ISSN 1807-1929
Revista Brasileira de Engenharia Agrícola e Ambiental
Brazilian Journal of Agricultural and Environmental Engineering
v.29, n.6, e287460, 2025
Campina Grande, PB – http://www.agriambi.com.br – http://www.scielo.br/rbeaa

DOI: http://dx.doi.org/10.1590/1807-1929/agriambi.v29n6e287460

Original Article

Estimation of soil water content based on simulated multi-spectral broadband reflectance and machine learning[1]

Estimativa do conteúdo de água do solo com base na reflectância simulada de banda larga multiespectral e aprendizado de máquina

Shaomin Chen[2], Zhilin Xu[2], Qiuyu Pu[2], Fangchuan Lou[2], Jiachen Gao[2], Shuai Tan[2*], Chao Gao[3*] & Xin Shen[4]

[1] Research developed at Kunming University of Science and Technology, Yunnan, Kunming, China
[2] Kunming University of Science and Technology/Faculty of Modern Agricultural Engineering, Yunnan, Kunming, China
[3] Henan Academy of Sciences/Institute of Geographical Sciences, Henan, Zhengzhou, China
[4] HuiJie Design Group co., ltd., Hunan, Changsha, China

HIGHLIGHTS:

Out-of-bag (OOB) importance analysis was applied to extract feature variables.
The nonlinear model random forest (RF) outperforms the linear model for soil water content retrieval.
The coupled OOB and RF model was demonstrated to be a reliable method for soil water content prediction.

ABSTRACT: The soil water content is a key evaluation factor in the agricultural and ecological fields. The objective of this study was to explore the combination of multi-satellite information to address the issue of low accuracy in soil water content retrieval from single-band, single-source information and to establish a model for the estimation of the soil water content based on a simulated multi-spectral method. An experiment involving the hyperspectral determination of soil samples (Oxisols) with varying values of the water content was conducted. The reflectance of eight multi-spectral satellite sensors was resampled based on the spectral response functions. The vegetation indices (VIs) were established by pairing all available sensor bands. The significant VIs and band reflectance were extracted using the correlation coefficient and out-of-bag (OOB) data importance analysis methods, and then linear models and nonlinear models were established for soil water content estimation. A significant correlation was achieved between the simulated multi-spectral reflectance, VIs, and soil water content. The nonlinear models had a better performance than the linear model. The combined OOB and random forest (OOB-RF) model achieved the highest prediction accuracy, with $R^2_{calibration}$ and $R^2_{prediction}$ values of 0.852 and 0.834, respectively. Overall, it was verified that the OOB-RF modeling method based on multi-spectral remote sensing was feasible for estimating the soil water content.

Key words: moisture state, broad spectrum, remote sensing, vegetation index, random forest algorithm

RESUMO: O conteúdo de água no solo é um fator-chave na avaliação nas áreas agrícola e ecológica. O objetivo deste trabalho foi explorar a combinação de informações multisatélites para abordar a questão da baixa precisão na recuperação do conteúdo de água do solo a partir de informações de banda única e fonte única, e estabelecer um modelo para estimativa do conteúdo de água do solo baseado no método multiespectral simulado. Foi realizado um experimento de determinação hiperespectral de amostras de solo (Latossolos) com valores variáveis de teor de água. A reflectância de oito sensores de satélite multiespectral foi remapeada com base nas funções de resposta espectral. Os índices de vegetação (VIs) foram estabelecidos combinando todas as bandas de sensores disponíveis. Os VIs significativos e a reflectância da banda foram extraídos pelos métodos de coeficiente de correlação e análise de importância de dados out-of-bag (OOB), e em seguida modelos lineares e não lineares foram estabelecidos para a estimativa do conteúdo de água no solo. Correlação significativa foi obtida entre refletância multiespectral simulada, VIs e teor de água no solo. Os modelos não lineares apresentaram melhor desempenho que o modelo linear. A combinação do modelo OOB e Random Forest (OOB-RF) obteve a maior acurácia de predição, com $R^2_{calibração}$ e $R^2_{predição}$ de 0,852 e 0,834, respectivamente. De forma geral, verificou-se que o método de modelagem OOB-RF baseado em sensoriamento remoto multiespectral foi viável para estimar o teor de água no solo.

Palavras-chave: estado de umidade, amplo espectro, sensoriamento remoto, índice de vegetação, algoritmo random forest

• Ref. 287460 – Received 11 Jun, 2024
* Corresponding author - E-mail: tans90@163.com; supergao2009@126.com
• Accepted 20 Nov, 2024 • Published 16 Dec, 2024
Editors: Toshik Iarley da Silva & Walter Esfrain Pereira

Introduction

The soil water content is a key indicator in many fields, such as farmland irrigation management, regional hydrological situation analysis, and watershed water balance calculation. In response to challenges such as the complexity, limited scale, and long-term field observation inherent in ground-based sampling or sensor-contact methods (Khanal et al., 2020), remote sensing technology is increasingly employed in soil moisture monitoring research due to its advantages of speed, non-invasiveness, and large-scale coverage (Li et al., 2021).

The quantitative estimation of the soil water content has been achieved by combining hyperspectral data with feature extraction methods, although the extracted feature wavelengths covered the visible light and near-infrared bands, and the specific wavelengths were greatly influenced by the experimental conditions. All of these factors limited the generalization ability of the model (Benedet et al., 2020). To address this, hyperspectral data were used to simulate multi-spectral broadband reflectance (LandSat8) and combined with simple linear regression to quantitatively estimate the soil water content in the Yellow River Delta with limited accuracy (Li et al., 2015b). Currently, research on monitoring the soil water content using multi-spectral remote sensing data mainly relies on vegetation indices, such as the normalized difference vegetation index (NDVI, NIR – R/NIR + R), which have a fixed construction format. Commonly, vegetation index (VI) construction methods pair all available sensor bands (visible light, near-infrared, and shortwave infrared) and select the best combination, obtained through band combination analysis (Jin et al., 2020). Alternatively, multiple VIs are correlated with crops, nutrient elements, etc., to filter suitable VIs (Li et al., 2016). The prediction accuracy of the model indicated that the VI constructed by band combination analysis performed better than existing VIs (Loozen et al., 2019).

Linear analysis models were often used in quantitative predictive modeling, such as multiple linear regression (MLR) and partial least squares regression (PLSR) (Ye et al., 2020). Compared to MLR, PLSR has advantages in regression modeling under conditions where the independent variables have multicollinearity (Wold et al., 2001). Machine learning methods can deal with large sample datasets, solve nonlinear problems, and possess high accuracy and robustness (Mouazen et al., 2010; Red et al., 2019; He et al., 2022). Although these methods can effectively enhance the inversion accuracy, there are also problems such as local optima and easy overfitting. Furthermore, random forest (RF) is a statistics-based machine learning method that neatly avoids these drawbacks (Qiao et al., 2022). However, previous studies adopted various VIs as the input factors to estimate the target parameter, based on the RF method, but the preselection of input factors was lacking (Fei et al., 2023).

