https://doi.org/10.17221/133/2022-SWR

Strategies and methods for predicting soil organic matter at the field scale based on the provincial near infrared spectral database

Shengyao Jia[1,2], Chunbo Hong[1,2], Hongyang Li[1,2], Yuchan Li[1,2], Siyuan Hu[3]*

[1]College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou, P.R. China
[2]Key Laboratory of Intelligent Manufacturing Quality Big Data Tracing and Analysis of Zhejiang Province, China Jiliang University, Hangzhou, P.R. China
[3]Zhejiang Provincial Emergency Management Science Research Institute, Hangzhou, P.R. China

**Corresponding author: lihongyang@cjlu.edu.cn*

Citation: Jia S.Y., Hong C.B., Li H.Y., Li Y.C., Hu S.Y. (2023): Strategies and methods for predicting soil organic matter at the field scale based on the provincial near infrared spectral database. Soil & Water Res., 18: 158–168.

Abstract: The development and provision of soil spectral library (SSL) could facilitate the application of near infrared (NIR) spectroscopy for economical, accurate, and efficient determination of soil organic matter (SOM). In this work, the performances of partial least squares regression (PLSR) and convolutional neural network (CNN) combined with the datasets of Zhejiang provincial SSL (ZSSL) and the feature subset (FS) were compared for the prediction of SOM at the target field. The FS dataset was chosen from ZSSL based on similarity to the spectral characteristics of the target samples. The results showed that compared with modelling using ZSSL, modelling using FS can greatly improve the prediction accuracy of the PLSR model, but the impact on the performance of the CNN model was limited. The method of mean squared Euclidean distance (MSD) was an effective way for determining the optimal spiking sample size for the PLSR model only using the spectral data of the spiking subset and the prediction set. The PLSR model combined with the FS dataset and the spiking subset determined by MSD achieved the optimal prediction results among all developed models, which is an accurate and easy-to-implement solution for the SOM determination based on ZSSL.

Keywords: convolutional neural network; soil organic content; soil spectral library; spiking sample size; strategy

Soil organic matter (SOM) is one of the most important indicators of cultivated soil, which plays a core role in soil nutrient cycling and transformation, providing nitrogen, phosphorus, potassium and other nutrients required for crop growth. Quickly and accurately obtaining the content and distribution of SOM, so as to fertilize reasonably and accurately, is of great significance for the sustainable development of agriculture and the successful implementation of precision agriculture (Li et al. 2022). During the last two decades, near infrared (NIR) spectroscopy has been widely employed as an effective tool for the quantitative analysis of soil attributes with the advantages of fast detection speed, no pollution, low cost, and simple operation (Seidel et al. 2019; Li et al. 2020; Davari et al. 2021). However, NIR spectroscopy

Supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21D010007 and Grant LR22F010001, the Science and Technology Program of the China's State Administration for Market Regulation under Grant 2022MK0364, and the Scientific Research Project of the Education Department of Zhejiang Province under Grant Y201737559, and the Zhejiang Provincial Emergency Management Science Research Institute Project under Grant 330000220130371007005.

https://doi.org/10.17221/133/2022-SWR

is susceptible to interference from soil type, moisture content, surface roughness and the nature of the compounds. Soil spectral predictive mechanisms may vary from one sample set to another, resulting in poor generalization ability of the calibration model (Voland & Emmerling 2011; Jia et al. 2016; Munnaf et al. 2021). This greatly limits the popularization and application of the technology. In order to enhance the calibration model applicability, more and more researchers tend to utilize soil spectral library (SSL) for modelling, including globe scale, national scale and regional or local scale (Guerrero et al. 2016; Nawar & Mouazen 2017; Yang et al. 2020). On the one hand, it increases the coverage of soil sample information for modelling. On the other hand, economic losses and waste of natural resources due to repeated sampling are reduced.

Nevertheless, developing an accurate spectroscopy model using SSL remains a challenging task. The main reason is that the SSL usually contain a large amount of sample information with different spectral characteristics, resulting in the unique characteristics of SOM in the target field can not be appropriately reflected in the calibration (Tsakiridis et al. 2019). To prevent degradation in the accuracy from the application of SSL-based models to local sample sets, two general strategies have been applied. One is to use a specified subset from the SSL for modelling which has similar spectral characteristics to the target soil samples. Seidel et al. (2019) selected 137 samples from the German SSL as a subset which were most similar to the target samples based on spectral cluster analysis. They considered that modelling with this subset can significantly improve the performance of the calibration model. The other strategy is spiking, that is adding a small number of the target samples to the SSL, thereby enhancing the impact of the calibration model on the target samples (Gogé et al. 2014; Jiang et al. 2017; Knadel et al. 2017). The spiking samples require laboratory chemical analysis of soil attributes which are costly. Optimizing the number of the spiking samples is thus very important. The methods of Kennard-Stone (Kennard & Stone 1969), fuzzy c-means (Havens et al. 2012) and Latin hypercube (McKay et al. 1979) are commonly used for selecting spiking samples. However, the values of soil attributes need to be known to determine the optimal spiking number, which are inconvenient in actual applications. The method by analysing the mean squared Euclidean distance (MSD) is able to identify the spiking number only using the spectral data of the spiking subset and the target samples (Ramirez-Lopez et al. 2014). Li et al. (2022) argued that MSD was a simple and effective method to determine an adequate spiking set.

