# Spatial Distribution Prediction of Available Iron in Farmland Soil of Hilly Areas Based on Topographic Factors and Random Forest: Postprint

**Authors:** Yang Qipo; Wu Wei; Liu Hongbin

**Date:** 2018-01-05T00:00:00+00:00

## Abstract

To investigate soil available iron content and its spatial distribution in hilly agricultural regions, this study selected a typical hilly area (2 km2) with homogeneous parent material within Yongxing Town, Jiangjin District, Chongqing, as the study area. A total of 309 soil samples were collected, and Ordinary Kriging (OK), Multiple Linear Regression (MLR), and Random Forest (RF) models were employed in combination with topographic factors including elevation, slope gradient, aspect, valley depth, plan curvature, profile curvature, convergence index, relative slope position index, and topographic wetness index to predict the spatial distribution of soil available iron. The prediction models were evaluated and selected using 85 validation points. The results indicated that: (1) soil available iron exhibited a highly significant positive correlation with valley depth and topographic wetness index, and a highly significant negative correlation with slope gradient, plan curvature, profile curvature, convergence index, and relative slope position index; (2) the Random Forest model demonstrated significantly higher prediction accuracy than Multiple Linear Regression and Ordinary Kriging interpolation, with a mean absolute error of 22.33 mg · kg-1, a root mean square error of 27.98 mg · kg-1, and a coefficient of determination of 0.76, establishing it as the optimal prediction model for spatial distribution of soil available iron content in the study area; (3) topographic wetness index and slope gradient were the primary topographic factors influencing the spatial distribution of soil available iron content in this region. Consequently, the Random Forest prediction model based on topographic factors can effectively explain the spatial variation of soil available iron content in hilly farmland, providing methodological reference and theoretical basis for predicting the content and spatial distribution of medium and trace elements in soils of hilly regions.

## Full Text

# Prediction of Spatial Distribution of Soil Available Iron in Hilly Farmland Based on Terrain Attributes and Random Forest Model

**YANG Qipo[13], WU Wei[23], LIU Hongbin[13]**

[1]College of Resources and Environment, Southwest University, Chongqing 400716, China

[2]College of Computer and Information Science, Southwest University, Chongqing 400715, China

[3]Chongqing Key Laboratory of Digital Agriculture, Chongqing 400716, China

### Abstract

Soil available iron is essential to plant growth, and detailed information about its spatial distribution is important for effective soil fertility management. To date, published works have mainly focused on investigating the spatial variability of soil available iron, with fewer studies predicting its spatial distribution. To understand the spatial distribution of soil available iron in hilly areas of southwest China, we conducted a study in 2014 in a typical hilly region with uniform soil parent material covering 2 $km^2$ in Yongxing Town, Jiangjin District, Chongqing. A total of 309 samples were collected from the cultivated soil layer at a depth of 20 cm and randomly divided into calibration (224 samples) and validation (85 samples) datasets. Nine terrain attributes—elevation, slope, aspect, valley depth, horizontal curvature, profile curvature, convergence index, relative position index, and topographic wetness index—were extracted from a 2 m resolution digital elevation model. Ordinary Kriging (OK), Multiple Linear Regression (MLR), and Random Forest (RF) models were used to predict soil available iron content based on these terrain attributes. Accuracy indicators including mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination ($R^2$) were applied to evaluate model performance using the validation dataset. Correlation results showed that topographic wetness index and valley depth had significant positive correlations with soil available iron, while slope, horizontal curvature, profile curvature, convergence index, and relative position index had significant negative correlations. Compared with OK and MLR methods, the RF model performed best with MAE of 22.33 $mg \cdot kg^{-1}$, RMSE of 27.98 $mg \cdot kg^{-1}$, and $R^2$ of 0.76. Additionally, RF model results indicated that topographic wetness index and slope were the major factors controlling the spatial distribution of soil available iron. Therefore, the Random Forest model combined with terrain attributes can effectively explain the spatial variability of soil available iron in this area. The outcomes of this work provide valuable information for predicting the spatial distribution of soil trace elements in hilly regions.

**Keywords:** terrain attribute; random forest model; soil available iron; spatial distribution prediction

Iron is an essential component of cytochromes and enzymes in plants, playing a vital role in metabolic processes such as photosynthesis and respiration, thereby affecting crop yield and quality [1-3]. Soil available iron, as the plant-available form of iron, serves as an important reference for soil iron supply capacity [4].

