Applied Environmental Research
- RESEARCH FOR SUSTAINABLE PLANET -

Research Article

Establishing Optimal Machine Learning Models for Monitoring Water Quality in Vietnam's Upper Ma River

Thanh-Son Ngo, Duc-Loc Nguyen*

Faculty of Natural Resources and Environment, Vietnam National University of Agriculture, Hanoi, Vietnam

*Correspondence Email: nguyenducloc@vnua.edu.vn

Abstract

This study aims to establish the optimal regression model for predicting total suspended solids (TSS) and Turbidity based on in situ data and spectral regions of Sentinel-2 images. Various machine learning models were evaluated, including Multilayer Perceptron Regression (MLPR), Random Forest Regression (RFR), AdaBoost Regression (ABR), Multiple Linear Regression (MLR), and K-Nearest Neighbors Regression (KNNR). These models were applied to different band combinations of spectral regions: visible (VIS), near-infrared (NIR), shortwave-infrared (SWIR), VIS+NIR (VNIR), and VIS+NIR+SWIR (VNIR+SWIR). The study results revealed that the MLR model, while not the best performer during training (R^2 = 0.89 for TSS and R^2 = 0.66 for turbidity), did not exhibit overfitting, with corresponding RI values in testing being 0.80 and 0.42, respectively. Variable selection for MLR models identified optimal spectral bands: B3, B5, B6, B8, B11, and B12 for TSS, and B4, B8, B11, and B12 for Turbidity. The final no-intercept multiple linear regression models achieved R^2 = 0.88 for TSS and R^2 = 0.62 for turbidity. Performance metrics for TSS were superior, with lower MAE, MSE, and RMSE compared to Turbidity. This study underscores the efficacy of using MLR models with selected spectral bands for accurate and generalizable predictions of TSS and turbidity.

ARTICLE HISTORY

Received: 13 Jul. 2024
Accepted: 28 Oct. 2024
Published: 18 Nov. 2024

KEYWORDS

Water quality monitoring;
Machine learning model;
Sentinel-2 imagery;
Turbidity;
Total suspended solids;
Upper Ma river

Introduction

Turbidity and total suspended solids (TSS) are critical parameters in assessing water quality and detecting water pollution [1–2]. Turbidity measures the cloudiness or haziness of water caused by large numbers of individual particles, while TSS quantifies the solid particles suspended in water. High levels of turbidity and TSS can impair water quality by reducing light penetration, disrupting aquatic ecosystems, and interfering with the life cycles of fish and other aquatic organisms [3–4]. These parameters also indicate the presence of pollutants such as sediments, organic matter, and microorganisms, making them essential for monitoring the health of water bodies and ensuring the safety of water for human use and ecological sustainability [2, 5].

Traditional methods for measuring TSS and turbidity primarily rely on in situ measurements and subsequent laboratory analyses. While these methods provide accurate and direct assessments, they have significant drawbacks. In situ sampling is often labor-intensive, time-consuming, and expensive, requiring personnel to visit sampling sites regularly [5]. Moreover, the transportation and handling of samples to laboratories introduce the risk of contamination and degradation, potentially affecting the reliability of results. The spatial and temporal coverage of in situ measurements is also limited, making it challenging to obtain a comprehensive understanding of water quality dynamics over large areas or extended periods [6]. Recent studies highlight these limitations. Boyd (2020) demonstrated that in situ methods, while accurate, are often logistically and economically

https://doi.org/10.35762/AER.2024053

infeasible for large-scale monitoring projects [4]. Sagan et al. (2020) discussed the limitations of spatial and temporal coverage in traditional methods, which often fail to capture dynamic changes in water quality [6].

The advent of remote sensing technology has revolutionized the monitoring of water quality parameters such as TSS and turbidity. Satellites equipped with advanced sensors can capture high-resolution spectral data over vast water bodies, providing extensive spatial and temporal coverage that is unattainable with traditional methods [7-8]. By analyzing reflectance data from various spectral bands, researchers can estimate the concentrations of TSS and turbidity across large regions in near real-time [7-8]. This approach allows for continuous monitoring, early detection of pollution events, and the assessment of long-term trends in water quality, ultimately enhancing the management and protection of aquatic environments [9-10]. Numerous studies have validated the effectiveness of satellite remote sensing in water quality monitoring. For example, [7] used Landsat-8 and Sentinel-2 data to model reservoir chlorophyll-a, TSS, and turbidity, achieving high accuracy in their predictions. Similarly, [8] confirmed the consistency of suspended particulate matter concentration retrievals from Sentinel-2 and Landsat-8 sensors, demonstrating the reliability of satellite data for water quality monitoring.

Machine learning techniques provide an effective tool for managing water quality using either field data [11] or satellite data [12–13]. By training machine learning models on spectral data and corresponding in situ measurements, it is possible to develop predictive models that accurately estimate water quality across different conditions and regions [14]. Machine learning algorithms can learn complex relationships between spectral reflectance and water quality parameters, improving the precision and reliability of predictions [15]. The integration of machine learning with satellite data not only automates the analysis process but also enhances the capacity to monitor and manage water quality more efficiently and effectively, providing valuable insights for environmental scientists and policymakers [12]. Several studies have demonstrated the advantages of integrating machine learning with satellite remote sensing for water quality monitoring. For instance, Granata et al. (2024) utilized a stacked MLP-RF algorithm to forecast evapotranspiration, achieving significant improvements in prediction accuracy [14]. Similarly, Khalifa et al. (2024) highlighted the robustness of Random Forest models in handling complex, nonlinear relationships in environmental data, underscoring their applicability in water quality monitoring [15].

This study is particularly important for Vietnam and the upper Ma River Basin for several reasons. The upper Ma River Basin, located in Northwest Vietnam, is a vital water resource supporting the livelihoods of millions of people. The basin is crucial for agricultural activities, fisheries, and as a source of drinking water. However, it faces significant environmental challenges due to deforestation, agricultural runoff, industrial discharges, and rapid urbanization, all of which contribute to increased levels of TSS and turbidity [16–17]. The upper Ma River Basin is also prone to seasonal variations in water quality, influenced by the tropical monsoon climate characterized by distinct dry and wet seasons. These variations can lead to fluctuations in pollutant levels, making continuous and comprehensive monitoring essential for effective water resource management [5]. Ensuring water quality in this region is critical for maintaining the health of aquatic ecosystems, protecting biodiversity, and securing safe water for human consumption.

By building on the foundations laid by previous research and incorporating advanced machine learning techniques, this study seeks to enhance the accuracy and efficiency of water quality monitoring in the upper Ma River. This will provide valuable insights for environmental management and contribute to the sustainable development of water resources in the region. Specifically, we aim to identify the most effective machine learning model for monitoring water quality in the upper Ma River, focusing on the selection of optimal spectral regions and model types. We will compare the performance of Multilayer Perceptron Regression (MLPR), Random Forest Regression (RFR), AdaBoost Regression (ABR), Multiple Linear Regression (MLR), and K-Nearest Neighbors Regression (KNNR) models in predicting TSS and turbidity based on spectral data from different regions, including the visible (VIS), near-infrared (NIR), and shortwave-infrared (SWIR) bands.

