

## Remote Sensing Inversion of Cropland Soil Nutrients in the Huangshui River Basin Based on XGBoost Model: Postprint

**Authors:** Liu Zunfang, Haochuan Lei, Sheng Haiyan

**Date:** 2023-11-13T00:00:00+00:00

### Abstract

The Huangshui River Basin constitutes an important component of the Hehuang Valley, where predicting the spatial distribution of soil nutrients in conjunction with environmental factors is particularly crucial for agricultural soil nutrient management. Research on soil nutrient inversion has seldom investigated the impact of parameters on model outcomes and model applicability. This study selected 28 factors including topographic variables, soil pH, and spectral reflectance from the study area, and constructed three machine learning models—Artificial Neural Network (ANN), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost)—combined with Bayesian optimization algorithm to predict the spatial distribution of cultivated soil nutrients. The coefficient of determination ( $R^2$ ), root mean square error (RMSE), and relative percent deviation (RPD) were calculated to evaluate the accuracy of the three models. The results indicate that: (1) The XGBoost model based on Bayesian-optimized hyperparameters exhibited superior prediction accuracy for total nitrogen (TN) content compared to other models ( $R^2=0.893$ ,  $RMSE=0.359$ ,  $RPD=2.470$ ). When predicting soil organic matter (SOM), available phosphorus (AP), and available potassium (AK) contents, the XGBoost model achieved validation set  $R^2$  values of 0.801, 0.509, and 0.442, respectively. (2) Comparing the optimization iterations and errors across the three models revealed that the BOA-XGBoost model required fewer parameter optimization iterations, demonstrated higher efficiency, and exhibited superior robustness. For different nutrients, the prediction accuracies of ANN and SVM models varied: the SVM model achieved higher accuracy in predicting SOM content ( $RPD=1.580$ ), while the ANN model attained optimal accuracy in predicting TN ( $RPD=2.460$ ). The XGBoost model constructed through hyperparameter optimization based on the Bayesian algorithm demonstrated high prediction accuracy and can achieve favorable prediction performance,

providing a valuable reference for precision agricultural fertilization in the Huangshui River Basin.

## Full Text

# Remote Sensing Inversion of Soil Nutrients in Cultivated Land of Huangshui River Basin Based on XGBoost Model

LIU Zunfang<sup>1</sup>, LEI Haochuan<sup>1</sup>, SHENG Haiyan<sup>2</sup>

<sup>1</sup> Department of Geological Engineering, Qinghai University, Xining 810016, Qinghai, China

<sup>2</sup> College of Agriculture and Animal Husbandry, Qinghai University, Xining 810016, Qinghai, China

---

## Abstract

The Huangshui River Basin constitutes a vital component of the Huangshui Valley, where predicting the spatial distribution of soil nutrients in conjunction with environmental factors is particularly crucial for agricultural soil nutrient management. However, few studies have investigated the influence of model parameters on inversion results and the applicability of different models in soil nutrient research. This study selected the Huangshui River Basin in Qinghai Province as the research area and utilized 28 factors including terrain variables, soil properties, and spectral reflectance. Combining these with the Bayesian optimization algorithm, we constructed three machine learning models—artificial neural network (ANN), support vector machine (SVM), and extreme gradient boosting (XGBoost)—to predict the spatial distribution of four farmland soil nutrients: soil organic matter (SOM), total nitrogen (TN), available phosphorus (AP), and available potassium (AK). Model accuracy was evaluated using the coefficient of determination ( $R^2$ ), root mean square error (RMSE), and relative percent deviation (RPD). The results revealed that all four nutrients exhibited moderate variability, with TN showing the highest coefficient of variation at 69.481%. The XGBoost model based on Bayesian-optimized hyperparameters outperformed other models in predicting TN content, achieving  $R^2$ , RMSE, and RPD values of 0.893, 0.359, and 2.470, respectively. For SOM, AP, and AK prediction, the XGBoost model validation set  $R^2$  values were 0.801, 0.509, and 0.442, with corresponding RPD values of 2.152, 1.274, and 1.210, respectively, demonstrating superior predictive capability. Comparative analysis of optimization iterations and errors across the three models indicated that the Bayesian-optimized XGBoost model required fewer parameter optimization steps, exhibited higher efficiency, and demonstrated better robustness. For different nutrients, model prediction accuracy varied; the SVM model showed high accuracy for SOM content prediction (RPD = 1.580), while the ANN model performed efficiently for TN prediction (RPD = 2.460). The XGBoost inversion