The content and spatial distribution of soil available iron are influenced by many factors, including soil parent material, climate, topography, soil nutrients, and tillage practices [5-7], resulting in large regional variations that affect normal plant growth. Hills are a widely distributed landform type, and accurately predicting the spatial distribution of soil available iron in hilly areas and understanding its spatial variability characteristics and relationship with topography are of great significance for soil fertility evaluation and agricultural land use planning. However, current research on spatial distribution prediction of soil properties in hilly areas has mainly focused on soil organic matter and macronutrients [8-9], while studies on soil available iron have concentrated primarily on spatial characteristic analysis. For example, Liao et al. [10] investigated the spatial variability and distribution patterns of soil available iron in farmland topsoil, Du et al. [11] studied the abundance and spatial variability characteristics of topsoil available iron, and Liu et al. [12] comprehensively evaluated the spatial pattern variability of soil available iron.

To date, methods for soil property spatial prediction have evolved from traditional soil mapping to digital soil mapping, mainly including geostatistical methods [13], linear regression models [14-15], and machine learning algorithms [16-18]. Geostatistical methods [19] in soil property spatial prediction are mathematical geological approaches based on regionalized variables and spatial correlation-variogram analysis [20]. However, actual conditions sometimes fail to meet second-order stationary or intrinsic assumptions, making it unreliable to apply geostatistics for soil property spatial prediction [21]. The prerequisite for constructing linear regression models is a linear relationship between soil properties and predictive variables, yet in reality their relationship is often nonlinear and complex [22]. Machine learning algorithms such as Decision Trees (DT) [23], Support Vector Machine (SVM) [24], and Back Propagation Neural Networks (BPNN) [25] have also been applied to soil property mapping in recent years. However, these models are prone to overfitting [26-27] and other defects. The Random Forest model, also a type of machine learning algorithm, overcomes these deficiencies to some extent, improves prediction accuracy, and provides an effective improvement and timely supplement to soil property spatial prediction methods. Guo et al. [28] used the Random Forest model combined with environmental variables to predict soil total nitrogen, achieving significantly higher accuracy than stepwise linear regression, generalized additive mixed models, and classification and regression tree models. Wang et al. [29] used remote sensing data to extract auxiliary environmental factors combined with Random Forest algorithm to predict topsoil organic matter in hilly areas, demonstrating that Random Forest models are more effective for prediction in complex geomorphic

regions.

At a certain regional scale, structural factors such as climate and soil parent material are relatively consistent, while topography redistributes water and heat conditions and affects soil development [30], becoming the main factor causing spatial variability of soil properties [31-32]. Therefore, many recent studies on soil property prediction have considered terrain factors as important predictive variables. Lian et al. [33] applied terrain factors and remote sensing indices to analyze the relationship between soil properties and environment and conduct spatial prediction. Grimm et al. [34] used terrain factors and soil parent material as auxiliary variables to predict soil organic carbon. Zhang et al. [35] established a spatial distribution prediction model for soil nutrients using terrain factors and vegetation indices.

This study aims to use only terrain factors as predictive variables and apply the Random Forest model to predict the spatial distribution of soil available iron content in hilly areas, providing methodological reference and theoretical basis for spatial distribution prediction of medium and trace elements in hilly region soils.

### 1.1 Study Area Description

This study was conducted in a typical hilly area within Yongxing Town, Jiangjin District, Chongqing, located between 106°09 27 -106°10 9 E and 29°00 10 -29°00 9 N, covering approximately 2 km². The region has a subtropical monsoon climate with an average annual temperature of 18 °C, annual sunshine hours of 1,253, annual precipitation of 900 mm, and a frost-free period of 335 days. The study area features hilly landforms with elevations ranging from 238 to 328 m. The soil parent material is Jurassic Shaximiao Formation purple sandstone and shale. Land use types consist of over 70% cultivated land, mainly dryland and paddy fields. According to soil genesis classification, dryland soils belong to the purple soil class, neutral purple soil subclass, and gray-brown purple soil genus; paddy field soils belong to the paddy soil class, gleyed paddy soil subclass, and gleyed purple soil genus.