The outcomes of this research will not only advance scientific understanding of water quality monitoring but also inform policy decisions and management strategies aimed at preserving the ecological integrity and water quality of the upper Ma River Basin. This is essential for ensuring the sustainability of water resources in Vietnam, thereby supporting the health and well-being of its population.

Materials and methods

1) Study area

The upper Ma river basin is located in Northwest (NW) Vietnam and covers a total area of 6,688 km^2 (Figure 1). The region is affected by the tropical monsoon climate which is characterized by the alternation between

dry and wet seasons. Annual rainfall averages 1,900 mm. The wet season lasts six months from May to October and accumulates approximately 85–95% of the total annual rainfall. Land use in the basin is dominated by forests (59% of the total area) and field crops (33%). The natural landscape of the basin has changed considerably since 1990s, with i.a. deforestation of large forest and protected areas by cutting and burning for the expansion of cultivated land mainly, which lead to decreased water quality [17]. In addition to deforestation, socio-economic development activities such as industry, mining, craft villages and agriculture have created increasing pressure on the natural environment of the basin [16], which is likely to be resulting in increased erosion and hence sediment load in the river system.

2) Data

2.1) TSS and turbidity in-situ measurement

TSS and turbidity measurements were carried out through extensive field campaigns conducted by the Vietnamese Center for Environment Monitoring. These surveys were performed at nine strategic locations along the upper Ma River, covering a wide geographical range that represents the diverse environmental conditions within the basin.

For TSS, water samples were collected at a depth of 0-50 cm using a specialized water sampler to ensure a representative sample of the surface layer. These samples were immediately preserved in 1 L dark-colored bottles to minimize exposure to light and prevent any alterations to the sample content. The bottles were then refrigerated and transported to the laboratory for analysis, where TSS was measured by filtering the water samples and drying the residue to a constant weight.

In contrast, turbidity was measured in situ using the Secchi disk method [18], a widely used approach to estimate the clarity of water. An 8-inch diameter disk, with alternating black and white quadrants, was lowered into the water column until it could no longer be seen from the surface. The depth at which the disk disappears is directly correlated with the turbidity of the water. Additionally, a turbidity meter was placed near the water surface (at a depth of less than 50 cm) for two minutes, and the readings were averaged to enhance accuracy.

To ensure precise data matching for modeling purposes, the dates of these in-situ measurements were synchronized with Sentinel-2 satellite imagery acquisitions. This synchronization minimized time discrepancies between the field data collection and the satellite observations, ensuring that both datasets captured similar environmental conditions during each campaign.

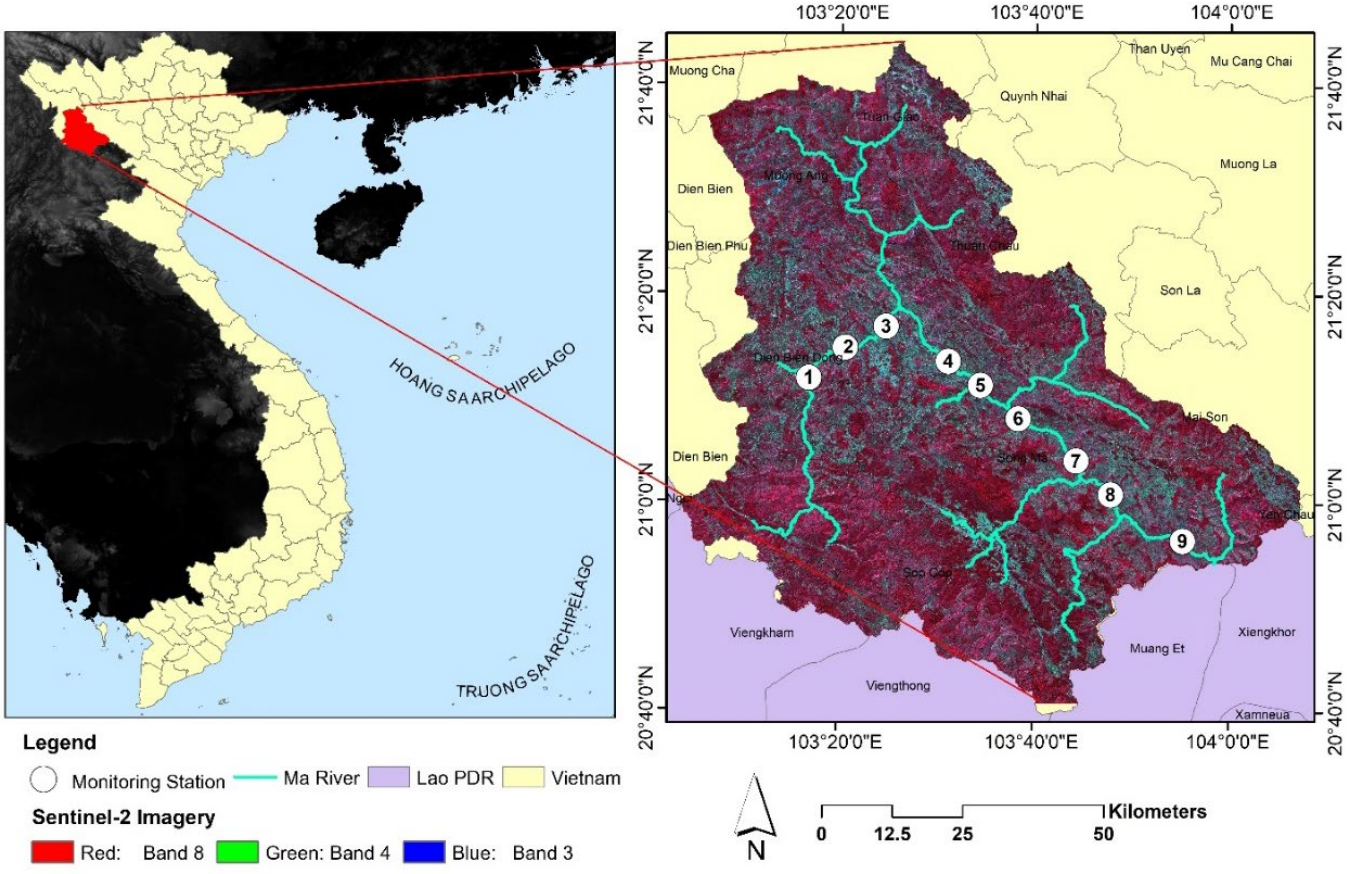

Figure 1 Map of the upper Ma river basin and the location of the monitoring stations.