model developed by integrating 28 environmental factors from the Huangshui River Basin with Landsat 8 remote sensing imagery proved more suitable for soil nutrient inversion research. This model can accurately characterize the spatial distribution patterns of soil nutrients in the basin, providing valuable support for precision agriculture fertilization strategies that improve fertilizer utilization efficiency and crop yields.

**Keywords:** soil nutrient; XGBoost; spatial distribution; environmental factor; Huangshui River Basin

---

## 1 Introduction

Soil organic matter (SOM), total nitrogen (TN), available phosphorus (AP), and available potassium (AK) are essential nutrients for vegetation growth and serve as critical indicators for evaluating soil fertility levels. Fertile soils provide abundant nutrients that promote crop growth, whereas nutrient-poor soils limit crop development and consequently affect yield. Therefore, spatial prediction and mapping of soil nutrients in cultivated land represent fundamental requirements for sustainable agriculture and play a significant role in guiding variable-rate fertilization and improving soil quality. Traditional soil nutrient detection methods are time-consuming, expensive, and unsuitable for large-scale spatial distribution mapping. With the advancement of remote sensing technology, widespread application in predicting nutrient spatial distribution and assessing soil fertility has enabled large-area, high-efficiency monitoring of soil nutrient content. This technology builds inversion models by analyzing spectral information and extracting sensitive spectral bands based on limited ground-measured data, thereby achieving spatial distribution prediction of nutrient content across extensive areas.

Numerous machine learning models have been combined with remote sensing imagery for soil nutrient spatial distribution inversion. For instance, Li et al. constructed an RF model for the Huangshui River Basin using different particle size fractions of Landsat imagery as objects, achieving higher prediction accuracy than partial least squares models with validation set RPD values exceeding 2, demonstrating suitability for TN content prediction in the basin. Lei et al. used Landsat imagery as a data source, applying reciprocal transformation to band reflectance as model independent variables to construct a neural network model for inverting the spatial distribution pattern of TN in Datong County of the Huangshui River Basin, obtaining a model  $R^2$  of 0.676 and RMSE of 0.359. Additionally, Xiao Yunfei inverted soil AP content in the Huangshui River Basin based on random forest models, achieving RPD values of 1.82. Beyond these models, extreme learning machines, multiple regression, and geographically weighted regression kriging have also demonstrated high prediction accuracy and found wide application in soil nutrient prediction research.

Current research employs both linear and nonlinear models for soil nutrient

spatial distribution studies, with machine learning models consistently outperforming linear models and thus proving more suitable for soil nutrient inversion research to improve prediction accuracy. Soil nutrient spatial distribution is influenced by multiple factors, and machine learning model parameter settings significantly affect prediction accuracy. However, existing studies predominantly determine model parameters through manual trial-and-error adjustment, introducing arbitrariness in parameter selection. The Bayesian optimization algorithm represents an efficient global intelligent optimization method that converges on optimal solutions faster than alternative optimization algorithms. While this algorithm has been applied to parameter optimization for soil salinity inversion studies—for example, Yang et al. optimized random forest model parameters for soil salinity inversion and Wang et al. optimized LightGBM models with high resulting accuracy—its application in soil nutrient spatial distribution research remains limited. Therefore, this study selected the Huangshui River Basin in Qinghai Province as the research area, used correlation analysis to select modeling factors, and employed Landsat 8 multispectral remote sensing imagery and field-measured farmland soil nutrient data to construct ANN, SVM, and XGBoost models for farmland soil nutrient inversion. Bayesian optimization was applied to model parameters to improve prediction accuracy, and the optimal model for nutrient spatial prediction in the basin was identified to provide foundational support and technical guidance for precision agriculture implementation.

