Research article: Spatial Prediction of Soil pH Values in Hunan Province Based on Multi-source Environmental Variables and Machine Learning Models.

Pengyang Li

Beijing No. 80 High School, Beijing, China

lipengyang2007@163.com

Abstract. This study aims to use multi-source environmental variables and machine learning models to spatially predict the soil pH values in Hunan Province, China. The study area is a main rice producing region of China, also known for its warm and humid climate, as well as its diverse soil types. The research employs three machine learning models: Random Forest(RF), Support Vector Machine(SVM), and Gradient Boosting Regressor Trees(GBRT), in conjunction with field soil sampling and remote sensing data, to predict soil pH values. The predictions are compared with the traditional Ordinary Kriging (OK) model. Through model evaluation, the best model is selected for spatial distribution prediction of soil pH, and SHAP analysis is used for importance analysis to determine the main factors affecting the pH of the red soil in the southern region. Our results showed that soils in the southeast and southwest of Hunan exhibited lower pH than those in northern regions, and land use is the most important factor affecting soil pH values in Hunan Province.

Keywords: Soil pH Value; Machine Learning; Spatial Prediction; Hunan Province.

1. Introduction

Soil is the material basis for ensuring people's basic livelihood, and the soil pH value is an important indicator reflecting the soil's acidity or alkalinity, which directly affects crop growth and soil fertility. Studying the spatial distribution of soil pH values is of great significance for agricultural production and soil health management. Many researchers choose spatial interpolation methods for predicting spatial information of soil properties, such as Kriging interpolation.

Hunan Province is located in the central part of China, with a complex and variable climate and topography, and a rich variety of soil types. Hunan is also the main rice producing region of China, In 2023, rice grain output was 61.3 billion catties, both rice growth area and rice production ranked the first in China. However, soil acidification has become an important factor limiting rice production in Hunan. This article takes Hunan Province in China as an example to compare the efficiency of three machine learning models—Random Forest, Support Vector Machine, and Gradient Boosting Regressor Trees—in learning the distribution of soil pH in Hunan Province. The best machine learning model is ultimately used to predict the spatial distribution of soil pH and is compared with the traditional Ordinary Kriging (OK) model. Subsequently, the importance of factors affecting the pH of red soil in the south is analyzed.

2. Materials and Methods

1. Regional overview:

Hunan Province (108°47' to 114°15'E, 24°38' to 30°08'N) is located in the central part of China, in the middle reaches of the Yangtze River. The total area of the province is 211,800 square kilometers, accounting for 2.2% of the total area of the country. Hunan has a subtropical monsoon climate, with distinct seasons, a warm and humid climate, and an annual precipitation of 1300-1800 mm, decreasing from southeast to northwest. The average annual temperature is 16-18°C[2]. The terrain and landforms of Hunan Province are complex, surrounded by mountains on the east, south, and west sides, with undulating hills in the central part, and the Dongting Lake Plain in the north. The parent materials of the soil in Hunan Province are complex and variable, resulting in a variety of soil types,

including red soil, paddy soil, tidal soil, lime soil, purple soil, and 13 other types of soil[1], among which paddy soil accounts for 16.50% of the total soil area in Hunan Province[1]. It is distributed in plains, low-lying hills, and some mountainous areas, mainly concentrated in the Dongting Lake Plain, Xiangjiang, Zijiang, Yuanjiang, Lishui, and other major river basins as well as valley plains of small and medium-sized rivers, providing unique basic resources for the agricultural economy of Hunan Province, such as rice planting and grain production.

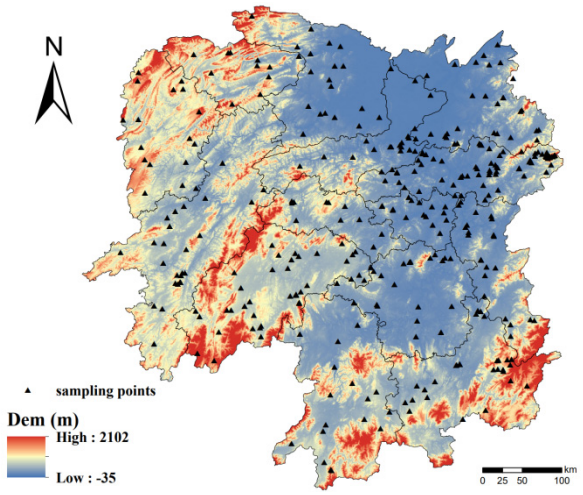


Fig 1. Sampling points in study area

2. Data source: The data for this study comes from field soil sampling and remote sensing data. The final database consists of 15 environmental variables and 1 target variable (soil pH). The distribution of sampling points is shown in Fig. 1.

