Revista Brasileira de Ciência do Solo

Division - Soil in Space and Time | Commission – Pedometrics

Estimation of soil organic carbon content by Vis-NIR spectroscopy combining feature selection algorithm and local regression method

Baoyang Liu[1,2], **Baofeng Guo**[1,2]*, **Renxiong Zhuo**[2] **and Fan Dai**[1]

[1] School of Automation, Hangzhou Dianzi University, Hangzhou, China.
[2] Key laboratory of IOT and Information Fusion Technology of Zhejiang Province, Hangzhou, China.

*** Corresponding author:**
E-mail: 694258721@qq.com

Received: June 09, 2023

Approved: September 21, 2023

How to cite: Liu B, Guo B, Zhuo R, Dai F. Estimation of soil organic carbon content by Vis-NIR spectroscopy combining feature selection algorithm and local regression method. Rev Bras Cienc Solo. 2023;47:e0230067
https://doi.org/10.36783/18069657rbcs20230067

Editors: José Miguel Reichert and Waldir de Carvalho Junior .

ABSTRACT: Soil organic carbon (SOC) content is a critical parameter for evaluating soil health. However, high redundancy and invalid information in soil hyperspectral data can reduce the accuracy and stability of SOC prediction models. This study developed a global partial least squares regression (PLSR) model and a local PLSR model for agricultural soils in the LUCAS 2015 database. Some variable selection methods were combined with the regression models and their effects on prediction accuracy were explored. In addition, when the genetic algorithm is utilized for spectral feature selection, we obtained a more representative spectral subset through a novel coding approach. The results illustrated that the best SOC estimation accuracy was achieved by the local PLSR combined with a coding-improved genetic algorithm (GA), with R^2 of 0.71, RMSEP of 5.7 g kg^{-1}, and RPD of 1.87. This study demonstrates that appropriate spectral band selection only slightly enhances the model performance of both global and local regressions, as PLSR models using the full spectrum show similar performance. Local PLSR models consistently outperform global ones using full spectrum or variable selection algorithms.

Keywords: local calibration, soil property, variable selection, LUCAS 2015 database.

INTRODUCTION

Soil organic carbon (SOC) is synthesized and decomposed by various organisms and is important for the physical, chemical and biological properties of soils (Gu et al., 2019). Monitoring the spatial distribution of SOC helps to analyze trends in carbon sequestration, thereby facilitating the stakeholders to plan how to deal with future climate change (Seely et al., 2010; Six and Paustian, 2014). However, traditional methods for estimating SOC levels, which rely primarily on manual sampling, are not feasible on a larger scale because traditional soil collection and analysis is costly and time-consuming (Sanchez et al., 2009; Conant et al., 2011; Viscarra-Rossel et al., 2016). Visible (Vis, 400-700 nm) to near-infrared (700-2500 nm) spectroscopy is therefore becoming an attractive alternative that can provide a rapid, simple, and non-destructive measurement to assess various soil properties while avoiding the drawbacks associated with traditional chemical analysis methods (Ward et al., 2019; Meng et al., 2022).

The use of spectral absorption to predict soil properties relies on the hypothesis that concentrations of specific soil properties have a linear correlation with absorption properties in the spectrum (Bellon-Maurel et al., 2011). These absorption features result from overtones and combination bands of fundamental vibrations of some of the functional groups of the molecules, such as the hydroxyl (OH), methyl (CH), and amino (NH) groups. Electron transitions (Ben-Dor et al., 1999) are responsible for absorption features in the Vis region (400-700 nm). The absorption features in the NIR region (700-2500 nm) are caused by molecular vibrations and rotations (Davies, 2005). Therefore, Vis-NIR spectra of soils hold promise for the quantification of soil properties (Viscarra-Rossel et al., 2010; Hong et al., 2019; Meng et al., 2020).

Reflectance spectra in the Vis-NIR region include strong and weak absorptions that partially overlap, resulting in many redundant spectral features that add complexity to the model. To extract the essential amount of potential quantitative information in soil hyperspectra, eliminating uninformative variables or selecting key variables is an effective process for building an accurate predictive model (Hong et al., 2020a; Tang et al., 2021). Li et al. (2009) estimated the moisture content of 80 corn samples and showed that the feature bands selected using the competing adaptive weighted sampling algorithm (CARS) method achieved better predictions than the full spectrum. Yu et al. (2016) applied the successive projection algorithm (SPA), uninformative variable elimination (UVE), and CARS methods and their combinations to develop a partial least squares regression (PLSR) model for estimating SOC based on 56 soil samples. The results showed that the CARS method was superior to the SPA and UVE methods. Li et al. (2019) compared five different variable selection algorithms and constructed SOC prediction models for 548 soil samples using PLSR and random forest, respectively. The results show that combining the random forest (RF) model with the improved CARS algorithm achieved the best prediction accuracy. Tang et al. (2021) used characteristic bands obtained from CARS, the digital elevation model, and spectral indices as input variables, and then applied the RF to build a SOC prediction model for 548 soil samples of different soil types, achieving good prediction results. These studies have demonstrated the effectiveness of CARS in feature band selection with small samples in a certain area. However, especially in large-scale soil databases, the performance of other potential spectral variable selection methods, such as heuristic algorithms, has yet to be investigated.

Spatial dependence due to soil type, land-use/land cover (LU/LC), climate, geology and other factors, as well as the increased variation in soil properties, can lead to significantly high estimation errors, as has been demonstrated in large databases (Araújo et al., 2001; Stenberg et al., 2010; Savvides et al., 2010; Nocita et al., 2014). However, according to Ramirez-Lopez et al. (2013), there may be a local stability of the spectral variation caused by the soil properties. Therefore, a local regression model is an effective method in massive databases (Nocita et al., 2014). Ramirez-Lopez et al. (2013) proposed a spectral-based learner (SBL), which uses an optimized principal component distance to retrieve the nearest neighbor of the predicted sample. Nocita et al. (2014) divided the LUCAS database into mineral and organic soils. Then, they applied local PLSR and obtained the best SOC predictions in mineral soils based on 250 nearest neighbors, using PLS distance as the spectral distance metric. Ward et al. (2019) used the Vis-NIR spectral band without the strong water absorption bands as an input variable and then built a local PLSR model to achieve the lowest SOC estimation error. Meng et al. (2022) used multiple stratification strategies to construct local random forest models on the basis of great group, genus, spectral similarity, and decision tree models. In most previous studies, the similarity between samples was calculated by local algorithms based on the whole Vis-NIR spectral bands. This can cause the nearest neighbor search to be influenced by the spectral response of other soil constituents, reducing the accuracy of the prediction of the soil property content.

Although some previous studies have shown that appropriate feature selection methods can improve prediction accuracy (Li et al., 2009; Yu et al., 2016; Wang et al., 2019), the soil samples in the above studies were very limited and all from the same region. The feature selection algorithms used were also not sufficiently diverse. However, for large-scale soil databases, there are few reports on whether feature selection algorithms can improve the prediction accuracy of SOC. This study aimed: (1) to compare the effects of different feature selection methods on the SOC prediction of global and local models and to determine a suitable feature selection method for a large-scale soil database; (2) to select more representative SOC feature bands using a genetic algorithm (GA) based on an improved coding approach, and thus improve the prediction accuracy of SOC.

MATERIALS AND METHODS

Database

European Land Use/Cover Area Frame Survey (LUCAS) topsoil database is currently the most extensive and consistent soil database on a continental scale (Tóth et al., 2013; Orgiazzi et al., 2017). In 2015, the LUCAS survey was carried out in all EU-28 Member States (MS) and included 21,859 topsoil samples (0.00-0.20 m) collected on different land-use. The database consists of 12 different soil properties, including SOC and spectral measurements in the Vis-NIR range (Tóth et al., 2013). Total carbon was measured by dry combustion using a VarioMax CN Analyzer (Elementar Analysensysteme GmbH, Germany) after heating the soil to 900 °C. The SOC content was then obtained by subtracting the carbonate content (measured according to ISO 10693:1995) from the total carbon. The Vis-NIR absorbance of the soil was measured using a FOSS XDS Rapid Content Analyzer (FOSS NIRSystems Inc. Denmark), operating in the 400-2500 nm wavelength range, with a spectral resolution of 2 nm and a spectral data interval of 0.5 nm, resulting in 4200 wavelengths.

In addition, we divided the LUCAS 2015 database into agricultural areas according to the LU/LC, consisting of about 12,817 soil samples. The agricultural areas are preferred in our research because vegetation interference is removed after harvesting. Therefore, they can be better used to map soil properties from airborne platforms in future research.

Data pre-processing

According to Nocita et al. (2014), the spectrum was affected by the step of absorbance value at the junction of the two spectrometer sensors at 1100 nm, so the spectral bands between 1052-1148 nm were removed in this study. In addition, as observed by Stevens et al. (2013), there were instrumental artifacts in the 400-500 nm spectral region, which were deleted from further analysis (Stevens et al., 2013).

We separately tested several preprocessing techniques, including standard normal variate (SNV), normalization, continuum removal (Clark and Roush, 1984), Savitzky-Golay smoothing (Savitzky and Golay, 1964), and first and second derivatives. However, we found that the best modeling results were obtained using only the first derivatives operation in the LUCAS agricultural subset, similar to other studies (Stevens et al., 2013; Araújo et al., 2014; Nocita et al., 2014). After the first-order derivative operation, the number of spectral bands is 949.