As the sample size in SSL increases, modelling between soil attributes and spectral data becomes increasingly complex. In recent years, convolutional neural network (CNN) has been widely employed in the field of big data processing, such as image classification (Wang et al. 2022), speech recognition (Mustaqeem & Kwon 2021), natural language processing (Hong et al. 2022), etc. With the characteristics of local connection and weight sharing, the CNN model can automatically identify and extract spectral characteristics from the original spectral data (Padarian et al. 2019). Several studies have successfully applied CNN for regression modelling using NIR spectroscopy data (Ng et al. 2020; Tsakiridis et al. 2020). The results showed that it has the capability to outperform PLSR for the prediction of various soil attributes. However, these conclusions are obtained based on the modelling of a large number of samples (> 10 000) in the SSL. When using a specified subset from the SSL for modelling or the SSL sample size is not large enough, the advantage of using CNN is uncertain.

Against the backdrop of the above research, this work aims to use the SSL of Zhejiang Province, P.R. China (ZSSL) to quantitative analysis SOM content in a target field for long-term soil fertility monitoring. In this work, the main objective is to investigate which strategy (feature subset, spiking, or a combination of both) is adopted without conducting laboratory chemical analysis of the target farmland SOM, and which method (PLSR or CNN) can be used to achieve accurate prediction of the target field SOM based on ZSSL.

The innovation points of this paper are as follows: First, using MSD method to determine the optimal spiking sample size based on spectral data only, thus reducing the workload of laboratory chemical analysis and improving the cost-effectiveness of NIR spectral detection. Secondly, based on ZSSL, it provides a low-cost and rapid solution measurement of SOM.

MATERIAL AND METHODS

Soil spectral library and target field. Zhejiang Province is located on the southeast coast of China, with an area of 105 500 km^2. The distribution map of sampling locations is shown in Figure 1A. Each

https://doi.org/10.17221/133/2022-SWR

point in Figure 1A represents the location of sampling points in a certain district of Zhejiang Province. There will be a different number of experimental fields in each sampling location. Different amounts of soil samples will be collected based on factors such as soil types and planting crops in different experimental fields. Finally, the total number of soil samples used in the ZSSL database for this study is 2 069. The soil samples were chosen from the topsoil (0–20 cm) during 2012–2021 and covered the main soil types of Zhejiang Province, such as paddy soil, red loam, yellow loam, moisture soil, coastal saline soil and so on. The target field was located in long-term cultivated farmland (30°11'N, 120°48'E) in Shangyu, Zhejiang Province, covering an area of 15.2 km^2, shown in Figure 1B. The main soil type is paddy soil. The field crops are generally rice, vegetables, soybeans, and potatoes. In various types of soils in Zhejiang Province, the primary minerals commonly occur are quartz and potassium feldspar, and the secondary minerals are illite. In general, the soil texture of red loam is clay loam, with kaolinite as the main clay mineral, followed by illite. The soil texture of yellow loam is generally silty loam or clay loam, and the clay minerals are mainly vermiculite, chlorite, and kaolinite, accompanied by illite and quartz. Compared with other types of soil, the content of secondary mineral illite in moisture soil, coastal saline soil, and various paddy soils significantly increased, while the content of montmorillonite, kaolinite, and chlorite slightly increased, with the occurrence of magnesium containing mineral vermiculite dolomite. In 2020, with the help of the land development centre of the target district, 120 topsoil samples were collected from this target field. The minimum spacing between sampling points is 100 meters. At each sampling point, collect 5 soil samples according to the "plum blossom method", with a sampling depth of 0–20 cm. Then mix the 5 soil samples evenly, pick out the straw and stones, lay them flat into a square and draw two diagonal lines. Take the two opposite pieces and discard the rest. Repeat the above operation until the required amount (about one kilogram) is obtained and use them as a soil sample for the experiment.

Spectral analysis and chemical analysis. The soil samples from the ZSSL and the target field were air-dried and sieved to pass through a < 2 mm mesh. The diffuse reflectance spectra were measured by a Fourier-type NIR spectrometer (Matrix-I, Bruker Optics Inc., Germany) under laboratory conditions. Absorbance, as log1/R where R is reflectance, was recorded in the wavelength range of 1 000 to 2 500 nm for a total of 1 555 wavelength variables per spectrum, with a spectral resolution of 8 cm^{-1}. For preprocessing, the spectra were first smoothed by averaging five successive wavelengths. Then, standard normalized variate (SNV) was used to reduce baseline offset and noise of the spectra. The pre-processed spectra were used for further analysis.

Chemical analyses of SOM content (g/kg) were performed by the agricultural testing centre of Zhejiang Provincial Academy of Agricultural Sciences. The content was measured colourimetrically after H_2SO_4-dichromate oxidation at 150 °C.

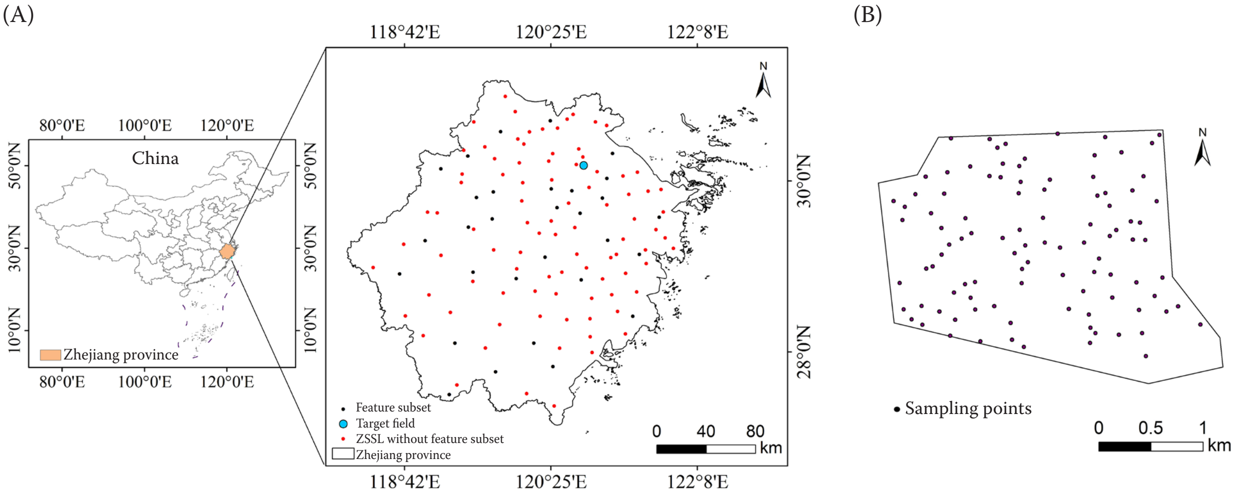

Figure 1. The sampling locations of the soil spectral library of Zhejiang Province and the feature subset (A) and the sampling points of the target field (B)

