Sensors and Materials, Vol. 36, No. 11 (2024) 5019–5029
MYU Tokyo

S & M 3847

High-precision Soil Ni Content Prediction Model Using Visible Near-infrared Spectroscopy Coupled with Recurrent Neural Networks

Cheng-Biao Fu,[1] Shuang Cao,[1] and An-Hong Tian[2,1*]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[2]Faculty of Land Resource Engineering, Kunming University of Science and Technology, Kunming 650093, China

(Received June 25, 2024; accepted October 31, 2024)

Keywords: visible near-infrared spectroscopy (Vis–NIR), soil nickel (Ni) content, preprocessing, recurrent neural network

Compared with traditional soil nickel (Ni) content determination methods, visible near-infrared spectroscopy (Vis–NIR) technology can achieve the fast and non-destructive prediction of soil Ni content. However, Vis–NIR spectroscopy data are susceptible to environmental factors during the collection process; thus, it is necessary to perform appropriate preprocessing operations before modeling to improve the data quality and modeling accuracy. In this study, we focus on the polluted farmland around the gold mine in Mojiang Hani Autonomous County, Yunnan Province. First, Savitzky–Golay smoothing was applied to the spectrum (R), and then the impact of using second-order derivative processing (R'') on modeling accuracy was investigated. The potentials of recurrent neural networks (RNNs), random forests (RFs), and partial least squares regression (PLSR) to predict soil Ni content were explored. The results indicated the following: (1) The model established by transforming R with second-order derivatives has shown a clear improvement in prediction accuracy. The use of second-order derivatives helps eliminate the effect of baseline drift on the spectra and also serves to remove noise and amplify differences between features. (2) RNN has the best performance among the three modeling methods, followed by RF and PLSR. Owing to the complex nonlinear relationships between spectral data, RNN has a greater advantage in coping with this situation, and RF has a limited capability to deal with this situation, which PLSR as a linear model does not have. (3) The best model for predicting soil Ni content in this study is R''-RNN, which has high prediction accuracy and generalization ability. Its validation set root mean square error (*RMSE*), coefficient of determination (R^2), relative analysis error (*RPD*), and ratio of performance to interquartile range (*RPIQ*) are 116.81 mg/kg, 0.85, 2.55, and 4.05, respectively. This study provides a new reference approach for monitoring heavy metals in contaminated farmland soil around gold mines.

*Corresponding author: e-mail: tah@kust.edu.cn
https://doi.org/10.18494/SAM5199

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

1. Introduction

Yunnan Province is located in the southwestern region of China, which contains abundant mineral resources. The Mojiang Gold Mine is one of the largest gold mines in Yunnan Province, dating back more than 300 years since the late Ming Dynasty. As a result of long-term mining, there has been some pollution of the surrounding environment, especially heavy-metal soil pollution. Nickel (Ni) is one of the eight major heavy metals causing soil pollution in China. Although Ni is an essential trace element for living organisms, excessive intake of Ni can have negative effects on animals and plants. Ni in soil is not easily transported and degraded, and long-term accumulation will affect the ecological balance of the soil.[1] However, the traditional method of measuring soil Ni content is complex and time-consuming, making it difficult to detect soil Ni content in large areas. Visible near-infrared spectroscopy (Vis–NIR) technology is widely used in agriculture, food, environmental detection, and other fields because of its high efficiency, reliability, and environmental friendliness. In recent years, a large number of scholars have utilized Vis–NIR spectroscopy to predict soil-related elements.[2,3]

However, the measurement of soil spectra is susceptible to natural light and other uncontrollable factors, leading to problems such as large amounts of noise and baseline drift in the spectral data. Appropriate preprocessing methods need to be adopted before modeling to improve data quality. Numerous scholars have proven that selecting relevant preprocessing techniques can improve model prediction accuracy.[4,5] The commonly used spectral preprocessing techniques currently include Savitzky–Golay smoothing, standard normal transformation, and multivariate scattering correction. Second-order derivative transformation is a classic preprocessing method, which can effectively reduce the impact of baseline drift on the spectrum. It also has the function of denoising and amplifying differences between features.[6] However, the potential of second-order derivative transformation for improving the accuracy of soil Ni content prediction models is still unclear.