In the agricultural subset of the LUCAS database, the distribution of the SOC content is highly skewed. Therefore, a natural logarithm was used to transform it into an approximately normal distribution. The Kennard-Stone technique (Kennard and Stone, 1969) was then applied to divide the data set into a calibration set (70 %) and a validation set (30 %). The evaluation of the model performance was performed only on the validation set.

Model calibration

Two different modelling strategies were tested (Figure 1): (i) global PLSR modeling based on pre-processed full spectral bands and feature bands selected using different variable selection algorithms; (ii) local PLSR modeling based on pre-processed full spectral bands and feature bands. We then compared the feature band-based SOC prediction model with a full-spectrum SOC prediction model, which was used to investigate the performance of the variable selection algorithm in improving prediction accuracy. All models were evaluated and calibrated on the same subset of LUCAS for comparison.

Spectral variable selection

Genetic Algorithm (GA) is an intelligent optimization method proposed by Holland in 1975 that simulates the Darwinian biological evolution process (Holland, 1975), and is based on the evolutionary process of biological chromosomes. In spectral analysis, the entire spectral band is usually divided into *m* intervals, each containing the same number of

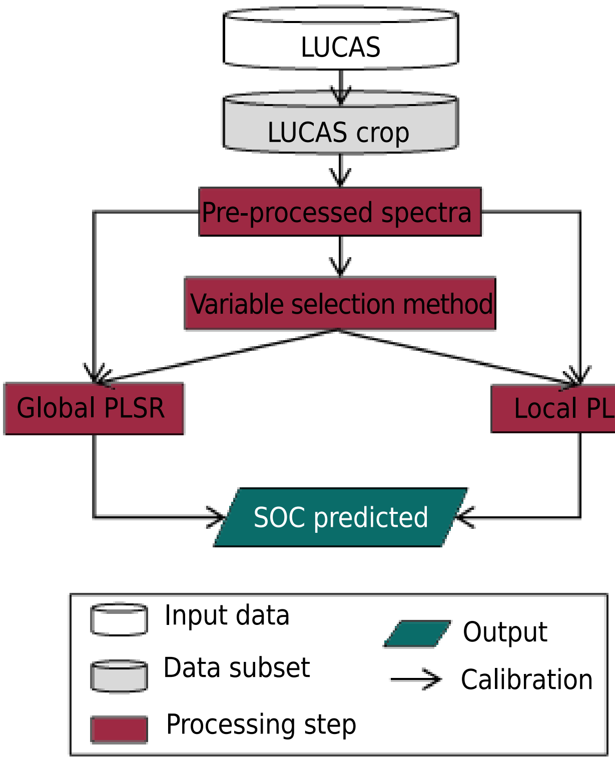

Figure 1. The overall flow chart of the study.

spectral bands. The *m* intervals are coded with a value of 0 or 1, which is the coefficient vector $A = [a_1, a_2, ... a_m]$. Then, the coefficient vector *A* is optimized by genetic algorithm, and the absorbance values of the spectral bands contained in the interval with the value of 1 are used as inputs to the SOC estimation model, and the objective function is the root mean square error (RMSE) of the SOC estimation model. Finally, when the objective function is minimized, the spectral bands contained in the interval with a value of 1 are the feature bands. There are two main problems with such an encoding approach: (1) an interval contains multiple spectral bands, and it is impossible to exclude invalid bands in the interval; (2) the importance of each spectral band cannot be accurately measured by using a binary encoding approach. To address these two problems, we used real numbers between 0 and 1 to encode each spectral band and obtained the weights of each spectral band that lead to the lowest RMSE of the SOC prediction values by using a genetic algorithm. Finally, the spectral band with weight >0.5 was selected as the characteristic band of SOC.

The SPA is a forward variable selection algorithm that uses basic manipulations in vector space to find a subset of variables with low collinearity (Araújo et al., 2001). In the SPA algorithm, one band is selected as the initial band, and the projections of that band onto the remaining bands are computed separately. Then, the band with the largest projection vector will be the selected band. The selected bands and the initial band are used as inputs to the multiple linear regression, and the RMSE values are recorded. Repeat the above steps *N* times (*N* is the number of bands to be selected). Finally, the subset of spectral variables corresponding to the smallest RMSE value is selected as the best set of variables.

Correlation analysis (CA) is performed to calculate the correlation between each wavelength in the spectral matrix and the SOC content. The higher the correlation coefficient, the more information about SOC is contained. Therefore, a threshold is set based on previous experience and bands with correlation coefficients above this threshold are used as feature bands. We chose the band with the highest correlation coefficient of 35 % as the characteristic band, i.e., a threshold of 0.26. The formula for the correlation coefficient is presented in equation 1:

$$r_i = \frac{\sum_{j=1}^{n}\left(R_{ij} - \bar{R}_i\right)\sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)}{\sqrt{\sum_{j=1}^{n}\left(R_{ij} - \bar{R}_i\right)^2 \sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2}} \ , i = 1,2,\cdots,m \qquad \text{Eq. 1}$$

in which: r_i denotes the correlation coefficient between the SOC and the the *i*th wavelength; R_{ij} denotes the absorption value of the *j*th sample at the *i*th wavelength; $\bar{R}_i$ denotes the average absorption value at the *i*th wavelength; Y_j denotes the SOC content of the *j*th sample; $\bar{Y}$ denotes the average SOC content of the sample; *i* denotes the number of wavelengths; and *j* denotes the number of samples.

The UVE is a variable selection method based on the analysis of the stability of the PLSR regression coefficients, which can effectively eliminate irrelevant variables. The UVE adds to the spectral matrix a matrix of random variables (i.e., noise information) containing the same number of variables as the spectral matrix. The regression coefficient matrix *B* is obtained by constructing the PLSR model, and then the reliability *C* of the regression coefficient vector *b* of each variable, which is the quotient of the mean and standard deviation of the regression coefficient vector *b*, is analyzed. Finally, the uninformative variables in the spectral matrix variables are removed one by one. The reliability *C* is shown in equation 2:

$$C_i = \frac{mean(b_i)}{S(b_i)} \quad \text{Eq. 2}$$

in which: *mean*(b_i) denotes the mean of the regression coefficient *b* for the *i*th column of spectral variables; *S*(b_i) denotes the standard deviation of the regression coefficient *b* for the *i*th column of spectral variables. The magnitude of the absolute value of C_i determines whether the *i*th column variable is used in the final PLSR model (Centner et al., 1996).

When the dimensionality of the input data is high, the CARS algorithm is a promising method to eliminate uninformative variables and/or perform wavelength selection to build a well-calibrated model (Li et al., 2009). The main steps of the CARS algorithm are to first build a PLSR model based on all the original spectral variables, record the absolute regression coefficients b_j, and calculate the weights $w_j = b_j/sum(b_j)$ for each band, where *j* is the number of spectral bands. Then, an exponentially decreasing function (EDF) was used to calculate the proportion of retained variables, i.e., $r_i = ae^{-ni}$. A subset of variables is selected from the retained spectral variables using adaptive reweighted sampling (ARS), and this subset of variables is used to build the PLSR and the root mean square error (RMSE) is recorded, and the above steps are repeated *N* times (*N* is the number of samples). Finally, the subset of variables with the smallest RMSE was selected as the best subset of variables.

Local PLSR

The PLSR has been used in a wide range of fields, including chemistry, agronomy, and biomedicine (Tiecher et al., 2021). We used 10-fold cross-validation to estimate the RMSE for different numbers of latent variables (LVs) and selected the minimum number of LVs within one standard deviation of the minimum RMSE, similar to Stevens et al. (2013). The adjusted-coefficient of determination (adjusted R^2; equation 3) was then calculated. Finally, the LV values obtained in both ways are averaged as the most optimal number of LVs in PLSR.

$$adj.\ R^2 = 1 - \left(1 - R^2\right)(n-1)/(n-k-1) \quad \text{Eq. 3}$$

in which: *n* is the number of samples and *k* is the number of LVs.

Local PLSR (LPLSR) is a memory-based learning method that outperforms many machine learning methods, such as neural networks, support vector machines, and decision trees, when dealing with large amounts of data (Ramirez-Lopez et al., 2013). The LPLSR selects the most similar sample set in terms of spectral features for each validation sample in the calibration set based on the similarity metric, and builds an independent PLSR model based on this sample set (Hong et al., 2020b). In local regression, prediction performance is greatly affected by the distance metric and the number of nearest neighbors. In this paper, to identify the model parameters for local regression, 30 % of the calibration set was randomly chosen as the test set and the remaining 70 % as the training set. We tested four possible distance measures in local methods: (i) Euclidean distance (euDist); (ii) Mahalanobis distance (MDist); (iii) Correlation distance (corDist) (equation 4); and (iv) Cosine distance (cosDist) (equation 5).

$$d_{st} = 1 - \frac{\left(x_s - \bar{x}_s\right)\left(y_t - \bar{y}_t\right)^T}{\sqrt{\left(x_s - \bar{x}_s\right)\left(x_s - \bar{x}_s\right)^T}\sqrt{\left(y_t - \bar{y}_t\right)\left(y_t - \bar{y}_t\right)^T}} \quad \text{Eq. 4}$$

$$d_{st} = 1 - \frac{x_s y_t^T}{\sqrt{\left(x_s x_s^T\right)\left(y_t y_t^T\right)}} \qquad \text{Eq. 5}$$

in which: d_{st} is the distance between samples *s* and *t*; x_s and y_t are the row vectors of samples *s* and *t*, respectively; $(\cdot)^T$ is the vector transpose; $\bar{x}_s$ and $\bar{y}_t$ are the averages of samples *s* and *t*.