Due to differences in the characteristics of the VIs, such as being resistant to saturation or reducing or eliminating soil background interference, it is necessary to optimize the input factors of the RF model to improve the prediction accuracy (Yue et al., 2016; Zhang et al., 2018b). Therefore, this study aimed to explore the combination of multi-satellite information to address the issue of low accuracy in soil water content retrieval from single-band, single-source information and to establish a model for the estimation of the soil water content based on a simulated multi-spectral method.

Material and Methods

The experiment was conducted at the College of Modern Agricultural Engineering, Kunming University of Science and Technology (102° 86’ E, 24° 85’ N, altitude of 1978 m). The soil used in this experiment (porosity: 61.65%, bulk density: 1.01g·cm^{-3}, clay: 20.03%, silt: 62.32%, sand: 17.65%, organic matter: 0.75%) can be classified as Oxisols (USDA, 2014; EMBRAPA, 2018). Through laboratory preparation, different soil water content samples were obtained. The collected soil samples were subjected to processes such as air-drying, grinding, sieving, and impurity removal. The prepared soil was placed in a disk with an inner diameter of 16 cm and a height of 1.7 cm (with mini-holes at the bottom), and all soils were saturated by water infiltration through the holes at the bottom (Figure 1). Subsequently, the water was drained through the holes to obtain soil samples with varying water contents.

As shown in Figure 2, the reflectance spectra of the soil samples were determined using the SR-2500 portable object

Figure 1. Image of the soil sample collected

Figure 2. SR-2500 portable object spectrometer

spectrometer (Spectral Evolution, Inc. 1 Canal St. Unit B-1 Lawrence, MA 01840 USA). The wavelength range of the instrument was 350–2500 nm, with a spectral resolution of 3.5 nm and a sampling interval of 1.5 nm for 350–1000 nm, and a spectral resolution of 22 nm and a sampling interval of 6 nm for 1000–2500 nm. The optical fiber had an 8° field of view (FOV). The output data with an interval of 1 nm were automatically interpolated by the instrument. Spectral reflectance data were collected when the weather was cloudless and the solar elevation angle (60 ~ 90 degrees) was appropriate. During sampling, the optical fiber was placed vertically above the sample at 15 cm to ensure that the FOV coverage did not exceed the edge of the sample. The spectral reflectance of each sample was collected 10 times, and the average value was taken as the spectral reflectance of the sample. The instrument calibration was done using a standard whiteboard before measurement, and recalibration was carried out every 10 min during the measurement process.

The soil water content was determined by the oven-drying method, which was carried out immediately after the collection of the spectral reflectance data of soil samples. The soil water content was calculated using Eq. 1:

$$\theta = \frac{W_1 - W_2}{W_2 - W_3} \times 100\% \tag{1}$$

where:

θ - represents the soil water content value;

W_1 - represents the mass of the wet soil plus the aluminum box, g;

W_2 - represents the mass of the dry soil plus the aluminum box, g; and,

W_3 - represents the mass of the empty aluminum box, g.

The field capacity of the experiment soil was measured by the Wilcox method; it was 31.63%. The samples were divided into two independent datasets by sample set partitioning based on the joint X-Y distance (SPXY) (Galvao et al., 2005), with a ratio of 2:1. The larger dataset was labeled as the calibration dataset, and the remainder of the samples were considered the prediction dataset.

In this paper, eight types of multi-spectral satellites, namely Landsat 8, Modis, Quickbird, GF-1-PMSA, GF-1-WFV1, GF-2, GF-4, and GF-6, were selected. The simulated spectra were obtained based on the spectral response function of each sensor band. The reflectance values of the multi-spectral satellite broadband data (Blue, Green, Red, Near-infrared (Nir), Short-wave infrared (Swir), Red edge 1, Red edge 2, and other bands) were simulated using Eq. 2 as follows (Trigg et al., 2000):

$$R = \sum_{\lambda=\lambda_{min}}^{\lambda_{max}} S_\lambda R_\lambda \Big/ \sum_{\lambda=\lambda_{min}}^{\lambda_{max}} S_\lambda \tag{2}$$

where:

R - represents the simulated reflectance of each sensor band;

λ_{min} and λ_{max} - represent the starting and ending wavelengths detected by the sensor band, nm;

S_λ - refers to the spectral response function of the sensor at wavelength λ; and,

R_λ - refers to the reflectance of the sample at wavelength λ.

The VI was the spectral transformations of two or more bands, aimed at enhancing the contribution of vegetation characteristics. As simple transformations of spectral bands, they are directly calculated without any bias or assumption about land cover categories, soil types, or climatic conditions. They enabled consistency in monitoring seasonal, interannual, and long-term changes in the target parameters. In this paper, band combinations based on the ratio vegetation index (RVI; Eq. 3), difference vegetation index (DVI; Eq. 4), and normalized difference vegetation index (NDVI; Eq. 5) have been established using a pairwise combination framework from the bands of eight satellites:

$$RVI = \frac{\rho_{\lambda_1}}{\rho_{\lambda_2}} \tag{3}$$

$$DVI = \rho_{\lambda_1} - \rho_{\lambda_2} \tag{4}$$

$$NDVI = \frac{\rho_{\lambda_1} - \rho_{\lambda_2}}{\rho_{\lambda_1} + \rho_{\lambda_2}} \tag{5}$$

where:

$\rho_{\lambda 1}$ - represents the reflectance of the band corresponding to wavelength λ_1; and,

$\rho_{\lambda 2}$ - represents the reflectance of the band corresponding to wavelength λ_2.

The correlation between each VI and the soil water content was evaluated using the Pearson correlation coefficient. This is based on the principle that when the absolute value of the correlation coefficient is closer to 1, the correlation between the VI and the soil water content is high, and vice versa.

The RF method is a tree classifier system proposed by Breiman (2001). The concept of out-of-bag (OOB) data emerges in ensemble learning algorithms based on the bootstrap method, such as RF. In this process, an average of approximately 63.2% of the samples would be sampled, with the remaining 36.8% known as OOB samples (Zhang et al., 2018b). This portion of data could be used to estimate the performance of the model, which is thus called OOB estimation. The importance of input data can be validated based on the OOB method. The principle and steps are as follows: first, obtain the predictive accuracy of a feature on the OOB data across all trees. Next, the values of features are randomly shuffled, while other features remain unchanged, and the predictive accuracy is calculated again. The difference between these two accuracies indicates the importance of the feature. The larger the difference in the feature importance, the greater the influence of this feature on the prediction results, indicating that it is more important (Yue et al., 2016).

Based on the analysis results of the correlation coefficient between the VIs and the soil water content, the VI with the highest correlation coefficient was selected. Regression models in the form of linear univariate, quadratic polynomial, and exponential functions were established.

PLSR is a model method that integrates principal component analysis, multivariate linear regression, and least squares regression (Ye et al., 2020). This method allows regression modeling when the number of samples is smaller than the number of variables. During the modeling process, the independent variables are compressed into several latent variables (LVs), and an appropriate number of LVs are chosen to establish a linear regression model between the dependent variables. To ensure that the PLSR model has the optimal predictive ability, the determination of the number of LVs is an issue that needs to be addressed. In this paper, 5-fold cross-validation was applied to determine the number of LVs. During this process, some data were treated as unknown to validate the performance of the model.

