Supporting Information for

Title: Evaluating Vis–NIR spectroscopy for laboratory and in-situ prediction of forest soil organic carbon fractions

List of Authors: Jiaxin Li [a, b, d], Pierre Roudier [c], Wenli Fei [a], Lidu Shen [a], Yage Liu [a], Anzhi Wang [a], Sam McNally [d], Yuan Zhang [a, *], Jiabing Wu [a, *]

Institutional affiliations:

a CAS Key Laboratory of Forest Ecology and Silviculture, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China

b University of Chinese Academy of Sciences, Beijing, China

c Bioeconomy Science Institute, Manaaki Whenua – Landcare Research, Palmerston North, New Zealand

d Bioeconomy Science Institute, Manaaki Whenua – Landcare Research group, Lincoln, New Zealand

* *Corresponding author.* E-mail address: zhangyuan@iae.ac.cn; wujb@iae.ac.cn

This word file includes:

Figure S1, S2, and S3

Supplementary Methods S1. Machine learning model specification

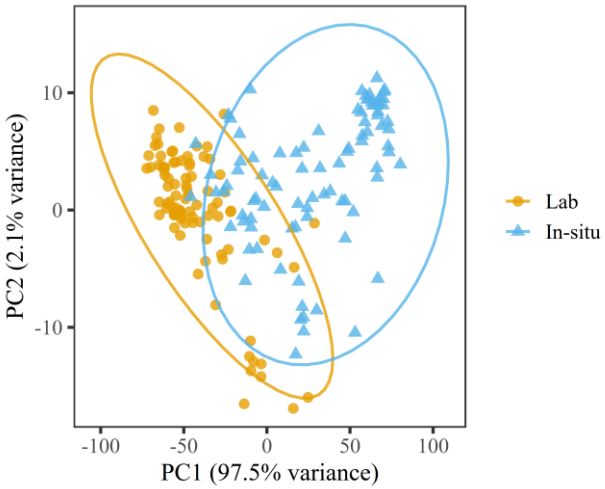

Fig S1. Principal component analysis (PCA) of soil spectra from laboratory (orange) and in-situ (blue) measurements. Training samples are shown as 95% confidence ellipses representing the PCA space defined by laboratory spectra. Test samples are plotted as points, with laboratory (Lab) samples as circles and in-situ samples as triangles.

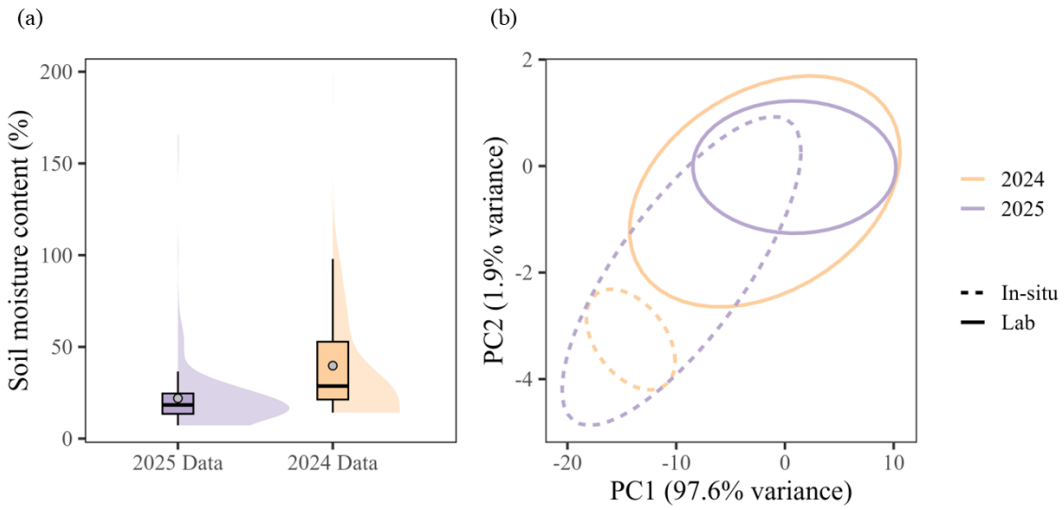

Fig. S2. (a) Distribution of in-situ soil moisture content (SMC) for the 2024 (orange) and 2025 (purple) datasets; (b) principal component analysis (PCA) of laboratory (solid) and in-situ (dashed) soil spectra.