Table 1 Dates of in-situ data collection and corresponding Sentinel-2 images

Data of in-situ campaign	Date of Sentinel-2 image acquisition	Difference time (days)	Image ID
17/03/2017	18/03/2017	1	T48QUJ_20170318T033531_B
16/05/2017	17/05/2017	1	T48QUJ_20170517T033541_B
17/07/2017	16/07/2017	1	T48QUJ_20170716T033541_B
20/8/2017	20/08/2017	0	T48QUJ_20170820T033529_B
12/03/2018	13/03/2018	1	T48QUJ_20180313T033531_B
25/05/2018	22/5/2018	3	T48QUJ_20180522T033541_B
12/07/2018	11/07/2018	1	T48QUJ_20180711T033541_B
20/08/2018	20/08/2018	0	T48QUJ_20180820T033531_B
19/10/2018	19/10/2018	0	T48QUJ_20181019T033741_B
03/03/2019	03/03/2019	0	T48QUJ_20190303T033639_B

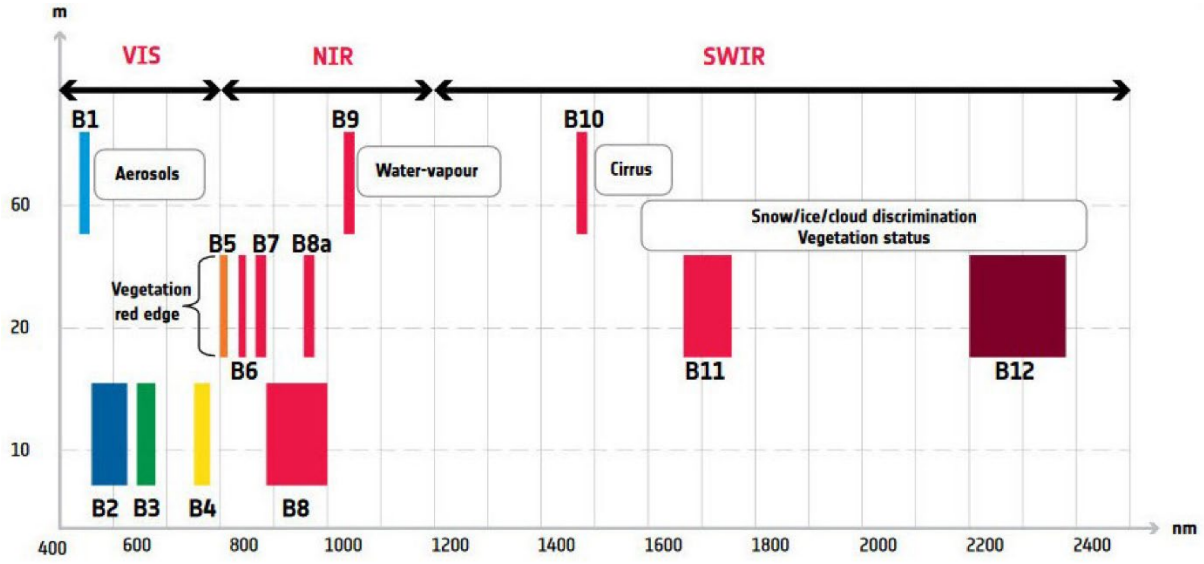

Figure 2 Sentinel-2 layout of spectral bands [19].

2.2) Sentinel-2 imagery

Ten level 1C top-of-atmosphere Sentinel-2 images covering the study area were acquired from the Sentinel Scientific Hub (https://scihub.copernicus.eu/). Of these, four images corresponded to the dates of TSS and turbidity data collection, five images were taken within one day of these dates, and one image was captured three days prior to the field data collection dates. Atmospheric correction was conducted using Sen2Cor, converting the images to Level 2A bottom-of-Atmosphere reflectance images. For each image, spectral bands in the visible region (bands 2, 3, and 4), near-infrared region (bands 5, 6, 7, and 8), and short-wave infrared region (bands 11 and 12) were utilized.

2.3) Machine learning models

To generate matchups between field measurements and spectral reflectance values, band reflectance in the pixels overlapping with the field sampling locations was retrieved. This data was used to build the input dataset for selecting the best regression model among various machine learning algorithms. Five regression algorithms were tried, including MLPR, RFR, ABR, MLR, and KNNR.

The MLPR, a type of artificial neural network, has emerged as a robust method for water quality modeling. MLPR is particularly well-suited for handling the non-linear relationships often present in environmental data [14]. It consists of an input layer, one or more hidden layers, and an output layer, with each layer comprising multiple interconnected neurons [20].

The RFR is a powerful technique for water quality modeling due to its ability to handle complex, nonlinear relationships in data [21]. It operates by constructing multiple decision trees during training and outputting the mean prediction of the individual trees [15]. This ensemble method improves accuracy and robustness, making it ideal for predicting water quality parameters from spectral data.

The ABR is a robust technique for water quality modeling, known for its ability to enhance predictive performance by combining multiple weak learners [22]. In ABR, sequential models are trained, with each new model focusing on the errors made by the previous ones [23]. This iterative process improves the overall accuracy and robustness of the predictions.

The MLR is a straightforward yet effective method for water quality modeling. It operates by modeling the linear relationships between multiple predictor variables (spectral bands) and response variables (TSS and turbidity) [24]. MLR calculates coefficients for each predictor, providing a clear equation that predicts water quality parameters based on spectral inputs. This method's simplicity and interpretability make it valuable for understanding how different factors influence water quality.

The KNNR is a versatile and intuitive method for water quality modeling. It operates by predicting the value of a water quality parameter, such as TSS or turbidity, based on the average values of the k-nearest data points in the feature space [25]. This non-parametric approach is particularly effective when the data exhibits complex, non-linear relationships.

In machine learning, models are defined by their parameters, but their performance is significantly influenced by hyperparameters. Tuning these hyperparameters is crucial for finding the optimal configuration that maximizes the model's performance. Various approaches exist for hyperparameter tuning, including manual search, grid search, and random search. Among these, grid search has been proven highly effective and is widely used in numerous studies. In this research, grid search was applied to optimize the hyperparameters of the RF, KNN, Adaboost, and MLP regression models. Multiple linear regression, being a relatively simple and straightforward algorithm with fewer hyperparameters, was used in its default form. The hyperparameters, search ranges, and optimal values for the four models are reported in Table 2.

2.4) Data standardization

When building regression models, it is common practice to standardize the data. Standardization is a preprocessing step that involves transforming the data so that it has a mean of zero and a standard deviation of one. This process ensures that all variables contribute equally to the analysis and helps to improve the performance and interpretability of the model.

Standardization transforms the original data D using the following Eq. 1.

$$D_{standardized} = \frac{D-\mu}{\sigma} \quad \text{(Eq. 1)}$$

Where D is original data, is the mean of the data, and is the standard deviation of the data.