The RF method is an ensemble learning algorithm based on multiple decision tree models (Breiman, 2001). The main steps used to implement this method are shown in Figure 3: 1) randomly select multiple samples from the original dataset using bootstrap sampling (repeatable sampling) to form multiple training sets; 2) for each training set, a decision tree model is constructed based on a subset of features (usually randomly selected); and 3) for each input sample, predictions are made on all decision trees, and the final predicted value is the average of the predicted values from all decision trees. The primary parameters of the RF include the number of decision trees (ntree) and the number of splits in internal nodes (mty). The optimal number for ntree was determined to be 100 after multiple rounds of training, while mty was given the default value (1/3), that is, one-third of the input variables.

The extreme learning machine (ELM) is a single-layer feed-forward neural network algorithm based on randomization. This method utilizes randomly initialized connection weights and biases between the input and hidden layers, which can achieve rapid model training. In this study, input samples were mapped to high-dimensional feature spaces through the "sigmoid" activation function. The equation $H\beta = T$ was solved to obtain the least squares solution β of connecting the hidden layer and the output layer, based on the output matrix of the hidden layer (H), the output matrix (T), and the output weight matrix (β), and the ELM learning process was completed (Huang et al., 2015b). The optimal ELM estimation performance can be achieved by adjusting the number of hidden layer nodes.

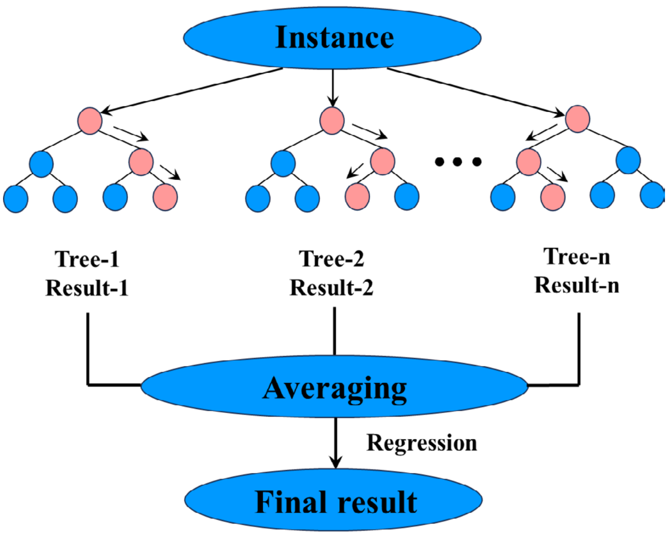


Figure 3. Flowchart of random forest algorithm

The coefficient of determination (R^2), root mean square error (RMSE), and relative percent deviation (RPD) were employed to quantify the prediction accuracy of models. A satisfactory model should possess a high R^2 and RPD, along with a low RMSE. If RPD ≤ 1.0, the model has no predictive ability; if 1.0 < RPD ≤ 1.4, the model can be used for qualitative analysis; if 1.4 < RPD ≤ 2.0, the model has a general predictive ability; and when RPD > 2.0, the model can be used for quantitative analysis. Statistical software such as MATLAB 2021b (The MathWorks, Natick, MA, USA), Microsoft Excel 2023, and Origin 2022 (OriginLab, Northampton, MA, USA) were used for data analysis and scientific drawing.

Results and Discussion

A total of 139 soil samples were measured. The calibration dataset contained 93 samples, with the soil water content ranging from 13.476 to 47.877%, and the average water content was 28.058%. The prediction dataset consisted of 46 samples, with the soil water content varying between 15.582 and 41.781%, and the average water content was 19.440% (Table 1). The range of the soil water content in the divided soil samples demonstrated adequate representativeness.

The spectral response functions of each sensor band from the eight satellites, namely Landsat 8, Modis, Quickbird, GF-1-PMSA, GF-1-WFV1, GF-2, GF-4, and GF-6, are shown in Figure 4. Among these, GF-6 possesses six bands (Blue, Green, Red, Nir, Red edge 1, and Red edge 2), while Landsat 8 and Modis each have five bands (Blue, Green, Red, Nir, and Swir). Each of the remaining five satellites incorporates four bands (Blue, Green, Red, and Nir). A similar distribution of the spectral response functions was observed across these eight satellites; it generally displayed an upwardly convex, steep parabolic shape.

Figure 5 shows the change in each band reflectance across different satellites with the same soil water content (13.476%). In general, the band reflectance changes of the eight satellites are similar. The bands are listed in descending order based on the reflectance value: Swir, Red edge 2, Red edge 1, Nir, Red, Green, and Blue.

Table 1. Descriptive statistics of the soil water content

Dataset	Samples	Minimum	Maximum	Average	SD	CV
				(%)		
Total sample	139	13.476	47.877	28.525	7.664	26.868
Calibration dataset	93	13.476	47.877	28.058	7.765	27.675
Prediction dataset	46	15.582	41.781	29.440	7.376	25.054