## 2 Materials and Methods

### 2.1 Soil Sample Collection and Data Sources

This study comprehensively considered terrain conditions, sampling accessibility, and transportation convenience across the research area. A total of 120 farmland soil samples were collected in spring 2021 before cultivation, targeting fields without crop residues or bare land. Sampling depth was 0–20 cm, with each sample comprising a mixture of soil collected from 5 subsampling points within a 10 m radius. Handheld GPS receivers recorded sampling point coordinates, and detailed environmental information near sampling sites was documented. During data preprocessing, 8 soil samples were eliminated due to outliers, resulting in 112 valid farmland soil samples for analysis. Laboratory analysis determined SOM, TN, AP, AK contents, and soil pH values.

Landsat 8 OLI remote sensing imagery was obtained from the Geospatial Data Cloud (<https://www.gscloud.cn/>) and USGS EarthExplorer (<https://earthexplorer.usgs.gov/>), with 30 m spatial resolution, acquisition dates in May 2021, and cloud cover less than 5%. Preprocessing included atmospheric correction, mosaicking, and clipping. ArcGIS 10.8 was used to extract six terrain factors: slope, aspect, relief degree of land surface, plane curvature, profile curvature, and topographic wetness index.

## 2.2 Research Methods

The technical route is illustrated in Figure 2. First, soil nutrient modeling factors were extracted from Landsat 8 data. Pearson correlation analysis was applied to screen factors based on absolute correlation coefficient values. Subsequently, ANN, SVM, and XGBoost models were constructed for farmland soil nutrient inversion, with Bayesian optimization applied to model parameters. The three models' inversion accuracies were compared to select the optimal model for predicting soil nutrient spatial distribution in the Huangshui River Basin.

**2.2.1 Inversion Models** **Artificial Neural Network (ANN)** simulates human brain neural network information transmission, processing fuzzy and nonlinear information through interactions among neurons. The model comprises an input layer, hidden layers, and output layer, with numerous neurons connected by weights between layers. The hidden layer can be multilayered, while the input and output layers each contain one layer. Input variables include environmental factors and band reflectance, while the output layer represents nutrient content values. Maximum iteration count, hidden layer neuron quantity, and activation functions were determined by Bayesian optimization.

**Support Vector Machine (SVM)** is a machine learning algorithm based on statistical theory that effectively addresses small-sample, nonlinear problems. Its core technology lies in kernel function selection, with four common types: polynomial kernel, Sigmoid kernel, linear kernel, and radial basis function (RBF). RBF is most commonly used in soil mapping research. This study employed RBF kernel, with Bayesian optimization determining penalty coefficient and gamma parameters.

**Extreme Gradient Boosting (XGBoost)** is an emerging ensemble learning algorithm whose core principle involves generating a tree at each iteration by learning a new function to fit residuals from the previous iteration. Building upon gradient boosting trees, XGBoost retains more objective function information, accelerates training speed, and incorporates regularization terms in the objective function to control model complexity and reduce overfitting, thereby effectively improving the gradient boosting tree algorithm. Model parameters including learning rate and tree depth were optimized using Bayesian optimization.

Compared with traditional random search and grid search algorithms, Bayesian optimization considers the performance of previously searched parameters, reducing optimization iterations and improving efficiency. The algorithm proceeds as follows: (1) Randomly select parameter combinations  $x$ , train models to obtain corresponding metrics  $y$ ; (2) Fit  $x$  and  $y$  using a surrogate function; (3) Select optimal parameter combination  $x^*$  using an acquisition function; (4) Input  $x^*$  into the model to obtain new  $y$ , repeating steps 2–4 until optimal parameters are determined.