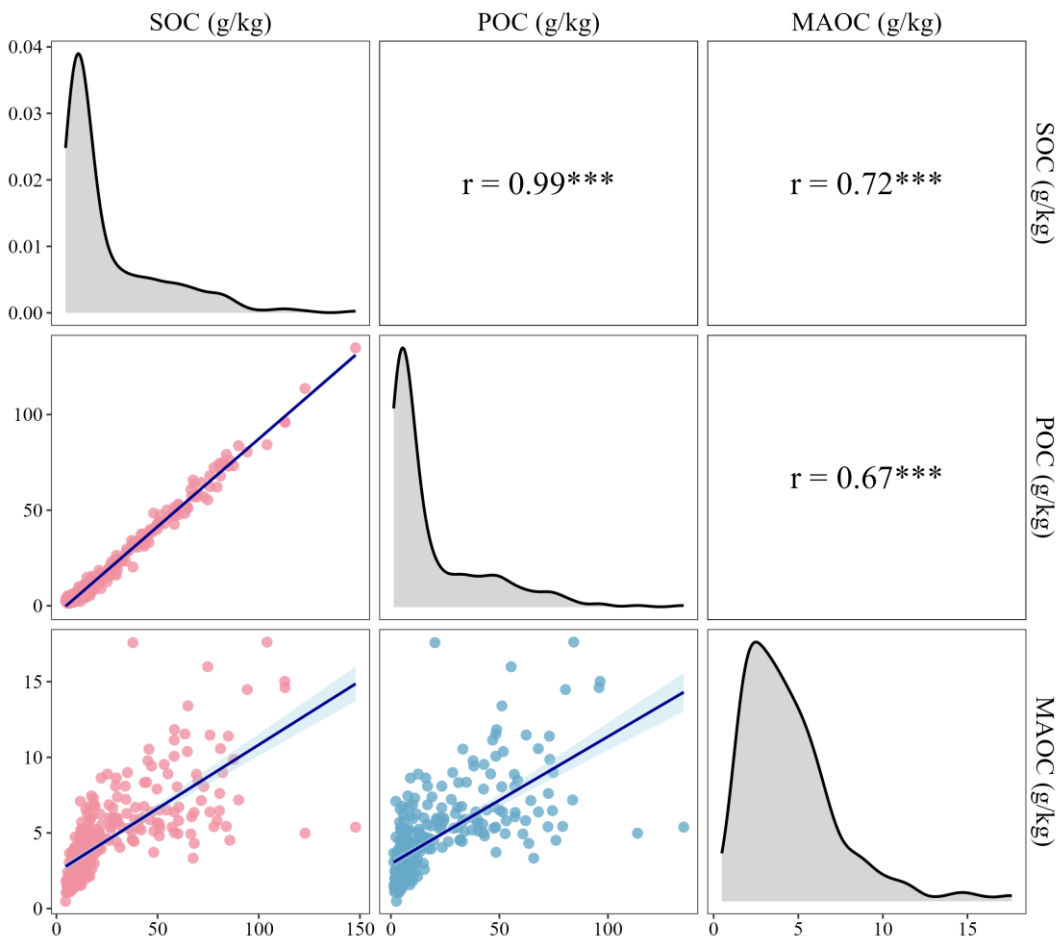

Fig. S3. Pairwise relationships among SOC, POC, and MAOC across all soil samples. Scatterplot matrix showing pairwise relationships among SOC, POC, and MAOC. The lower panels show scatterplots with linear regression lines and 95% confidence intervals. The upper panels display Pearson correlation coefficients (r) with significance levels (* $p < 0.001$). Diagonal panels represent the density distributions of each variable.

Supplementary Methods S1. Machine learning model specifications

S1.1 Principal Component Regression (PCR)

PCR employs Principal Component Analysis (PCA) to reduce the dimensionality of spectral data. This simplifies the data structure and eliminates the influence of multicollinearity among variables (Conforti et al., 2018). When constructing a PCR model, the extracted principal components are taken as independent variables, and the soil properties to be predicted are used as dependent variables. The key factor affecting the accuracy of this model lies in determining the number of principal components. To determine the optimal number of principal components, we use a five-fold cross-validation algorithm and select the appropriate number of principal components with the goal of minimizing the RMSE. In this process, we utilize the 'pls' package (Liland et al., 2024).