SD - Standard deviation; CV - Coefficient of variation

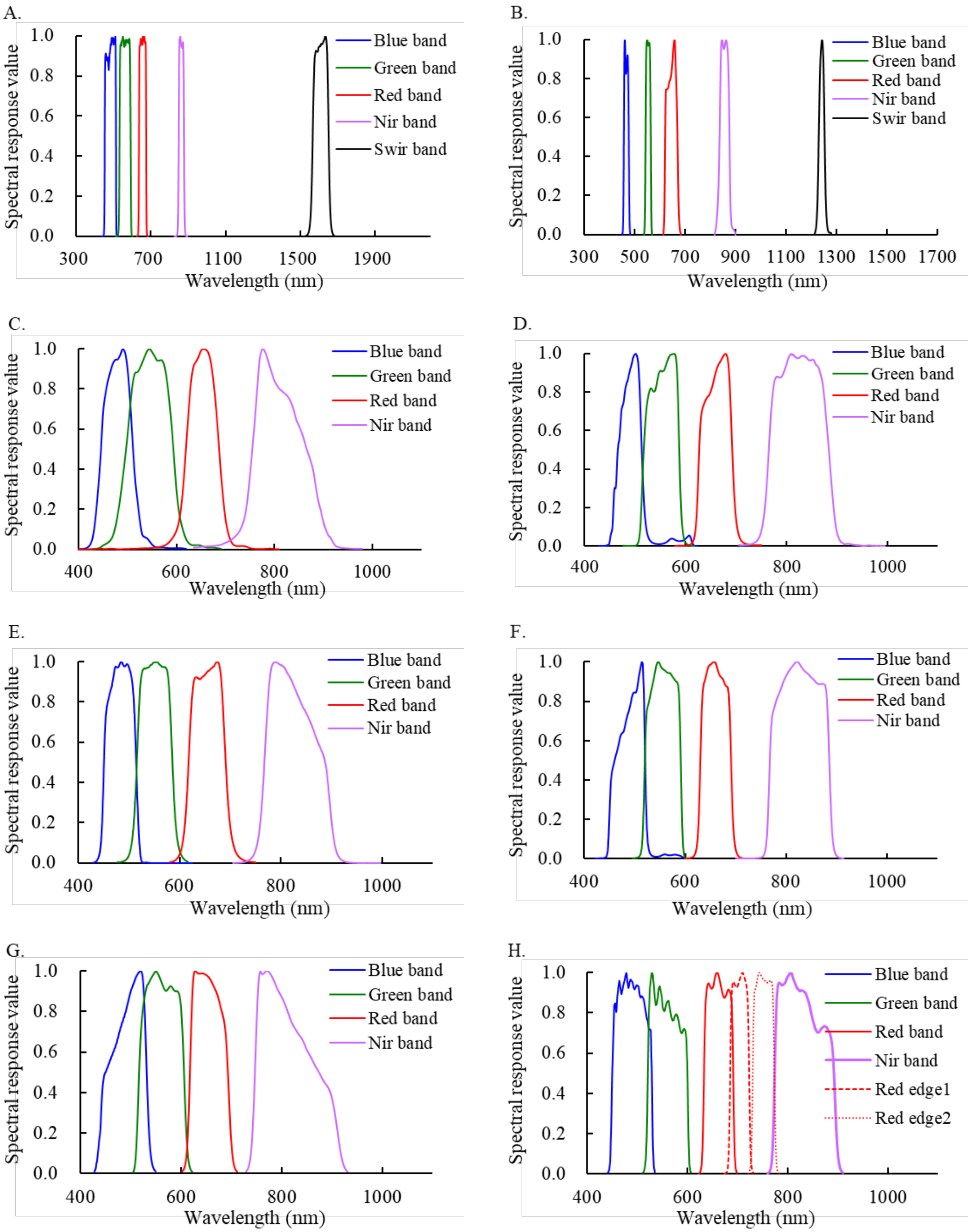


Figure 4. Spectral response functions of different satellite sensors (A: Landsat 8, B: Modis, C: Quickbird, D: GF-1-PMSA, E: GF-1-WFV1, F: GF-2, G: GF-4, H: GF-6)

Figure 6 shows the reflectance changes of each band of Landsat 8 at different soil water contents. The variation characteristics across the reflectance values of different bands under varying soil water conditions were similar, overall presenting a trend of initially decreasing and then increasing as the soil water content increased. The lowest reflectance value of all bands was observed when the soil sample water content was near the field capacity.

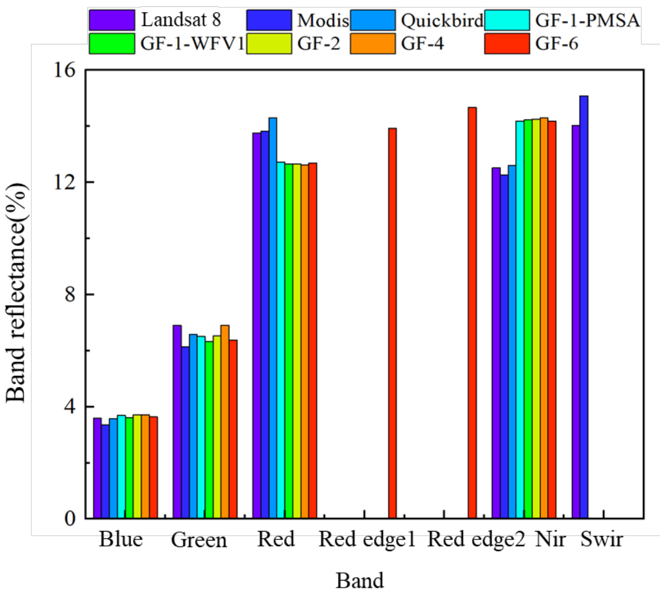


Figure 5. Reflectance of each satellite band for the same soil sample with water content of 13.476%

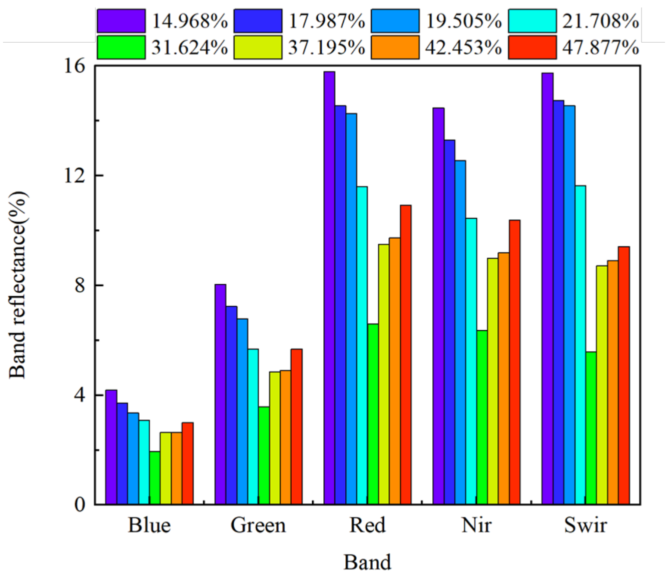
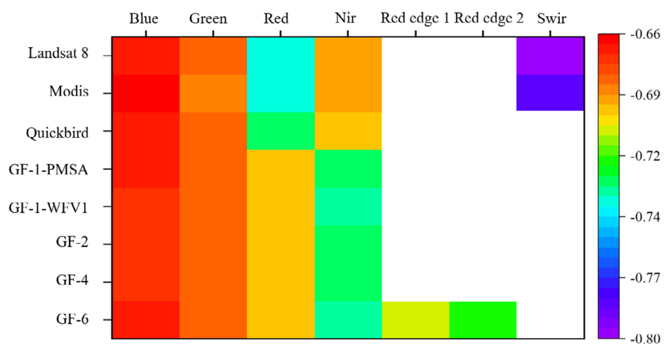


Figure 6. Reflectance of Landsat 8 bands at different soil water contents

Overall, the correlation coefficients between six bands of the eight satellites and the soil water content were negative (Figure 7). The highest absolute value of the correlation coefficient between the Swir band of Landsat 8 and the soil water content was -0.797.

The correlation coefficients between the three forms of VIs and the soil water content are shown in Figure 8, with each band-containing satellite listed from top to bottom as follows: Landsat 8, Modis, Quickbird, GF-1-PMSA, GF-1-WFV1, GF-2, GF-4, and GF-6. The correlation coefficients of the RVI, DVI, and NDVI are symmetric along the diagonal. Among these VIs, for the RVI, the Red edge 2 band of GF-6 and the Swir band of Landsat 8 achieved the highest correlation coefficient (0.813); for the DVI, the Swir and Green bands of Landsat 8 obtained the highest correlation coefficient (0.826); and for the NDVI, the Swir band of Landsat 8 and the Red edge 2 band of GF-6 obtained the highest correlation coefficient (0.808).