Table 2 Hyperparameters of regression models

Regressor	Hyperparameter	Search value	Optimal value
KNN	n_neighbors	[1, 3, 5, 7, 9]	3
	Weight	[uniform, distance]	Uniform
	Metric	[minkowski, euclidean, manhattan]	Euclidean
Multilayer perceptron	hidden_layer_sizes	[1, 5, 10, 20]	2
	activation	[identity, logistic, tanh, relu]	Relu
	solver	[lbfgs, sgd, adam]	Adam
	learning_rate	[constant, invscaling, adaptive]	Constant
Random forest	learning_rate_init	[0.001, 0.01, 0.1, 1]	0.001
	n_estimators	[1, 5, 10, 20]	20
	max_features	[sqrt, log2, None]	None
	max_depth	[1, 5, 10, 20]	10
AdaBoost	n_estimators	[1, 5, 10, 20]	10
	learning_rate	[0.001, 0.01, 0.1, 1]	0.1

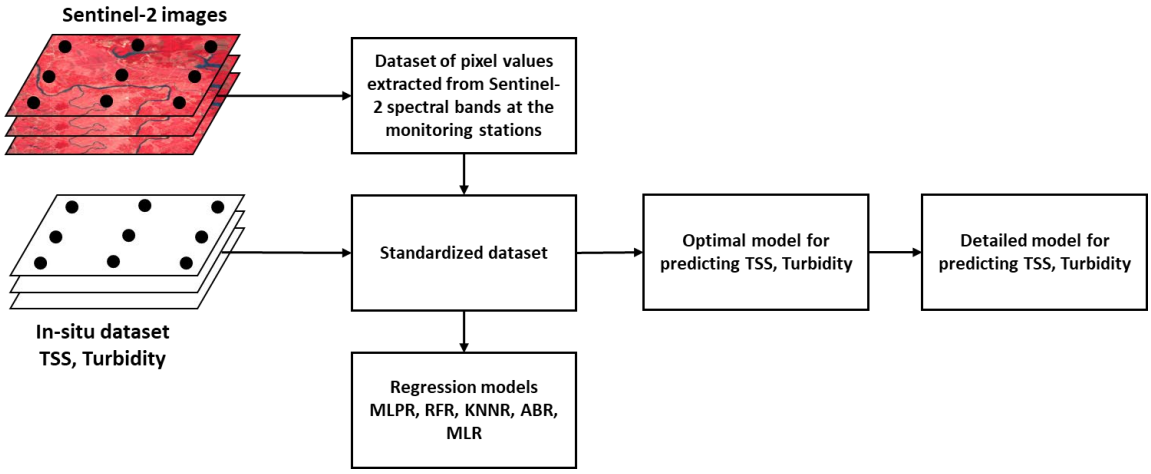

Figure 3 Flowchart of the study.

2.5) Accuracy metrics

To evaluate and compare different models to determine the most appropriate one for predicting water quality parameters, several accuracy metrics were used.

R^2 is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where a value closer to 1 suggests a better fit of the model to the data. The R^2 is calculated by Eq. 2.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad \text{(Eq. 2)}$$

Where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

Adjusted R^2 adjusts the R^2 value for the number of predictors in the model, providing a more accurate measure of model performance when multiple variables are involved. It can decrease if the added predictors do not improve the model significantly. The adjusted R^2 is calculated by Eq. 3. Where n is the number of observations and k is the number of predictors.

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

AIC (Akaike information criterion) is a measure of the relative quality of a statistical model for a given dataset. It balances model fit and complexity by penalizing the inclusion of unnecessary parameters. Lower AIC values indicate better models. The AIC is calculated by Eq. 4. Where L is the likelihood of the model.

$$AIC = 2k - 2\ln(L) \quad \text{(Eq. 4)}$$

BIC (Bayesian information criterion) is similar to AIC but includes a stricter penalty for models with more parameters. It is derived from Bayesian probability and is useful for model selection among a finite set of models. The BIC is calculated by Eq. 5.

$$BIC = kln(n) - 2\ln(L) \quad \text{(Eq. 5)}$$

Results and discussion

1) Data standardization and correlation

Tables 3 and 4 display the values of in situ turbidity, in situ TSS, and spectral bands before and after standardization, respectively. As shown in Table 3, their means are equal to 0 and their standard deviations are equal to 1. Post-standardization, Turbidity ranges from -1.38 to 3.02, while TSS ranges from -1.06 to 2.97.

Table 3 Values of variables before standardization

Index	Turbidity	TSS	B2	B3	B4	B5	B6	B7	B8	B11	B12
Count	90	90	90	90	90	90	90	90	90	90	90
Mean	104.04	260.92	0.27	0.29	0.30	0.26	0.31	0.34	0.32	0.20	0.13
Std	68.19	186.35	0.25	0.23	0.23	0.19	0.19	0.19	0.23	0.12	0.1
Min	9.80	64.00	0.04	0.06	0.04	0.08	0.10	0.11	0.05	0.04	0.02
25%	60.48	133.25	0.09	0.14	0.14	0.14	0.17	0.19	0.13	0.12	0.06
50%	88.50	202.00	0.15	0.19	0.21	0.18	0.24	0.28	0.26	0.16	0.09
75%	141.50	324.50	0.38	0.37	0.39	0.32	0.37	0.42	0.46	0.27	0.19
Max	310.00	815.00	1.08	0.99	0.94	0.80	0.82	0.85	0.91	0.63	0.51

Table 4 Values of variables after standardization

Index	Turbidity	TSS	B2	B3	B4	B5	B6	B7	B8	B11	B12
Count	90	90	90	90	90	90	90	90	90	90	90
Mean	0	0	0	0	0	0	0	0	0	0	0
Std	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Min	-1.38	-1.06	-0.92	-1.0	-1.14	-0.95	-1.08	-1.18	-1.16	-1.33	-1.20
25%	-0.64	-0.69	-0.72	-0.65	-0.70	-0.64	-0.71	-0.77	-0.81	-0.67	-0.78
50%	-0.23	-0.32	-0.48	-0.44	-0.39	-0.43	-0.37	-0.33	-0.29	-0.34	-0.47
75%	0.55	0.34	0.44	0.34	0.40	0.29	0.32	0.39	0.59	0.55	0.58
Max	3.02	2.97	3.26	3.01	2.82	2.78	2.69	2.62	2.51	3.49	3.90

Figure 4 illustrates the Pearson correlations among Turbidity, TSS, and spectral bands. In the visible and near-infrared regions (VNIR), the correlation values between TSS and nine selected spectral bands are notably higher than those between turbidity and bands. Specifically, Pearson correlation values between turbidity and bands in the VNIR region range from 0.70 to 0.75, whereas those between TSS and bands range from 0.74 to 0.82. Conversely, in the shortwave infrared region, Pearson correlation values between Turbidity and spectral bands are higher than those of TSS, ranging from 0.54 to 0.57 for Turbidity and 0.50 to 0.53 for TSS.

2) Select optimal regression model and spectral region

Figure 5 illustrates the RI values for different regression models of TSS during the training phase. The RFR and ABR models demonstrated superior performance across all spectral regions. Specifically, RFR's RI values ranged from 0.90 (shortwave infrared: SWIR) to 0.96 (VNIR +SWIR), while ABR's RI values ranged from 0.83 (SWIR) to 0.95 (VNIR+SWIR). The MLPR followed, with RI values ranging from 0.58 (SWIR) to 0.93 (VNIR+SWIR). In the VIS, NIR, and SWIR spectral regions, the MLR model performed the worst, with RI values fluctuating from 0.31 (SWIR) to 0.68 (NIR). However, in the VNIR and VNIR+SWIR regions, the MLR model performed better, with RI values fluctuating from 0.85 (VNIR) to 0.89 (VNIR+SWIR).