According to Ward et al. (2019), after determining the distance metric, the optimal number of training samples is found by testing different numbers of nearest neighbors. We used 400 nearest neighbors in the local PLSR.

Model assessment

To assess the model accuracy, the coefficient of determination (R^2) (Equation 6), root mean square error of prediction (RMSEP) (Equation 7), relative RMSEP (rRMSEP) (equation 8), the ratio of performance deviation (RPD) (Equation 9), and the ratio of performance to interquartile range (RPIQ) (equation 10) were used as model performance evaluation metrics (Baumgardner et al., 1985; Tiecher et al., 2021).

$$R^2 = 1 - \sum_{i=1}^{n}\left(yp_i - yo_i\right)^2 / \sum_{i=1}^{n}\left(yo_i - \bar{y}o\right)^2 \qquad \text{Eq. 6}$$

in which: yo_i is the observed value of sample *i*; yp_i is the predicted value of sample *i*; and $\bar{y}o$ is the mean of the observed SOC value.

$$\text{RMSEP} = \left(\sum_{i=1}^{n}\left(yp_i - yo_i\right)^2 / n\right)^{1/2} \qquad \text{Eq. 7}$$

in which: *n* is the number of samples.

$$\text{rRMSEP} = 100 \times RMSEP / \bar{y}o \qquad \text{Eq. 8}$$

$$\text{RPD} = sd\left(yo\right) / RMSEP \qquad \text{Eq. 9}$$

in which: *sd* is the standard deviation.

$$\text{RPIQ} = IQ\left(yo\right) / RMSEP \qquad \text{Eq. 10}$$

in which: *IQ* is the interquartile range.

Table 1. Statistical summary of the LUCAS agricultural subset soil organic carbon

Dataset	N	SOC							
		Min	Max	Q25	Median	Q75	Mean	Std	Skew
		g kg^{-1}							
Total	12817	0.1	534.8	11.2	16.8	26.7	25.1	35.0	7.4
Calibration	8971	0.1	534.8	11.4	17.6	29.2	27.9	40.7	2.4
Validation	3846	0.1	147.7	11.0	15.2	22.1	18.4	11.9	2.7

Q25: the first quartile split of the bottom 25 %; Q75: the third quartile split of the top 75 %.

RESULTS

LUCAS database and pretreatment

We divided the LUCAS 2015 database into agricultural areas and reduced the sample size to 12,817. Within the agricultural subset of LUCAS, the SOC content ranged from 0.1-534.8 g kg^{-1} with a mean of 25.1 g kg^{-1}. The clay content ranges from 2 to 620 g kg^{-1}, with a mean of 210 g kg^{-1}. The $CaCO_3$ content ranges from 0-976 g kg^{-1} with a mean of 68.2 g kg^{-1}. All data sets showed a skewed distribution of SOC (Table 1).

Due to the SOC content and the mineral components, the absorption spectra showed large differences. Figure 2 shows the average and first-order derivative absorption spectra of different classes of SOC content in the LUCAS agricultural subset. Figure 2a shows that the spectral absorption increases with the increase of SOC. The difference in absorption spectra was more pronounced for higher SOC content in the Vis region compared to the NIR region (Baumgardner et al., 1985). This coincides with the observations of Stenberg et al. (2010), who reported an increase in absorption for organic soils in the NIR region (Stenberg et al., 2010). The first-order differential operation has the property of reflecting subtle changes in the spectrum, and the first-order differential absorption spectrum in the Vis region is dramatically affected by the SOC (Figure 2b). We found two absorption peaks near 600 nm that appear to be related to the SOC. The shoulder at 650 nm may be associated with a small concentration of hematite (Viscarra-Rossel et al., 2010). Water absorption bands were found in all SOC content classes at 1455 and 1915 nm. At 2050 nm, the absorption is determined by the nitrogen content, and we found

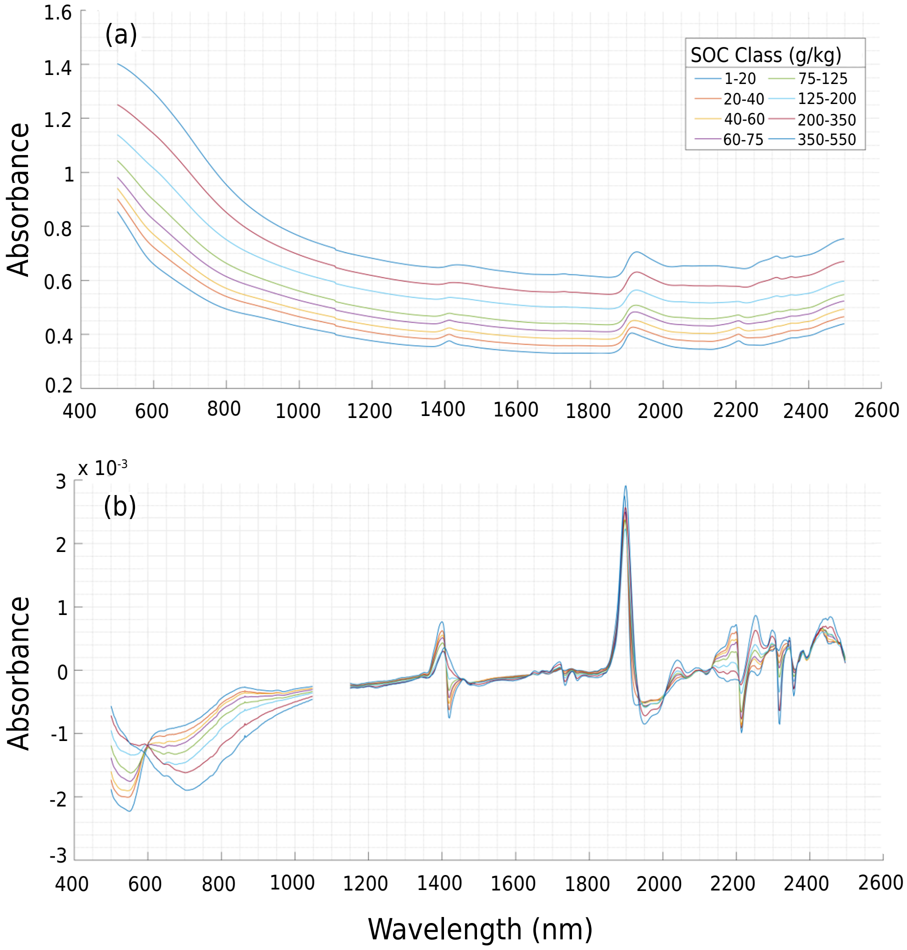

Figure 2. Original absorption spectrum (a) and first-order derivative absorption spectrum with 1052-1148 nm removed (b).

that the absorption depth increases with increasing SOC content due to a correlation between nitrogen and organic carbon (Post and Noble, 1993). The absorption feature at 2204 nm is usually caused by a combination of Al-OH bending and O-H stretching vibrations, which are related to clay mineralogy and are essential for low SOC soils. Between 2300 and 2400 nm, there is a characteristic C-H peak associated with organic matter. Despite the diversity of large soil databases and the multiple interactions of absorption spectra with other soil properties (e.g., particle size, carbonate, etc.), the average and first-order difference spectra classified according to SOC content classes showed the same patterns as in earlier research (Viscarra-Rossel et al., 2006).

Spectral variable selection

SPA

A plot of the RMSE values of the PLSR model built using different subsets of spectral band variables in the SPA variable selection process is shown in figure 3. As shown in figure 3, with the number of variables below 59, the RMSE showed a consistent decline with only local fluctuations as the number of variables increased, indicating that the added spectral bands made an effective contribution to the prediction of SOC. As the number of variables surpassed 59, the RMSE values rose dramatically and then dropped rapidly. The marked red stars in figure 3 indicate the number of bands in the optimal subset of spectral variables chosen by the SPA, i.e., containing 59 significant spectral band variables with an RMSE value of 27.5 g kg^{-1}. After spectral feature selection by SPA, the original spectral data was highly compressed, with the number of feature bands being only 6.2 % of the original bands.