### 1.2 Soil Sample Collection and Chemical Analysis

Based on topographic maps and 1:1,000 land use status maps, 309 soil sampling points (258 dryland and 51 paddy field) were established following the principles of uniformity and representativity according to the National Technical Regulations for Cultivated Land Quality Survey and Evaluation, with sampling points distributed across different topographic positions and aspects [Figure 1: see original paper]. Soil samples were collected in November 2014, with coordinates and other information recorded for each point. At each sampling site, soil was collected along an "S" pattern from the 0–20 cm tillage layer using 10 soil cores and mixed into approximately 1 kg samples using the quartering method. Samples were air-dried, ground, and sieved in the laboratory, and

soil available iron content was determined using the DTPA extraction-atomic absorption spectrometry method [36].

### 1.3 Terrain Attribute Extraction

Based on the study area scale and landform characteristics, 6,373 elevation points and 1 m contour interval topographic maps were obtained through field measurements. In ArcGIS, the 3D Analyst module was used to create a TIN surface from elevation points and contour lines, which was then used to generate a 2 m resolution Digital Elevation Model (DEM). Drawing on relevant research regarding terrain and soil nutrients [37], nine terrain attributes were selected: Elevation (Ele), Slope (Slo), Aspect (Asp), Valley depth (Vd), Horizontal curvature (Hc), Profile curvature (Pc), Convergence index (Ci), Relative position index (Rpi), and Topographic wetness index (Twi). The meanings of these terrain attributes are described in [38], and they were extracted using SimDTA software [39]. Referencing studies on the relationship between DEM spatial resolution and terrain information [40-41], the original DEM was resampled to compare the effects of different DEM resolutions (2 m, 5 m, 10 m, 15 m, 30 m) on prediction results. When the DEM spatial resolution was 2 m, the model achieved the highest prediction accuracy; therefore, a 2 m resolution DEM was selected for this study.

#### 1.4.1 Ordinary Kriging

Ordinary Kriging (OK) is a fundamental geostatistical method based on variogram theory and structural analysis. It provides unbiased optimal estimation of regionalized variables within a certain area through linear combination of known sampling point data [42].

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

where $Z^*(x_0)$ represents the estimated value at point $x_0$, $Z(x_i)$ represents the $i$-th effective observation value ($i = 1, 2, 3, \ldots, n$), and $\lambda_i$ are weights generated through the semivariogram with $\sum_{i=1}^{n} \lambda_i = 1$.

#### 1.4.2 Multiple Linear Regression

Multiple Linear Regression (MLR) is based on ordinary least squares and examines the linear relationship between a dependent variable and multiple independent variables under given data conditions [43]. It is widely applied in soil nutrient spatial distribution prediction, with the expression:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_n x_{ni}$$

where $x_i$ $(i = 1, 2, 3, \ldots, n)$ represent independent variables, and $y_i$ $(i = 0, 1, 2, \ldots, n)$ and $\beta_j$ $(j = 0, 1, 2, \ldots, n)$ represent dependent variables and regression coefficients, respectively.

### 1.4.3 Random Forest Model

Random Forest (RF) is a versatile machine learning algorithm based on classification trees. It uses bootstrap resampling to randomly extract multiple samples from the original dataset for decision tree model construction, with final results obtained through voting across multiple trees. Random Forest regression has no requirements for data distribution, type, or structure, demonstrates good tolerance to noise and outliers, and is not prone to overfitting [44].

## 1.5 Semivariogram Function

The semivariogram function studies soil property variability and can determine the spatial correlation of soil attributes [45], expressed as:

$$r(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i + h) - z(x_i)]^2$$

where $r(h)$ represents the average semivariance between point pairs at distance $h$, $N(h)$ represents the number of point pairs at distance $h$, and $z(x_i)$ denotes the observed value at point $x_i$.

## 1.6 Model Construction and Accuracy Evaluation

In this study, 224 sampling points (190 dryland and 34 paddy field) were randomly selected as the training set for model construction, while the remaining 85 points (68 dryland and 17 paddy field) served as the validation set for accuracy evaluation. The Ordinary Kriging method estimates unknown points based on known data without requiring predictive variables, and interpolation was performed in ArcGIS using the exponential model obtained from semivariance analysis. Multiple Linear Regression was calculated using SPSS software, while the Random Forest model was implemented using the RandomForest package [46] in R software, with parameters ntree (number of decision trees) and mtry (number of splits per tree node) set to 150 and 3, respectively.