In contrast, during the testing phase, MLR was the best model in the VNIR and VNIR+SWIR regions, with RI values of 0.75 and 0.80, respectively. In the NIR and SWIR spectral regions, KNN performed the best, with RI values of 0.56 and 0.39, respectively. The RI values of MLPR, RFR, and ABR were significantly lower in the testing phase compared to the training phase, indicating potential overfitting during training. Overall, the SWIR region, which includes only bands 11 and 12, resulted in the lowest RI values for all models. Conversely, the VNIR+SWIR region, which includes nine bands, resulted in the highest RI values for all models because the VNIR bands are sensitive to surface reflectance properties, particularly water clarity and particle scattering. The presence of suspended solids in water leads to increased scattering of light, which is strongly captured in the visible and near-infrared regions, especially in bands corresponding to wavelengths between 400 and 900 nm. Secondly, the SWIR bands provide additional information related to water content and the absorption properties of suspended particles. Bands in the SWIR region (1,000 –2,500 nm) are particularly effective at detecting moisture and finer sediments that are typically not as detectable in the VNIR bands. The combination of VNIR and SWIR suspended solids, and other environmental variables, resulting in a more comprehensive and accurate prediction of TSS concentrations. Empirically, numerous studies have validated the utility of combining VNIR and SWIR bands for water quality monitoring, particularly for TSS estimation. For example, Wang et al. (2020) demonstrated that the combination of VNIR and SWIR bands produced higher accuracy in retrieving suspended particulate matter compared to using either region alone [26]. Additionally, the SWIR bands are less affected by atmospheric scattering, further improving the robustness of predictions in diverse environmental conditions. This combination thus enhances the ability to capture both the optical properties of suspended solids and their interactions with water, leading to the superior performance observed across all machine learning models.

The analysis revealed that RFR and ABR models, despite their high RI values during training, suffered from overfitting, as indicated by the significant drop in RI values during testing. This highlights the importance of evaluating model performance across both phases to ensure generalizability. The MLR model, while not the top performer in the training phase, exhibited a better balance between training and testing RI values, particularly in the VNIR and VNIR+SWIR regions, indicating better generalizability.

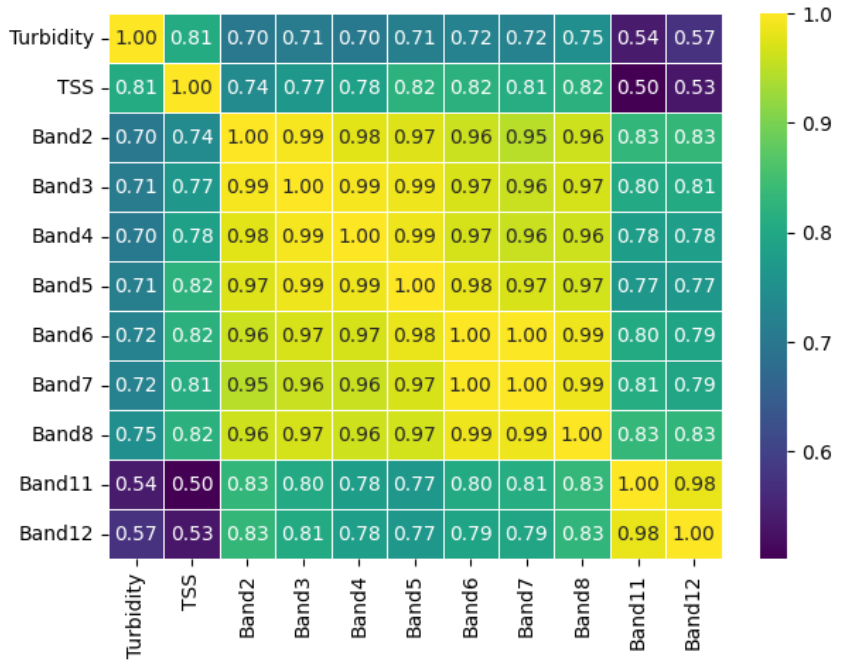

Figure 4 Correlation among turbidity, TSS, and spectral bands.

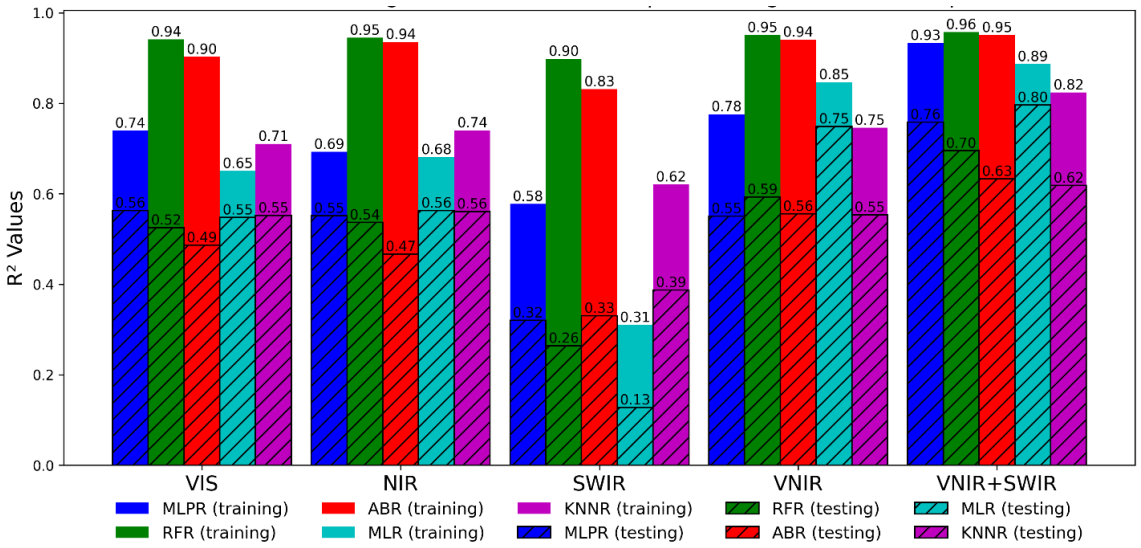

Figure 5 R^2 values for different regression models and spectral regions (TSS).

The VNIR+SWIR spectral region consistently provided the highest RI values across all models, suggesting that this region captures the most relevant information for TSS prediction. The combination of VNIR and SWIR bands appears to enhance the model's ability to predict TSS accurately by integrating information from both spectral regions. In contrast, the SWIR region alone, limited to bands 11 and 12, resulted in the lowest RI values, highlighting its limited utility in isolation.