Taking the RVI, DVI, and NDVI with the highest correlation coefficients as independent variables, three regression models (linear, quadratic polynomial, and exponential functions) were established to estimate the soil water content. Figure 9 shows the scatter plots and regression equations for different regression models, and the results show that the DVI has a superior regression performance compared to the RVI and NDVI. Through regression models and predictive accuracy (Figures 9 and 10), the linear estimation model (y = 8.686x + 22.552) based on the DVI obtained the best performance, with an R^2 of the calibration dataset (R^2_c) of 0.654 and an R^2 of the prediction dataset (R^2_p) of 0.872. However, the results obtained by the regression method showed that the accuracy in the calibration dataset was lower than that in the prediction dataset, indicating an evident underfitting phenomenon.



The white areas in the figure indicate the absence of corresponding bands for the satellite.

Figure 7. Correlations between simulated satellite reflectance and soil water content

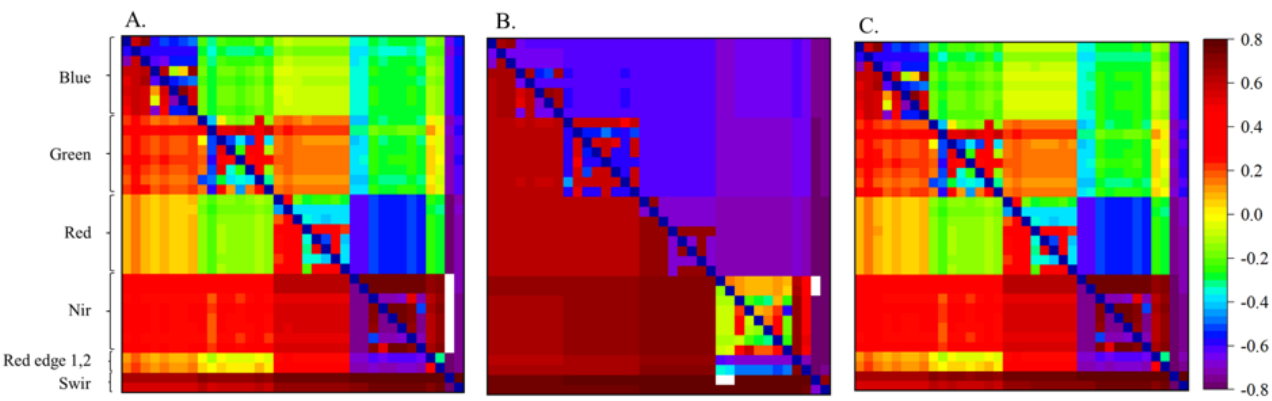


Figure 8. Correlations between ratio vegetation index (RVI) (A), difference vegetation index (DVI) (B), and normalized difference vegetation index (NDVI) (C) constructed by pairing the simulated band reflectance of eight satellites and soil water content

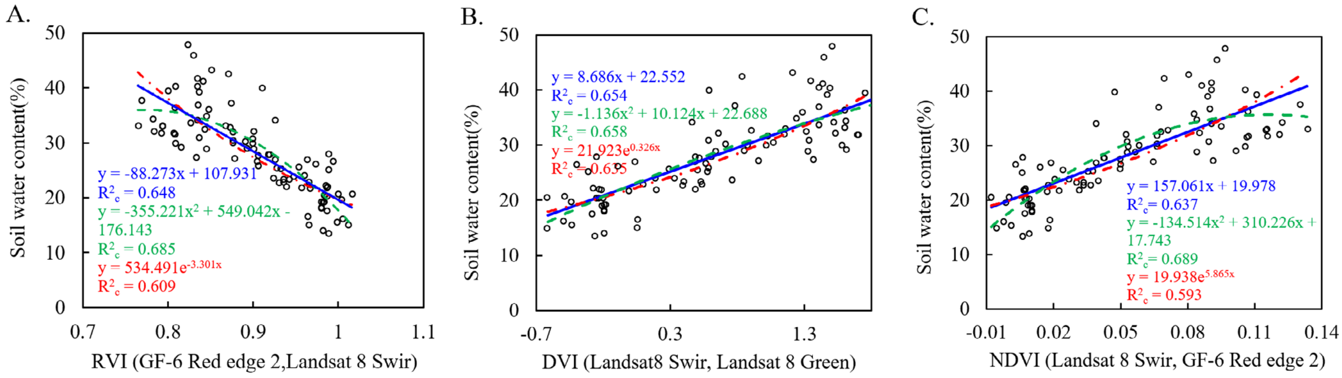


Figure 9. Soil water content as a function of ratio vegetation index (RVI) (A), difference vegetation index (DVI) (B), and normalized difference vegetation index (NDVI) (C)

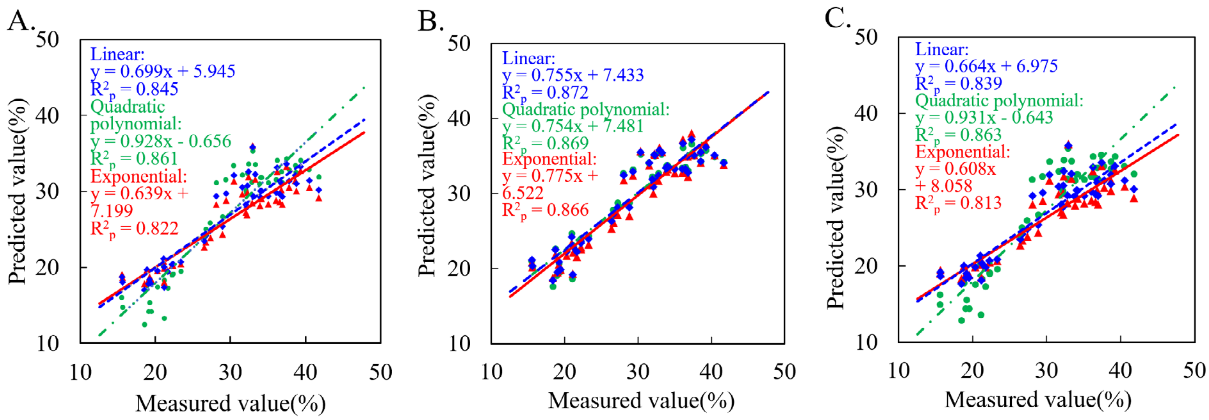


Figure 10. Prediction accuracy and scatter plot of the regression model based on RVI (GF-6 Red edge 2, Landsat 8 Swir) (A), DVI (Landsat 8 Swir, Landsat 8 Green) (B), and NDVI (Landsat 8 Swir, GF-6 Red edge 2) (C)

Previous studies were based on a single VI, such as the NDVI, to establish soil water estimation models (Liu et al., 2002; Schnur et al., 2010; Yueh et al., 2022). However, compared to complex models or the integration of multiple data sources, a single VI might not accurately reflect the water status below the ground surface, or the accuracy might be reduced in areas with extreme differences in vegetation coverage; in addition, it was affected by various factors such as seasonal changes, vegetation types, and climatic conditions. Therefore, the single VI prediction model has certain limitations, while the use of too many VIs may lead to overfitting.