UVE

Figure 4 shows the stability analysis of the spectral variable selection using the UVE method. The black vertical line was the variable split line. The blue curve to the left of the split line showed the stability *C* distribution curves of the 949 spectral variables after a first-order derivative operation. The red curve to the right side of the split line was the stability C distribution curve of the 949 random noise information variables. The two black dashed lines indicate the maximum and minimum limits of the stability (the criterion for the threshold was the peak value of the stability of the noise variable). The spectral variables, whose stability values lay outside the two horizontal black dashed lines, were retained as useful information variables and other spectral variables were excluded. In figure 4, we found that more wavelength variables were selected in the

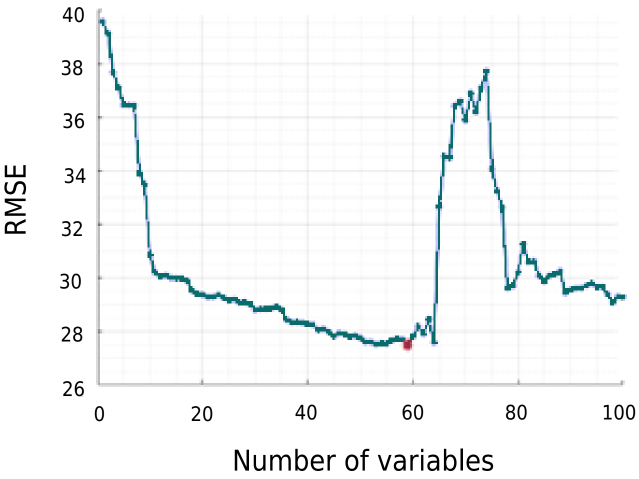

Figure 3. The RMSE for different subsets of variables using SPA.

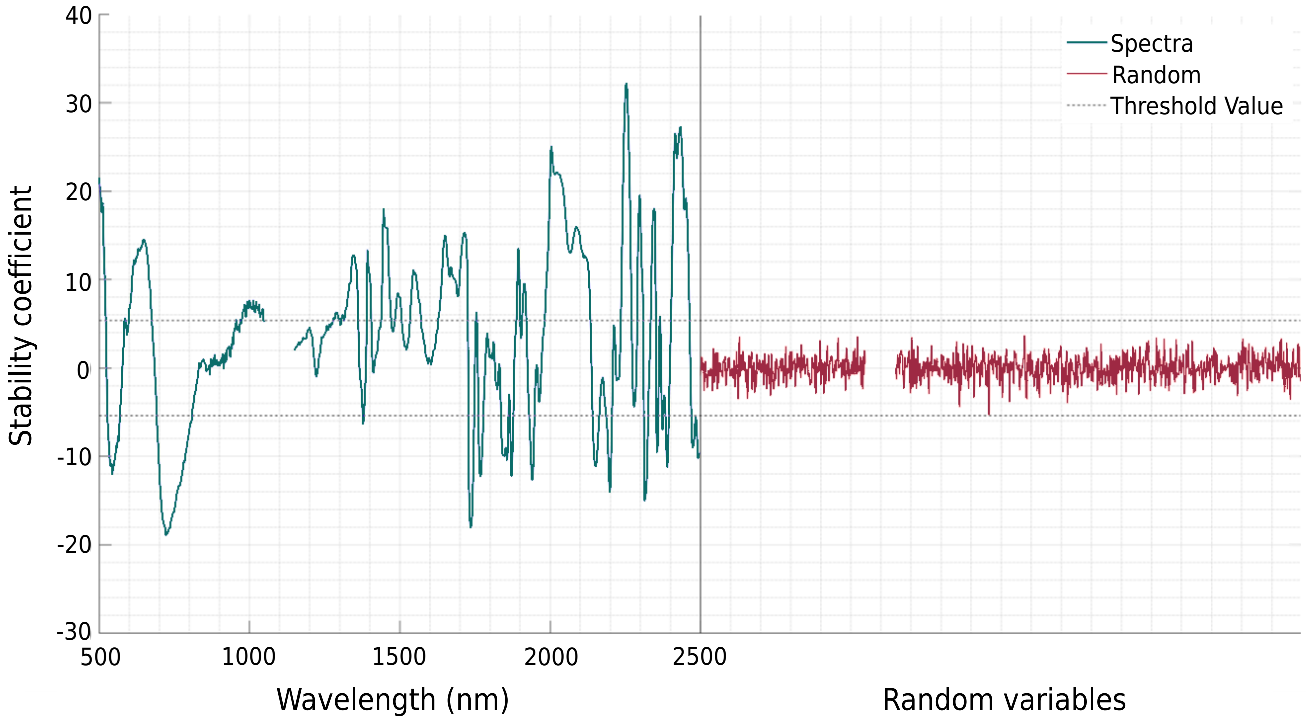

Figure 4. Distribution of stability value C for spectral and random variables in UVE variable selection.

500-760, 1300-1800, and 1900-2500 nm ranges. After variable selection using UVE, 570 spectral variables were retained.

CARS

Results of the feature band selection using the CARS method are shown in figure 5. Figure 5a displays the number of spectral bands screened out in the CARS variable selection process. With the exponential decay function, the number of spectral bands progressively declined as the number of runs grew, and the rate of the decrease gradually tended to level off. This shows that there are two stages of “coarse screening” and “fine screening” in the selection of variables by the CARS algorithm.

Figure 5b displays the RMSECV obtained from the PLSR model based on the spectral variables retained after each sampling. The RMSECV values flattened out and then spiked. The RMSECV value reached its lowest value of 19.7 g kg^{-1} at a sampling count 18, indicating that spectral information unrelated to SOC was removed in the first 18 variable selections. After the 18th screening, the RMSECV gradually increased.

Figure 5c shows the regression coefficient plots for all spectral variables during each sampling. Combined with the analysis in figure 5b, the selected spectral variables are the best subset containing 110 spectral bands when the number of sampling times is 18.

Comparison of spectral feature selection approaches

Spectral bands obtained by different feature selection approaches, including GA, correlation analysis (CA), SPA, UVE, and CARS, were each used as input variables for the PLSR modeling analysis. To better analyze the effects of different spectral band selection approaches, the performance of PLSR based on the full spectrum was also evaluated (Table 2). In table 2, the GA-PLSR (PLSR based on GA with improved coding approach) and UVE-PLSR outperformed the Full-PLSR, with R^2 improvement of 0.04 and 0.01 and RMSEP reduction of 0.4 and 0.2 g kg^{-1}, respectively, in the validation set. The CARS-PLSR and Full-PLSR achieved similar SOC prediction accuracy. However, only the RPD of GA-PLSR exceeded 1.4 in the validation set. The number of feature wavelengths filtered

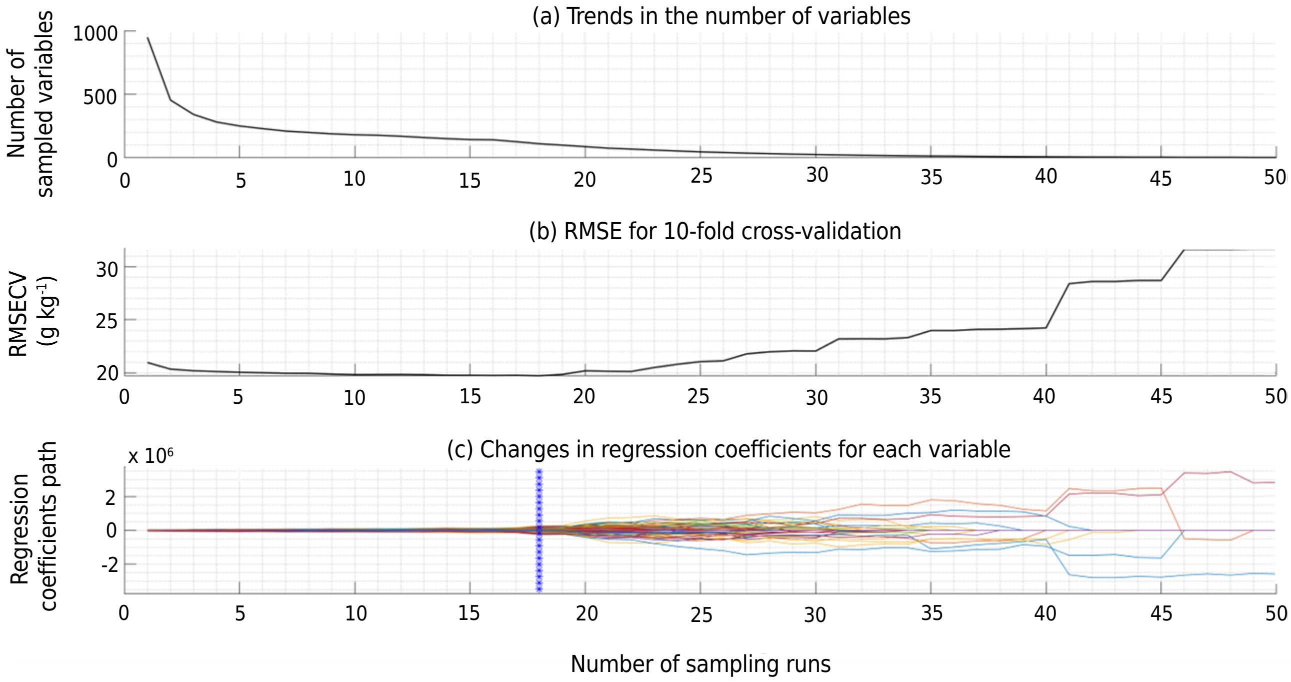

Figure 5. Results of feature selection using the CARS method.

by the different variable selection methods varied widely. The SPA screened the least number of bands (59 bands), representing 6.2 % of the total number of bands. The UVE method screened the highest number of bands (570 bands), accounting for 60.1 % of the total number of bands. The GA based on binary coding and the GA based on an improved coding method selected the same number of feature bands, but the model performance of oGA-PLSR (PLSR model based on GA with binary coding approach) and GA-PLSR differed significantly. Therefore, considering all factors together, the GA based on an improved coding approach was the optimal spectral variable selection method for the agricultural subset of LUCAS in the calibrated global model. This provided an empirical basis for future global modeling on a large scale. Corresponding to the results in table 2, GA-PLSR achieved the best fit with a correlation coefficient of 0.79 (Figure 6).