Based on the findings, the MLR model emerged as the most reliable choice, offering balanced performance with high RI values and minimal overfitting. Therefore, selecting MLR model in conjunction with the VNIR+SWIR spectral bands is recommended for achieving accurate and generalizable TSS predictions.

Figure 6 illustrates the RI values for various regression models of turbidity during the training phase. Similar to TSS, the RFR and ABR models exhibited superior performance across all spectral regions. Specifically, the RI values for RFR model ranged from 0.85 (SWIR) to 0.93 (VNIR+SWIR), while ABR model's RI values spanned from 0.74 (SWIR) to 0.89 (VNIR+SWIR). The MLPR model followed, with RI values ranging from 0.50 (SWIR) to 0.85 (VNIR+SWIR). Conversely, the MLR model had the poorest performance across all spectral regions, with RI values varying from 0.35 (SWIR) to 0.66 (VNIR +SWIR). The KNNR model performed better than MLR model, with RI values ranging from 0.56 (SWIR) to 0.66 (VNIR+SWIR).

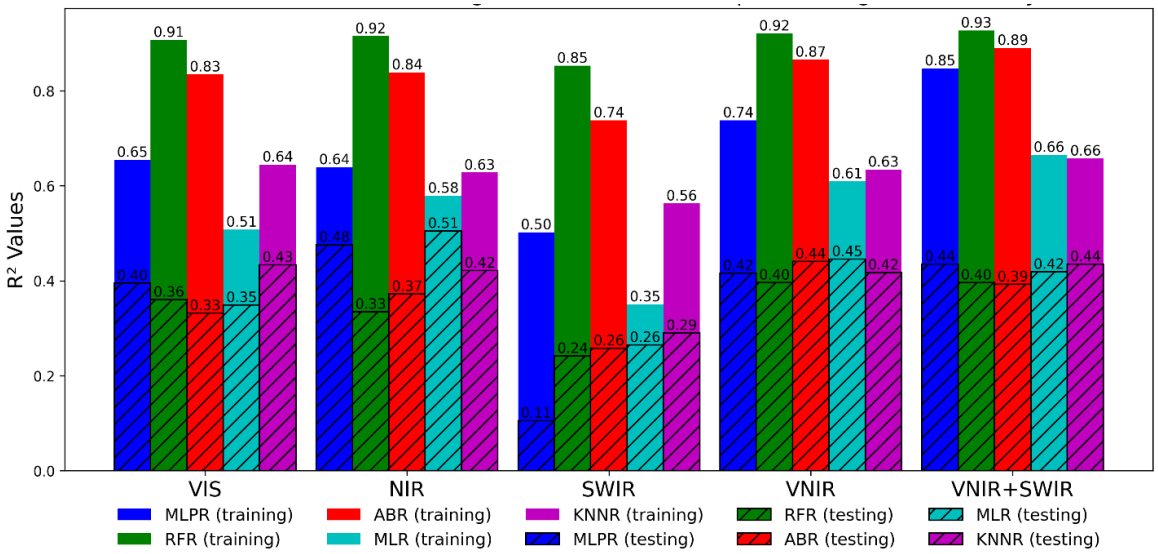

Figure 6 R^2 values for different regression models and spectral regions (turbidity).

During the testing phase, all models exhibited low performance, with most RI values below 0.5. The MLR model performed the best, achieving the highest RI value of 0.51 in the NIR region. The lowest RI value was 0.11, recorded by the MLPR model in the SWIR region. The performance of the RFR and ABR models was also reduced, with RI values fluctuating from 0.24 and 0.26 (SWIR) to 0.40 and 0.44 (VNIR), respectively. The RI value for the KNNR model ranged from 0.29 (SWIR) to 0.44 (VNIR+SWIR).

The analysis reveals that during the training phase, the RFR and ABR models deliver high R^2 values, indicating strong model fit to the training data. However, their reduced performance in the testing phase suggests overfitting. In contrast, the MLR model, while not the top performer during training, exhibits a better balance between training and testing RI values, particularly VNIR+SWIR regions, indicating better generalizability. This mirrors the findings from TSS prediction, where MLR was also identified as a robust model with minimal overfitting.

3) Determining optimal configuration of multiple linear regression models for predicting TSS and turbidity

3.1) Variable selection

In regression analysis, the number of possible combinations of n factors refers to the number of different subsets of the n factors that can be formed. This includes all possible subsets, ranging from the empty set (no factors) to the set containing all n factors. The total number of subsets of a set with n factors is given by 2^n. This is because each factor can either be included or excluded from a subset, resulting in 2 possibilities per factor, and thus 2^n combinations in total. However, since the empty does not provide any factors for the model, it is excluded. Therefore, the number of meaningful combinations is 2^n-1.

In this study, nine bands were nine factors for TSS and turbidity models. Hence, 2^9-1 = 511 combinations were created. These combinations were examined to determine which bands contribute valuable information to the models using adjusted R^2. Additionally, indices such as R^2, AIC, and BIC were reported to compare models. Table 5 and table 6 report the five best combinations based on adjusted R^2 for TSS model and turbidity model, respectively.

Table 5 Best combinations for TSS model based on adjusted R^2

ID	B2	B3	B4	B5	B6	B7	B8	B11	B12	R2	R2a	AIC	BIC
246	0	1	1	1	1	0	1	1	1	0.880	0.870	-176.039	-156.041
190	0	1	0	1	1	1	1	1	1	0.880	0.870	-176.196	-156.197
430	1	1	0	1	0	1	1	1	1	0.881	0.871	-176.330	-156.332
174	0	1	0	1	0	1	1	1	1	0.879	0.871	-177.400	-159.902
182	0	1	0	1	1	0	1	1	1	0.880	0.872	-178.019	-160.521

Table 6 Best combinations for turbidity model based on adjusted R^2

ID	B2	B3	B4	B5	B6	B7	B8	B11	B12	R2	R2a	AIC	BIC
78	0	0	1	0	0	1	1	1	1	0.64	0.619	-80.91	-65.911
390	1	1	0	0	0	0	1	1	1	0.64	0.619	-80.991	-65.992
142	0	1	0	0	0	1	1	1	1	0.64	0.619	-80.996	-65.997
134	0	1	0	0	0	0	1	1	1	0.635	0.619	-81.830	-69.331
70	0	0	1	0	0	0	1	1	1	0.637	0.620	-82.145	-69.646

The tables present the ID for each combination, the presence (1) or absence (0) of each band (B2 to B12), and the corresponding statistical measures (R^2, adjusted R^2, AIC, and BIC). According to Table 4 and Table 5, the combination of bands that offer the best predictive power for TSS is B3, B5, B6, B8, B11, and B12, while the combination of bands that offer the best predictive power for Turbidity is B4, B8, B11, and B12. The corresponding R2, adjusted R2, AIC, and BIC for the TSS model are 0.880, 0.872, -178.019, and -160.521, respectively, and for the turbidity model are 0.637, 0.620, -82.145, and -69.646, respectively.

