Kuqa River Oasis; (2) obtain preselected significant variables from preprocessed spectra using 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 using PLSR and RF based on preselected and optimal variables. These results provide methodological guidance for fast and efficient estimation of SOM content in arid regions using vis-NIR spectroscopy.

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, covering a total area of 9.5$×10^{3}$ km^2 [Figure 1: see original paper]. The region experiences large diurnal temperature variations, low rainfall, and high evaporation. The annual average temperature ranges from 10.5°C to 14.4°C, with a maximum temperature of 41.1°C. Average annual precipitation is only 43.1 mm, and evaporation is relatively high, making this a typical arid to extremely arid area (Han et al., 2022). Soil texture is primarily clay loam, chalky clay loam, loamy clay, and chalky clay. Land cover and land use types mainly include agricultural land, grassland, bare land, woodland, and saline land.

2.2 Soil Sample Collection and Chemical Analysis

From July 7 to July 19, 2017, surface soil (0–20 cm) was collected from the oasis area using the five-point sampling method within 30 m$×30$m plots, with five subsamples mixed into a single composite sample. A total of 144 soil samples were collected, dried, ground, and sieved($0.15 mm) in the laboratory after removing debris (stones, plant roots, and humus). Samples were prepared in two portions for spectroscopic measurements and SOM analysis. SOM content was determined using the potassium dichromate oxidation method heated with an electric sand bath (Jin et al., 2016).

2.3 Soil Spectra Collection and Preprocessing

Soil reflectance spectra were measured using an ASD FieldSpec®3 portable spectrometer (Analytical Spectral Devices, Boulder, Colorado, USA) with a wavelength range of 350.0–2500.0 nm. The sampling interval was 1.4 nm for 350.0–1000.0 nm and 2.0 nm for 1000.0–2500.0 nm, yielding 2151 output wavelengths. Spectra were measured in a dark environment using a 50-W halogen lamp positioned 30 cm from the soil surface at a zenith angle of 5°. A reference white board was used to obtain absolute reflectance before measurements. Each soil sample was measured five times, and the arithmetic mean was calculated as the final reflectance spectrum.

Spectral data contain both chemical information about the sample and irrelevant information and noise, including linear or nonlinear transformations and signal noise problems caused by absorption and scattering of signal intensity at the soil surface (Jin et al., 2016). Therefore, edge wavelengths from 350.0 to 399.0 nm and from 2401.0 to 2500.0 nm were removed from the original spectrum. The original reflectance spectra were processed using Savitzky-Golay (SG) smoothing and first derivative (FD) processing. The SG smoothing method reduces noise to enhance the signal-to-noise ratio (Savitzky and Golay, 1964). FD processing differentiates overlapping peaks, attenuates background interference, corrects baseline drift, sharpens spectral features, and captures minute details of spectral

curves (Wang et al., 2018; Ge et al., 2022b). SG smoothing and FD processing were implemented in R software using the "prospectr" package.

To avoid the impact of outlier sample values on model performance, the Monte Carlo outlier detection (MCOD) method was applied to remove sample outliers prior to modeling (Schomberg et al., 2018). The MCOD method was implemented using a toolbox in MATLAB software. The outlier plot for 144 soil samples generated through the MCOD method is shown in Figure 2 [Figure 2: see original paper]. The plot was divided into four areas, with approximately 34 points identified as outliers and excluded from subsequent analysis. The remaining 110 points were used as valid samples for the follow-up study.

2.4 Feature Selection Methods

2.4.1 Competitive Adaptive Reweighted Sampling (CARS) The CARS algorithm selects characteristic wavelengths from soil spectra by emulating the "survival of the fittest" principle from Darwin' s evolutionary theory. In each iteration, wavelength variables with large absolute regression coefficient values in the PLSR model are retained, while those with small absolute values are removed using adaptive reweighted sampling to obtain a series of wavelength variable subsets. An exponential decay function and adaptive reweighted sampling method are then used to achieve competitive variable selection. The root mean square error of cross-validation (RMSECV) is calculated using cross-validation, and the optimal wavelength subset is selected based on minimization of RMSECV values (Li et al., 2019; Xing et al., 2021).