Model prediction accuracy was evaluated using mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination ($R^2$). Smaller MAE and RMSE values and larger $R^2$ values indicate higher model prediction accuracy. The calculation formulas are as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |M_i - P_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_i - P_i)^2}$$

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \bar{M})^2}{\sum_{i=1}^{n}(M_i - \bar{M})^2}$$

where $M$ represents measured values, $P$ represents predicted values, $\bar{M}$ is the mean of measured values, and $n$ is the number of validation samples.

### 2.1 Descriptive Statistics of Soil Available Iron

The statistical results of soil available iron content in the study area are presented in Table 1 . The training set showed a slightly larger range of soil available iron content (3.00–276.97 mg · kg $^{1}$) compared to the validation set (4.20–208.38 mg · kg $^{1}$), and a higher standard deviation (57.35 mg · kg $^{1}$ versus 53.85 mg · kg $^{1}$). Analysis of variance indicated no significant difference between training and validation sets, demonstrating that both datasets were representative of soil available iron content in the study area. The coefficient of variation for both training and validation sets exceeded 1, indicating significant differences in soil available iron content among sampling points and strong variability, consistent with results from Xu et al. [47] who studied spatial distribution of topsoil available iron in the middle reaches of the Tuojiang River using geostatistical methods. Skewness values for both datasets were greater than 1, and K-S test results showed p-values less than 0.05, indicating that soil available iron content did not follow a normal distribution. However, log-transformed soil available iron data followed a normal distribution and could be used for multiple linear regression simulation and Ordinary Kriging interpolation.

### 2.2 Semivariance Analysis of Soil Available Iron

Based on the statistical results of soil available iron content, semivariance analysis was performed using GS+ software, with the optimal fitting model and characteristic parameters presented in Table 2 . The nugget-to-sill ratio (basal effect) indicates the strength of spatial correlation, with values less than 25% representing strong spatial correlation, greater than 75% representing weak correlation, and 25%–75% indicating moderate spatial autocorrelation. The basal effect for soil available iron was 14.5%, demonstrating strong spatial correlation and indicating that differences in soil available iron content among sampling points were mainly influenced by structural factors such as climate, soil parent material, and topography. Since this study area is a small-scale hilly region with identical soil parent material and consistent climatic conditions, topography emerges as the primary factor influencing the spatial distribution of soil available iron.

### 2.3 Correlation Between Soil Available Iron and Terrain Attributes

Table 3 presents correlations between soil available iron and terrain attributes. Soil available iron showed highly significant correlations with slope, valley depth, horizontal curvature, profile curvature, convergence index, relative position index, and topographic wetness index, indicating that soil available iron in the study area was significantly affected by terrain factors. Specifically, soil available iron had highly significant positive correlations with valley depth and topographic wetness index (correlation coefficients of 0.298 and 0.592, respectively), suggesting that areas with low elevation and high soil moisture facilitate accumulation of soil available iron. Conversely, highly significant negative correlations were observed with slope, horizontal curvature, profile curvature, convergence index, and relative position index (correlation coefficients of -0.371, -0.327, -0.228, -0.174, and -0.428, respectively), indicating that areas with high elevation and rugged terrain had lower soil available iron content.

### 2.4 Comparison of Soil Available Iron Content Across Land Use Types

Table 4 compares soil available iron content across different land use types in the study area. Dryland samples (190) far outnumbered paddy field samples (34), likely because dryland terrain is more complex with smaller, fragmented plots, making soil available iron content more susceptible to terrain influences. Analysis of variance revealed that soil available iron content in paddy fields ($151.04 \pm 11.54$ mg $\cdot$ kg $^{1}$) was significantly higher than in dryland ($31.85 \pm 2.20$ mg $\cdot$ kg $^{1}$). The coefficient of variation for dryland soil available iron (0.95) was greater than for paddy fields (0.45), though both showed moderate variability (coefficient of variation between 0.1 and 1 indicates moderate variation), indicating certain differences in soil available iron content among sampling points within the same land use type.