Partial least squares regression (PLSR) as a classical linear regression model is still chosen by many scholars.[7] With the advancement of science and technology, PLSR is gradually replaced by more complex models such as support vector machine regression and random forest (RF) in machine learning. These models can capture the nonlinear relationship between data and better fit the data, thereby improving the prediction accuracy of the model.[8,9] The emergence of deep learning has driven science and technology further. The convolutional neural network (CNN), deep belief network (DBN), and recurrent neural network (RNN) are common deep learning algorithms. Deep learning algorithms have a complex network structure. They enable the automatic extraction of data features and better capture the intrinsic structure and patterns of the data by abstracting the features. Deep learning models typically have the characteristics of high accuracy and high generalization ability.[10,11] Recently, some scholars have used Vis–NIR spectroscopy and deep learning algorithms to predict soil properties. For example, Bai *et al.* used four deep learning algorithms, namely, 1D-CNN, 2D-CNN, long short-term memory network (LSTM), and DBN, to combine five feature selection algorithms to predict the content of soil inorganic carbon.[12] Among the four modeling methods, LSTM performed the best, with an R^2 of 0.92 for the optimal model validation set.[12] However, there are few reports on deep learning algorithms for soil monitoring in Yunnan Province.

Overall, the goal of this study is to seek a high-precision soil Ni content prediction model. In the following, the impacts of different preprocessing and modeling methods on model prediction accuracy are discussed and analyzed.

2. Materials and Methods

2.1 Study area overview and data collection

The research area is located in Mojiang Hani Autonomous County, Yunnan Province. Mojiang County is located at the southwest edge of the Yunnan-Guizhou Plateau, with a total area of 5312 square kilometers. The Tropic of Cancer passes through this area. The territory of Mojiang County is mountainous and contains rich mining resources. Here, the rainfall is abundant, with a clear distinction between dry and wet seasons. Owing to the mountainous terrain, the farmland within the area is mainly terraced, and the soil is mostly acidic. The main crops include rice, corn, and tea. The distribution of sampling points in the study area is shown in Fig. 1.

The contaminated farmland around the gold mine in Mojiang Hani Autonomous County, Yunnan Province was used as the research site. A total of 122 samples were collected from the farmland near the gold mine from February 11 to 15, 2022. We collected soil samples from 0 to 20 cm below the soil surface, with each sample weighing approximately 1 kg, and recorded the specific location information of the samples during collection. The samples were brought back to the laboratory, and dried naturally to remove the effect of moisture on the spectra. The samples were ground and sieved through a 2 mm fine sieve and divided into two equal parts. One part was used to measure the Vis–NIR spectroscopy reflectance, and the other part was used to determine the soil Ni content. Vis–NIR spectroscopy was conducted using an ASD

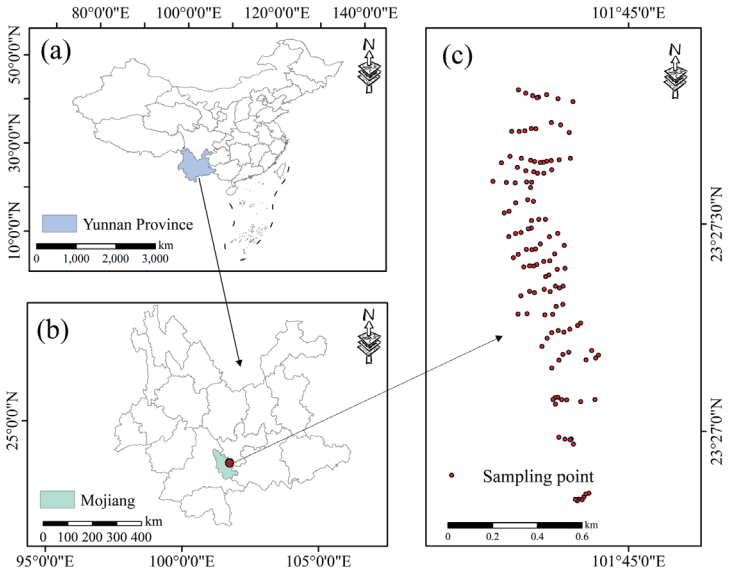

Fig. 1. (Color online) Research sampling area.

FieldSpec® 3 portable ground spectrometer (350–2500 nm). To avoid the effect of natural light, the spectrum measurement experiment was conducted in a dark room. Each sample was measured three times, and the average value was finally taken as the spectral reflectance of the sample.