ZSSL – soil spectral library of Zhejiang Province

Modelling methods. Partial least squares regression (PLSR) is one of the most widely used algorithms for modelling soil NIR spectroscopy. It is a linear regression model that projects spectra into latent variables explaining the variances within the spectra and the response variables. The optimum number of latent variables for PLSR models is determined by minimization of the root mean square error of cross-validation after leave-one-out cross-validation.

The CNN model consists of several convolutional layers, pooling layers, and fully-connected (or dense) layers. In the convolutional layer, the 1-dimension depth-wise convolution structure was used to filter a given input and extract different local features from the input spectrum. The convolution operation can be expressed as follows:

$$y_i^k = f\left(\sum_{j=1}^{n^{k-1}} x_j^{k-1} w_{i,j}^k + b_i^k\right), i = 1,2,...,n^k$$

where:

y_i^k – the i^{th} feature map on the k^{th} layer;

x_j^{k-1} – the j^{th} input feature map on the k–1^{th} layer;

$w_{i,j}^k$ – the convolution kernel weight between the j^{th} input feature map on the k–1^{th} layer and the i^{th} feature map on the k^{th} layer;

b_i^k – the bias;

n^k – the number of output feature maps on k^{th} layer;

n^{k-1} – the number of input feature maps on k–1^{th} layer.

Following each convolution layer, batch normalization was used to normalize a layer by shifting and scaling the activations to prevent the internal covariate shift problem. The rectified linear unit (RELU) was applied as the activation function. Then, the max pooling layer was adopted to reduce the risk of overfitting by providing a more abstract representation of a layer. The fully-connected layer was used to connect all outputs of the previous layer to all inputs of the next layer. The CNN module hyperparameters are summarized in Table 1. Figure 2 depicts the overall network architecture.

To train the CNN model, the calibration set was divided into two parts based on the five-fold cross-validation technique. The training set accounted for 80% samples was used to tune the model parameters. The validation set accounted for 20% samples was used evaluate the model accuracy. The root mean square error between the measured and predicted values of SOM was applied as the loss function. During the training stage, the weights of the model were adjusted based on the Adam optimizer with an initial learning rate of 0.001. The mini-batch sizes and the maximum number of training iterations were set to 32 samples and 400 epochs respectively. The training was stopped if no improvement in the accuracy of the loss function occurred, or the number of the training epochs reached the maximum. The CNN model was implemented in Matlab2019 Deep Learning Toolbox (Mathworks Co., MA., USA).

Modelling strategies. Since the dataset of ZSSL covered a large amount of soil information which might not share edaphic characteristics with the targe field samples, the feature subset (FS) was chosen from ZSSL for modelling based on spectral similarity metrics. The steps for the determination of FS were as follows. First, principal component analysis (PCA) was performed on the ZSSL samples and the target field samples. The number of principal components

Table 1. Hyperparameter settings of the convolutional neural network (CNN) model

Type	Kernel size	Filters	Output size	Activation
Convolutional + batch norm	8	8	1 548.8	RELUs
Max-pooling	2	–	774.8	–
Convolutional + batch norm	8	16	767.16	RELUs
Max-pooling	2	–	383.16	–
Convolutional + batch norm	8	32	376.32	RELUs
Max-pooling	2	–	188.32	–
Convolutional + batch norm	8	64	181.64	RELUs
Flatten	–	–	11 584.1	–
Fully-connected	–	32	1.32	ReLUs
Fully-connected	–	–	1	linear

RELUs – rectified linear units

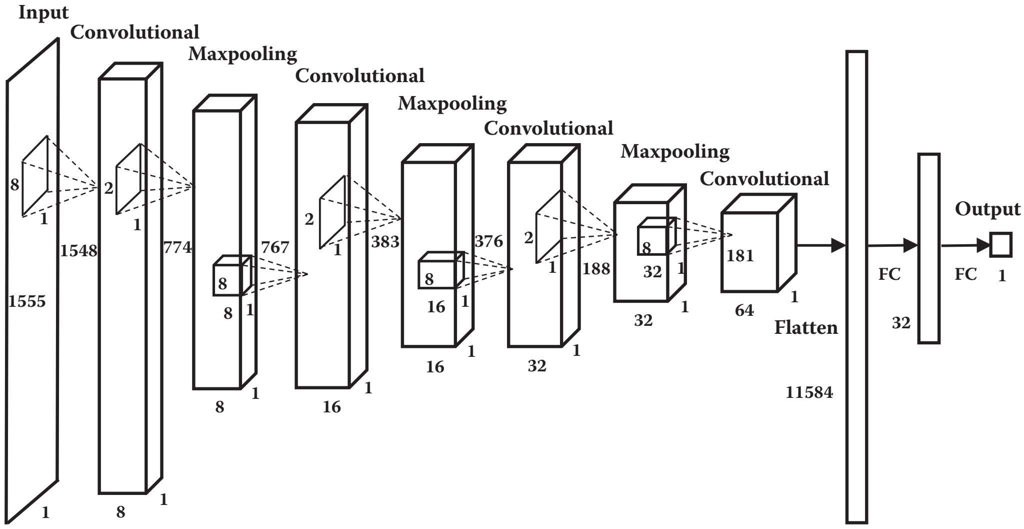

Figure 2. The architecture of the convolutional neural network (CNN) model
FC – fully-connected

(PCs) was determined according to the threshold of explained variance (≥ 99%). Then, for each soil sample in the target field, 50 samples from ZSSL were selected according to the minimum Mahalanobis distance in the PC feature space of the combined ZSSL and target sample spectra. Finally, the repeatedly selected samples were removed to form FS, and the total number of soil samples used for this study in the FS database is 174.