**2.2.2 Evaluation Metrics** This study employed  $R^2$ , RMSE, and RPD as evaluation metrics. Smaller RMSE and larger  $R^2$  values indicate smaller deviations and higher correlations between predicted and measured nutrient values, reflecting higher model prediction accuracy and better representation of farmland soil nutrient spatial distribution characteristics. When  $RPD > 2$ , prediction accuracy is high; when  $1.4 < RPD < 2$ , prediction accuracy is moderate; when  $RPD < 1.4$ , prediction accuracy is low. The formulas are:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$
$$RPD = \frac{SD}{RMSE}$$

where  $n$  represents the total number of soil samples,  $y_i$  is the measured nutrient content,  $\hat{y}_i$  is the model-predicted nutrient content,  $\bar{y}$  is the mean nutrient value, and  $SD$  is the validation set standard deviation.

## 3 Results

### 3.1 Conventional Statistical Analysis of Soil Nutrients

Descriptive statistical analysis of soil nutrients in the Huangshui River Basin (Table 1) revealed that SOM content ranged from 4.310 to 59.789  $\text{g} \cdot \text{kg}^{-1}$  with a mean of 28.377  $\text{g} \cdot \text{kg}^{-1}$ ; TN content ranged from 0.372 to 5.874  $\text{g} \cdot \text{kg}^{-1}$  with a mean of 1.232  $\text{g} \cdot \text{kg}^{-1}$ ; AP content ranged from 0.001 to 0.162  $\text{g} \cdot \text{kg}^{-1}$  with a mean of 0.055  $\text{g} \cdot \text{kg}^{-1}$ ; and AK content ranged from 0.048 to 0.488  $\text{g} \cdot \text{kg}^{-1}$  with a mean of 0.204  $\text{g} \cdot \text{kg}^{-1}$ . Coefficients of variation were 39.356%, 69.481%, 58.182%, and 42.647% for SOM, TN, AP, and AK, respectively, indicating moderate variability for all four nutrients.

### 3.2 Correlation Analysis Between Soil Nutrients and Environmental Variables/Band Reflectance

Correlation analysis between nutrients and environmental variables showed that SOM and TN contents were significantly correlated with elevation, aspect, profile curvature, topographic wetness index, and relief degree ( $P < 0.05$ ), but not significantly correlated with slope or plane curvature. AP and AK showed higher correlations with environmental factors, exhibiting similar correlation patterns. Higher elevation areas in the study region contained greater SOM and TN contents. AK content showed significant negative correlation with slope—as slope increased and soil acidity decreased, AK content decreased, consistent with findings from Dai Zijun and others.

Correlation analysis between nutrients and Landsat 8 band reflectance (Table 3) indicated that all nutrients correlated with both original reflectance (b1–b7) and mathematically transformed reflectance (logarithmic and reciprocal processing). Compared with original band reflectance, mathematically transformed reflectance showed improved correlations with nutrients. For example, the correlation coefficient between SOM and b1 was -0.267, which increased to -0.495 after reciprocal transformation—a 57.88% improvement in absolute correlation value. Overall, TN showed the best correlation with reflectance, with significant correlations ( $r = -0.169$  to  $-0.151$ ,  $P < 0.05$ ) observed between multiple bands and TN content.

### 3.3 Model Construction and Optimization

Based on the above modeling factors and measured farmland nutrient data, inversion models were constructed and optimized using Bayesian optimization. The optimal parameter combination was selected to minimize model prediction error. As optimization iterations increased, model errors stabilized, indicating optimal parameter values. Comparative analysis revealed that the XGBoost model required the fewest optimization iterations to find optimal parameter combinations and achieved the lowest final prediction error for soil nutrient spatial distribution (Figure 3).