2.2 Spectral preprocessing and dataset partitioning

The entire band of the original spectrum was subjected to Savitzky–Golay (S–G) second-order 15-point smoothing, and then second-order derivative transformation was performed.[13] Min-max normalization was performed on the observed values, and the data were denormalized after the model prediction was completed.

The dataset was divided into the calibration set (67%, 82 samples) and validation set (33%, 40 samples) according to the concentration ranking scheme. The specific steps for dividing the dataset are the following: sort the samples in ascending order of the Ni content, and every three samples from a group. The first two samples in a group are assigned to the calibration set, and the remaining one is assigned to the validation set. This partitioning method can evenly partition the dataset, making the distribution of the calibration set and validation set consistent.

2.3 Modeling methodology

RNN is a special type of neural network and has the function of short-term memory. Its hidden layer is a cyclic structure, and the output at the current moment will affect the input at the future moment. Specifically, the input of the hidden layer at the current moment includes not only the output of the input layer at the previous moment, but also the output of the hidden layer at the previous moment. Thus, RNN can learn long-term dependences between data.[14] Meanwhile, RNN can achieve end-to-end learning. The structure diagram of RNN is shown in Fig. 2.

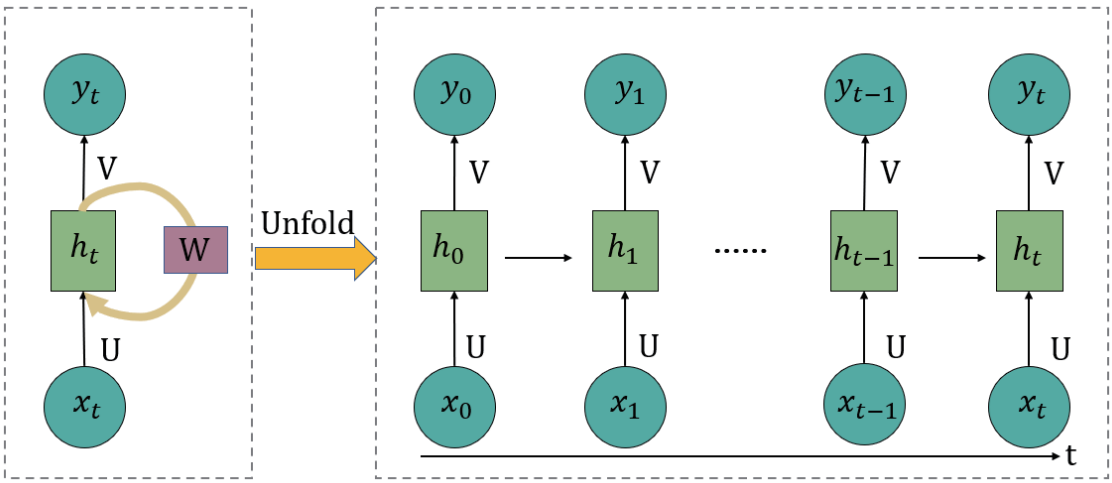

Fig. 2. (Color online) Structure of RNN.

RF is a type of ensemble learning that consists of multiple decision trees, each of which works independently. RF considers all the decisions of the decision tree to obtain the final prediction result using certain rules, which largely reduces the risk of overfitting. Owing to the simultaneous operation of multiple decision trees, it considerably shortens the time for model training and prediction, so RF is often used to process high-dimensional data.[15]

PLSR is a classic linear model that exhibits strong predictive performance when there are multicollinearity relationships between data. It uses the idea of principal component analysis to reduce the dimensionality of data, overcoming the challenge of multivariate collinearity by creating new low-dimensional representations of independent and dependent variables. On these new dimensions, PLSR seeks models to reduce the covariance between predictor and response variables, thereby improving the predictive ability and stability of the model.[16]

2.4 Model evaluation

In this study, we comprehensively evaluated the performance of a model using the root mean square error (*RMSE*), the coefficient of determination (R^2), the relative analysis error (*RPD*), and the ratio of performance to the interquartile range (*RPIQ*).[17] *RMSE* measures the gap between the predicted and observed values of the model; the smaller the *RMSE* value, the more accurate the prediction of the model. R^2 describes the correlation between the predicted values and the observed values. The value of R^2 is between 0 and 1, and the larger the R^2 value, the better the performance of the model. *RPD* and *RPIQ* are used to evaluate the prediction accuracy of the model. When *RPD* is above 2.0 and *RPIQ* is above 2.2, it suggests that the model has excellent predictive ability. When the *RPD* is between 1.8 and 2.0, it shows that the model has good predictive ability. When the *RPD* is between 1.4 and 1.8 and the *RPIQ* is between 1.7 and 2.2, it demonstrates that the model has average predictive ability. When *RPD* is less than 1.4 and *RPIQ* is less than 1.7, it indicates that the model is unable to make predictions. The formulas for the four evaluation indexes are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}, \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2}, \tag{2}$$