Out of the 120 samples of the target field, up to 30 samples were randomly selected to spike the initial calibration models. To evaluate the optimal number of spiking samples, 10 to 30 samples were randomly selected as a spiking subset with a step of 5 samples. For each spiking subset, Gaussian kernel density estimates of the probability density function were calculated based on the first PC of the spectral data, which were called as P_S. Then, Gaussian kernel density estimates of the probability density function of the prediction set were calculated in the same way, which were called as P_V. Finally, the mean squared Euclidean distance (MSD) between P_S and P_V was calculated. According to the MSD, the optimal spiking subset was determined. The detailed description of the MSD can be found in Ramirez-Lopez et al.(2014).

In this work, the methods of PLSR and CNN combined with the datasets of ZSSL and FS were utilized to establish calibration models for the prediction of SOM in the prediction set. The established models were denoted as ZSSL_PLSR, ZSSL_CNN, FS_PLSR and FS_CNN, respectively. After subtracting the spiking samples, there were 90 samples left in the target field, which were used as the prediction set. Because the ZSSL dataset was much larger than the FS dataset, in order to balance the leverage of the spiking samples between ZSSL and FS, the spiking samples added to the ZSSL were extra-weighted by a copy of λ, then the number of spiking samples in the ZSSL would become the original λ times. In this study, the spiking samples added to the ZSLL were extra-weighted by λ = 12, which was approximately equal to the ratio of the number of samples in the ZSSL dataset (2 069) to the feature set FS (174). The root mean squared error of prediction in the prediction set (RMSE), the bias and the coefficient of determination (R^2) have been applied to evaluate the prediction accuracy. Generally, large value of R^2 and small values of bias and RMSE indicate good predictions.

RESULTS AND DISCUSSION

Statistical analysis and soil spectral characteristics. Table 2 summarizes the SOM content for each dataset used for calibration and prediction, and Figure 3 shows the distribution of SOM values in the core datasets.

The ZSSL dataset covered the largest range (2.4 to 96.5 g/kg) and variation of SOM content, but the mean value was the lowest (26.0 g/kg). Derived from ZSSL based on spectral similarity to the target field

Table 2. Statistics of soil organic matter for each soil dataset

Dataset	n	Min	Max	Mean	SD
		(g/kg)			
ZSSL	2 069	2.4	96.5	26.0	11.5
Feature subset	174	9.1	68.7	29.5	10.2
Prediction set	90	14.5	48.2	32.9	8.5
Spiking subset	30	13.4	56.2	33.9	8.7

ZSSL – soil spectral library of Zhejiang Province; n – number of samples; SD – standard deviation

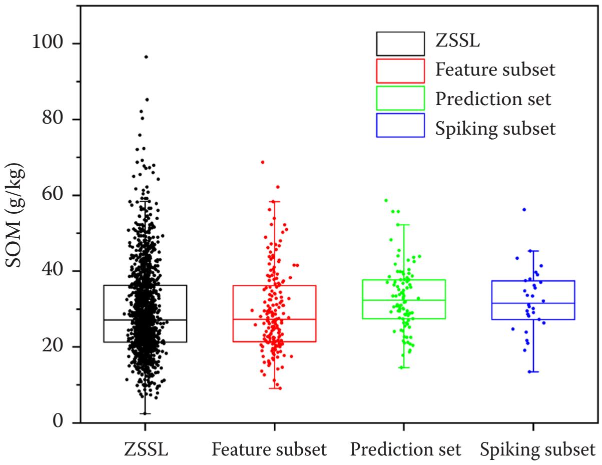

Figure 3. The distribution of soil organic matter (SOM) values in the core datasets

ZSSL – soil spectral library of Zhejiang Province

samples, the FS dataset still covered a larger range than the prediction set. Compared with ZSSL, the range and the mean value of SOM content of the FS dataset were much closer to those of the prediction set. The SOM distribution of the spiking subset was comparable to the prediction set.

Figure 4A showed the averaged spectrum of each dataset in the range of 1 000–2 500 nm with similar trends. The significant peaks around 1 400 and 2 000 nm were attributed to the absorption of water in the soil, while the crests around 2 200 nm were related to the absorptions of clay minerals (Viscarra Rossel & Behrens 2010). The absorbance of the averaged ZSSL spectrum was slightly lower than those of the other three datasets which may be caused by the overall lower SOM content. The score plot of each dataset sample in the spectral feature space of the first two PCs was shown in Figure 4B. The two leading principal components accounted for 97.5% of the total variations. As can be seen, the ZSSL dataset exhibited the greatest spectral variability and covered most parts of the scores of the prediction set and spiking subset. The samples whose scores located around the prediction set and spiking subset were selected as the FS dataset, which had much less variation than the ZSSL dataset.

The optimal number of spiking samples. Figure 5A showed the density distributions of the probability density function for the spiking subset and the prediction set, which were determined by the PC1 of the spectral data. As the spiking samples were randomly selected from the target field, the density distributions between the spiking subset and the prediction set did not exhibit regularity as the sample size increased. This irregularity also occurred in the MSD variations, seen in Figure 5B.