### 2.4 Comparison of Model Prediction Accuracy for Soil Available Iron Content

Model prediction accuracy was evaluated using 85 sampling points (68 dryland and 17 paddy field), with results shown in Table 5 . The Random Forest (RF) model achieved significantly lower mean absolute error (MAE = 22.33 mg $\cdot$ kg $^{1}$) and root mean square error (RMSE = 27.98 mg $\cdot$ kg $^{1}$), and higher coefficient of determination ($R^2 = 0.76$) compared to Ordinary Kriging (OK) and Multiple Linear Regression (MLR). The MLR model could only fit basic linear relationships between soil available iron and terrain factors, limiting its predictive accuracy for nonlinear relationships. The OK model showed lower prediction accuracy because it relied solely on the spatial autocorrelation of soil available iron without considering environmental factors such as topography. In contrast, the RF model fully considered terrain influences on soil available iron and captured complex nonlinear relationships, resulting in higher prediction accuracy. These results demonstrate the superiority of the RF model and the feasibility

of using terrain factors to predict soil available iron spatial distribution at small watershed scales in hilly farmland.

Given the significant differences in soil available iron content between land use types (dryland versus paddy field), it was necessary to consider land use effects in model construction. Since land use type is a qualitative variable, the Random Forest model, which better accommodates qualitative variables than MLR, can capture more effective information between land use type and soil available iron. To test whether including land use type could further improve RF model prediction accuracy, land use type (dryland, paddy field) was added as a predictive variable to the RF model. The resulting terrain-and-land-use-based RF model showed decreased prediction accuracy (MAE = 23.77 $mg \cdot kg^{1}$, RMSE = 32.95 $mg \cdot kg^{1}$, $R^2$ = 0.66). This decline occurred because paddy field and dryland distributions are closely related to terrain [48-50]; land use type represents a comprehensive, holistic, qualitative expression of terrain that interferes with, masks, and creates collinearity with other terrain factors, preventing full expression of each terrain factor' s contribution to the model and increasing prediction errors. Based on comprehensive accuracy evaluation, the optimal model for predicting soil available iron content in this study area is the Random Forest model based solely on terrain factors.

### 2.5 Influence of Terrain Attributes on Spatial Variation of Soil Available Iron Content

Figure 2 [Figure 2: see original paper] illustrates the influence of terrain attributes on the Random Forest prediction model for soil available iron content. As one of the five soil-forming factors, topography affects material and energy exchange between soil and environment, causing changes in soil physicochemical properties and nutrients [51-52]. The results show that topographic wetness index (Twi) and slope (Slo) are the primary terrain factors influencing soil available iron spatial distribution. Slope indirectly affects soil fertility characteristics by influencing rainfall and infiltration time, surface soil particle movement, runoff sediment-carrying capacity, and erosion patterns [53-55]. Topographic wetness index quantitatively simulates soil moisture content and runoff generation potential, with higher values indicating greater soil moisture [56-58]. Correlation analysis revealed that soil available iron content was highly significantly positively correlated with topographic wetness index (0.592) and highly significantly negatively correlated with slope (-0.371). These findings indicate that as slope increases, terrain steepness intensifies, rainfall erosion on soil intensifies, and soil nutrients (including available iron, total iron, and organic matter) are severely lost through leaching. Low total iron content directly leads to insufficient available iron, while lack of organic matter also causes available iron deficiency. Conversely, as topographic wetness index increases, soil moisture content rises, creating relatively anaerobic conditions. Under such poorly aerated conditions, soil reducibility increases, facilitating conversion of $Fe^{3}$ to $Fe^{2}$ and promoting soil available iron accumulation [59-60].

### 2.6 Spatial Distribution Prediction of Soil Available Iron

Applying the Random Forest prediction model with terrain attributes as independent variables, a spatial distribution map of soil available iron content was generated in ArcGIS [Figure 3: see original paper]. The distribution pattern of soil available iron content is closely related to study area topography. Steep terrain areas show lower soil available iron content due to intensified soil erosion and leaching caused by rainfall, leading to loss of soil available iron and other nutrients. Low-lying areas exhibit higher soil available iron content because lost soil nutrients accumulate in depressions, and high soil moisture content creates reducing conditions that further enhance soil available iron content.

## Conclusions

This study first used terrain attributes as predictive variables to predict the spatial distribution of soil available iron in the study area using Ordinary Kriging, Multiple Linear Regression, and Random Forest models, selecting the superior Random Forest model. By comparing a terrain-and-land-use-based Random Forest model, we concluded that the optimal prediction model for soil available iron content in this study area is the Random Forest model based on terrain attributes.

1. Study area topography is closely related to the spatial distribution of soil available iron content. Soil available iron showed highly significant correlations with slope, valley depth, horizontal curvature, profile curvature, convergence index, relative position index, and topographic wetness index. Spatial variation of soil available iron content is mainly influenced by structural factors.