$$RPD = \frac{SD}{RMSE}, \tag{3}$$

$$RPIQ = \frac{Q3 - Q1}{RMSE}, \tag{4}$$

where $\widehat{y_i}$ is the predicted value, $\overline{y_i}$ is the mean of the observed value, y_i is the observed value, n is the number of samples, SD is the standard deviation of the predicted value, $Q1$ is the lower quartile of the predicted value, and $Q3$ is the upper quartile of the predicted value.

3. Results and Discussion

3.1 Modeling effects of the original spectra and the transformed second-order derivatives

Six quantitative prediction models for soil Ni content were constructed using RNN, RF, and PLSR, with spectral R and spectral R'' transformed by the second-order derivative as independent variables and soil Ni content as the dependent variable. The performance of each model is shown in Table 1. It can be seen that among the three models established using R, only RNN showed good predictive ability (validation *RPD* greater than 1.8). The *RMSE* of the RF and PLSR validation sets is above 200 mg/kg, R^2 is around 0.5, *RPD* is around 1.4, and *RPIQ* is around 2.2. This indicates that the stability and prediction accuracy of R-RF and R-PLSR are relatively low, and they do not qualify as quantitative prediction models. The prediction accuracy of the model built after R was transformed by the second-order derivative has been clearly improved. Specifically, all models based on R'' qualify as predictive models. Among them, R''-RNN has an excellent ability to predict Ni, and R''-RF and R''-PLSR can also predict the soil Ni content well. Compared with the model based on R, the *RMSE* of the validation set of the model based on R'' is within the range of 116.81–163.41 mg/kg, illustrating that the error between the observed and predicted values is reduced. The R^2 values are all above 0.7, indicating that the ability of the model to fit the data has been improved. The values of *RPD* and *RPIQ* have also been improved by one level, further demonstrating that the model prediction accuracy has been clearly improved. This is attributed to several benefits of using the second-order derivative transformation for spectral data: first, this helps eliminate baseline drift in the spectra, thus improving the quality and reliability of the spectral data. Second, second-order derivative processing can highlight feature peaks and valleys, that is, amplify the key features, based on which the model can show better prediction ability. Finally, second-order derivative processing removes a portion of noise, which reduces the effect of noise on the spectrum and makes the spectral data more suitable for quantitative modeling.[18]

Table 1
Performance of each quantitative prediction model for soil Ni content.

Model	Calibration			Validation			
	RMSE (mg/kg)	R^2	*RPD*	*RMSE* (mg/kg)	R^2	*RPD*	*RPIQ*
R-RNN	141.20	0.79	2.18	155.26	0.73	1.92	3.04
R-RF	156.55	0.74	1.97	209.02	0.51	1.43	2.26
R-PLSR	200.94	0.57	1.53	218.92	0.46	1.36	2.16
R''-RNN	77.41	0.94	3.98	116.81	0.85	2.55	4.05
R''-RF	135.56	0.81	2.27	155.10	0.73	1.92	3.05
R''-PLSR	164.52	0.71	1.87	163.41	0.70	1.82	2.89

3.2 Predictive performance of different modeling approaches

From the perspective of different modeling methods, RNN performs the best among the three models under R and R''. RNN can also predict Ni well by using spectrum R as an input variable without second-order derivative processing, and the prediction accuracy is even higher than that of R''-PLSR. Among all soil Ni content prediction models, R''-RNN has the highest prediction accuracy, with validation sets of *RMSE*, R^2, *RPD*, and *RPIQ* of 116.81 mg/kg, 0.85, 2.55, and 4.05, respectively. First, this is facilitated by the fact that RNN is good at processing time series data, which means that RNN can combine the context of the data to fully mine the information of minute details in the data. Usually, it is this detailed information that determines the prediction quality of a model. Second, the hidden layer in RNN is a cyclic structure. RNN can remember the output of the data at the previous time and affect the output at future times, allowing RNN to learn the long-term dependences present in the data and better fit the observed values during prediction. Third, as a deep learning algorithm, RNN can achieve end-to-end learning without the need for additional feature extraction operations. This is also an important reason why deep learning algorithms are generally superior to machine learning algorithms and linear models. RNN also has another advantage of deep learning algorithms: a keen ability to capture complex nonlinear relationships between data.[19]