3.2) Detailed MLR models for predicting TSS and turbidity

The detailed MLR models for predicting TSS and turbidity (Table 7) include the coefficients, standard errors, t-values, and p-values for each selected band, and their intercept values. Notably, the intercept values for both models are zero because the input data was standardized. Thus, the resulted regression models are no-intercept multiple linear regression (NIMLR) models.

For the TSS model, all p-values of the variables in the TSS model are less than 0.05, meaning that each variable (B3, B5, B6, B8, B11, and B12) is statistically significant at the 5% significance level. This indicates a strong relationship between these bands and TSS, suggesting that changes in these bands are significantly associated with changes in TSS. This enhances confidence in the model's reliability and validity. The coefficients reveal that B3, B6, and B11 have inverse relationships with TSS, while B5, B8, and B12 have direct relationships.

For the turbidity model, all p-values of the variables are less than 0.05, except for B4, meaning that predictors B8, B11, and B12 are statistically significant at the 5% significance level. The p-value for B4 is 0.066, which is greater than 0.05, indicating that B4 is not statistically significant. This suggests that changes in B4 are not significantly associated with changes in Turbidity, and its inclusion may not contribute meaningful predictive power. Hence, B4 will be removed to improve the model's performance and eliminate unnecessary complexity.

In conclusion, the NIMLR for predicting TSS following Eq. 6 and the NIMLR model for predicting turbidity following Eq. 7.

With six variables, the NIMLR model for TSS outperforms than that for turbidity model. In fact, the performance metrics of TSS are $R^2 = 0.88$, MAE = 0.26 (mg L^{-1}), MSE = 0.12 (mg L^{-1}), and RMSE = 0.34 (mg L^{-1}). In contrast, the metrics of NIMLR model for turbidity, which uses 3 variables, are $R^2 = 0.62$, MAE = 0.47 (NTU), MSE = 0.37 (NTU), RMSE = 0.61 (NTU). Figure 7 shows the scatter plot of predicted values by models and the correspond observed values.

$$TSS = -2.264*B2+2.449*B5-1.305*B6+2.240*B8-1.164*B11+0.804*B12 \text{ (mg } L^{-1}) \quad \text{(Eq. 6)}$$

$$Turbidity = 0.958*B8-1.332*B11+1.089*B12 \text{ (NTU)} \quad \text{(Eq. 7)}$$

Table 7 Detailed MLR models for predicting TSS and turbidity

Detailed TSS model						Detailed turbidity model					
ID	Var.	Coef.	Std. err.	t-val	p-val	ID	Var.	Coef.	Std. err.	t-val	p-val
1	B3	-2.264	0.321	-7.061	0.000	1	B4	-0.458	0.246	-1.863	0.066
2	B5	2.449	0.449	5.460	0.000	2	B8	1.429	0.279	5.124	0.000
3	B6	-1.305	0.464	-2.814	0.006	3	B11	-1.464	0.375	-3.899	0.000
4	B8	2.240	0.327	6.852	0.000	4	B12	1.185	0.366	3.240	0.002
5	B11	-1.164	0.248	-4.696	0.000						
6	B12	0.804	0.238	3.377	0.001						

Table 8 Final detailed MLR model for predicting Turbidity

ID	Variable	Coef.	Std. err.	t-val	p-val
1	B8	0.958	0.12	7.995	0.0
2	B11	-1.332	0.374	-3.563	0.001
3	B12	1.089	0.367	2.964	0.004

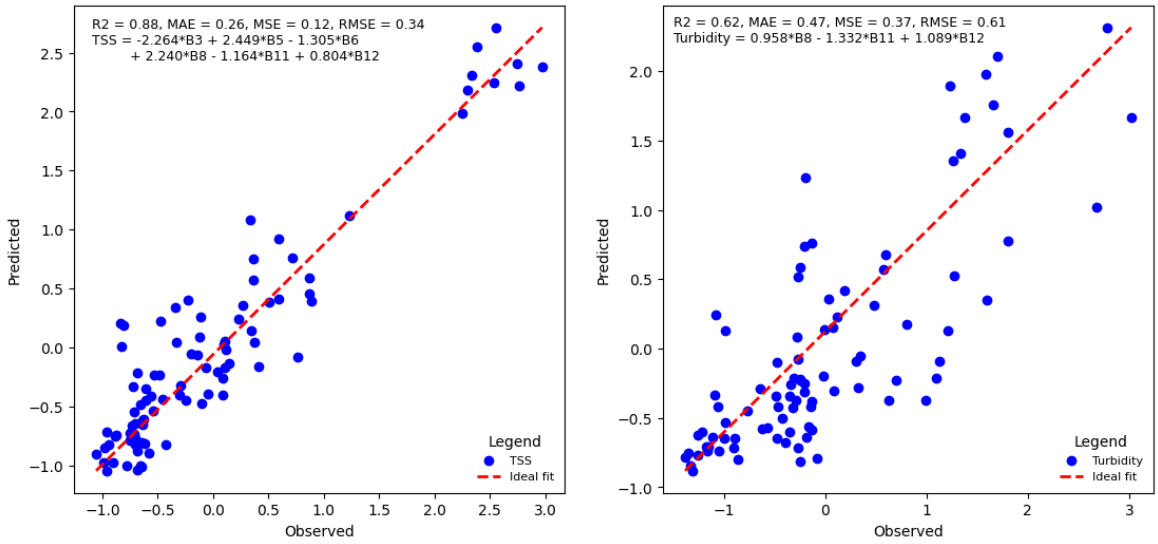

Figure 7 Scatter plots of NIMLR for TSS and turbidity.

Discussion

The results highlight that the VNIR+SWIR spectral region provided the highest RI values across all models, confirming its superior ability to capture information relevant to predicting TSS. This result is consistent with prior studies, which have shown that the combination of visible and shortwave infrared bands improves the accuracy of water quality models. The high performance of the MLR model in testing suggests it may be the most reliable model for generalizing predictions, despite other models like Random Forest and AdaBoost showing better performance during training. The balance between training and testing results for MLR indicates that it is less prone to overfitting, which is critical in predictive modeling applications where unseen data must be handled effectively.

The study demonstrates that integrating Sentinel-2 imagery with machine learning models is an effective approach for monitoring water quality, especially in regions where frequent in-situ measurements are impractical. The findings support the wider use of MLR and VNIR+SWIR spectral bands for predicting TSS in different water bodies, not only in Vietnam but in other regions facing similar environmental challenges. The use of spectral data for large-scale monitoring offers a scalable and efficient solution for managing water resources and detecting pollution in real-time.

While the VNIR+SWIR region was effective, the study's reliance on data from a single region (upper Ma River) limits the generalizability of the results to other geographical areas. Additionally, the models were trained on a relatively small dataset, which may affect their ability to generalize to significantly different environmental conditions. Future work could expand the dataset, incorporate more environmental variables, and apply more advanced techniques like regularization to mitigate overfitting in tree-based models like Random Forest and AdaBoost.