Soil nutrient spatial distribution is influenced by both anthropogenic factors (fertilization, irrigation) and natural factors (terrain, soil type). Modeling factor selection significantly affects inversion accuracy. Therefore, considering terrain, original band reflectance, and mathematically transformed reflectance, factors showing extremely significant correlations with the four nutrients were prioritized as modeling inputs. To avoid factor redundancy within the same category, the segmented maximum value method was applied to select modeling factors for each nutrient—that is, selecting the most highly correlated factors from each of the three categories (environmental, original reflectance, transformed reflectance). Due to differences in nutrient content, properties, and correlations with the three factor categories, modeling factor selection varied among nutrients (Table 4).

During Bayesian optimization, 5-fold cross-validation determined final parameter values. Table 5 presents the optimal hyperparameter combinations for the three models.

### 3.4 Model Accuracy Comparison

Based on optimal hyperparameter values, 70% of samples were randomly selected for model training, with the remaining 30% used for validation. Table 6 compares model prediction accuracy. The XGBoost model demonstrated significantly higher inversion accuracy than ANN and SVM models. For TN content prediction, XGBoost achieved  $R^2 = 0.893$ ,  $RMSE = 0.359$ , and  $RPD = 2.470$ , representing improvements of 0.092 and 0.141 in  $R^2$  compared with ANN and

SVM models, respectively. For SOM, AP, and AK prediction, XGBoost validation set  $R^2$  values were 0.801, 0.509, and 0.442, with corresponding RPD values of 2.152, 1.274, and 1.210, respectively. Although all three models showed relatively low validation set  $R^2$  values for AP and AK, XGBoost maintained the highest  $R^2$  values, indicating significantly superior accuracy.

### 3.5 Spatial Prediction of Soil Nutrients

The optimal inversion model was applied to spatially predict the four farmland soil nutrients, with distribution maps generated (Figure 4). The predictions revealed differences among models in inverting the same nutrient content. SOM and TN showed high mutual correlation with similar overall spatial distribution trends—high contents in the northwest and low contents in the southeast. Using the XGBoost model, SOM content ranged from 7.91 to 55.74  $\text{g} \cdot \text{kg}^{-1}$ , with the lowest value of 7.91  $\text{g} \cdot \text{kg}^{-1}$  and highest value of 55.74  $\text{g} \cdot \text{kg}^{-1}$ . TN content ranged from 0.07 to 0.42  $\text{g} \cdot \text{kg}^{-1}$ , with the lowest value of 0.07  $\text{g} \cdot \text{kg}^{-1}$  and highest value of 0.42  $\text{g} \cdot \text{kg}^{-1}$ . AP and AK contents were relatively low overall, with minimal spatial variation. AP content ranged from 0.04 to 0.11  $\text{g} \cdot \text{kg}^{-1}$ . In summary, XGBoost predictions closely matched measured data ranges and effectively reproduced the actual spatial distribution of farmland soil nutrients in the study area.

## 4 Discussion

The XGBoost model is an emerging ensemble learning algorithm characterized by fast training speed and high fitting accuracy, which has been successfully applied in salinity inversion and winter wheat TN content estimation. However, its application in soil nutrient inversion research remains limited. This study utilized XGBoost to invert farmland soil nutrient content in the Huangshui River Basin and evaluated its applicability. Results showed that different algorithms produced varying modeling effects for the same nutrient, with XGBoost achieving the highest prediction accuracy. The model effectively captured local spatial distribution characteristics, as demonstrated in detailed subregion maps (Figure 5), where XGBoost-produced images showed clearer details and better reflected SOM spatial distribution compared with true-color composite remote sensing images.