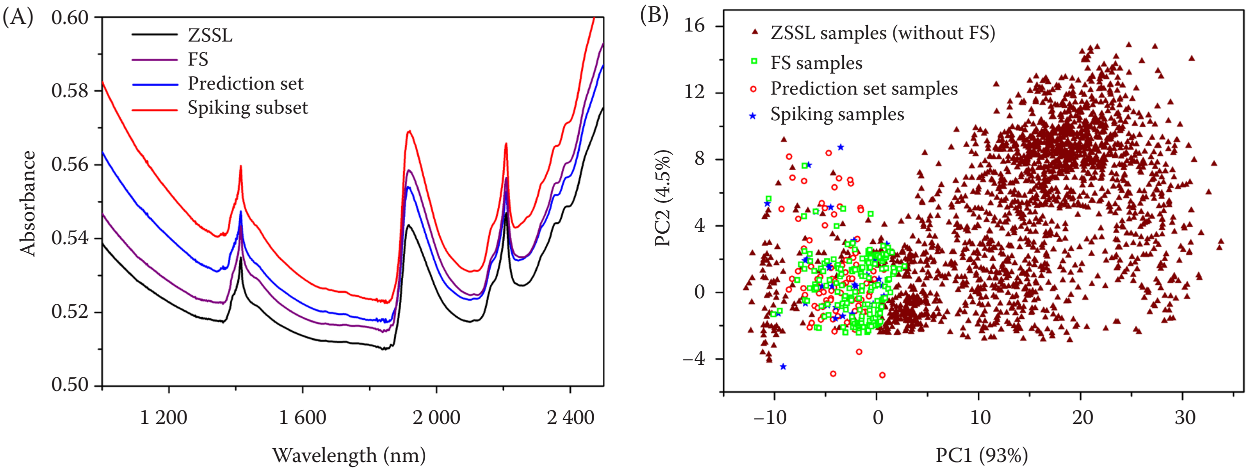

Figure 4. Mean spectra of the different datasets (A) and score plot of the two leading principal components (PC) of the spectra data with points coloured and shaped according to the different datasets (B)

ZSSL – soil spectral library of Zhejiang Province; FS – feature subset

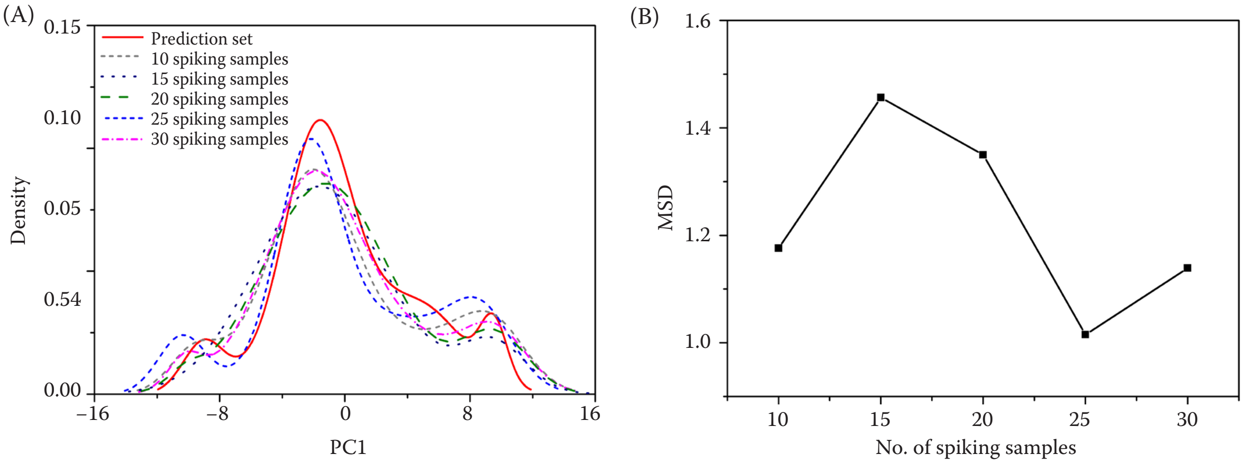

Figure 5. Density distributions for the spiking subsets and the prediction set based on the first principal component (PC) of the spectral data (A) and the mean squared Euclidean distance (MSD) between the density estimates of the spiking subsets and the prediction set (B)

Figure 5A showed that when the number of spiking samples is 25, the density distribution of the spiking subset is most similar to the prediction set. In addition, by calculating the MSD between the density estimates of the spiking subset and the prediction set, it was found that when the number of spiking samples increased from 10 to 30, the MSD reached its minimum value when the number of spiking samples was 25, as shown in Figure 5B. It means that the subset of 25 samples performed better in terms of the replication of the density distribution of the PC1 in the prediction set than the other subsets. In general, a good spiking subset needs to ensure both a good coverage of the predictor space and a good replication of the distribution of the predictor variables (Ramirez-Lopez et al. 2014; Li et al. 2022). Therefore, the spiking subset of 25 samples was considered to be the best representative of the prediction set.

The above method to determine the optimal number of spiking samples by MSD can avoid the chemical analysis of soil samples in advance to determine their SOM values, thus effectively improving the cost-effectiveness of NIR spectral detection.

