land

Article

Prediction of Spatial Distribution of Soil Organic Carbon in Helan Farmland Based on Different Prediction Models

Yuhan Zhang [1,2], Youqi Wang [2,3], Yiru Bai [1,2,*], Ruiyuan Zhang [1,2], Xu Liu [1,2] and Xian Ma [1,2]

[1] School of Geography and Planning, Ningxia University, Yinchuan 750021, China; 12021130813@stu.nxu.edu.cn (Y.Z.); 12021130805@stu.nxu.edu.cn (R.Z.); liuxu0506@163.com (X.L.); maxian2236@163.com (X.M.)

[2] Breeding Base for State Key Laboratory of Land Degradation and Ecological Restoration in Northwestern China, Ningxia University, Yinchuan 750021, China; wangyouqi@nxu.edu.cn

[3] School of Ecology and Environment, Ningxia University, Yinchuan 750021, China

[*] Correspondence: baiyiru@nxu.edu.cn; Tel.: +86-182-9501-7031

Abstract: Soil organic carbon (SOC) is widely recognized as an essential indicator of the quality of arable soils and the health of ecosystems. In addition, an accurate understanding of the spatial distribution of soil organic carbon content for precision digital agriculture is important. In this study, the spatial distribution of organic carbon in topsoil was determined using four common machine learning methods, namely the back-propagation neural network model (BPNN), random forest algorithm model (RF), geographically weighted regression model (GWR), and ordinary Kriging interpolation method (OK), with Helan County as the study area. The prediction accuracies of the four different models were compared in conjunction with multiple sources of auxiliary variables. The prediction accuracies for the four models were BPNN (MRE = 0.066, RMSE = 0.257) > RF (MRE = 0.186, RMSE = 3.320) > GWR (MRE = 0.193, RMSE = 3.595) > OK (MRE = 0.198, RMSE = 4.248). Moreover, the spatial distribution trends for the SOC content predicted with the four different models were similar: high in the western area and low in the eastern area of the study region. The BPNN model better handled the nonlinear relationship between the SOC content and multisource auxiliary variables and presented finer information for spatial differentiation. These results provide an important theoretical basis and data support to explore the spatial distribution trend for SOC content.

Keywords: environmental auxiliary variables; machine learning; spatial distribution; soil organic carbon

check for updates

Citation: Zhang, Y.; Wang, Y.; Bai, Y.; Zhang, R.; Liu, X.; Ma, X. Prediction of Spatial Distribution of Soil Organic Carbon in Helan Farmland Based on Different Prediction Models. *Land* **2023**, *12*, 1984. https://doi.org/10.3390/land12111984

Academic Editor: Nick B. Comerford

Received: 13 September 2023
Revised: 16 October 2023
Accepted: 25 October 2023
Published: 27 October 2023

1. Introduction

Soil is the largest carbon reservoir in terrestrial ecosystems, holding approximately three times as much carbon as the atmospheric carbon pool and two and a half times as much carbon as the terrestrial vegetation carbon pool [1–3]. Agricultural soils in particular are a huge carbon reservoir. By storing huge amounts of organic carbon (SOC), agricultural soils play an important role in sustaining soil fertility for promoting plant growth and mitigating climate change [4,5]; they also hold more than 10% of global organic carbon stocks. SOC not only helps maintain soil fertility and soil microbial activity, but it also has a marked impact on soil health in that it affects soil water, nutrient retention capacity, and soil structure [6,7]. As the most active and important component of the global carbon pool, agricultural SOC is highly spatially heterogeneous and vulnerable to human activities and various natural factors, including climate, soil properties, and topography [8–11]. Therefore, the exploration of the spatial distribution of agricultural SOC is necessary for carbon balance sustainment and soil health improvement [12].

Remote sensing technology based on satellites has been increasingly used to detect and classify objects on Earth [13]. Remote sensing data are obtained by applying complex prediction pipelines [14]. Currently, the combination of remote sensing technologies and geographic information systems (GISs) has opened up new opportunities for soil science research [15]. For example, using remote sensing techniques, Abdoli et al. [16] predicted soil organic carbon (SOC) in selected agricultural soils in parts of Iran, Dhiman et al. [17] predicted soil nutrients in North India, and Jia et al. [18] predicted the soil potential of hydrogen (pH) in the north of Yinchuan based on remote sensing data and machine learning algorithms. Remote sensing technology and machine learning (ML) have made a new contribution to the research progress of soil properties.

ML technology is known to be effective in organizing and processing large amounts of data from different sources [19]. Zhao et al. [20] compared the prediction of pH and soil texture using the random forest algorithm (RF) model and other models; they concluded that the local uncertainty observed by the other models was overestimated to a greater extent than that of the RF model, which was better able to quantify the predicted uncertainty. Hao et al. [21] used a BPNN to predict the best measure of land tillage protection and the regression coefficient between the predicted and actual output values. In modeling soil attributes, such as the geographical distribution of physical properties and nutrients, strength of furrow erosion, and influence of environmental factors, neural network models were remarkably accurate [22,23]. Lu et al. [24] predicted soil heavy metals using the back-propagation neural network (BPNN) and RF models; the prediction results showed that the BPNN model had a better prediction performance. Wang et al. [25] compared the geographically weighted regression (GWR), ordinary Kriging (OK) interpolation, and multiple linear regression (MLR) models for SOC prediction and concluded that GWR had the highest prediction accuracy. Yuan et al. [26] used GWR to reveal an effective tool for the spatial variability of environmental variables, which allowed for a better spatial understanding of the complex relationships between environmental parameters, which was difficult to achieve with traditional statistical analyses. In summary, it is clear that various algorithms produced different performances and results.