Figure 7 shows the spectral bands filtered by the five feature selection approaches. The feature spectral bands were mainly focused on 400-700, 1200-1600 and 1800-2400 nm. The spectral properties in the Vis range were mostly caused by electron leaps, which were manifested in the soil by constituents such as acanthite and hematite (Yu et al., 2016). The spectral properties in the NIR range were mainly caused by Al-OH, C-H, O-H and C=O groups, which were manifested in the soil by constituents such as kaolinite, aliphatic compounds, and carbohydrates (Viscarra-Rossel et al., 2010). Based on the model performance of GA-PLSR, CARS-PLSR, and UVE-PLSR (Table 2), it was shown that GA, UVE, and CARS are effective spectral variable selection methods capable of accurately extracting the spectral bands associated with SOC and improving the prediction performance of the global model.

Local PLSR

Test results generated from all distance measurements are shown in figure 8. The local PLSR based on different variable selection methods showed large variations in

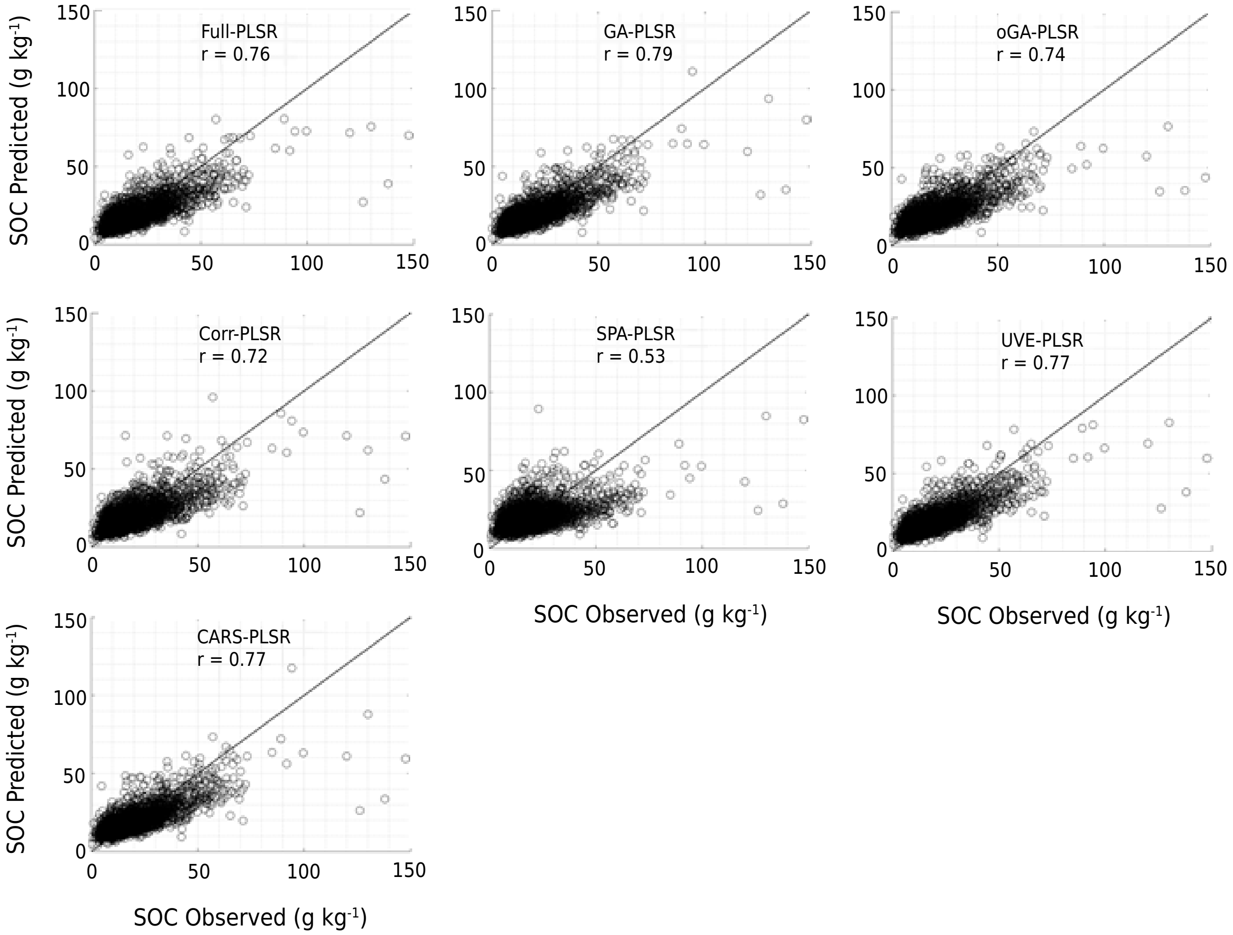

Figure 6. Predicted vs. measured SOC values for different PLSR models in the validation set. Full-PLSR is a PLSR based on the whole spectrum; GA-PLSR is a PLSR based on a genetic algorithm with an improved coding approach; oGA-PLSR is a PLSR based on a genetic algorithm with a binary coding approach; Corr-PLSR is a PLSR based on CA; SPA-PLSR is a PLSR based on SPA; UVE-PLSR is a PLSR based on UVE. CARS-PLSR is a PLSR based on CARS.

model performance due to various distance metrics. Euclidean distance obtained better predictions in all models except for the highest RMSE in the SPA-LPLSR. In contrast to Ward et al. (2019), MDist achieved the highest RMSE in all six local models except SPA-LPLSR. We found that cosine distance and correlation distance achieved good predictions in all local regression models and may serve as a more general alternative option for distance metrics in the local algorithm.

Table 2. Validation performance of SOC based on global PLSR and different variable selection approaches

Model	Number of variables	Model performance (Validation sets)				
		R^2	RMSEP	rRMSEP	RPD	RPIQ
Full-PLSR				g kg^{-1}		
GA-PLSR	949	0.47	7.8	14.86	1.37	1.70
GA-PLSR	448	0.51	7.4	14.27	1.42	1.77
Corr-PLSR	449	0.43	8.1	15.28	1.33	1.66
SPA-PLSR	332	0.42	8.3	15.49	1.31	1.63
UVE-PLSR	59	0.17	10.2	18.48	1.10	1.37
CARS-PLSR	110	0.46	7.8	14.88	1.37	1.70

Full-PLSR is a PLSR based on the whole spectrum; GA-PLSR is a PLSR based on a genetic algorithm with an improved coding approach; oGA-PLSR is a PLSR based on a genetic algorithm with a binary coding approach; Corr-PLSR is a PLSR based on CA; SPA-PLSR is a PLSR based on SPA; UVE-PLSR is a PLSR based on UVE; CARS-PLSR is a PLSR based on CARS.

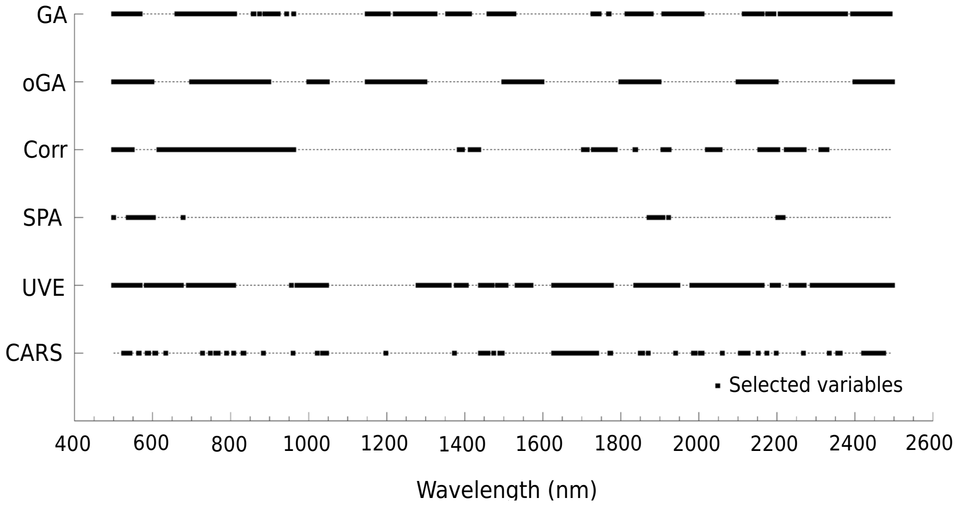

Figure 7. Spectral bands selected by various feature selection techniques.