2.1 Soil Data:

The sampling depth is 0 to 20 cm. Sourced from "Soil Series of China: Central and Western Regions, Hunan Volume", a total of 188 sample points were collected in the study area from 2014 to 2018, following the principles of randomness, uniformity, and representativeness. Additionally, soil data was retrieved from the Web of Science and CNKI for Hunan, which includes: (pH, CEC, SOC, total nitrogen, total potassium, total phosphorus, soil type, soil use type).

2.2 Topographic Data:

The Digital Elevation Model (DEM) was derived from SRTM data with a spatial resolution of 30 meters. ArcGIS 10.2 was used to calculate the slope, aspect, and Terrain Wetness Index (TWI).

2.3 Normalized Differential Vegetation Index (NDVI):

from National Ecosystem Science Data Center (nesdc.org.cn), with a spatial Resolution of 30m.

2.4 Climate Data:

It mainly includes the mean annual temperature (MAT) and the mean annual precipitation (MAP). The data comes from the National Earth System Science Data Center (www.geodata.cn/main/), with a spatial resolution of 1 km.

2.5 Environmental Quality Index:

The concentration of SO2, National Tibetan Plateau Data Center (https://data.tpdc.ac.cn/home), with a spatial resolution of 1km.

All environmental variables and soil property spatial prediction results are unified to a resolution of 1km using the nearest neighbor interpolation method.

1. Research method: Based on the collected data from 368 soil sample points, they were divided into training and testing sets in a ratio of 8:2. Three machine learning models, Random Forest(RF), Support Vector Machine(SVM), and Gradient Boosting Regressor Trees(GBRT), were used to predict the spatial distribution of soil pH in Hunan Province. Using the three models for data modeling, and then inputting the normalized features and pH into the three models to select the best model.

The performance of the model was evaluated using Coefficient of Determination (R2) and Root Mean Square Error (RMSE). Higher R2 and RMSE values close to zero indicate that the model's predictions have higher accuracy. The calculation formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y - y_i)^2}{\sum_{i=1}^{n} (y - \bar{y})^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y - y_i)^2}{n}} \tag{2}$$

In this formula, n is the number of samples, y is the true value of the data, y_i is the predicted value, $\bar{y}$ is the average of the true value.

After evaluation, the best model was selected to predict the spatial distribution map, and the prediction results were compared with those from the OK model.

Using SHAP analysis for importance analysis, the most influential factors affecting the pH value were identified.

3. Results and Analyses

3.1 Environmental Factors Selection

By comparing the correlation coefficients between soil pH and environmental factors as well as the correlation coefficients between environmental factors and environmental factors, 11 influencing factors of SOC, TN, TP, NDVI max, TW, precipitation, temperature, land use, land type, and SO_2 were finally identified as the final environmental factors.