The three models exhibited different prediction accuracies for different nutrients. XGBoost showed high accuracy for TN (RPD = 2.470) and SOM (RPD = 2.152), while SVM performed well for SOM (RPD = 1.580) and ANN for TN (RPD = 2.460). Other scholars have estimated nutrients in the Huangshui River Basin using different methods. For example, Hu et al. used random forest to predict SOM with validation  $R^2 = 0.676$  and RPD = 1.82, lower than the XGBoost accuracy in this study ( $R^2 = 0.801$ , RMSE = 4.321, RPD = 2.152). Spatial interpolation methods produce nutrient distribution maps that lack detail, whereas remote sensing combined with machine learning can compensate

for this limitation, though inversion accuracy depends on image spatial resolution.

Soil nutrient distribution exhibits spatial autocorrelation and heterogeneity, influenced by multiple factors. This study analyzed correlations between nutrients and environmental factors, finding that elevation, aspect, and relief degree were the primary factors affecting nutrient spatial distribution. Previous research indicates that prediction accuracy is low when using only environmental variables or only multispectral band reflectance. Zhang et al. found that adding environmental factors to reflectance data improved model accuracy. Mathematical transformation of spectral reflectance can effectively enhance correlations with soil nutrients, facilitating higher-precision model development. This study's reciprocal transformation of b1 increased its correlation with SOM from -0.267 to -0.495, consistent with findings from Sun Mingyue and Du Jun.

Multiple factors collectively determine soil nutrient spatial distribution. This study's selection of auxiliary factors was limited, and future work should focus on 协同 inversion under combined environmental and anthropogenic conditions, as well as the advantages and disadvantages of different optimization algorithms for parameter selection. To accurately predict micronutrient spatial distribution, hyperspectral remote sensing imagery should be considered to extract bands significantly correlated with nutrients, thereby improving prediction capability and providing references for precision agriculture fertilization in the Huangshui River Basin.

## 5 Conclusions

1. All four nutrients in the Huangshui River Basin showed moderate variability, with TN exhibiting the highest variation (69.481%). SOM and AK showed the greatest variability, while TN and AP were at moderate levels.
2. Integrating 28 environmental factors, band reflectance, and derived reflectance from Landsat 8 imagery, the study compared three machine learning models and found XGBoost achieved the highest accuracy for TN content prediction ( $R^2 = 0.893$ , RMSE = 0.359, RPD = 2.470). Prediction accuracy for SOM, AP, and AK was also superior with XGBoost, with validation set  $R^2$  values of 0.801, 0.509, and 0.442, and RPD values of 2.152, 1.274, and 1.210, respectively. The ANN and SVM models showed lower predictive capability.
3. The spatial distribution patterns of SOM and TN predicted by the XGBoost model were consistent, showing generally high values in the northwest and low values in the southeast, with relatively low values in the central region. AP and AK contents were low overall with minimal variation trends. The Bayesian-optimized XGBoost model achieved high accuracy in inverting soil nutrient spatial distribution patterns in the Huangshui River Basin, effectively reflecting local distribution characteristics and proving more suitable for soil nutrient inversion research.