The distance metric that yields the lowest RMSE in the test set was adopted for different local PLSR models. Table 3 shows the model performance for different local PLSRs in the validation set. In comparison with Full-LPLSR, the performance of GA-LPLSR and UVE-LPLSR was improved, with the R^2 increased by 0.05 and 0.03, the RPD increased by 0.14 and 0.07, and the RMSEP decreased by 0.5 and 0.3 g kg^{-1}, respectively. The optimal model performance was obtained by GA-LPLSR. However, the SOC prediction accuracy of Corr-LPLSR, SPA-LPLSR, and CARS-LPLSR was lower than that of Full-LPLSR. SPA-LPLSR exhibited the worst model performance, with an RPD of only 1.37. It was the only local model that failed to achieve a wide range of SOC predictions. According to the results in table 3, the GA-LPLSR model obtained the best fit with a correlation coefficient of 0.88, as shown in figure 9.

DISCUSSION

Most of the SOC-related characteristic wavelengths detected by the spectral selection technique are distributed in the Vis-NIR region. In the previous literature, Li et al. (2019) applied the CARS approach to select the sensitive wavebands of soil organic carbon in chestnut-calcium, black-calcium, gray-calcium, and mountain meadow soils, which were mainly distributed in 1900-2400 nm. Bao et al. (2020) screened the optimal spectral subsets of SOC for black soil, black calcareous soil, wind-sand soil and meadow soil by CARS algorithm, mainly at 1350-2400 nm, with a few at 400-1200 nm. In contrast to the aforementioned research, we found that the feature bands in the Vis region screened by the GA and UVE methods were significantly more than those of the SPA and CARS methods. This suggests that the spectral bands in the Vis region are critical for SOC prediction. This idea is also supported by the absorption peak near 600 nm according to the first-order derivative absorption spectrum in figure 2 (Bartholomeus et al., 2008). The feature bands detected by correlation analysis were mainly concentrated in the Vis region and less in the NIR region, which might be the major factor contributing to the poor performance of Corr-PLSR. The feature bands selected by the genetic algorithm based on binary coding were distributed in the form of regions, which resulted in some redundant bands being selected as well as some bands with critical SOC information being removed, making the model performance worse than the full-spectrum-based model.

Among the global models, the performance of Corr-PLSR, SPA-PLSR, and oGA-PLSR was worse than that of Full-PLSR. The SPA method greatly avoided the overlap of information between different spectral bands, but the SPA-PLSR model had the lowest RPD value of 1.16 in the validation set among all the models. This may be because the effective

Table 3. Validation performance of SOC based on local PLSR and different feature selection approaches

Model	Distance metric	Model performance (Validation sets)				
		R^2	RMSEP	rRMSEP	RPD	RPIQ
			g kg^{-1}			
Full-LPLSR	euDist	0.67	6.2	11.72	1.74	2.16
GA-LPLSR	euDist	0.71	5.7	10.86	1.87	2.33
oGA-LPLSR	euDist	0.64	6.8	12.31	1.65	2.06
Corr-LPLSR	euDist	0.65	6.6	11.97	1.70	2.11
SPA-LPLSR	MDist	0.46	8.2	14.88	1.37	1.70
UVE-LPLSR	euDist	0.69	5.9	11.23	1.81	2.25
CARS-LPLSR	euDist	0.66	6.3	11.80	1.72	2.14

Full-LPLSR is a LPLSR based on the whole spectrum; GA-LPLSR is a LPLSR based on a genetic algorithm with an improved coding approach; oGA-LPLSR is a LPLSR based on a genetic algorithm with a binary coding approach; Corr-LPLSR is a LPLSR based on CA; SPA-LPLSR is a LPLSR based on SPA; UVE-LPLSR is a LPLSR based on UVE; CARS-LPLSR is a LPLSR based on CARS.

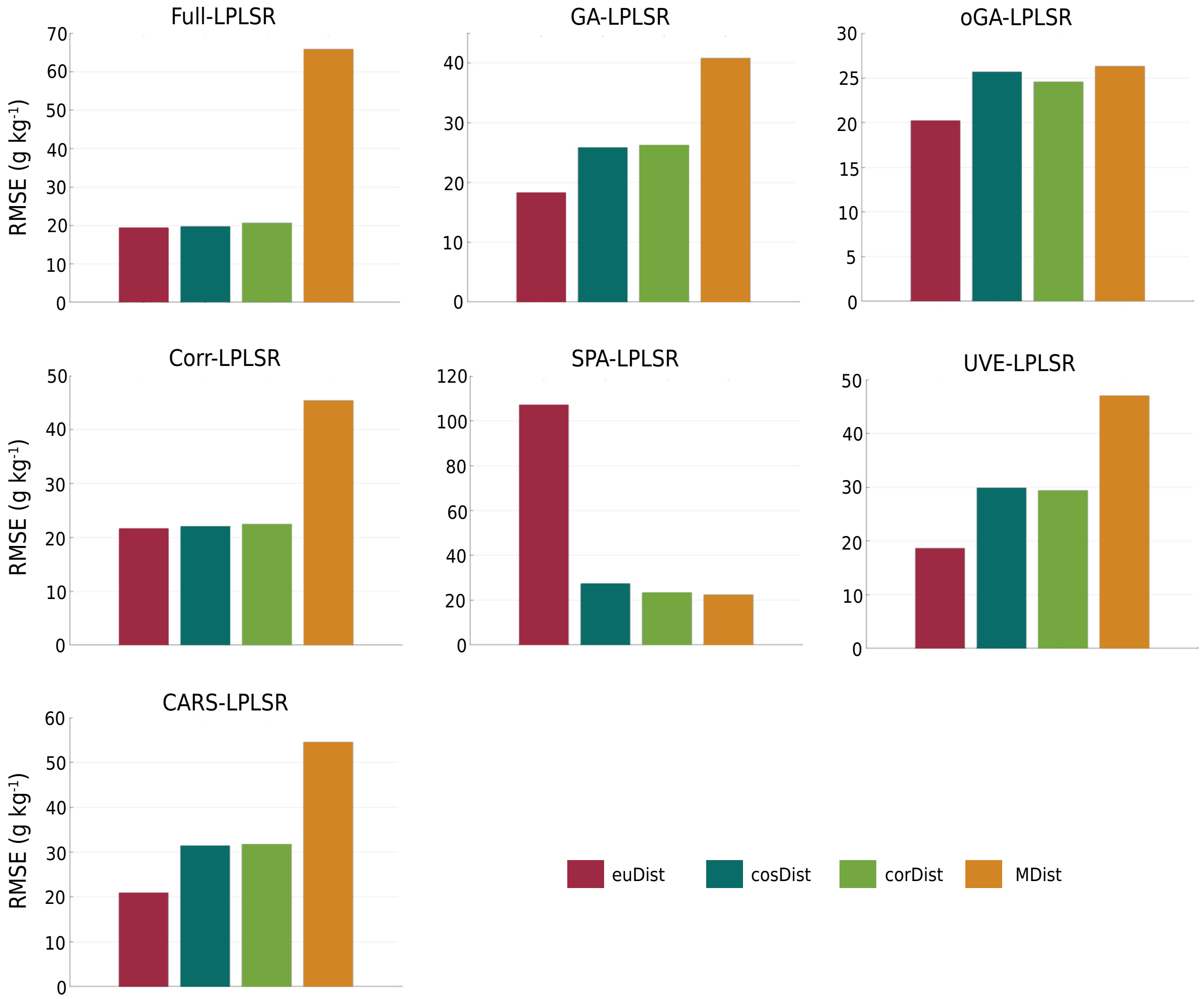

Figure 8. Model quality of local PLSR methods measured at different distances in an independent test set. Full-LPLSR is a LPLSR based on the whole spectrum; GA-LPLSR is a LPLSR based on a genetic algorithm with an improved coding approach; oGA-LPLSR is a LPLSR based on a genetic algorithm with a binary coding approach; Corr-LPLSR is a LPLSR based on CA; SPA-LPLSR is a LPLSR based on SPA; UVE-LPLSR is a LPLSR based on UVE; CARS-LPLSR is a LPLSR based on CARS.