The performance of RF is moderate. Since RF is composed of multiple decision trees, it has a natural advantage in dealing with high-dimensional data such as Vis–NIR spectroscopy data. As a type of ensemble learning, RF considers the judgment of multiple decision trees, which reduces the risk of data overfitting and improves the model's generalization ability. In addition, RF divides the data into multiple subsets and constructs decision tree analysis for each subset, resulting in feature selection operations that are not necessary for RF.[20] This is also the reason why RF was chosen in machine learning algorithms in this study. However, RF has limited ability in handling complex nonlinear relationships in data. From the above facts, RF achieved moderate predictive performance in this study.

The worst performance of PLSR is due to the complex nonlinear relationships between hyperspectral data, whereas PLSR is only good at dealing with multicollinear relationships between the data. In addition, PLSR does not have feature extraction capabilities, and facing such a large number of data, PLSR cannot fully learn them, resulting in unsatisfactory modeling results. In this study, the performances of the three modeling methods were ranked in the following order: RNN > RF > PLSR. This gradient relationship can be clearly and intuitively seen in Fig. 3.

To more intuitively demonstrate the predictive performance of each model, scatter plots of predicted and observed values for six model calibration and validation sets were drawn (Fig. 4). The red line in the figure represents the fitting line of the predicted value points in the validation set. The closer the fitting line is to the 1:1 line, the better the predictive performance of the model. Meanwhile, the more points in the graph are clustered on the 1:1 line, the smaller the error between the predicted value and the observed value. As can be seen from the left side of Fig. 4, the modeling effect based on R is not ideal. The points of the three models are relatively scattered, and the angle between the fitting line and the 1:1 line is relatively large, implying that