Evaluation of the established models without spiking. The calibration models provided different prediction results for SOM using the datasets of ZSSL and FS without spiking, listed in Table 3. The performances of the models established based on the dataset of FS were better than those built from the dataset of ZSSL in terms of higher R^2 and lower RMSE values, especially for the PLSR models. This is mainly due to the large differences in spectral characteristics for the ZSSL samples, resulting in the inability to adequately characterize the unique spectral characteristics of SOM in the prediction set. Compared with the ZSSL dataset, the spectral characteristics of the FS dataset were much closer to the prediction set, and the variabilities were smaller than those of ZSSL (seen in Figure 4B), which made it easier for the FS-based models to identify the spectral characteristics of SOM in the prediction set. Araújo et al. (2014) and Shi et al. (2014) obtained similar conclusions. They both considered that it was necessary to select feature dataset from the SSL for modelling, so as to improve the prediction accuracy.

The performance of CNN was better than that of PLSR when using ZSSL for modelling. When the sample size was large, the spectral data presented

Table 3. Comparison of prediction results using different calibration models

Models	*n*	R^2	RMSE	Bias
			(g/kg)	
ZSSL_PLSR	2 069	0.69	4.69	0.55
ZSSL_CNN	2 069	0.73	4.44	0.45
FS_PLSR	174	0.80	3.76	–0.31
FS_CNN	174	0.74	4.31	–0.43

ZSSL – soil spectral library of Zhejiang Province; PLSR – partial least squares regression; CNN – convolutional neural network; FS – feature subset; R^2 – coefficient of determination; RMSE – root mean squared error of prediction in the prediction set

https://doi.org/10.17221/133/2022-SWR

more and more nonlinear information. By adopting different convolution structures and learning rules, CNN can automatically perceive the local information in spectral data and obtain the characteristic information of SOM, avoiding the complex feature extraction and data reconstruction process in PLSR. However, the advantage of CNN on a small number of samples was minimal, as it required lots of data to train the parameters. When the FS dataset was adopted for modelling, the prediction accuracy of CNN was weaker than that of PLSR. The conclusion was supported by Ng et al. (2020). They utilized NIR spectroscopy for the prediction of soil organic carbon and investigated the effect of the training sample size on the prediction accuracy of CNN and PLSR. The results showed that at a lower number of samples (< 1 000), PLSR performed better than CNN. When the sample size exceeded 2 000, the performance of CNN outweighed PLSR. Padarian et al. (2019) obtained a similar conclusion and argued that the efficiency of the CNN model increased with the sample size.

Evaluation of the established models with spiking and extra-weighted. The calibration models provided different prediction results for SOM using the datasets of ZSSL and FS with different numbers of spiking samples, showed in Figure 6. It is worth noting that the spiking samples added to the ZSSL were extra-weighted by a copy of 12, which was approximately equal to the ratio of the number of samples in the ZSSL dataset and the feature set. As can be seen, when the spiking number was greater than 15, the RMSE values of ZSSL_PLSR and FS_PLSR decreased first and then increased slowly. They both achieved the lowest RMSE values when the spiking number was 25. The RMSE values of ZSSL_CNN and FS_CNN decreased gradually along with the increase of the spiking number over the entire range. The indicators of R^2 showed the corresponding trends with RMSE, while the indicators of bias demonstrated similar patterns to RMSE when the spiking number was 15–30.

Compared with the calibration models without spiking (Table 2), the models with spiking samples of more than 20 with and without copies had larger values of R^2, and lower values of RMSE and bias, which showed that the strategies of spiking and extra weighting with copies were efficient. After adding