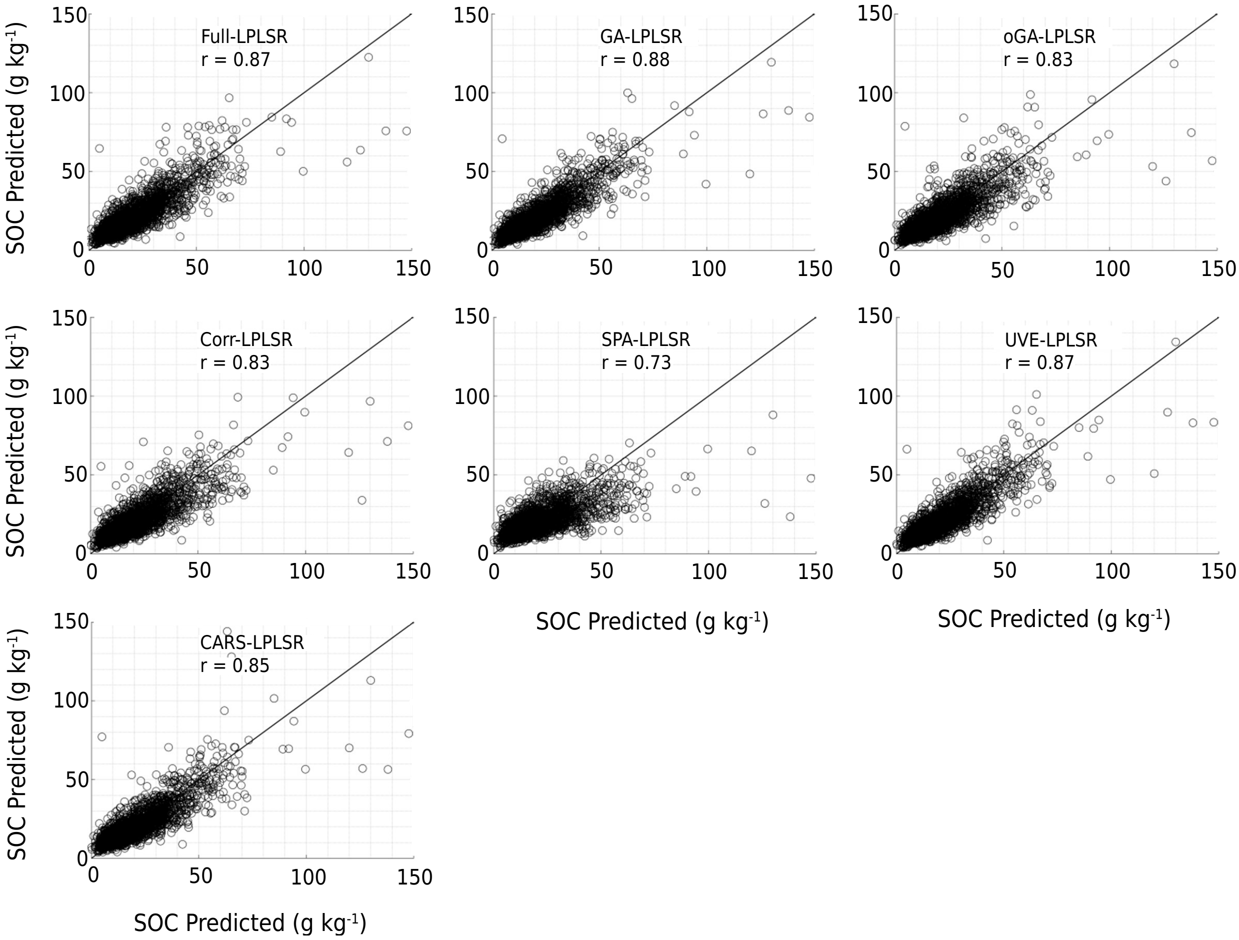

Figure 9. Predicted vs. measured SOC values for the validation sample of various local regression models. Full-LPLSR is a LPLSR based on the whole spectrum; GA-LPLSR is a LPLSR based on a genetic algorithm with an improved coding approach; oGA-LPLSR is a LPLSR based on a genetic algorithm with a binary coding approach; Corr-LPLSR is a LPLSR based on CA; SPA-LPLSR is a LPLSR based on SPA; UVE-LPLSR is a LPLSR based on UVE; CARS-LPLSR is a LPLSR based on CARS.

information related to SOC in soil hyperspectral data is not co-linear and the extracted feature bands fail to express all the band information, leading to the worst performance of the SPA-PLSR model. Previous studies have showed that the CARS method is an effective tool for accurately screening feature bands associated with the attributes to be predicted (Li et al., 2019; Tang et al., 2021). Similarly, Li et al. (2014) compared the GA, UVE and CARS variable selection methods for predicting the soluble solids content of 160 “Ya” pears using PLSR, and the CARS method obtained the optimal results. Yu et al. (2016) combined the SPA, UVE and CARS methods with PLSR to estimate the SOC of 56 soil samples and showed that the CARS method was superior to the SPA and UVE methods. Tang et al. (2021) studied 548 soil samples from three different soil types and combined the CARS method and random forest to estimate SOC with good results. However, in this research, we observed different results from previous studies, where the performance of SOC estimation models built with the CARS method was similar to that of the Full-PLSR model. This may be due to the large variation in the content of materials such as sand, calcium carbonate, and nitrogen in a large-scale soil database, which in turn affects the response of the SOC absorption characteristics (Stenberg et al., 2010). This leads to the inability of the CARS algorithm to detect the sensitive bands associated with the SOC accurately. The UVE algorithm is able to remove spectral information irrelevant to the SOC, thereby improving prediction accuracy (Centner et al., 1996). The GA is also able to select sensitive bands associated with SOC by optimizing the weighting coefficients of each spectral band. Among the global models based on six variable selection methods, the GA-PLSR showed the best performance with R^2 of 0.51 and RPD of 1.42, which achieved only a rough estimation of SOC.

We found that the local PLSR approach greatly enhanced the model performance in contrast to the global PLSR. The results of Full-LPLSR showed that the model performance improved by +43 % R^2, -20 % RMSEP and +27 % RPIQ compared to Full-PLSR. This is due to the fact that, in large soil databases, there may be local stabilization of spectral changes associated with soil properties, leading to better results with local methods. For local PLSR methods, the Euclidean distance is a more appropriate distance metric.

Comparing our results with those of Ward et al. (2019), who used the local PLSR on the LUCAS 2012 database with R^2 = 0.67, RMSEP = 5.2 g kg^{-1}, RPD = 1.74 and RPIQ = 1.96, a slight improvement in model performance was observed in our study. In this paper, the RMSEP for GA-LPLSR is slightly higher than the results of Ward et al. (2019). This is due to the higher standard deviation of 35.0 g kg^{-1} for SOC in the LUCAS 2015 database compared to the LUCAS 2012 database (14.1 g kg^{-1}). Stenberg et al. (2010), Nocita et al. (2014), and Ward et al. (2019) concluded that the prediction error of the spectral model rises as the standard deviation of the predicted soil properties increases.

CONCLUSION

Feature selection-based prediction models achieve similar performance to full-spectrum-based prediction models in both local and global models, with the exception of SPA (successive projection algorithm). Genetic Algorithm, with improved coding approaches and UVE, are able to improve the SOC prediction accuracy slightly. Therefore, if we want to improve the prediction accuracy of SOC in large soil databases, we should not put too much effort into the commonly used variable selection algorithms.

Prediction performance of local PLSR is better than that of global PLSR, regardless of the variable selection method used. There are significant differences in the RMSE of the distance metrics used for local PLSR, and they depend on the variable selection method used. Among the local PLSR algorithms, the euclidean distance achieves the lowest RMSE, except for the SPA-based local PLSR. The performance of cosine distance and correlation distance is more stable.

This study provides a reference for selecting appropriate feature selection methods for large scale SOC prediction and appropriate distance metrics in local PLSR, and will also contribute to the development of more robust organic carbon quantification models based on satellite hyperspectral data.

ACKNOWLEDGMENTS

Thank you for the LUCAS 2015 soil database provided free of charge by the European Soil Data Center managed by the Joint Research Center (JRC) of the European Commission. Thank Prof. Baofeng Guo for his financial support.

AUTHOR CONTRIBUTIONS

Conceptualization: Baoyang Liu (equal) and Baofeng Guo (equal).

Data curation: Baoyang Liu (lead).

Formal analysis: Baoyang Liu (equal), Renxiong Zhuo (equal) and Fan Dai (equal).

Investigation: Baoyang Liu (equal) and Baofeng Guo (equal).

Methodology: Baoyang Liu (equal) and Baofeng Guo (equal).

Resource: Baoyang Liu (equal), Baofeng Guo (equal), Renxiong Zhuo (equal) and Fan Dai (equal).

Software: Baoyang Liu (lead).

Supervision: Baofeng Guo (lead).

Validation: Baoyang Liu (equal), Renxiong Zhuo (equal) and Fan Dai (equal).

Visualization: Baoyang Liu (equal), Renxiong Zhuo (equal) and Fan Dai (equal).

Writing - original draft: Baoyang Liu (lead).

Writing - review & editing: Baoyang Liu (equal), Baofeng Guo (equal), Renxiong Zhuo (equal) and Fan Dai (equal).

REFERENCES

Araújo MCU, Saldanha TCB, Galvão RKH, Yoneyama T, Chame RC, Visani V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr Intell Lab Syst. 2001;57:65-73. https://doi.org/10.1016/S0169-7439(01)00119-8

Araújo SR, Wetterlind J, Demattê JAM, Stenberg B. 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. Eur J Soil Sci. 2014;65:718-29. https://doi.org/10.1111/ejss.12165

Bao Y, Meng X, Ustin S, Wang X, Zhang X, Liu H, Tang H. Vis-SWIR spectral prediction model for soil organic matter with different grouping strategies. Catena. 2020;195:104703. https://doi.org/10.1016/j.catena.2020.104703

Bartholomeus HM, Schaepman ME, Kooistra L, Stevens A, Hoogmoed WB, Spaargaren OSP. Spectral reflectance based indices for soil organic carbon quantification. Geoderma. 2008;145:28-36. https://doi.org/10.1016/j.geoderma.2008.01.010

Baumgardner MF, Silva LF, Biehl LL, Stoner ER. Reflectance properties of soils. Adv Agron. 1986;38:1-44. https://doi.org/10.1016/S0065-2113(08)60672-0

Bellon-Maurel V, McBratney A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils-critical review and research perspectives. Soil Biol Biochem. 2011;43:1398-410. https://doi.org/10.1016/j.soilbio.2011.02.019

Ben-Dor E, Irons J, Epema GF. Soil reflectance. In: Rencz AN, editor. Remote sensing for the earth science. New York: Wiley; 1999. p. 111-88.