2. The Random Forest model using terrain attributes as independent variables achieved MAE of 22.33 $mg \cdot kg^{-1}$, RMSE of 27.98 $mg \cdot kg^{-1}$, and $R^2$ of 0.76, significantly outperforming other prediction models. This approach can serve as a new model for predicting the spatial distribution of medium and trace elements in soil using terrain attributes.

3. Topographic wetness index and slope are the main terrain factors affecting the spatial distribution of soil available iron content in this region.

The Random Forest model effectively captured the relationship between terrain attributes and soil available iron content in the study area, explaining 76% of the spatial variability in soil available iron using terrain factors alone, though prediction accuracy could be further improved. Future research should consider incorporating additional environmental variables and screening predictive variables to enhance Random Forest model accuracy for spatial distribution prediction of medium and trace elements in soil.

## References

[1] Wang L P. Review on iron nutrition and control of iron deficiency in plants[J]. Journal of Anhui Agricultural University, 1995, (01): 17-22.

[2] Shen H Y, Xiong H C, Guo X T, et al. Progress of molecular and physiological mechanism of iron uptake and translocation in plants[J]. Plant Nutrition & Fertilizer Science, 2011, 17(6): 1522-1530.

[3] Li J C, Yu H, Yang S X, et al. Research progress of molecular regulation of iron uptake in plants[J]. Plant Physiology Journal, 2016(6): 835-842.

[4] Zhang C M. Analysis and evaluation of available Zn, Mn, Cu and Fe contents of topsoil in Gulang Irrigation Region[J]. Pratacultural Science, 2011, 28(6): 1221-1225.

[5] Li Q, Zhou J H, Yang R S, et al. Soil nutrients spatial variability and soil fertility suitability in Qujing tobacco-planting area[J]. Chinese Journal of Applied Ecology, 2011, 22(4): 950-956.

[6] Wu J, Li Y H, Li Z B, et al. Spatial distribution and influencing factors of farmland soil organic matter and trace elements in the Nansihu Region[J]. Acta Ecologica Sinica, 2014, 34(6): 1596-1605.

[7] Jenny H. The soil resource. Origin and behavior[J]. Vegetatio, 1984, 57(2-3): 102-102.

[8] Huang A, Yang L A, Du T, et al. Spatial distribution of the soil organic matter based on multiple soil factors[J]. Arid Land Geography, 2015, 38(5): 994-1003.

[9] Li Q Q, Wang C Q, Zhang W J, et al. Prediction of soil nutrients spatial distribution based on neural network model combined with geostatistics[J]. Chinese Journal of Applied Ecology, 2013, 24(2): 459-466.

[10] Liao Q, Nan Z R, Wang S L, et al. Spatial distribution characteristics of available microelement contents in oasis cropland soils of arid areas[J]. Research of Environmental Sciences, 2011, 24(3): 273-280.

[11] Du J, Zhang Y Q, Zhou J C, Li M. Spatial distribution characteristics of available microelements contents in irrigation-silted soil in Yaodu District[J]. Chinese Agricultural Science Bulletin, 2014, 30(3): 162-167.

[12] Liu Y H, Ni Z Y, Xie G X, Xu L J, Zhong L B, Ma L Q. Spatial variability and impacting factors of trace elements in hilly region of cropland in northwestern Zhejiang Province[J]. Journal of Plant Nutrition and Fertilizer, 2016, (06): 1710-1718.

[13] Shi W J, Yue T X, Shi X L, et al. Research progress in soil property interpolators and their accuracy[J]. Journal of Natural Resources, 2012, (01): 163-175.

[14] Zhang G P, Guo P T, Wang Z Y, Liu H B, et al. Prediction of spatial distribution of hilly farmland with purple soil nutrient[J]. Transactions of the Chinese Society of Agricultural Engineering, 2013, 29(6): 113-120.

[15] WANG J, CUI L, GAO W, et al. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy[J]. Geoderma, 2014, 216(4): 1-9.

[16] Xu J B, Song L S, Xia Z, et al. Spatial variability of available phosphorus for cultivated soil based on GARBF neural network[J]. Transactions of the Chinese Society of Agricultural Engineering, 2012, 28(16): 158-165.

[17] Shi W, Nan Z T, Li R, Zhan L, Zhang X M, Zhan Y H. Support vector machine based soil mapping of a typical permafrost area in the Qinghai-Tibet plateau[J]. Acta Pedologica Sinica, 2011, (03): 461-469.