The correlation between the three VIs, namely the RVI, DVI, and NDVI, and the soil water content was analyzed; it indicated that the VI construction method used in this study was beneficial for maximizing the multi-spectral information. The basic principle of VIs lies in utilizing the spectral reflectance characteristics of target objects in different bands and forming the specific indices that are sensitive to target information by calculating the reflectance ratio or the difference between different bands (Loozen et al., 2019). The DVI exhibited a superior prediction accuracy compared to the RVI and NDVI. Although the RVI and NDVI also achieved a significant prediction accuracy, the range of independent variables in the models was small (Figures 9 and 10), which meant that the dispersion or variability allowed in the input data was low. Consequently, the RVI and NDVI regression models may not accurately predict scenarios with large variations in independent variables.

A large number of VIs were obtained by pairing the reflectance values of each band of eight multi-spectral satellites, and the OOB importance analysis was conducted to reduce redundant data. Based on the ranking of the OOB importance analysis results from the three VIs and the reflectance values of all spectral bands from the eight satellites, the top two most important types of data were selected (Table 2). The VIs and spectral band reflectance selected through the OOB method demonstrated a significant correlation with the soil water content, with an absolute correlation coefficient (|r|) above 0.722. The highest correlation coefficient with the soil water content was observed for the vegetation indices constructed by pairing the DVI for Landsat 8 Green and Landsat 8 Swir and the DVI for Landsat 8 Swir and Landsat 8 Nir, which both reached a correlation of 0.808. The lowest correlation was found in the Swir band of Modis, with a correlation coefficient of 0.722.

The VIs and band reflectance selected through the OOB method showed a high correlation with the soil water content ($|r| > 0.722$, $p < 0.01$) (Table 2). The most frequent band that contributed to the VIs was Swir, followed by Nir, and only the Green band in the visible spectrum, which is similar to previous studies (Babaeian et al., 2021). This is because water molecules have strong absorption characteristics in the Swir band, leading to the partial absorption of Swir as the water content increases gradually. Therefore, the Swir band is often used to monitor the soil water in remote sensing (Wang et al., 2008). Although the Nir band is also absorbed by water, the absorption is relatively weak.

Table 2. OOB importance analysis and correlation between VIs, spectral band reflectance, and soil water content

Number	VI/reflectance	Importance	Absolute value of correlation coefficient |r|
1	RVI (GF-6 Nir, Landsat 8 Swir)	0.988	0.789**
2	RVI (Landsat 8 Swir, GF-6 Nir)	0.924	0.805**
3	DVI (Landsat 8 Green, Landsat 8 Swir)	1.132	0.808**
4	DVI (Landsat 8 Swir, Landsat 8 Nir)	1.076	0.808**
5	NDVI (GF-6 Nir, Landsat 8 Swir)	1.012	0.798**
6	NDVI (Landsat 8 Swir, Landsat 8 Nir)	0.995	0.795**
7	Swir (Landsat 8)	0.905	0.744**
8	Swir (Modis)	0.893	0.722**

** - Significant at the p ≤ 0.01 by t - test

Based on the dataset selected through the OOB method (Table 2), the three models (PLSR, RF, ELM) were established to predict the soil water content (Table 3). According to the results, the generalization ability of the PLSR model was inferior to that of the RF and ELM models. The PLSR model showed an obvious underfitting phenomenon, as indicated by R^2_c being less than R^2_p, which was considered unsuitable for the soil water content prediction model in this study. The RF and ELM models yielded a high accuracy for both the calibration and prediction datasets. The scatter plots between the measured and predicted values for the calibration dataset and prediction dataset for the RF and ELM are shown in Figures 11b and c. Additionally, the slopes of the fitting line were all close to 1, indicating that the models had achieved satisfactory prediction results. Moreover, the R^2_c values obtained from the RF and ELM models were significantly better than the unitary regression model based on the VI in Figures 9 and 10. The R^2_c and R^2_p values of the RF and ELM models were both greater than 0.820, and the RPD was greater than 2, which indicated that the model can achieve a high-precision quantitative estimation of the soil water content. RF was the most optimal model among the two machine learning models; the R^2_c, R^2_p, and RPD values were 0.852, 0.834, and 2.144, respectively.

Table 3. Regression models of soil water content using partial least squares regression (PLSR), random forest (RF), and extreme learning machine (ELM)

Modeling methods	Calibration dataset		Prediction dataset		
	R^2_c	RMSE (%)	R^2_p	RMSE (%)	RPD
PLSR	0.718	4.120	0.877	2.966	2.053
RF	0.852	3.013	0.834	3.317	2.144
ELM	0.874	2.751	0.820	3.752	2.176

Figure 12 shows the R^2_p and RMSE box plots of the RF and ELM, which were run multiple times based on the prediction dataset; these plots indicate that the combined OOB and RF (OOB-RF) model had stronger stability, higher accuracy, and less deviation, making it the best model for estimating the soil water content in this study.

The RF algorithm is easy to parallelize and has high accuracy and strong resistance to overfitting. It is also more stable and serves as an ideal benchmark model for processing nonlinear data. In this study, the RF method was used to perform OOB importance analysis. For each feature, the predictive performance difference based on OOB data was compared between including or excluding each feature, enabling the determination of its importance. Different variable extraction methods coupled with various statistical regression models yield a differentiated modeling ability (Wang et al., 2021). Commonly used variable selection methods include correlation coefficient analysis (r), variable importance in the projection (VIP) analysis, and OOB importance analysis. The range of soil water content values was relatively wide (Table 1, with samples with soil water content values above and below the field capacity), and considering the change in characteristics between the soil spectral reflectance and soil water content (Peng et al., 2009), linear regression (PLSR) methods have certain limitations. In terms of the nonlinear model, the stability of the ELM was weaker than RF in the prediction of the soil water content in this study (Figure 8).

The OOB method fitted well with the RF model, due to the fact that the determination of the variable importance in the OOB method is based on the contributions of variables participating in the modeling process of the RF regression

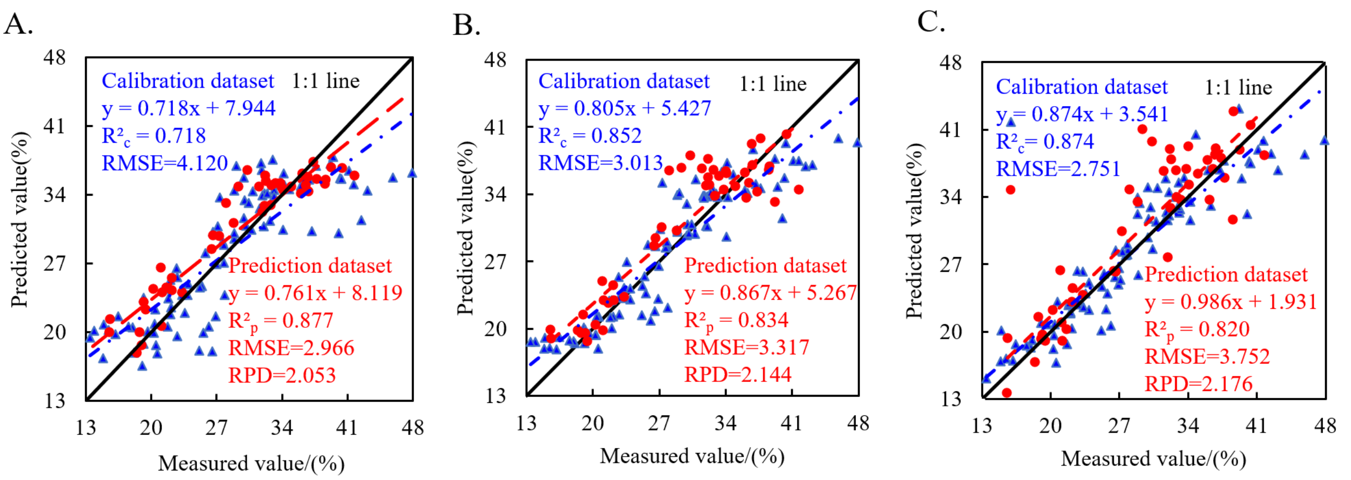


Figure 11. Scatter plots of predicted and measured values based on the partial least squares regression (PLSR) (A), random forest (RF) (B), and extreme learning machine (ELM) (C)