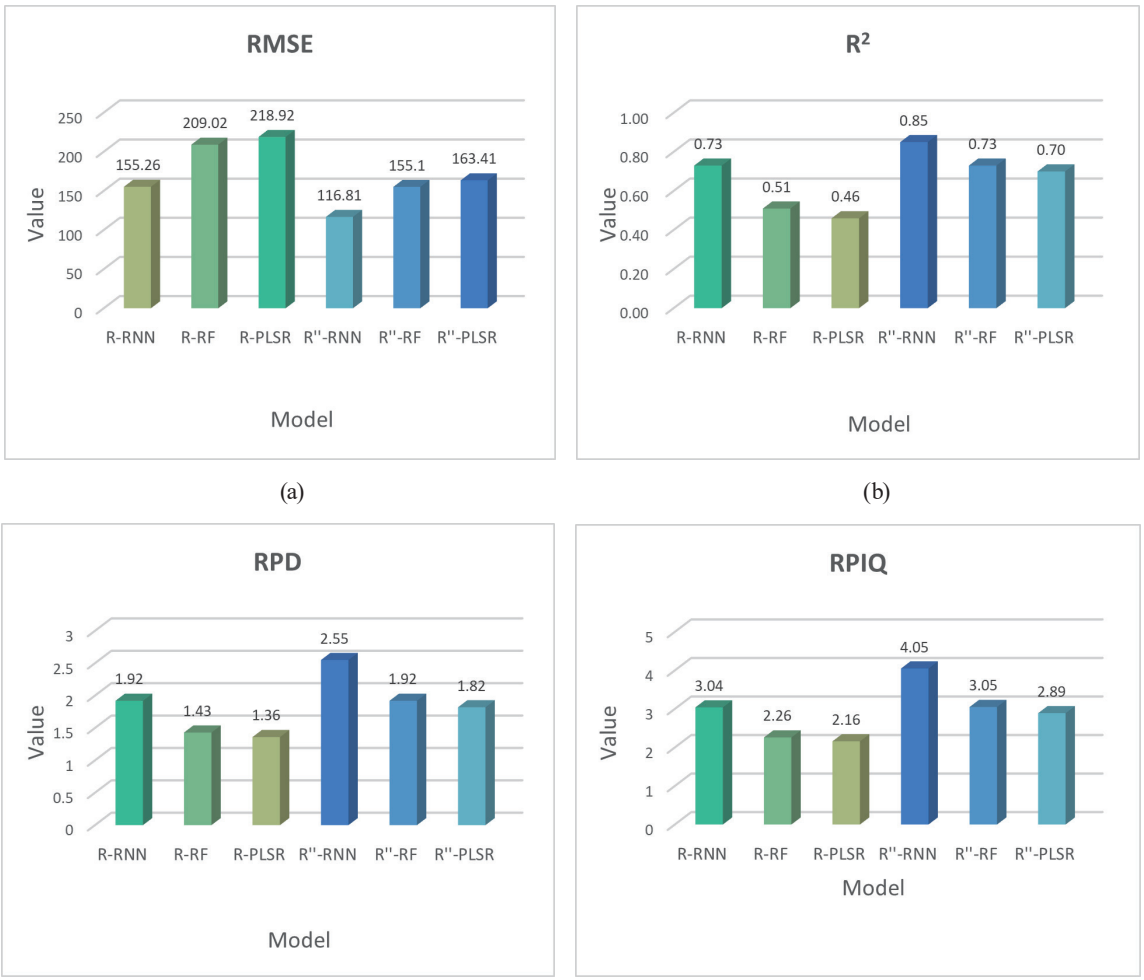

Fig. 3. (Color online) Quantitative prediction models for soil Ni content: (a) *RMSE* value, (b) R^2 value, (c) *RPD* value, and (d) *RPIQ* value.

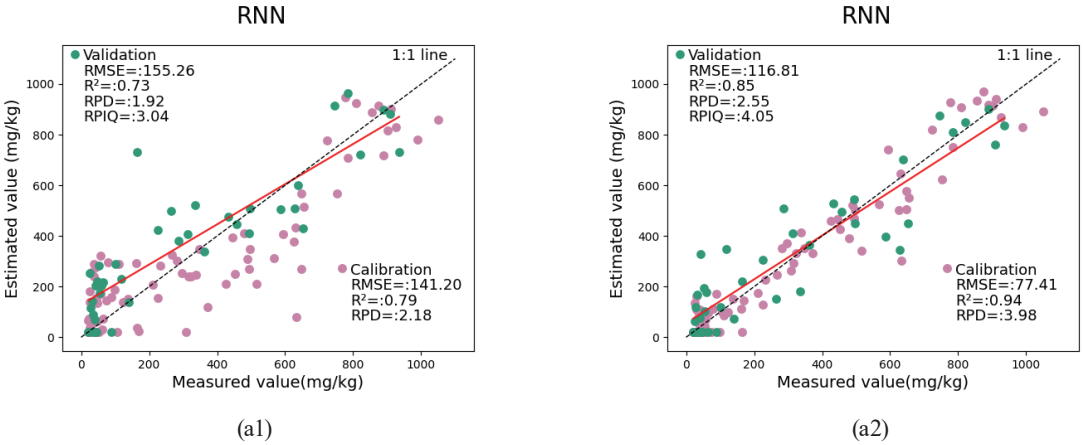

Fig. 4. (Color online) Scatter plots of Ni content predicted by the models: (a1) R-RNN, (a2) R''-RNN, (b1) R-RF, (b2) R''-RF, (c1) R-PLSR, and (c2) R''-PLSR.