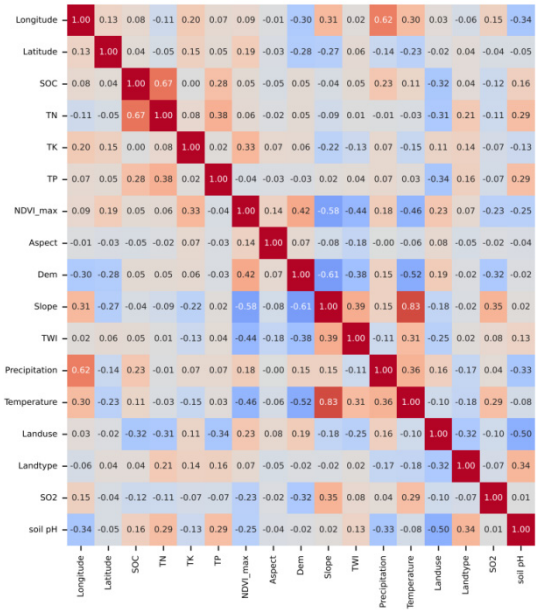


Fig 2. Correlation coefficient between soil pH and environmental factors

3.2 Optimal Hyperparameters Analysis of Three Models

Using grid search, the optimal hyperparameters for each model were determined during the training phase through 5-fold cross-validation (CV). The optimal values for the parameters n_estimators and max_depth in the Random Forest (RF) model were rf = RandomForestRegressor(n_estimators=487, max_depth=8, random_state=0). The optimal parameter values for the Support Vector Machine (SVM) model were SVM = SVR(C=5.6, epsilon=0.1). The optimal parameter values for the Gradient Boosting Regressor Trees (GBRT) model were GBRT = GradientBoostingRegressor(learning_rate=0.1, n_estimators=31, random_state=0).

3.3 Model Evaluation

Based on the three models that have been established, 80% of the training set and 20% of the test set were imported to determine their respective R^2 and RMSE values.

Table 1. Values of R^2 and RMSE of training set and testing set of the three models

model	training set R^2	training set RMSE	test set R^2	test set RMSE
RF	0.88	0.34	0.53	0.80
SVM	0.73	0.51	0.52	0.85
GBRT	0.68	0.56	0.52	0.80

After comparison, as Tab.1 and Fig. 3 show, the RF model's predictive accuracy was greater than that of SVM and GBRT, with the highest R^2 and the lowest RMSE. Therefore, the RF model was ultimately selected.

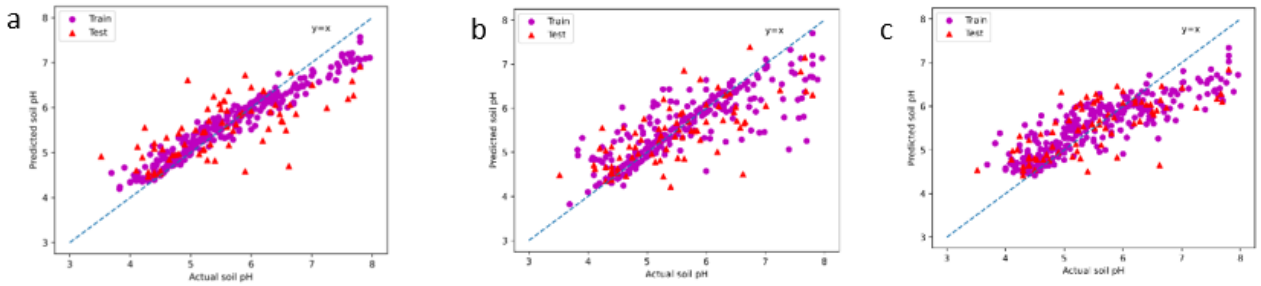


Fig 3. The performance of (a) Random Forest (RF), (b) Support Vector Regression (SVR), and (c) Gradient Boosting Regressor Trees (GBRT) in predicting the spatial distribution of soil pH in Hunan Province.

3.4 Prediction of Soil pH Spatial Distribution

The spatial distribution of soil pH values in the study area predicted by the RF model was shown in Fig. 4, where areas with lower pH are distributed on the southeast and southwest sides of Hunan Province, while the areas with higher pH are distributed on the northern side.

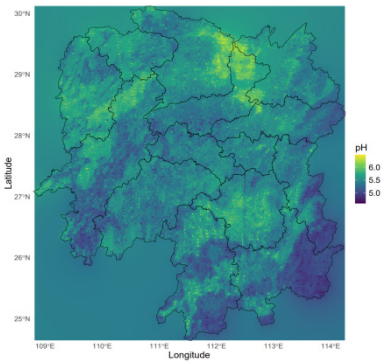


Fig 4. RF prediction pH spatial distribution map

3.5 Model Comparison

Compared to the traditional Ordinary Kriging (OK) model, the RF model shows higher predictive accuracy in this study, as reflected in the values of R^2 and RMSE. In the RF model, RMSE=0.8, R^2=0.53, while in the OK model, RMSE=0.92, R^2=0.45. The RF model's RMSE is significantly lower than the OK model's RMSE, and its R^2 is higher than that of the OK model.