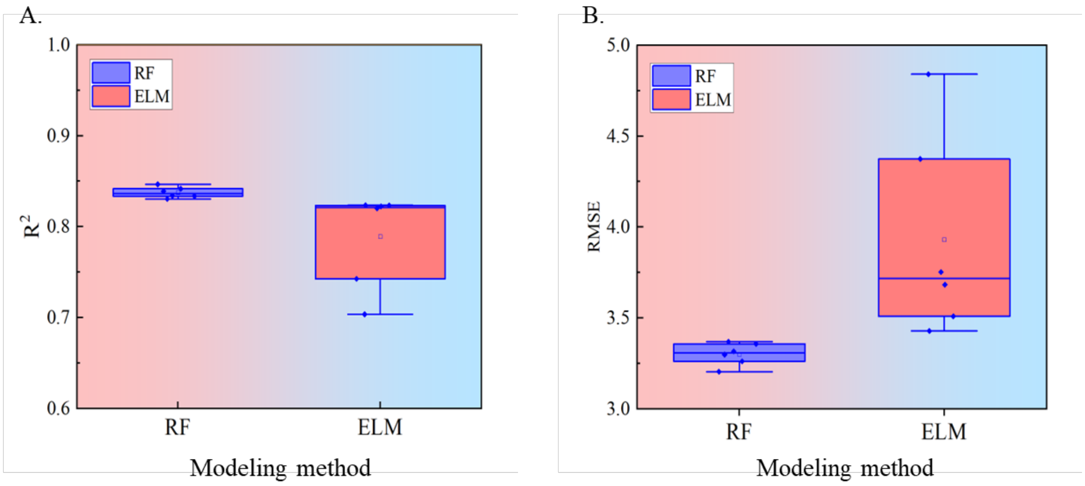


Figure 12. The coefficient of determination of prediction (R^2_p) (A) and root mean square error (RMSE) (B) box plots for random forest (RF) and extreme learning machine (ELM), which were run six times

model (Wold et al., 2001). Wang et al. (2021) also found that the predictive ability of the OOB-RF models was satisfactory. The OOB-RF model yielded R^2_c, R^2_p, and RPD values of 0.852, 0.834, and 2.144, respectively, indicating that the OOB-RF model could accurately predict the soil water content. Therefore, this study suggested using OOB-RF as a prediction model for the soil water content. Multi-spectral remote sensing technology can provide a scientific basis for regional soil water dynamic monitoring and ecological environment management.

Conclusion

1. Overall, the reflectance in all satellite bands exhibited a nonlinear trend; it initially decreased and then increased as the soil water content increased.

2. The vegetation indices, namely the ratio vegetation index (GF-6 Red edge 2, Landsat 8 Swir), difference vegetation index (Landsat 8 Swir, Landsat 8 Green), and normalized difference vegetation index (GF-6 Red edge 2, Landsat 8 Swir), showed a high correlation (r) with the soil water content, with values of 0.813, 0.826, and 0.808, respectively. However, the regression model showed an underfitting phenomenon.

3. The nonlinear models (extreme learning machine and random forest) had better performances than the linear model (partial least squares regression). The combined out-of-bag and random forest (OOB-RF) model achieved the highest accuracy and stability for both the calibration and prediction datasets. The coefficient of determination of the calibration dataset (R^2_c) and the coefficient of determination of the prediction dataset (R^2_p) of the OOB-RF model were 0.852 and 0.834, respectively; the root mean square error values of the calibration and prediction datasets were 3.013 and 3.317%, respectively, and the relative percent deviation was 2.144.

4. Overall, the OOB-RF modeling method based on multi-spectral remote sensing technology was feasible for estimating the soil water content.

Contribution of authors: Participation in the design of the research, Shaomin Chen and Zhilin Xu; Collection, Qiuyu Pu and Fangchuan Lou; Analysis and interpretation of data, Zhilin Xu and Jiachen Gao; Preparation of the manuscript, Shaomin Chen and Zhilin Xu; Literature review, Shaomin Chen and Shuai Tan; Administration and acquisition of financing, Chao Gao and Shuai Tan; Supervision of the work, Chao Gao and Xin Shen.

Supplementary documents: Data will be made available on reasonable request.

Conflicts of interest: The authors declare no conflicts of interest.

Financing statement: This study was funded by the National Natural Science Foundation of China (52209056), the Yunnan Fundamental Research Projects (202401AT070325), and the Yunnan Province Undergraduate Innovation and Entrepreneurship Training Plan Program (S202410674184), the Scientific Research Fund Project of Yunnan Provincial Department of Education (2022J0061).

Acknowledgments: The authors would like to thank Kailun Peng and Shuai Zhang of Kunming University of Science and Technology for their help with the experiment.

Literature Cited

Babaeian, E.; Sadeghi, M.; Gohardoust, M. R.; Arthur, E.; Effati, M.; Jones, S. B.; Tuller, M. The feasibility of shortwave infrared imaging and inverse numerical modeling for rapid estimation of soil hydraulic properties. Vadose Zone Journal, v.20, e20089, 2021. https://doi.org/10.1002/vzj2.20089

Benedet, L.; Faria, W. M.; Silva, S. H. G.; Mancini, M.; Guilherme, L. R. G.; Dematte, J. A. M.; Curi, N. Soil subgroup prediction via portable X-ray fluorescence and visible near-infrared spectoscopy. Geoderma, v.365, e114212, 2020. https://doi.org/10.1016/j.geoderma.2020.114212

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

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. Sistema Brasileiro de Classificação de Solos, 5.ed. Embrapa, Rio de Janeiro, Brazil, 2018, 356p.

Fei, S.; Hassan, M. A.; Xiao, Y.; Su, X.; Chen, Z.; Cheng, Q.; Duan, F.; Chen, R.; Ma, Y. Uav-Based Multi-Sensor Data Fusion and Machine Learning Algorithm for Yield Prediction in Wheat. Precision Agriculture, v.24, p.187-212, 2023. https://doi.org/10.1007/s11119-022-09938-8

Galvao, R. K. H.; Araujo, M. C. U.; José, G. E.; Pontes, M. J. C.; Silva, E. C.;Saldanha, T. C. B. A Method for Calibration and Validation Subset Partitioning. Talanta, v.67, p.736-740, 2005. https://doi.org/10.1016/j.talanta.2005.03.025.