[18] Lu Y Y, Zhang G L, Zhao Y G, et al. Extracting and mapping of soil depth distribution rules in complex landscape environment[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, 30(18): 132-141.

[19] Lark R M. Towards soil geostatistics[J]. Spatial Statistics, 2012, 1: 92-99.

[20] Timothy C. Coburn. Geostatistics for Natural Resources Evaluation[J]. Journal of Environmental Quality, 2012, 42(4): 437-438.

[21] Xu J B, Song L S, Peng L. Research review on methods of spatial prediction of soil nutrients[J]. Ecology & Environmental Sciences, 2011, (Z2): 1379-1386.

[22] Ding J L, Wang F. Environmental modeling of large-scale soil salinity information in an arid region: A case study of the low and middle altitude alluvial plain north and south of the Tianshan Mountains, Xinjiang[J]. Acta Geographica Sinica, 2017, 72(1): 64-78.

[23] Breiman L I, Friedman J H, Olshen R A, et al. Classification and Regression Trees (CART)[J]. Biometrics, 2015, 40(3): 358.

[24] Besalatpour A, Hajabbasi M, Ayoubi S, et al. Prediction of soil physical properties by optimized support vector machines[J]. International Agrophysics, 2012, 26(2): 109-115.

[25] Koley S. Machine Learning for Soil Fertility and Plant Nutrient Management Using Back Propagation Neural Networks[J]. Social Science Electronic Publishing, 2016, 2(2): 292-297.

[26] Dietterich T. Overfitting and undercomputing in machine learning[J]. Acm Computing Surveys, 1995, 27(3): 326-327.

[27] Hawkins D M. The problem of overfitting[J]. Cheminform, 2004, 44(1): 1.

[28] Guo P T, Li M F, Luo L, et al. Prediction of soil total nitrogen for rubber plantation at regional scale based on environmental variables and random forest approach[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(5): 194-202.

[29] Wang Y Y, Qi Y B, Chen Y, et al. Prediction of Soil Organic Matter Based on Multi-resolution Remote Sensing Data and Random Forest Algorithm[J]. Acta Pedologica Sinica, 2016, (02): 342-354.

[30] Huang C Y. Soil Science[M]. Beijing: China Agriculture Press, 1999, 141-142.

[31] Qin S, Fan Y, Liu H B, et al. Study on the Relations Between Topographical Factors and the Spatial Distributions of Soil Nutrients[J]. Research of Soil & Water Conservation, 2008, 15(1): 275-279.

[32] Song X, Li L D, Kou C L, et al. Soil nutrient distribution and its relations with topography in Huangshui River drainage basin[J]. Chinese Journal of Applied Ecology, 2011, 22(12): 3163-3168.

[33] Lian G, Guo X D, Fu B J, et al. Prediction of the spatial distribution of soil properties based on environmental correlation and geostatistics[J]. Transactions of the Chinese Society of Agricultural Engineering, 2009, 25(7): 237-242.

[34] Grimm R, Behrens T, Märker M, et al. Soil organic carbon concentrations and stocks on Barro Colorado Island-Digital soil mapping using Random Forest analysis[J]. Geoderma, 2008, 146(1-2): 102-113.

[35] Zhang S M, Wang Z M, Zhang B, et al. Prediction of spatial distribution of soil nutrients using terrain attributes and remote sensing data[J]. Transactions of the Chinese Society of Agricultural Engineering, 2010, 26(5): 188-194.

[36] Bao S D. Soil and Agricultural Chemistry Analysis[M]. Beijing: China Agriculture Press, 2002, 47-56.

[37] Deng O P, Zhou X, Huang P P, et al. Correlations between spatial variability of soil nutrients and topographic factors in the purple hilly region of Sichuan[J]. Resources Science, 2013, (12): 2434-2443.

[38] Moore I D, Gessler P E, Nieslen G A, et al. Soil attribute prediction using terrain analysis[J]. Soil Sci Soc Am J, 1993, 57: 443-452.

[39] Qin C Z, Lu Y J, Bao L L, Zhu A X, et al. Simple Digital Terrain Analysis Software (SimDTA 1.0) and Its Application in Fuzzy Classification of Slope Positions[J]. Journal of Geo-Information Science, 2009, 11(6): 737-743.