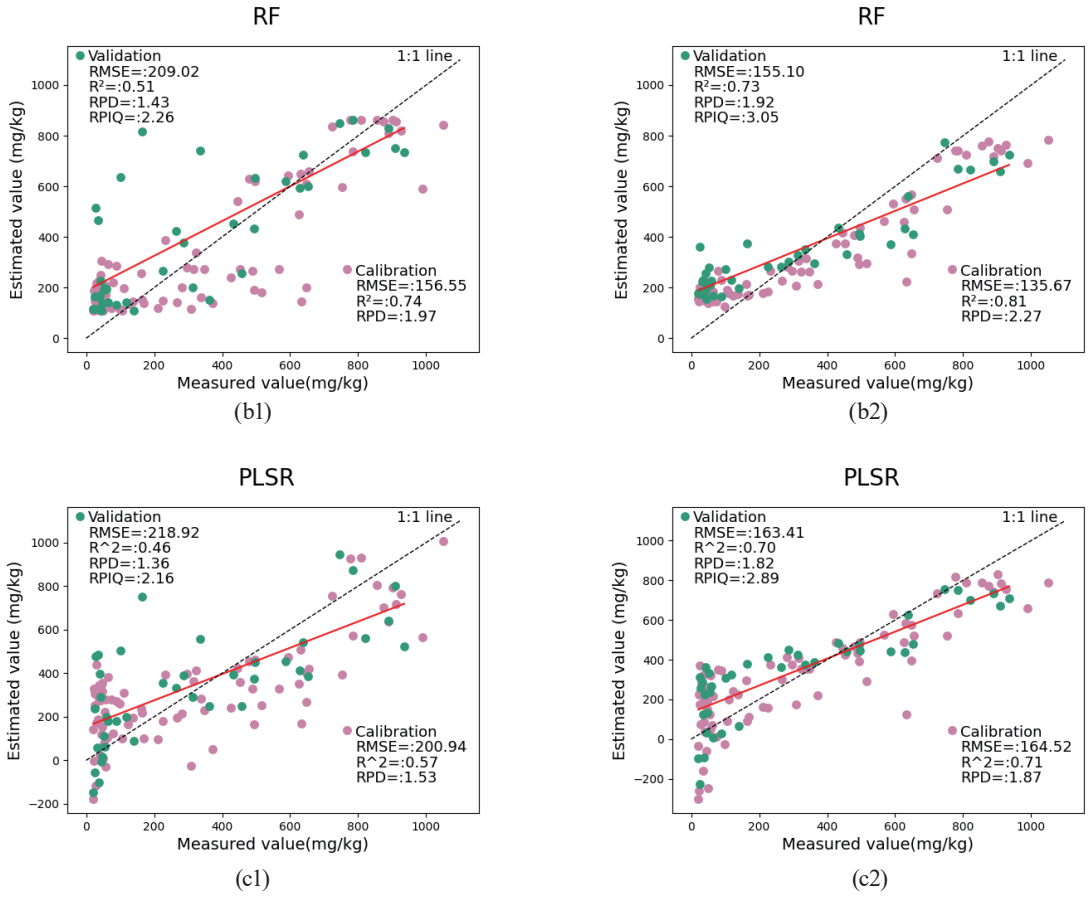

Fig. 4. (Color online) (Continued) Scatter plots of Ni content predicted by the models: (a1) R-RNN, (a2) R''-RNN, (b1) R-RF, (b2) R''-RF, (c1) R-PLSR, and (c2) R''-PLSR.

the prediction accuracy of the model is low. The prediction accuracy of the model built based on R'' is clearly improved (right panels of Fig. 4), and the points in the panels are more clustered on a 1:1 line. Among all the models, only the predicted values of the PLSR model have negative values. This is because PLSR cannot learn the nonlinear relationship between hyperspectral data, resulting in low prediction accuracy and generalization ability of the model, and even negative values.

4. Conclusions

In this study, 122 soil Vis–NIR spectroscopy data were first processed using Savitzky–Golay smoothing, and then the effect of second-order derivative processing on predicting soil Ni content was analyzed. Second-order derivative processing can highlight important features in the spectrum while reducing the impact of noise and baseline drift, thus improving the quality and reliability of spectral data. Subsequently, the prediction accuracies of different modeling methods were compared, and the performances of the three modeling methods were ranked in

the following order: RNN > RF > PLSR. In addition to learning long-term dependences in data, RNN can also sensitively capture nonlinear relationships between data and better fit observed values in prediction. Finally, we developed a high-precision and stable soil Ni content prediction model (R''-RNN) and provided a new approach for the large-scale detection of soil elements.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (42067029,42361007), the Reserve Talents Project for Young and Middle-aged Academic and Technical Leaders in Yunnan Province (202205AC160005), and the Yunnan Province "Xing Dian Ying Talent Support Programme" Young Talents Project (KKXX202303001).