S1.2 Partial Least Squares Regression (PLSR)

PLSR as a multivariate statistical method, integrates the characteristics of principal component analysis, canonical correlation analysis, and linear regression (Wold et al., 2001).The difference between PLSR and PCR lies in that when extracting principal components, PLSR takes into account the information of both independent variables and dependent variables simultaneously. It projects the independent variables onto the direction with the highest correlation to the dependent variables to construct a regression model, thus emphasizing the association between them. Its working principle is to extract a set of latent variables from the hyperspectral data matrix (Abdi, 2010). This process can effectively eliminate the collinearity among independent variables. These latent variables not only represent the key features of the hyperspectral data but also demonstrate the highest covariance with the target variable. Subsequently, a linear regression model is established between the latent variables and the target variable to predict the target variable. We use 5-fold cross-validation to determine the minimum RMSE to determine the optimal number of latent variables (ncomp; ranging from 1-10) for spectral data from different pretreatment methods. PLSR was performed using the R package 'pls' (Liland et al., 2024).

S1.3 Lasso Regression

Lasso Regression extends the ordinary least squares method by introducing an additional shrinkage regularization term. This term enables the selection of important wavelength bands based on the data (Tibshirani, 1996). Specifically, the regularization term shrinks the coefficients that are uncorrelated with the target variable towards zero. As a result, the bands with significantly non - zero weights can be identified as relevant bands. In LASSO regression, α is set to a fixed value of 1, which means that LASSO regression is L1 regularization. Therefore, the LASSO model has only one hyper-parameter λ, which controls the strength of L1 regularization and thus affects the complexity of the model. In this study, five-fold cross-validation is used to select the optimal λ. This allows for efficient selection of hyper-parameter and reduces the number of times a model needs to be trained. The 'glmnet' package (Rahman et al., 2022) in R to implement the Lasso regression model.

S1.4 Cubist

Cubist recursively partitions the hyperspectral data matrix using a decision - tree - like structure. Starting from the entire matrix, it divides the matrix into multiple non - overlapping regions according to specific rules, such as the spectral intensity values of key bands. Subsequently, a linear model is fitted for each subset. Due to factors such as soil texture and moisture, the relationship between soil properties and spectral features is complex and non - linear (Breiman et al., 2017). Cubist takes these local variations fully into account by constructing local linear models, enabling the linear models in each region to accurately adapt to the relationship between the two. The performance of the Cubist model is affected by multiple hyper-parameters. Among them, the number of committees determines the number of regression trees, influencing the comprehensiveness of data pattern exploration, computational cost, and sensitivity to overfitting. The number of neighbors determines the number of nearest neighbors when generating piece-wise models, affecting the utilization of local data, model smoothness, robustness to noise, and sensitivity to local fluctuations. In the current study, 5 - fold cross - validation was utilized to determine the optimal values for the number of committees (10, 50) and the number of neighbors (ranging from 1 to 6, incrementing by 1). The cross - validation procedure was exclusively conducted on the calibration set, with the folds being randomly assigned. We utilized the ‘Cubist’ R package (Kuhn and Quinlan, 2025)to implement Cubist.

S1.5 Random Forest (RF)

Random Forest, a widely - applied ensemble learning algorithm, boasts several advantages such as high training efficiency, the ability to handle large - scale datasets, and the natural capacity to evaluate feature importance (Breiman, 2001). Its principle involves constructing multiple decision trees through the bootstrap method of sampling with replacement. The diversity among these trees reduces the model's variance and enhances its generalization ability (Douglas et al., 2018). RF has two hyperparameters, namely mtry (the number of variables per tree) and ntree (the number of trees). The mtry parameter, which influences the randomness and diversity of the trees, may lead to high bias or low diversity when taking inappropriate values. The ntree parameter, which determines the size of the forest, generally allows for improved model performance with an increase in its value, yet improper values of it may result in overfitting (Nawar and Mouazen, 2019). We set mtry as a variable ranging from 1 to 15 and fix ntree at 500. Subsequently, five - fold cross - validation is employed to select the optimal parameters. The ‘randomForest’ package (Liaw, 2002) in R was used for this purpose.

S1.6 Gradient Boosting Decision Trees (GBDT)

GBDT is an ensemble algorithm that combines multiple decision trees to construct a powerful predictive model. Its learning process is based on the residuals of the preceding decision trees. Initially, training samples are assigned initial weights. After the first decision tree is fitted, the error between the predicted values and the true values is calculated. Subsequent decision trees use this error as a new target, assigning higher weights to samples with larger errors to iteratively optimize the predictive capability

(Hastie et al., 2009; Krishnapuram et al., 2016). In terms of specific model settings, the squared error loss function is used as the objective function for regression tasks, with a learning rate set to 0.1, a maximum tree depth of 6, and 2 threads are used. Both the sample and feature sampling ratios are set to 0.8. The evaluation metric is RMSE, with the number of iterations set to 100, and training is set to stop early if there is no improvement in the evaluation metric within 10 rounds. We utilized the 'xgboost' (Chen et al., 2014) in R to accomplish this.