[40] Hu X M, Qin C Z. Effects of different topographic attributes on determining appropriate DEM resolution[J]. Progress in Geography, 2014, 33(1): 50-56.

[41] Huang X L, Tang G A, Liu K. The Influence of DEM resolution on the extraction of terrain texture feature[J]. Journal of Geo-Information Science, 2015, 17(7): 822-829.

[42] Dale L. Zimmerman, M. Bridget Zimmerman. A Comparison of Spatial Semivariogram Estimators and Corresponding Ordinary Kriging Predictors[J]. Technometrics, 2012, 33(1): 77-91.

[43] Fedotova O, Teixeira L, Alvelos H. Software Effort Estimation with Multiple Linear Regression: review and practical application[J]. Journal of Information Science & Engineering, 2013, 29(5): 925-945.

[44] Lindner C, Bromiley P A, Ionita M C, et al. Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1862-1874.

[45] Goovaerts P. Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties[J]. Biology & Fertility of Soils, 1998, 27(4): 315-334.

[46] Breiman L, Cutler A. Breiman and Cutler' s random forests for classification and regression. R Package Version 4.6-7, 2013.

[47] Xu X X, Zhang S R, Yu N N, et al. Soil available iron spatial distribution and influencing factors analysis based on GIS in middle reaches of Tuojiang[J]. Southwest China Journal of Agricultural Sciences, 2012, (03): 977-981.

[48] Han J P, Jia N F. Relationship between topographic factor and land use - a case study of Zhuanyaogou watershed[J]. Chinese Journal of Eco-Agriculture, 2010, 18(5): 1071-1075.

[49] Ha K, Ding Q L, Men M X, et al. Spatial distribution of land use and its relationship with terrain factors in hilly area[J]. Geographical Research, 2015, 34(5): 909-921.

[50] Wu A B, Liu X, Zhao Y X. Influences of topographic on distribution and change of land use types in hilly region-taking Yanshan hilly region as an example[J]. Research of Agricultural Modernization, 2014, 35(1): 103-107.

[51] Zhang L P, Wang X Y, Zhang H S. Evolution of physical and chemical characteristics of loess with different landforms in slope field under sand cover[J]. Scientia Geographica Sinica, 2011, (2): 178-183.

[52] Xu G C, Li Z B, Li P, et al. Quantitative analysis of soil erosion and nutrient loss in Yingwugou watershed of the Dan River[J]. Transactions of the Chinese Society of Agricultural Engineering, 2013, 29(10): 160-167.

[53] Wu H. The Relationship between terrain factors and spatial variability of soil nutrients for Pine-Oak mixed forest in Qinling Mountains[J]. Journal of Natural Resources, 2015, (05): 858-869.

[54] Wang L, Wang L, Wang Q Q. The processes of nitrogen and phosphorus loss and migration in slope cropland under different slopes[J]. Journal of Soil and Water Conservation, 2015, 29(2): 69-75.

[55] Zheng Z Z, Qin F, Li T X. Changes in soil surface microrelief of purple soil under different slope gradients and its effects on soil erosion[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(8): 168-175.

[56] Zhang C X, Yang Q K, Li R. Advancement in Topographic Wetness Index and Its Application[J]. Progress in Geography, 2005, 24(6): 116-123.

[57] Lewis G L, Holden N M. The Modification of Soil Moisture Deficit Calculation Using Topographic Wetness Index to Account For the Effect of Slope and Landscape Position[C]// 2012 Dallas, Texas, July 29 - August 1, 2012. 2012.

[58] Maduako I N, Ndukwu R I, Ifeanyichukwu C, et al. Multi-Index Soil Moisture Estimation from Satellite Earth Observations: Comparative Evaluation of the Topographic Wetness Index (TWI), the Temperature Vegetation Dryness Index (TVDI) and the Improved TVDI (iTVDI)[J]. Journal of the Indian Society of Remote Sensing, 2017, 45(4): 631-642.

[59] Zhu H, Hu W, Bi R, et al. Scale- and location-specific relationships between soil available micronutrients and environmental factors in the Fen River basin on the Chinese Loess Plateau[J]. Catena, 2016, 147: 764-772.

[60] Zhu H, Zhao Y, Nan F, et al. Relative influence of soil chemistry and topography on soil available micronutrients by structural equation modeling[J]. Journal of Soil Science & Plant Nutrition, 2016, 16(4): 1038-1051.

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