References

1 M. Rizwan, K. Usman, and M. Alsafran: Chemosphere **357** (2024) 142028. https://doi.org/10.1016/j.chemosphere.2024.142028
2 D. Zhao, M. Arshad, and J. Wang, J. Triantafilis: Comput. Electron. Agric. **182** (2021) 105990. https://doi.org/10.1016/j.compag.2021.105990
3 Z. Zhang, J. Ding, J. Wang, and X. Ge: Catena **185** (2020) 104257. https://doi.org/10.1016/j.catena.2019.104257
4 W. Yang, Y. Xiong, Z. Xu, L. Li, and Y. Du: Infrared Phys. Technol. **126** (2022) 104359. https://doi.org/10.1016/j.infrared.2022.104359
5 R. Khodabakhshian, B. Emadi, M. Khojastehpour, M. R. Golzarian, and A. Sazgarnia: Int. J. Food Prop. **20** (2017) 4152. https://doi.org/10.1080/10942912.2015.1126725
6 X. Xu, S. Chen, Z. Xu, Y. Yu, S. Zhang, and R. Dai: Remote Sens. **12** (2020) 3765. https://doi.org/10.3390/rs12223765
7 E. Afriyie, A. Verdoodt, and A. M. Mouazen: Soil Tillage Res. **215** (2022) 105218. https://doi.org/10.1016/j.still.2021.105218
8 J. Ssali Nantongo, E. Serunkuma, G. Burgos, M. Nakitto, F. Davrieux, and R. Ssali: Spectrochim. Acta, Part A (2024) 124406. https://doi.org/10.1016/j.saa.2024.124406
9 F. Nyarko, F. M. G. Tack, and A. M. Mouazen: Sci. Total Environ. **841** (2022) 156582. https://doi.org/10.1016/j.scitotenv.2022.156582
10 J. Pyo, S. M. Hong, Y. S. Kwon, M. S. Kim, and K. H. Cho: Sci. Total Environ. **741** (2020) 140162. https://doi.org/10.1016/j.scitotenv.2020.140162
11 D. Xiao, Q. H. Vu, and B. T. Le: Microchem. J. **165** (2021) 106182. https://doi.org/10.1016/j.microc.2021.106182
12 Z. Bai, S. Chen, Y. Hong, B. Hu, D. Luo, J. Peng, and Z. Shi: Geoderma **437** (2023) 116589. https://doi.org/10.1016/j.geoderma.2023.116589
13 Z. Zhang, J. Ding, C. Zhu, and J. Wang: Spectrochim. Acta, Part A **240** (2020) 118553. https://doi.org/10.1016/j.saa.2020.118553
14 M. Zhao, H. Cang, H. Chen, C. Zhang, T. Yan, Y. Zhang, P. Gao, and W. Xu: LWT **183** (2023) 114861. https://doi.org/10.1016/j.lwt.2023.114861
15 J. Fernández-Habas, M. Carriere Cañada, A. M. García Moreno, J. R. Leal-Murillo, M. P. González-Dugo, B. Abellanas Oar, P. J. Gómez-Giráldez, and P. Fernández-Rebollo: Comput. Electron. Agric. **192** (2022) 106614. https://doi.org/10.1016/j.compag.2021.106614
16 A. Tian, J. Zhao, C. Fu, and H. Xiong: Spectrochim. Acta, Part A **282** (2022) 121647. https://doi.org/10.1016/j.saa.2022.121647
17 Y. Wang, S. Chen, Y. Hong, B. Hu, J. Peng, and Z. Shi: Comput. Electron. Agric. **212** (2023) 108067. https://doi.org/10.1016/j.compag.2023.108067
18 V. Khosravi, F. Doulati Ardejani, S. Yousefi, and A. Aryafar: Geoderma **318** (2018) 2941. https://doi.org/10.1016/j.geoderma.2017.12.025
19 Z. Pang, F. Niu, and Z. O'Neill: Renewable Energy **156** (2020) 279289. https://doi.org/10.1016/j.renene.2020.04.042
20 F. B. De Santana, A. M. De Souza, and R. J. Poppi: Spectrochim. Acta, Part A **191** (2018) 454462. https://doi.org/10.1016/j.saa.2017.10.052

About the Authors

Cheng-Biao Fu received his B.S. and M.S. degrees from Chongqing University of Posts and Telecommunications, China, in 2005 and 2009, respectively. He received his Ph.D. degree from Kunming University of Science and Technology, China, in 2021. Since 2023, he has been an associate professor at Kunming University of Science and Technology. His research interests are in information identification and processing, hyperspectral remote sensing, and machine learning. (fcb@kust.edu.cn)

Shuang Cao is currently a graduate student at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, focusing on hyperspectral remote sensing and deep learning. (20232204158@stu.kust.edu.cn)

An-Hong Tian received her B.S. and M.S. degrees from Chongqing University of Posts and Telecommunications, China, in 2007 and 2010, respectively. She received her Ph.D. degree from Kunming University of Science and Technology, China, in 2020. Since 2023, she has been a professor at Kunming University of Science and Technology. Her research interests are in hyperspectral remote sensing, nonlinear systems, and artificial intelligence. (tah@kust.edu.cn)