## References

- [1] Wang R M, Zou R Y, Liu J M, et al. Spatial distribution of soil nutrients in farmland in a hilly region of the Pearl River Delta in China based on geostatistics and the inverse distance weighting method[J]. *Agriculture*, 2021, 11(1): 50, doi: 10.3390/agriculture11010050.
- [2] Alemu L, Mesfin B. Performance of mid infrared spectroscopy to predict nutrients for agricultural soils in selected areas of Ethiopia[J]. *Heliyon*, 2022, 8(3): e09050, doi: 10.1016/j.heliyon.2022.e09050.
- [3] Li Guanwen, Gao Xiaohong, Xiao Nengwen, et al. Estimation soil organic matter contents with hyperspectra based on sCARS and RF algorithms[J]. *Chinese Journal of Luminescence*, 2019, 40(8): 1030-1039.
- [4] Yang Xiaoyu, Bao Nisha, Cao Yue, et al. Estimation and mapping of soil nutrient in farmland based on UAV imaging spectrometry[J]. *Geography and Geo-information Science*, 2021, 37(5): 38-45.
- [5] Zheng Mandi, Xiong Heigang, Qiao Juanfeng, et al. Hyperspectral based estimation model about organic matter in desert soil at different levels of human disturbance[J]. *Arid Land Geography*, 2018, 41(2): 384-392.
- [6] Liu Huanjun, Zhang Meiwei, Yang Haoxuan, et al. Inversion of cultivated soil organic matter content combining multi-spectral remote sensing and random forest algorithm[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2020, 36(10): 134-140.
- [7] Miran N, Rasouli Sadaghiani M H, Feiziasl V, et al. Predicting soil nutrient contents using Landsat OLI satellite images in rain-fed agricultural lands, northwest of Iran[J]. *Environmental Monitoring and Assessment*, 2021, 193(9): 607, doi: 10.1007/s10661-021-09067-1.
- [8] Zheng Miao, Wang Xiang, Li Sijia, et al. Remote sensing inversion of soil organic matter and total nitrogen in black soil region[J]. *Scientia Geographica Sinica*, 2022, 42(8): 1336-1347.
- [9] Gulhane V A, Rode S V, Pande C B. Wavelet for predicting soil nutrients using remotely sensed satellite images[J]. *International Journal of Computer Applications*, 2017, 174(4): 35-38.
- [10] Li Guanwen, Gao Xiaohong, Yang Lingyu, et al. Estimating soil organic matter contents from different soil particle size using visible and near infrared reflectance spectrum: A case study of the Huangshui Basin[J]. *Chinese Journal of Soil Science*, 2017, 48(6): 1360-1370.
- [11] Zhang Huan, Gao Xiaohong. Analysis of spatial variability and mapping of soil organic matter contents in complex terrain areas[J]. *Research of Soil and Water Conservation*, 2020, 27(5): 93-100.
- [12] Liu Z F, Lei H C, Lei L, et al. Spatial prediction of total nitrogen in soil

surface layer based on machine learning[J]. *Sustainability*, 2022, 14(19): 11998, doi: 10.3390/su141911998.

[13] Dharumarajan S, Lalitha M, Niranjana K, et al. Evaluation of digital soil mapping approach for predicting soil fertility parameters: A case study from Karnataka Plateau, India[J]. *Arabian Journal of Geosciences*, 2022, 15(5): 386, doi: 10.1007/s12517-022-09629-8.

[14] Yang Lianbing, Chen Chunbo, Zheng Hongwei, et al. Retrieval of soil salinity content based on random forests regression optimized by Bayesian optimization algorithm and genetic algorithm[J]. *Journal of Geo-information Science*, 2021, 23(9): 1662-1674.

[15] Lei Haochuan, Liu Zunfang, Yu Xiaojing, et al. Spatial pattern inversion of soil surface total nitrogen in Datong County of Qinghai Province based on Landsat 5 image[J]. *Journal of Qinghai University*, 2021, 39(6): 79-86.

[16] Xiao Yunfei. Estimation of soil organic matter, total carbon and total nitrogen by hyperspectral remote sensing in Qinghai Province[D]. Xining: Qinghai Normal University, 2019.

[17] Liu F, Qin T L, Yan D H, et al. Classification of instream ecological water demand and crucial values in a semi-arid river basin[J]. *Science of the Total Environment*, 2020, 712: 136409, doi: 10.1016/j.scitotenv.2019.136409.

[18] Dong B Q, Qin T L, Wang Y, et al. Spatiotemporal variation of nitrogen and phosphorus and its main influencing factors in Huangshui River Basin[J]. *Environmental Monitoring and Assessment*, 2021, 193(5): 292, doi: 10.1007/s10661-021-09067-1.

[19] Ruhollah T M, Ram N, Sood K, et al. Artificial bee colony feature selection algorithm combined with machine learning algorithms to predict vertical and lateral distribution of soil organic matter in South Dakota, USA[J]. *Carbon Management*, 2017, 8(3): 277-291.