S1.7 Memory-Based Learning (MBL)

MBL is a spectral inversion method (Ramírez-López et al., 2013). It makes use of historical spectral data and corresponding target parameters, retrieving relevant information and conducting inferences through similarity measures (such as distance functions). We employed PLSR to establish a local model. Principal Component Analysis (PCA) was utilized to extract the first two principal components. These components were then used to compute the Mahalanobis distance, which served as the basis for the nearest - neighbor search. The number of nearest neighbors (k) was systematically varied from 50 to 140, with an increment of 10 at each step, to explore the optimal parameter setting for the model.

S1.8 Support Vector Machine Regression (SVMR)

SVMR is a machine learning algorithm based on statistical learning theory. It has excellent generalization ability and performs remarkably well when dealing with small sample data (Vapnik, 2000). Its principle is to find an optimal hyperplane in the feature space, minimizing the distance from all training data points to this hyperplane. SVMR uses a kernel function to map the data from a low-dimensional space to a high-dimensional space, transforming nonlinear relationships into linear relationships in the high-dimensional space. Meanwhile, slack variables are introduced to handle noise and outliers in the data (Smola and Schölkopf, 2004). In this study, we employ a linear kernel function, set the ε value to 0.1, and select different penalty parameter values (specifically 0.001, 0.01, 0.1, and 1). Based on the RMSE, the optimal penalty parameter value is screened out from these parameter values through 10-fold cross-validation. R package ‘e1071’ (Meyer et al., 2024) and ‘caret’ (Kuhn, 2008) was used to implement the SVMR model.

S1.9 Gaussian Process Regression (GPR)

GPR is a non-parametric regression method based on Bayesian statistics and Gaussian process theory (Rasmussen and Williams, 2005). First, the Gaussian process prior described by the mean function and the covariance function is set, the likelihood function is calculated according to the training data, and then the posterior distribution of the objective function is obtained with the help of Bayes' theorem for prediction. The kernel function is the core of the GPR model, which determines how the model captures nonlinear relationships in the data. Due to the smoothness of spectral data, this paper chooses Radial Basis Function (RBF) as the kernel function of the model, which has local characteristics and can effectively balance the similarity between data points. For smooth spectral data, the RBF kernel function can well capture the correlation between adjacent wavelength data points. Improve model performance

by adjusting the scale parameter sigma (range from 0.0005 to 0.1, length = 50) representing the RBF kernel. We used the R package 'kernlab' (Karatzoglou et al., 2004) to implement the GPR model.

S1.10 Artificial Neural Network (ANN)

Artificial neural network is a computational model that simulates the structure and function of biological neural networks (Rossel and Behrens, 2010). It is composed of interconnected neurons and possesses powerful learning ability and adaptability, especially excelling in solving nonlinear problems (McCulloch and Pitts, 1943). Its network structure includes an input layer, hidden layers, and an output layer. In the practical application scenario of soil spectral model inversion, this model can adjust neurons and weights to focus on the characteristic bands related to target variables (such as soil organic carbon content, etc.), thereby improving the prediction accuracy and applicability. The artificial neural network constructed in this study has two hidden layers. The first hidden layer is equipped with 64 neurons, and the second hidden layer is equipped with 32 neurons. Both layers adopt the Rectified Linear Unit (ReLU) (Heaton, 2018) as the activation function. The output layer has only one neuron and uses a linear activation function. To optimize the model training process, the Adam optimizer is used, which can adaptively adjust the learning rate according to the gradient situations of different parameters. In this study, the learning rate is set to 0.001. The entire construction work of the model is completed with the 'TensorFlow' package (Abadi et al., 2016) in Python.

S1.11 Recurrent Neural Network (RNN)

RNN is a type of neural network specifically designed for processing sequential data. Different from traditional feed - forward neural networks, RNN has recurrent connections between its recursive neurons, which enables the output signals to be fed back to the input layer (Yang et al., 2020). By capturing long - term dependencies in sequences, RNN can improve the prediction accuracy of target variables. However, when dealing with long spectral sequences, RNN models may encounter the problems of gradient vanishing or gradient explosion. Therefore, Long Short - Term Memory networks (LSTM) and Gated Recurrent Units (GRU) have been developed to address these issues. In this study, we attempt to use the GRU model set by (Yang et al., 2020).