The CARS algorithm used in this study was run in MATLAB software. Optimal variables were selected using the MCOD method, with the number of Monte Carlo samples set to 50 and iterative sampling performed. By comparing RMSECV values across samples, the variables corresponding to the sampling times with minimal RMSECV values were selected as the optimal variable set.

2.4.2 Successive Projections Algorithm (SPA) The SPA is a vector space covariance minimization algorithm for forward variable selection. It aims to improve covariance between variables by quickly filtering multiple characteristic wavelengths from the full spectrum using simple projection operations, thereby enhancing inter-variable covariance and greatly reducing computational effort to increase modeling speed. Details of SPA operations are provided in the literature (Araújo et al., 2001). The SPA was run in MATLAB software.

2.4.3 Boruta Algorithm The Boruta algorithm determines the importance of all features in a dataset with respect to the target variable, selects important features, and removes redundant feature variables (Keskin et al., 2019). This algorithm employs a black-box prediction model with good prediction accuracy to obtain importance indices related to target variables. The essential idea of the Boruta algorithm is to evaluate the importance of each feature variable through a circular method. By replicating the original feature set, a random mixture of

each feature value is used to construct a shadow feature with randomness. The final sample dataset for the model is a new feature set created by combining original features and shadow features. In each iteration of the RF algorithm, the importance scores of original features and shadow features are compared to select the optimal feature set for modeling (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, calculated as follows:

$$Z\text{-score} = \frac{\mathrm{MSE}_{\mathrm{OOB}}}{\mathrm{SD}_{\mathrm{MSE}_{\mathrm{OOB}}}}$$

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

$$\mathrm{MSE}_{\mathrm{OOB}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The final result uses the maximum Z score of the shadow feature (shadowMax) as the filtering indicator. When the Z score of a feature variable is larger than shadowMax, the feature is considered important; otherwise, it is considered unimportant and excluded from modeling (Ge et al., 2022a).

2.5 Modeling Methods

2.5.1 Partial Least Squares Regression (PLSR) The PLSR model combines the advantages of principal component analysis, canonical correlation analysis, and multiple linear regression, and is used to better address strong covariance and situations where the number of variables exceeds 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) and identify the optimal number of latent variables for the PLSR model. The "libPLS" package in R software was used for model implementation.

2.5.2 Random Forest (RF) Model The RF model is a decision tree-based classification and regression algorithm that uses bootstrap sampling to randomly select subsets of samples from the original data for decision tree modeling, where each decision tree is independent. The final model prediction is obtained by combining the voting results from all decision trees (Zhang et al., 2019; Ma et al., 2021). The RF model performs well across many datasets, does not easily overfit, and offers advantages in data modeling. Model parameters must be optimized before application, as they significantly impact performance. When running the RF model, three parameters require definition: the number of trees ("ntree"), the minimum node size ("nodeSize"), and the number of input

variables randomly selected as candidates at each split (“mtry”). The “ntree” was set to 1000 after repeated testing. A grid search technique with ten-fold cross-validation was then used to optimize “mtry” and “nodeSize” , with “mtry” set from 2 to 30 in steps of 2, and “nodeSize” set from 1 to 10 in steps of 1.

2.5.3 Assessment of Prediction Quality In this study, 110 samples were divided into three groups using the Kennard–Stone algorithm, with two groups serving as the training set (74 samples) and one as the validation set (36 samples). Model performance was evaluated using the coefficient of determination (R^2), RMSE, and residual prediction deviation (RPD) (Chang et al., 2001). Smaller RMSE for the validation set, larger R^2, and greater RPD indicate better model prediction. According to previous studies (Nocita et al., 2014; Bao et al., 2017; Luo et al., 2022), RPD < 1.4 indicates poor model performance and inability to predict real samples; $1.4 \leq$ RPD ≤ 2.0 indicates barely acceptable predictions requiring further improvement; and RPD > 2.0 demonstrates good model performance. The formulae for the three evaluation indicators are as follows:

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

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

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSE}}$$

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

3 Results

3.1 Descriptive Statistics of Soil Organic Matter (SOM) Content

The statistical characteristics of SOM content are shown in Table 1 . SOM content ranged from 5.49 to 59.86 g/kg, with a mean of 29.05 g/kg and standard deviation of 11.34 g/kg. The mean 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, indicating moderate variation and reasonable sample division.