He, B.; Jia, B.; Zhao, Y.; Wang, X.; Wei, M.; Dietzel, R. Estimate soil moisture of maize by combining support vector machine and chaotic whale optimization algorithm. Agricultural Water Management, v.267, e107618, 2022. https://doi.org/10.1016/j.agwat.2022.107618

Huang, G.; Huang, G.; Song, S.; You, K. Trends in extreme learning machines: A review. Neural Networks, v.61, p.32-48, 2015b. https://doi.org/10.1016/j.neunet.2014.10.001

Jin, N.; Zhang, D.; Li, Z.; He, L. Evaluation of water status of winter wheat based on simulated reflectance of multispectral satellites. Transactions of the Chinese Society for Agricultural Machinery, v.51, p.243-252, 2020. https://doi.org/10.21203/rs.3.rs-3936097/v1

Khanal, S.; Kushal, K. C.; Fulton, J. P.; Shearer, S.; Ozkan, E. Remote Sensing in Agriculture-Accomplishments, Limitations, and Opportunities. Remote Sensing, v.12, e3783, 2020. https://doi.org/10.3390/rs12223783

Li, F.; Chang, Q.; Shen, J.; Wang, L. Estimation of wheat leaf nitrogen content based on simulated multi-spectral broadband reflectance. Transactions of the Chinese Society for Agricultural Machinery, v.47, p.302-308, 2016. https://doi.org/2016.10.6041/j.issn.1000-1298.2016.02.040

Li, P.; Zhao, G.; Gao, M.; Chang, C.; Wang, Z.; Zhang, T.; An, D.; Jia, J. Hyperspectral estimation and remote sensing retrieval of soil water regime in the Yellow River delta. Acta Pedologica Sinica, v.52, p.1262-1272, 2015. https://doi.org/10.11766/trxb201408270429

Li, Z.; Leng, P.; Zhou, C.; Chen, K.; Zhou, F.; Shang, G. Soil moisture retrieval from remote sensing measurements: Current knowledge and directions for the future. Earth-Science Reviews, v.218, e1036732021, 2021. https://doi.org/10.1016/j.earscirev.2021.103673

Liu, L.; Zhang, B.; Zheng, L.; Tong, Q.; Liu, Y.; Xue, Y.; Yang, M.; Zhao, C. Target classification and soil water content regression using land surface temperature (LST) and vegetation index(VI). Journal of Infrared and Millimeter Waves, v.21, p.269-273, 2002.

Loozen, Y.; Karssenberg, D.; De Jong, S. M.; Wang, S.; Van Dijk, J.; Wassen, M. J.; Rebel, K. T. Exploring the use of vegetation indices to sense canopy nitrogen to phosphorous ratio in grasses. International Journal of Applied Earth Observation and Geoinformation, v.75, p.1-14, 2019. https://doi.org/10.1016/j.jag.2018.08.012

Mouazen, A. M.; Kuang, B.; De Baerdemaeker, J.; Ramon, H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma, v.158, p.23-31, 2010. https://doi.org/10.1016/j.geoderma.2010.03.001

Peng, J.; Zhang, Y.; Zhou, Q.; Pang, X.; Wu, W. The progress on the relationship physics-chemistry properties with spectrum characteristic of the soil. Chinese Journal of Soil Science, v.40, p.1204-1208, 2009. https://doi.org/10.19336/j.cnki.trtb.2009.05.049

Qiao, L.; Tang, W.; Gao, D.; Zhao, R.; An, L.; Li, M.; Sun, H.; Song, D. Uav-Based Chlorophyll Content Estimation by Evaluating Vegetation Index Responses under Different Crop Coverages. Computers and Electronics in Agriculture, v.196, e106775, 2022. https://doi.org/10.1016/j.compag.2022.106775

Red, R.; Saffaj, T.; Ilham, B.; Saidi, O.; Issam, K.; Brahim, L.; El Hadrami, E. A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. Chemometrics and Intelligent Laboratory Systems, v.195, e103873, 2019. https://doi.org/10.1016/j.chemolab.2019.103873

Schnur, M. T.; Xie, H.; Wang, X. Estimating root zone soil moisture at distant sites using MODIS NDVI and EVI in a semi-arid region of southwestern USA. Ecological Informatics, v.5, p.400-409, 2010. https://doi.org/10.1016/j.ecoinf.2010.05.001

Trigg, S.; Flasse, S. Characterizing the spectral-temporal response of burned savannah using in situ spectroradiometry and infrared thermometry. International Journal of Remote Sensing, v.21, p.3161-3168, 2000. https://doi.org/10.1080/01431160050145045

United States. Soil Survey Staff. Keys to Soil Taxonomy (12th ed.) USDA NRCS. 2014. Available at: <https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/survey/>. Accessed on: Nov 13, 2020.

Wang, L.; Qu, J.; Hao, X.; Zhu, Q. Sensitivity studies of the moisture effects on MODIS SWIR reflectance and vegetation water indices. International Journal of Remote Sensing, v.29, p.7065-7075, 2008. https://doi.org/10.1080/01431160802226034

Wang, Q.; Li, J.; Jin, T.; Chang, X.; Zhu, Y.; Li, Y.; Sun, J.; Li, D. Comparative analysis of Landsat-8, Sentinel-2, and GF-1 Data for retrieving soil moisture over wheat farmlands. Remote Sensing, v.12, e2708, 2020. https://doi.org/10.3390/rs12172708

Wang, Y.; Chen, H.; Chen, J.; Wang, H.; Xing, Z.; Zhang, Z. Comparation of rice yield estimation model combining spectral index screening method and statistical regression algorithm. Transactions of the Chinese Society of Agricultural Engineering, v.37, p.208-216, 2021. https://doi.org/10.11975/j.issn.1002-6819.2021.21.024

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

Ye, X. J.; Abe, S.; Zhang, S. Estimation and mapping of nitrogen content in apple trees at leaf and canopy levels using hyperspectral imaging. Precision Agriculture, v.21, p.198-225, 2020. https://doi.org/10.1007/s11119-019-09661-x

Yue, J.; Yang, G.; Feng, H. Comparative of remote sensing estimation models of winter wheat biomass based on random forest algorithm. Transactions of the Chinese Society of Agricultural Engineering, v.32, p.175-182, 2016. https://doi.org/10.11975/j.issn.1002-6819.2016.18.024

Yueh, S. H.; Shah, R.; Chaubell, M. J.; Hayashi, A.; Xu, X.; Colliander, A. A Semiempirical Modeling of Soil Moisture, Vegetation, and Surface Roughness Impact on Cygnss Reflectometry Data. Ieee Transactions on Geoscience and Remote Sensing, v.60, e5800117, 2022. https://doi.org/10.1109/tgrs.2020.3035989.

Zhang, C.; Yang, G.; Li, H.; Tang, F.; Liu, C.; Zhang, L. Remote sensing inversion of leaf area index of winter wheat based on random forest algorithm. Scientia Agricultura Sinica, v.51, p.855-867, 2018b. https://doi.org/10.3864/j.issn.0578-1752.2018.05.005