Brown DJ, Bricklemyer RS, Miller PR. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma. 2005;129(3):251-67. https://doi.org/10.1016/j.geoderma.2005.01.001.

Centner V, Massart D-L, Noord OE, Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. Anal Chem. 1996;68:3851-8. https://doi.org/10.1021/ac960321m

Clark RN, Roush TL. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. J Geophys Res. 1984;89:6329-40. https://doi.org/10.1029/JB089iB07p06329

Conant RT, Ogle SM, Paul EA, Paustian K. Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. Front Ecol Environ. 2011;9:169-73. https://doi.org/10.1890/090153

Davies AMC. An introduction to near infrared spectroscopy. NIR News. 2005;16:9-11.

Gu X, Wang Y, Sun Q, Yang G, Zhang C. Hyperspectral inversion of soil organic matter content in cultivated land based on wavelet transform. Comput Electron Agr. 2019;167:105053. https://doi.org/10.1016/j.compag.2019.105053

Holland JH. Adaptation in natural and artificial systems. Ann Arbor: University of Michigan Press; 1975.

Hong Y, Chen S, Liu Y, Zhang Y, Yu L, Chen Y, Liu Y, Cheng H, Liu Y. Combination of fractional order derivative and memory-based learning algorithm to improve the estimation accuracy of soil organic matter by visible and near-infrared spectroscopy. Catena. 2019;174:104-16. https://doi.org/10.1016/j.catena.2018.10.051

Hong Y, Chen S, Chen Y, Linderman M, Mouazen AM, Liu Y, Guo L, Yu L, Liu Y, Cheng H, Liu Y. Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest. Soil Till Res. 2020a;199:104589. https://doi.org/10.1016/j.still.2020.104589

Hong Y, Guo L, Chen S, Linderman M, Mouazem AM, Yu L, Chen Y, Liu Y, Liu Y, Cheng H, Liu Y. Exploring the potential of airborne hyperspectral image for estimating topsoil organic carbon: Effects of fractional-order derivative and optimal band combination algorithm. Geoderma. 2020b;365:114228. https://doi.org/10.1016/j.geoderma.2020.114228

Kennard RW, Stone LA. Computer aided design of experiments. Technometrics. 1969;11:137-48. https://doi.org/10.1080/00401706.1969.10490666

Li H, Liang Y, Xu Q, Cao D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal Chim Acta. 2009;648:77-84. https://doi.org/10.1016/j.aca.2009.06.046

Li J, Peng Y, Chen L, Huang W. Near-infrared hyperspectral imaging combined with CARS algorithm to quantitatively determine soluble solids content in “Ya” pear. Spectrosc Spect Anal. 2014;34:1264-9. https://doi.org/10.3964/j.issn.1000-0593(2014)05-1264-06

Li W, Gao X, Xiao N, Xiao Y. Estimation soil organic matter contents with hyperspectra based on sCARS and RF algorithms. J Lumin. 2019;40:1030-9.

Meng X, Bao Y, Liu J, Liu H, Zhang X, Zhang Y, Wang P, Tang H, Kong F. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. Int J Appl Earth Obs. 2020;89:102111. https://doi.org/10.1016/j.jag.2020.102111

Meng X, Bao Y, Zhang X, Wang X, Liu H. Prediction of soil organic matter using different soil classification hierarchical level stratification strategies and spectral characteristic parameters. Geoderma. 2022;411:115696. https://doi.org/10.1016/j.geoderma.2022.115696

Nocita M, Stevens A, Toth A, Panagos G, van Wesemael B, Montanarella L. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol Biochem. 2014;68:337-47. https://doi.org/10.1016/j.soilbio.2013.10.022

Orgiazzi A, Ballabio C, Panagos P, Jones A, Fernández-Ugalde O. LUCAS Soil, the largest expandable soil dataset for Europe: A review. Eur J Soil Sci. 2017;69:140-53. https://doi.org/10.1111/ejss.12499

Post JL, Noble PN. The near-infrared combination band frequencies of dioctahedral smectites, micas, and illites. Clays Clay Miner. 1993;41:639-44. https://doi.org/10.1346/CCMN.1993.0410601

Ramirez-Lopez L, Behrens T, Schmidt K, Stevens A, Demattê JAM, Scholten T. The spectrum-based learner: A new local approach for modeling soil Vis-NIR spectra of complex datasets. Geoderma. 2013;195-196:268-79. https://doi.org/10.1016/j.geoderma.2012.12.014

Sanchez PA, Ahamed S, Carré F, Hartemink AE, Hempel J, Huising J, Lagacherie P, Mcbratney AB, Mckenzie NJ, Mendonça-Santos ML, Minasny B, Montanarella L, Okoth P, Palm CA, Sachs JD, Shepherd KD, Vågen T-G, Vanlauwe B, Walsh MG, Winowiecki LA, Zhang GL. Digital soil map of the world. Science. 2009;325:680-1. https://doi.org/10.1126/science.1175084

Savitzky A, Golay M. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem. 1964;36:1627-39. https://doi.org/10.1021/ac60214a047

Savvides A, Corstanje R, Baxter SJ, Rawlins BG, Lark RM. The relationship between diffuse spectral reflectance of the soil and its cation exchange capacity is scale-dependent. Geoderma. 2010;154:353-8. https://doi.org/10.1016/j.geoderma.2009.11.007

Seely B, Welham C, Blanco JA. Towards the application of soil organic matter as an indicator of forest ecosystem productivity: Deriving thresholds, developing monitoring systems, and evaluating practices. Ecol Indic. 2010;10:999-1008. https://doi.org/10.1016/j.ecolind.2010.02.008

Six J, Paustian K. Aggregate-associated soil organic matter as an ecosystem property and a measurement tool. Soil Biol Biochem. 2014;68:A4-9. https://doi.org/10.1016/j.soilbio.2013.06.014

Stenberg B, Viscarra-Rossel RA, Mouazen AM, Wetterlind J. Visible and near infrared spectroscopy in soil science. Adv Agron. 2010;107:163-215. https://doi.org/10.1016/S0065-2113(10)07005-7

Stevens A, Nocita M, Tóth G, Montanarella L, van Wesemael B. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS One. 2013;8:66409. https://doi.org/10.1371/journal.pone.0066409

Tiecher T, Moura-Bueno JM, Caner L, Minella JPG, Evrard O, Ramon R, Naibo G, Barros CAP, Silva YJAB, Amorim FF, Rheinheimer DS. Improving the quantification of sediment source contributions using different mathematical models and spectral preprocessing techniques for individual or combined spectra of ultraviolet-visible, near-and middle-infrared spectroscopy. Geoderma. 2021;384:114815. https://doi.org/10.1016/j.geoderma.2020.114815

Tang H, Meng X, Su X, Ma T, Liu H, Bao Y, Zhang M, Zhang X, Huo H. Hyperspectral prediction on soil organic matter of different types using CARS algorithm. Transactions of the CSAE. 2021;37:106-13.

Toth G, Jones A, Montanarella L, Alewell C, Ballabio C, Carre F, Brogniez D, Guicharnaud R, Gardi C, Hermann T, Meusburger K, Nocita M, Panagos P, Rusco E, Stevens A, van Liedekerke M, Van Wesemael B, Weynants M, Yigini Y. LUCAS Topoil Survey - methodology, data and results. EUR 26102. Luxembourg: Publications Office of the European Union; 2013. https://doi.org/10.2788/97922

Viscarra-Rossel RA, Behrens T, Ben-Dor E, Brown DJ, Demattê JAM, Shepherd KD, Shi Z, Stenberg B, Stevens A, Adamchuk V, Aïchi H, Barthès BG, Bartholomeus HM, Bayer AD, Bernoux M, Böttcher K, Brodský L, Du CW, Chappell A, Fouad Y, Genot V, Gomez C, Grunwald S, Gubler A, Guerrero C, Hedley CB, Knadel M, Morrás HJM, Nocita M, Ramirez-Lopez L, Roudier P, Campos EMR, Sanborn P, Sellitto VM, Sudduth KA, Rawlins BG, Walter C, Winowiecki LA, Hong SY, Ji W. A global spectral library to characterize the world's soil. Earth-Sci Rev. 2016;155:198-230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra-Rossel RA, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma. 2010;158:46-54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra-Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma. 2006;131:59-75. https://doi.org/10.1016/j.geoderma.2005.03.007

Wang X, Zhang XK, Li HX, Zhang X, Liu H, Dou X, Yu Z. The minimum level for soil allocation using topsoil reflectance spectra: Genus or species? Catena. 2019;174:36-47. https://doi.org/10.1016/j.catena.2018.11.001

Ward KJ, Chabrillat S, Neumann C, Foerster S. A remote sensing adapted approach for soil organic carbon prediction based on the spectrally clustered LUCAS soil database. Geoderma. 2019;353:297-307. https://doi.org/10.1016/j.geoderma.2019.07.010

Yu L, Hong Y, Zhou Y. Wavelength variable selection methods for estimation of soil organic matter content using hyperspectral technique. Transactions of the CSAE. 2016;32:95-102.