3.2 Soil Spectral Analysis

The measured soil spectra showed that reflectance spectral curves of all soil samples followed roughly the same trend. In the 400.0–800.0 nm interval, curves increased with increasing reflectance; after 800.0 nm, curves were generally smooth except for moisture absorption valleys. Compared with original spectral curves, SG smoothing produced minimal changes, only making the spectral curves smoother. Therefore, FD preprocessing was implemented on the basis of SG smoothing in this study. As shown in Figure 3 [Figure 3: see original paper], FD spectral curves exhibited reduced spacing, increased density, and significantly enhanced spectral feature regions compared with original spectral curves.

3.3 Correlation Analysis of SOM Content with Original and Preprocessed Soil Spectra

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

3.4 Feature Variable Selection

3.4.1 CARS Algorithm to Extract Feature Variables Figure 5 [Figure 5: see original paper] shows the variable selection process of the CARS algorithm. The number of retained wavelengths gradually decreased as iterations increased, with the rate of decrease changing from fast to slow (Fig. 5a). RMSECV showed a trend from large to small and then from small to large, reaching its minimum (9.44) at 26 iterations (Fig. 5b). This occurred because during iterations 1–26, RMSECV decreased by continuously eliminating wavelengths less correlated with SOM content that had minimal impact on modeling results. After 26 iterations, wavelengths strongly correlated with SOM content began to be removed, causing RMSECV to increase. Figure 5c presents the stability trajectory of wavelength variables, with each curve showing the trend of stability for each variable across iterations. The optimal variable subset with the smallest RMSECV is marked with an asterisk. Thus, the variable set corresponding to the 26th sampling was the optimal subset of 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 feature variables from the spectral data. The range of feature variables to be selected was set from 1 to 10 (Fig. 6 [Figure 6: see original paper]), with calibration and prediction set sample settings kept constant. Figure 6a shows the RMSE trend with the number of variables included in the model. As the number of variables increased, the minimum RMSE gradually decreased, reaching a minimum (9.47) when five variables were included. When the number of variables approached six, further increases introduced wavelength variables unrelated to predicted values or with greater noise, causing RMSE to increase. Figure 6b shows the distribution of 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.

3.4.3 Boruta Algorithm to Select Feature Variables When the Z score of a feature variable exceeds shadowMax, the feature is considered important. As shown in Figure 7 [Figure 7: see original paper], the maximum shadowMax value was 3.15, and 23 feature wavelengths had Z scores larger than the maximum shadowMax value: 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 were selected for subsequent modeling.

3.5 Model Construction and Comparative Analysis

Table 2 shows the results of PLSR and RF models using preselected and optimal variables. In the PLSR model, prediction results based on optimal wavelengths were superior to those based on preselected wavelengths. Among these, the model based on the CARS algorithm performed best, with an R^2 of 0.67 and RPD of 2.12 in the validation set, while the Boruta-PLSR model ranked second. Compared to the PLSR model, the RF model based on preselected variables achieved an R^2 of 0.54 and RPD of 1.64 for the validation set, showing slight improvement for rough sample prediction. The best-performing model was Boruta-RF, which achieved an R^2 of 0.78 and RPD of 2.38 for the validation set. Models built using feature wavelengths selected by CARS and SPA showed slightly worse performance, though their validation set R^2 values were higher than those of preselected variables, and calibration set R^2 values were closer to validation set values, indicating better model stability.