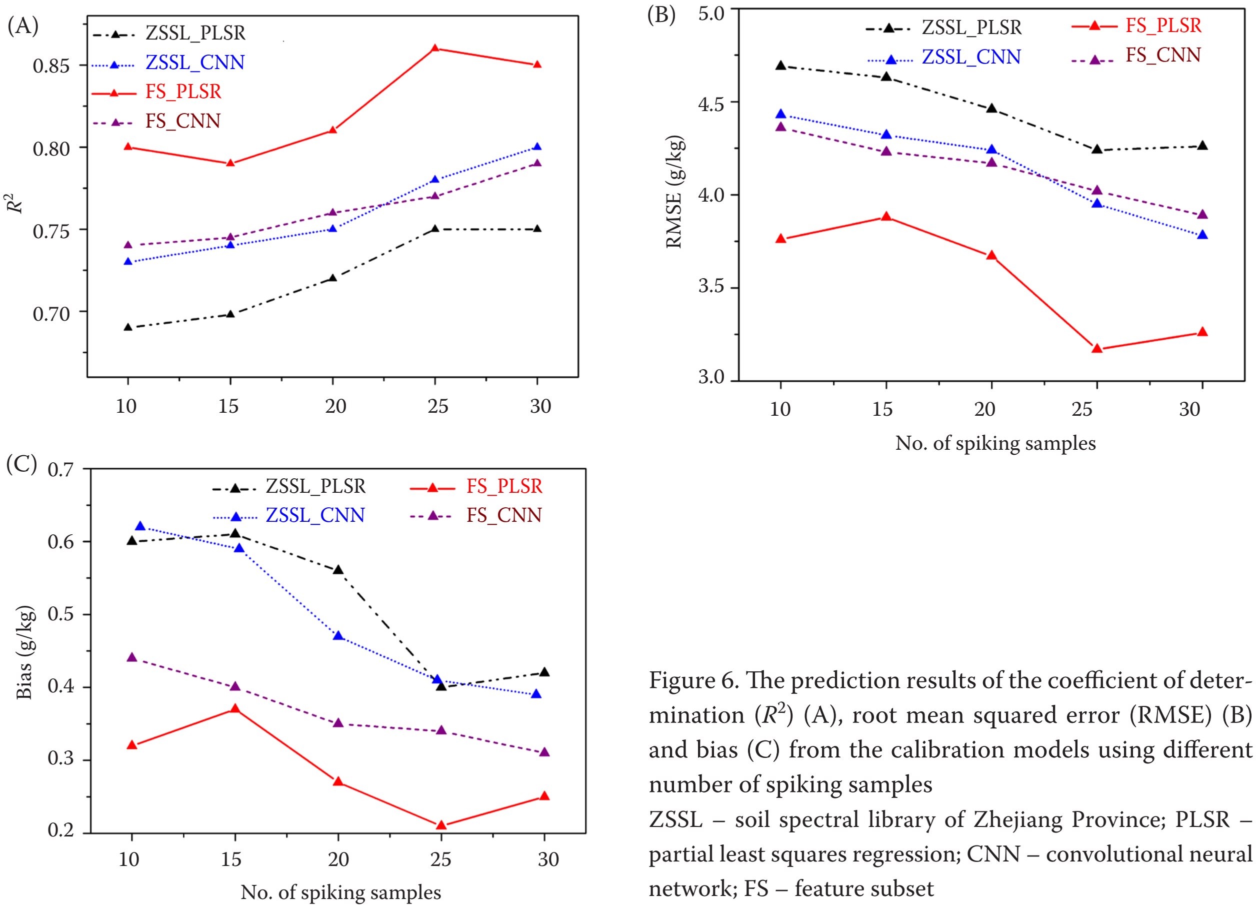

Figure 6. The prediction results of the coefficient of determination (R^2) (A), root mean squared error (RMSE) (B) and bias (C) from the calibration models using different number of spiking samples

ZSSL – soil spectral library of Zhejiang Province; PLSR – partial least squares regression; CNN – convolutional neural network; FS – feature subset the spiking samples, the prediction models can better adapt to the characteristics of soils in the target field, thus improving the prediction ability of the model. Hong et al. (2018) found similar results that the accuracy of the prediction models could be improved after adding the spiking samples of the target area to the original calibration model and re-modelling. Therefore, in the actual work, how to determine the optimal number of spiking samples is very important.

Seidel et al. (2019) adopted the Kennard-Stone algorithm to select up to 30 samples from each of the two target fields as the spiking samples according to the maximum difference principle of Mahalanobis distance of the spectral data. By comparing the model performance under different spiking sample sizes, it was found that an optimal spiking sample size of 15–20 can achieve a good prediction of SOM in the target fields. However, these methods of determining the optimal size of spiking samples by comparing the performance of prediction models under different numbers of spiking samples require the physical and chemical values of all spiking samples. Therefore, when making choices with a large number of spiking samples, it will bring a lot of laboratory chemical analysis and measurement work.

In this study, we adopted the MSD method to determine the optimal number of spiking samples as 25 only through the spectral data of the prediction set and spiking subset. Compared with the calibration models without spiking (Table 3), both the ZSSL_PLSR model and the FS_PLSR model obtained the best prediction results with 25 spiking samples. It indicated that the MSD method can provide some useful help for selecting the optimal number of spiking samples, which can reduce the workload of laboratory analysis relatively, and improve the prediction ability of the model through spiking.

Among all developed models, the FS_PLSR model achieved the best prediction accuracy at each spiking number and outperformed the ZSSL_PLSR model by a large margin. When the number was 25, the optimal prediction results were obtained with R^2 value of 0.86, RMSE value of 3.17 g/kg, and bias value of 0.21 g per kg for SOM. However, the performance of FS_CNN has not been effectively improved compared with ZSSL_CNN, although the spectral characteristics of FS were much closer to the prediction set than those of ZSSL. Furthermore, when the spiking sample size was greater than 25, the ZSSL_CNN model performed better than that of FS_CNN due to larger values of R^2 and smaller values of RMSE. The results indicated that modelling with the FS dataset was effective for improving the performance of the PLSR model, but the impact on the performance of the CNN model was limited.

The RMSE and bias trends of ZSSL_PLSR and FS_PLSR (Figure 6B and C) were similar to that of MSD (Figure 5B), and all of them obtained the lowest values at the spiking number of 25, which showed that 25 was the optimal spiking sample number. The results indicated that the method of MSD was feasible for the determination of the spiking sample size for the ZSSL_PLSR and FS_PLSR models. However, for the ZSSL_CNN and FS_CNN models, the method of MSD was invalid. The larger the spiking sample size, the higher the prediction accuracy of ZSSL_CNN and FS_CNN. The FS dataset selected samples based on the similarity to the spectral characteristics of the prediction set, while MSD determined the spiking sample size based on the similarity to the spectral density distributions of the prediction set. Compared with the CNN-based models, the PLSR-based models can better utilize these spectral similarity information for the prediction of SOM. For CNN, how it works remains largely a mystery, as it is hidden in layers of computation. Large amounts of training data can better reflect the advantages of this method.

CONCLUSION

In this work, the ZSSL dataset and the FS dataset which were chosen from ZSSL based on the similarity to the spectral characteristics of the prediction set were used for the prediction of SOM in the target field. The comparison among different strategies and methods allowed the following conclusions to be drawn.

First, compared with calibration using the ZSSL dataset, calibration using the FS dataset can greatly improve the prediction accuracy of the PLSR-based models, but the impact on the performance of the CNN-based models was limited.

Second, compared with the calibration models without spiking, spiking and extra weighting with copies were efficient for improving the prediction accuracy.

Third, the MSD method was an effective way for determining the optimal spiking sample size for the PLSR-based models only using the spectral data.

Fourth, the performance of CNN was better than that of PLSR when using ZSSL for modelling. When the FS dataset was adopted for modelling, the performance of CNN was weaker than that of PLSR. The advantages of CNN mainly focused on processing

https://doi.org/10.17221/133/2022-SWR

big data. The FS_PLSR model achieved the best prediction performance for SOM compared with the models of ZSSL_PLSR, ZSSL_CNN, and FS_CNN at each spiking number.