3.6 Evaluation of the Importance of Environmental Factors

Through SHAP analysis, as shown in Figures 5, the relationship between land use and pH value was significantly higher than that of other features with pH value. Land use is the most visual manifestation of the impact of human activities on land[3]. Different land uses (e.g., cropland, grassland, and woodland) can significantly affect soil pH for the following main reasons:

1. Fertilization strategy: Different land use is usually accompanied by different fertilization strategies. Long-term application of nitrogen fertilizers can lead to soil acidification, while the use of potash fertilizers may neutralize some of the acidity[3].

2. Organic matter content: Land use practices affect the amount of organic matter in the soil. Woodlands and grasslands usually have a high organic matter content, while arable land is relatively low. Decomposition of organic matter produces organic acids, which affect soil pH [3-5].

3. Other factors: Different land use practices also lead to differences in soil texture, rainfall and irrigation practices, which work together to affect soil pH.

In summary, different land use practices significantly alter soil pH by influencing factors such as fertilization strategies, organic matter content, soil texture and water management.

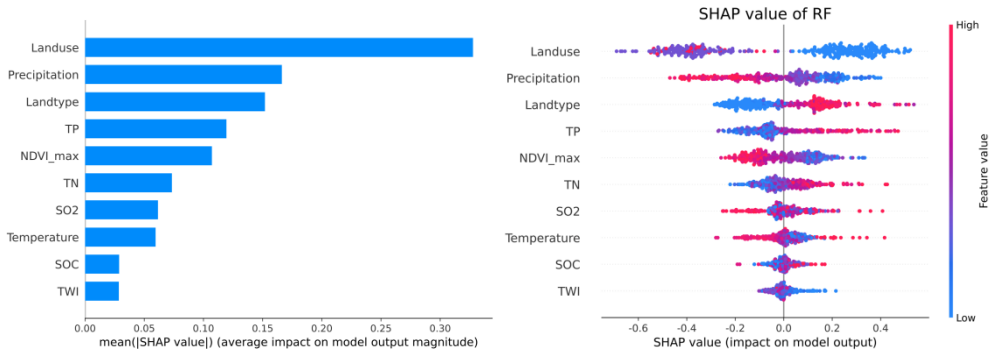


Fig 5. The impact of the main driving factors on soil pH content: Figure (a) is a bar chart of the mean absolute SHAP values of the driving factors affecting soil pH content, and Figure (b) is a set of swarm plots where each point corresponds to an individual sample in the study. The position of the points on the x-axis shows the influence of that feature on the model's prediction results for that sample. The positive SHAP value represents a positive influence on the target, and vice versa. When the color is close to red, the eigenvalue is higher, and when it is close to blue, the eigenvalue is lower

4. Conclusion

This study, based on a large number of sampling point data, used three models for training and predicting the spatial distribution of pH, and the following conclusions were drawn:

1) The Random Forest (RF) model predicts soil pH values with higher accuracy than the Support Vector Machine (SVM), Gradient Boosting Regression Trees (GBRT), and traditional Ordinary Kriging (OK) models.

Land use is the most important factor affecting soil pH values in Hunan Province.

References

[1] Wei, L., Study on the content and distribution characteristics of antimony in paddy soils of Hunan [D]. Hunan Normal University,2022.DOI:10.27137/d.cnki.ghusu.2021.002019.

[2] Ouyang, N., Attribution of Typical Red Earth of Hunan Province in China Soil Taxonomy and Their Genetic Characteristics[D].Hunan Agricultural University,2024. DOI:10.27136/ d.cnki. ghunu. 2021. 001274.

[3] Yang, H., Xu, C., Sai, M., Cao, J., Cao, L., Zhang, A., Liu, H., Effects of land use on soil moisture, pH and electrical conductivity[J].Acta Agriculturae Zhejiangensis,2016,28(11):1922-1927.

[4] Land Use and Soil Type Exert Strongly Interactive Impacts on the pH Buffering Capacity of Acidic Soils in South China (mdpi.com).

[5] Han, T., Liu, K., Huang, J., Ma, C., Zheng, L., Wang, H., Qu, X., Ren, Y., Yu, Z., Zhang, H., Spatio-temporal evolution of soil pH and its driving factors in the main Chinese farmland during past 30 years[J]. Journal of Plant Nutrition and Fertilizers, 2020, 26(12): 2137-2149. DOI: 10.11674/zwyf.20399.