4 Discussion

Figure 8 [Figure 8: see original paper] shows the distribution of feature variables selected by the three variable selection methods. The number of variables selected by all three algorithms was significantly reduced compared with preselected variables, with the smallest accounting for only 1.4% of preselected

variables. Additionally, the optimal variables obtained by the three methods had similar distribution ranges, primarily located in the near-infrared spectral regions of 1200.0–1600.0, 1700.0–2000.0, and 2200.0–2400.0 nm. The fundamental and overtone vibrational absorption of carbonyl (C=O), carbon-hydrogen (C-H), aluminium-hydroxy (Al-OH), and hydroxide (O-H) bonds are the main manifestations in the near-infrared spectral range (Jin et al., 2016), which explains 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 near 1900.0 nm is dominated by interlayer water. The absorption near 2000.0 nm represents a combination of -OH stretching vibrations with Al-OH and magnesium hydroxyl (Mg-OH) bending vibrations. The CARS and Boruta algorithms also selected a small number of SOM spectral features in the 400.0–780.0 nm visible range, consistent with previous studies (Araújo et al., 2001; Nocita et al., 2014; Li et al., 2019). This suggests that the selected wavelengths in this study are reasonable.

Conventional soil spectral variable selection methods using Pearson correlation analysis only consider simple linear patterns between independent variables and the dependent variable, while exploration of deeper nonlinear relationships and elimination of information redundancy appear weak (Wang et al., 2019; Ge et al., 2021). Therefore, we suggest using correlation analysis for variable preselection. As shown in Table 2, the PLSR and RF models using significant wavelengths from Pearson correlation analysis achieved RPD values of 1.24 and 1.64, respectively, indicating only coarse estimation of soil information. This may be due to redundant or irrelevant information among selected variables, resulting in lower model accuracy (Nocita et al., 2014). However, the accuracy of PLSR and RF models based on the three feature variable selection algorithms was further improved compared to preselected wavelength models, with validation set R^2 improving by an average of 25%, demonstrating the importance of optimal variable selection.

Compared to traditional linear regression models, machine learning algorithms offer 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 showed increased R^2 and RPD values and decreased RMSE in the test set. However, variable selection methods performed differently across modeling schemes. In the PLSR model, the CARS algorithm showed greater competitiveness, while the Boruta algorithm ranked second. The CARS algorithm is a linear method, while the PLSR model can better handle linear information between spectra and SOM. The combination of PLSR and CARS can effectively improve model accuracy, consistent with previous research (Vohland et al., 2014). Among nonlinear models, Boruta combined with RF achieved the best prediction accuracy across all 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 and RF are nonlinear algorithms, and the Boruta algorithm is based on the RF classifier, enabling better predic-

tion accuracy (Hong et al., 2021). The poor performance of SPA in both models may be because SPA aims to eliminate covariance between variables through projection without incorporating soil property information, causing some spectrally rich wavelengths to be excluded and resulting in lower model performance (Araújo et al., 2001). Additionally, as noted 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 similar, different models showed very different results. Therefore, we suggest implementing a suitable modeling scheme according to different variable selection strategies when building SOM content prediction models. This method was effective and fast for estimating SOM content but lacked spatial expressiveness. Furthermore, soil type was not considered in this study, despite different soil textures and compositions affecting spectral characteristics. Further research is needed on improving the spatial expressivity of SOM content and combining SOM predictions across different soil types to improve model accuracy.

5 Conclusions

Original spectra were preprocessed and preselected using Pearson correlation analysis, after which CARS, SPA, and Boruta algorithms were used to select spectral feature wavelengths. 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 achieved the best accuracy for SOM content prediction, improving R^2 to 0.78 and RPD to 2.38, thus achieving accurate SOM content prediction. The regression model coupled with variable selection algorithms greatly reduced model complexity while ensuring accuracy, providing technical support for rapid and nondestructive estimation of SOM content in arid lands 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 helped 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, Northeast 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: 390–399.

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.