Conclusions

This study has demonstrated the potential of machine learning models for predicting TSS and turbidity using Sentinel-2 spectral data. The MLR model, particularly when combined with VNIR+SWIR bands, provided the most reliable predictions due to its balanced performance and minimal overfitting. These findings underscore the importance of selecting appropriate spectral regions for accurate water quality monitoring.

The use of MLR models with VNIR+SWIR bands can greatly enhance water quality monitoring efforts in Vietnam, especially in regions where traditional in-situ measurements are limited by cost and logistics. The ability to apply remote sensing data for continuous, large-scale environmental monitoring has significant implications for water resource management, pollution control, and early detection of water quality issues. Local environmental agencies and policy makers could incorporate these models to improve the management of river basins, mitigate the impacts of urbanization and industrialization, and respond to the challenges posed by climate change.

Future studies should focus on expanding the dataset to include a wider variety of environmental conditions and geographical regions. This will help to improve the generalizability of the models and further validate the findings across different ecosystems. Additionally, exploring hybrid models that integrate additional environmental variables, such as chlorophyll-a or water temperature, could enhance the predictive power of machine learning algorithms. Incorporating advanced techniques like regularization or deep learning methods may also mitigate overfitting in tree-based models, offering more robust predictions.

References

[1] Tomperi, J., Isokangas, A., Tuuttila, T., Paavola, M. Functionality of turbidity measurement under changing water quality and environmental conditions. Environmental Technology, 2022, 43(7), 1093–1101.

[2] Wu, J.L., Ho, C.R., Huang, C.C., Srivastav, A.L., Tzeng, J.H., Lin, Y.T. Hyperspectral sensing for turbid water quality monitoring in freshwater rivers: Empirical relationship between reflectance and Tturbidity and total solids. Sensors, 2014, 14(12), 22670–22688

[3] Adjovu, G.E., Stephen, H., James, D., Ahmad, S. Measurement of total dissolved solids and total suspended solids in water systems: A review of the issues, conventional, and remote sensing techniques. Remote Sensing, 2023, 15(14), 3534.

[4] Boyd, C.E., Water quality protection. *In*: Boyd, C.E. (ed.). Water quality: An introduction, Cham: Springer International Publishing, 2020, 379–409.

[5] Beaton, A.D., Schaap, A.M., Pascal, R., Hanz, R., Martincic, U., Cardwell, C.L., …, Mowlem, M.C. Lab-on-chip for in situ analysis of nutrients in the deep sea. ACS Sensors, 2022, 7(1), 89–98.

[6] Sagan, V., Peterson, K.T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B.A., …, Adams, C. Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. Earth-Science Reviews, 2020, 205, 103187.

[7] Ouma, Y.O., Noor, K., Herbert, K. Modelling reservoir chlorophyll-a, TSS, and turbidity using Sentinel-2A MSI and Landsat-8 OLI Satellite sensors with empirical multivariate regression. Journal of Sensors, 2020, 2020(1), 8858408.

[8] Wang, H., Wang, J., Cui, Y., Yan, S. Consistency of suspended particulate matter concentration in turbid water retrieved from Sentinel-2 MSI and Landsat-8 OLI sensors. Sensors, 2021, 21(5), 1662.

[9] Khan, R.M., Salehi, B., Mahdianpari, M., Mohammadimanesh, F. Water quality monitoring over finger lakes region using sentinel-2 imagery on google earth engine cloud computing platform. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2021, V-3-2021, 279–283.

[10] Yang, Y., Jin, S., Long-time water quality variations in the Yangtze river from Landsat-8 and Sentinel-2 images based on neural networks. Water, 2023, 15(21), 3802.

[11] Krohkaew, J., Nilaphruek, P., Witthayawiroj, N., Uapipatanakul, S., Thwe, Y., Crisnapati, P.N. Thailand raw water quality dataset analysis and evaluation. Data, 2023, 8(9).

[12] Ma, C., Zhao, J., Ai, B., Sun, S., Yang, Z. Machine learning based long-term water quality in the turbid Pearl river estuary, China. Journal of Geophysical Research: Oceans, 2022, 127(1), e2021JC018017.

[13] Ma, Y., Song, K., Wen, Z., Liu, G., Shang, Y., Lyu, L., …, Hou, J. Remote sensing of turbidity for lakes in northeast China using Sentinel-2 images with machine learning algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14, 9132–9146.

[14] Granata, F., Di Nunno, F., de Marinis, G. Advanced evapotranspiration forecasting in central Italy: Stacked MLP-RF algorithm and correlated Nystrom views with feature selection strategies. Computers and Electronics in Agriculture, 2024, 220, 108887.

[15] Khalifa, F.A., Abdelkader, H.M., Elsaid, A.H. An analysis of ensemble pruning methods under the explanation of Random Forest. Information Systems, 2024, 120, 102310.

[16] Son La Provincial People's Committee. Report on Socio-Economic Development of Son La Province. Son La. 2018

[17] Mello, K.D., Valente, R.A., Randhir, T.O., Vettorazzi, C.A. Impacts of tropical forest cover on water quality in agricultural watersheds in southeastern Brazil. Ecological Indicators, 2018, 93, 1293–1301.

[18] Bowers, D.G., Roberts, E.M., Hoguane, A.M., Fall, K.A., Massey, G.M., Friedrichs, C.T. Secchi disk measurements in turbid water. Journal of Geophysical Research: Oceans, 2020, 125, 5, e2020JC016172.

[19] Alparone, L., Arienzo, A., Garzelli, A. Spatial resolution enhancement of vegetation indexes via fusion of hyperspectral and multispectral satellite data. Remote Sensing, 2024, 16(5), 875.

[20] Patthi, S., Murali Krishna, V.B., Reddy, L., Arandhakar, S. Photovoltaic string fault optimization using multi-layer neural network technique. Results in Engineering, 2024, 22, 102299.

[21] Liu, S., Liu, Y., Xia, R., Using random forest to disentangle the effects of environmental conditions on height-to-diameter ratio of Engelmann spruce. New Forests, 2024, 55(2), 213–229.

[22] Hussain, S.S., Zaidi, S.S. AdaBoost Ensemble approach with weak classifiers for gear fault diagnosis and prognosis in DC Motors. Applied Sciences, 2024, 14(7), 3015.

[23] Sui, Q., Ghosh, S.K. Active learning for stacking and AdaBoost-related models. Stats, 2024, 7(1), 110–137.

[24] Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., Wang, X., Pu, L. Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. Ecological Indicators, 2021, 120, 106925.

[25] Li, S., Zhang, K., Chen, Q., Wang, S., Zhang, S. Feature selection for high dimensional data using weighted K-Nearest neighbors and genetic algorithm. IEEE Access, 2020, 8, 139512–139528.

[26] Oltra-Carriy, R., Baup, F., Fabre, S., Fieuzal, R. Briottet, X. Improvement of soil moisture retrieval from hyperspectral VNIR-SWIR data using clay content information: From laboratory to field experiments. Remote Sensing, 2015, 7(3), 3184–3205.