Therefore, when using NIR technology to detect SOM in a new target area, it is not necessarily necessary to base on large-scale SSL or conduct a large amount of chemical analysis on the spiking samples of the target area. Instead, it is possible to effectively predict SOM in the new target area based on the FS that is similar to the spectral characteristics of the prediction set and adding the optimal spiking samples determined by the MSD method.

REFERENCES

Araújo S.R., Wetterlind J., Demattê J.A.M., Stenberg B. (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: 718–729.

Davari M., Karimi S.A., Bahrami H.A., Hossaini S.M.T., Fahmideh S. (2021): Simultaneous prediction of several soil properties related to engineering uses based on laboratory Vis-NIR reflectance spectroscopy. Catena, 179: 104987.

Gogé F., Gomez C., Jolivet C., Joffre R. (2014): Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma, 213: 1–9.

Guerrero C., Wetterlind J., Stenberg B., Mouazen A.M., Gabarron-Galeote M.A., Ruiz-Sinoga J.D., Zornoza R., Rossel R.A.V. (2016): Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil & Tillage Research, 155: 501–509.

Havens T.C., Bezdek J.C., Leckie C., Hall L.O., Palaniswami M. (2012): Fuzzy c-means algorithms for very large data. IEEE Transactions on Fuzzy Systems, 20: 1130–1146.

Hong R.C., Liu D.Q., Mo X.Y., He X.N., Zhang H.W. (2022): Learning to compose and reason with language tree structures for visual grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44: 684–696.

Hong Y., Chen Y., Zhang Y., Liu Y., Liu Y., Yu L., Liu Y., Cheng H. (2018): Transferability of Vis-NIR models for soil organic carbon estimation between two study areas by using spiking. Soil Science Society of America Journal, 82: 1231–1242.

Jia S., Li H., Wang Y., Tong R., Li Q. (2016): Recursive variable selection to update near-infrared spectroscopy model for the determination of soil nitrogen and organic carbon. Geoderma, 268: 92–99.

Jiang Q.H., Li Q.X., Wang X.G., Wu Y., Yang X.L., Liu F. (2017): Estimation of soil organic carbon and total nitrogen in different soil layers using VNIR spectroscopy: Effects of spiking on model applicability. Geoderma, 293: 54–63.

Kennard R.W., Stone L.A. (1969): Computer aided design of experiments. Technometrics, 11: 137–148.

Knadel M., Gislum R., Hermansen C., Peng Y., Moldrup P., de Jonge L.W., Greve M.H. (2017): Comparing predictive ability of laser-induced breakdown spectroscopy to visible near-infrared spectroscopy for soil property determination. Biosystems Engineering, 156: 157–172.

Li H., Jia S., Le Z. (2020): Prediction of soil organic carbon in a new target area by near-infrared spectroscopy: Comparison of the effects of spiking in different scale soil spectral libraries. Sensors, 20: 4357.

Li H., Li Y., Yang M., Chen S., Shi Z. (2022): Strategies for efficient estimation of soil organic content at the local scale based on a national spectral database. Land Degradation & Development, 33: 1649–1661.

McKay M.D., Beckman R.J., Conover W.J. (1979): A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21: 239–245.

Munnaf M.A., Guerrero A., Nawar S., Haesaert G., Van Meirvenne M., Mouazen A.M. (2021): A combined data mining approach for on-line prediction of key soil quality indicators by Vis-NIR spectroscopy. Soil & Tillage Research, 205: 104808.

Mustaqeem, Kwon S. (2021): Att-Net: Enhanced emotion recognition system using lightweight self-attention module. Applied Soft Computing, 102: 107101.

Nawar S., Mouazen A.M. (2017): Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. Catena, 151: 118–129.

Ng W., Minasny B., Mendes W.D., Dematt J.A.M. (2020): The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. Soil, 6: 565–578.

Padarian J., Minasny B., McBratney A.B. (2019): Using deep learning to predict soil properties from regional spectral data. Geoderma Regional, 16: e00198.

Ramirez-Lopez L., Schmidt K., Behrens T., van Wesemael B., Dematte J.A.M., Scholten T. (2014): Sampling optimal calibration sets in soil infrared spectroscopy. Geoderma, 226: 140–150.

Seidel M., Hutengs C., Ludwig B., Thiele-Bruhn S., Vohland M. (2019): Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations. Geoderma, 354: 113856.

Shi Z., Wang Q., Peng J., Ji W., Liu H., Li X., Rossel R.A.V. (2014): Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Science China-Earth Sciences, 57: 1671–1680.

Tsakiridis N.L., Tziolas N.V., Theocharis J.B., Zalidis G.C. (2019): A genetic algorithm-based stacking algorithm for predicting soil organic matter from vis-NIR spectral data. European Journal of Soil Science, 70: 578–590.

Tsakiridis N.L., Keramaris K.D., Theocharis J.B., Zalidis G.C. (2020): Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multichannel 1-D convolutional neural network. Geoderma, 367: 114208.

Viscarra Rossel R.A., Behrens T. (2010): Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158: 46–54.

Vohland M., Emmerling C. (2011): Determination of total soil organic C and hot water-extractable C from VIS-NIR soil reflectance with partial least squares regression and spectral feature selection techniques. European Journal of Soil Science, 62: 598–606.

Wang W.L., Ma X.H., Leng L.C., Wang Y.J., Liu B.D., Sun J.F. (2022): A Hybrid CNN based on global reasoning for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 19: 6012605.

Yang J.C., Wang X.L., Wang R.H., Wang H.J. (2020): Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis-NIR spectroscopy. Geoderma, 380: 114616.

Received: September 23, 2022
Accepted: June 14, 2023
Published online: June 29, 2023