S1.12 Convolutional Neural Network (CNN)

CNN is a type of deep learning network designed to process grid-structured data. Its network architecture is mainly composed of convolutional layers, pooling layers, and fully connected layers (Lecun et al., 1998). In the field of soil spectral research, soil spectral data is used as the input data. In the convolutional layer, multiple convolutional kernels slide point by point over the spectral data to precisely capture the characteristic bands within the spectral data. The captured feature data will then undergo down sampling in the pooling layer, aiming to reduce the computational load and minimize the risk of overfitting. Finally, the fully connected layer, leveraging its powerful nonlinear mapping ability, conducts in-depth analysis and integration of the previously processed data, and then performs regression analysis tasks (Malek et al., 2018). In this study, the setting of the CNN network architecture mainly

refers to the method of (Haghi et al., 2021). Since the current grid structure, from a theoretical perspective and based on previous practices, is suitable for processing soil spectral data, we have not made any modifications to it.

Reference

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A system for large-scale machine learning. https://doi.org/10.48550/ARXIV.1605.08695

Abdi, H., 2010. Partial least squares regression and projection on latent structure regression (PLS regression). WIREs Computational Statistics 2, 97–106. https://doi.org/10.1002/wics.51

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 2017. Classification and regression trees. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781315139470

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2014. xgboost: Extreme gradient boosting. https://doi.org/10.32614/CRAN.package.xgboost

Conforti, M., Matteucci, G., Buttafuoco, G., 2018. Using laboratory vis-NIR spectroscopy for monitoring some forest soil properties. J Soils Sediments 18, 1009–1019. https://doi.org/10.1007/s11368-017-1766-5

Douglas, R.K., Nawar, S., Alamar, M.C., Mouazen, A.M., Coulon, F., 2018. Rapid prediction of total petroleum hydrocarbons concentration in contaminated soil using vis-NIR spectroscopy and regression techniques. Science of The Total Environment 616–617, 147–155. https://doi.org/10.1016/j.scitotenv.2017.10.323

Haghi, K.R., Pérez-Fernández, E., Robertson, J., 2021. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. Geoderma 396, 115071. https://doi.org/10.1016/j.geoderma.2021.115071

Hastie, T., Tibshirani, R., Friedman, J., 2009. Boosting and additive trees, in: Hastie, T., Tibshirani, R., Friedman, J. (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, pp. 337–387. https://doi.org/10.1007/978-0-387-84858-7_10

Heaton, J., 2018. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning. Genet Program Evolvable Mach 19, 305–307. https://doi.org/10.1007/s10710-017-9314-z

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. **kernlab** - an *S4* package for kernel methods in *R*. J. Stat. Soft. 11. https://doi.org/10.18637/jss.v011.i09

Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (Eds.), 2016. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, CA, USA, august 13-17, 2016. ACM. https://doi.org/10.1145/2939672

Kuhn, M., 2008. Building predictive models in *R* using the **caret** package. J. Stat. Soft. 28. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., Quinlan, R., 2025. Cubist: Rule- and instance-based regression modeling.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791

Liaw, A., 2002. Classification and regression by randomForest. R news.

Liland, K.H., Mevik, B.-H., Wehrens, R., 2024. pls: Partial least squares and principal component regression.

Malek, S., Melgani, F., Bazi, Y., 2018. One-dimensional convolutional neural networks for spectroscopic signal regression. Journal of Chemometrics 32, e2977. https://doi.org/10.1002/cem.2977

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133. https://doi.org/10.1007/BF02478259

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2024. e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien.

Nawar, S., Mouazen, A.M., 2019. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. Soil and Tillage Research 190, 120–127. https://doi.org/10.1016/j.still.2019.03.006

Rahman, S.I., Jamal-Eddine, Z., Xia, Z., Awwad, M., Armitage, R., Rajan, S., 2022. Simulation of GaN-based light emitting diodes incorporating composition fluctuation effects. https://doi.org/10.48550/ARXIV.2211.05704

Ramírez-López, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013. The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma 195–196. https://doi.org/10.1016/j.geoderma.2012.12.014

Rasmussen, C.E., Williams, C.K.I., 2005. Gaussian processes for machine learning.

Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, Diffuse reflectance spectroscopy in soil science and land resource assessment 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) Volume 58, Pages 267-288.

Vapnik, V.N., 2000. The vicinal risk minimization principle and the SVMs, in: Vapnik, V.N. (Ed.), The Nature of Statistical Learning Theory. Springer, New York, NY, pp. 267–290. https://doi.org/10.1007/978-1-4757-3264-1_9

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, PLS Methods 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Yang, J., Wang, X., Wang, R., Wang, H., 2020. Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using vis–NIR spectroscopy. Geoderma 380, 114616. https://doi.org/10.1016/j.geoderma.2020.114616