[20] Yang Xin, Yuan Ziran, Ye Yin, et al. Winter wheat total nitrogen content estimation based on UAV hyperspectral remote sensing[J]. *Spectroscopy and Spectral Analysis*, 2022, 42(10): 3269-3274.

[21] Miao J, Zhen J N, Wang J J, et al. Mapping seasonal leaf nutrients of mangrove with Sentinel-2 images and XGBoost method[J]. *Remote Sensing*, 2022, 14(15): 3679, doi: 10.3390/rs14153679.

[22] Komolafe A A, Olorunfemi I E, Oloruntoba C, et al. Spatial prediction of soil nutrients from soil, topography and environmental attributes in the northern part of Ekiti State, Nigeria[J]. *Remote Sensing Applications: Society and Environment*, 2021, 21: 100450, doi: 10.1016/j.rsase.2020.100450.

[23] Ji W S, Liu Y Q. Research on quantitative evaluation of remote sensing and statistics based on wireless sensors and farmland soil nutrient variability[J]. *Computational Intelligence and Neuroscience*, 2022, 2022: 3646264, doi:

10.1155/2022/3646264.

[24] Hu Guigui, Yang Fenli, Yang Lian-an, et al. Spatial prediction modeling of soil organic matter content based on principal components and machine learning[J]. *Arid Land Geography*, 2021, 44(4): 1114-1124.

[25] Dai Zijun. Spatio-temporal variation of soil nutrients in Huangshui River Basin and its affecting factors in the past 30 years[D]. Xining: Qinghai Normal University, 2018.

[26] Fang Huiting, Meng Jihua, Cheng Zhiqiang. Spatio-temporal variability of soil available nutrients based on remote sensing and crop model[J]. *Scientia Agricultura Sinica*, 2019, 52(3): 478-490.

[27] Liu Zunfang, Lei Haochuan, Lei Lei. Analysis on spatial distribution of soil organic matter and available phosphorus in Huangshui Basin[J]. *Science Technology and Engineering*, 2022, 22(34): 15095-15102.

[28] Li Yingying, Zhao Zhengyong, Yang Qi, et al. Prediction of soil organic matter content based on artificial neural network model and GF-1 remote sensing data[J]. *Soils*, 2022, 54(1): 191-197.

[29] Sun Mingyue. Inversion of soil salinity and organic matter on arable land in the Yellow River Delta based on multi-source remote sensing[D]. Tai'an: Shandong Agricultural University, 2022.

[30] Du Jun, Gu Junwei, Qiu Shike, et al. Prediction of total nitrogen by using hyperspectral data based on support vector regression[J]. *Henan Science*, 2020, 38(10): 1585-1590.

[31] Chen B L, Zheng H W, Luo G P, et al. Adaptive estimation of multi-regional soil salinization using extreme gradient boosting with Bayesian TPE optimization[J]. *International Journal of Remote Sensing*, 2022, 43(3): 778-811.

[32] Wang L Y, Hu P, Zheng H W, et al. Integrative modeling of heterogeneous soil salinity using sparse ground samples and remote sensing images[J]. *Geoderma*, 2023, 430: 116321, doi: 10.1016/j.geoderma.2022.116321.

---

### Figure and Table Captions:

[Figure 1: see original paper] Distribution of soil sampling points in the study area

[Figure 2: see original paper] Flowchart of the nutrient inversion and analysis

[Figure 3: see original paper] Model optimization results

[Figure 4: see original paper] Spatial distributions of soil nutrients by three models

[Figure 5: see original paper] Local spatial distribution characteristics of SOM

Routine statistical analysis results of soil nutrients in the Huangshui River Basin

Correlation analysis between soil nutrients and environmental variables

Correlation analysis between soil nutrients and band reflectance

Soil nutrient modeling factors

Optimal hyperparametric combination of the three models

Precision comparison of different nutrient inversion models

*Note: Figure translations are in progress. See original paper for figures.*

*Source: ChinaXiv — Machine translation. Verify with original.*