Meanwhile, there was a mostly nonlinear characteristic between SOC content (as an output variable) and remotely sensed data (as an input variable) [16], so classical linear models may not be suitable for identifying complex nonlinearities. ML models, such as the RF and BPNN models, can help capture nonlinearities and further improve estimation accuracy [27,28]. Additionally, different environmental factors could have an impact on ML performance. The relative importance of remote sensing indices for SOC prediction was higher in Zeraatpisheh's study [29]. However, Mahmoudzadeh [30] argued that elevation had a greater impact on SOC predictions. Therefore, the choice of different environmental variables could also have an impact on the prediction results; this is also crucial to understand the effects of environmental factors on SOC [31]. Therefore, two classical models and two ML algorithms were selected and used to understand the optimal model and effect of environmental factors on SOC in our study area.

Helan County is located in the Ningxia Hui Autonomous Region of Northwest China. It is a major grain producing zone with arid and semiarid regions [32], and there is an important arable land reserve in Ningxia [18]. Therefore, an accurate and effective analysis of the spatial distribution of SOC content in local agricultural soils is essential to assess soil health and optimize land management in the area [33]. However, the spatial distribution and prediction of organic carbon in the agricultural soils of Helan County, based on multisource variables and ML algorithms, have been little studied. The objectives of this study were to (1) understand the characteristics and influencing factors of the spatial distribution of SOC content; (2) identify a set of auxiliary variable combinations with a high impact on SOC prediction; and (3) determine the best machine learning algorithm for SOC spatial prediction performance in the study area.

2. Materials and Methods

2.1. Geologic Setting

Located in Helan County in the Ningxia Hui Autonomous Region of Northwest China (Figure 1), the research region belongs to the Yellow River Diversion and Irrigation District. There are hilly regions in the west and plains in the east, the average altitude being 1250 m.

The research area has a temperate continental dry climate zone, with an average annual evaporation of 2000 mm from soils and rivers, a mean annual precipitation of 113.3 mm, a frost-free period of 188 days, a mean annual temperature of 8.5 °C, a maximum temperature of 35 °C, and a minimum temperature of approximately −20 °C [34]. The soil type was mainly light grey calcium soil. The soil here is low in clay and sand content and is mainly cultivated for maize, wheat, and rice.

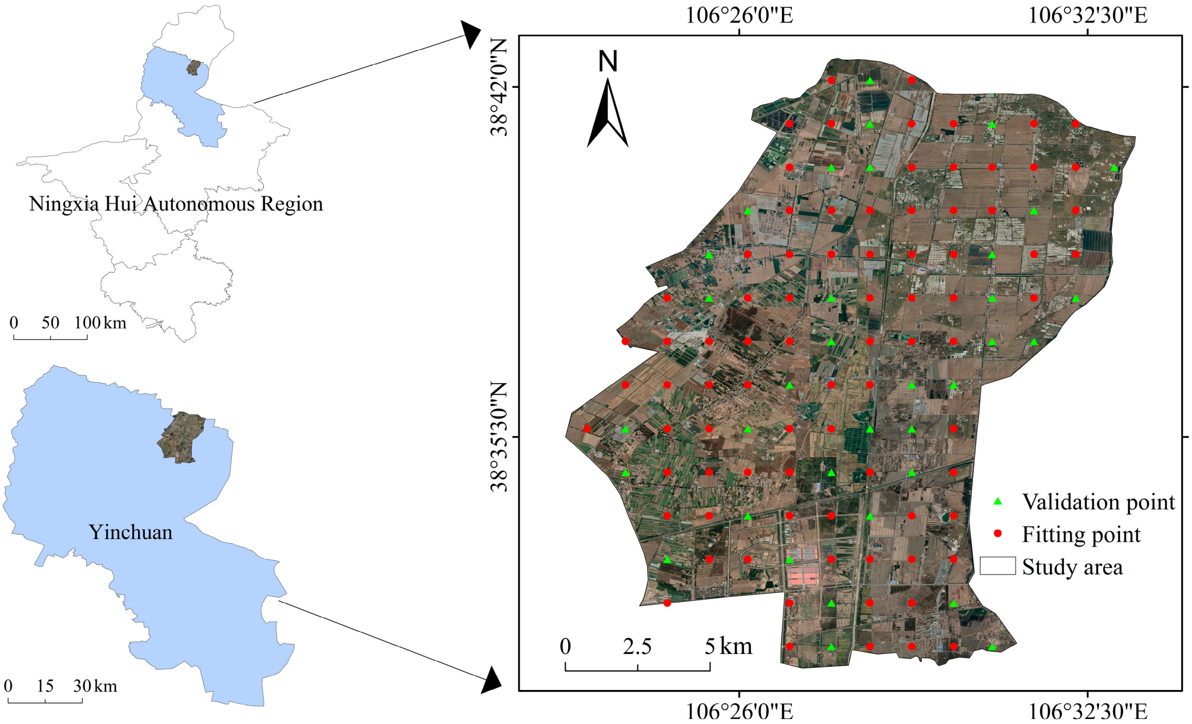

Figure 1. Location of the soil-sampling sites in the study area.

2.2. Sampling and Chemical Analysis

The sampling points were laid out in a grid format, with the surface cover removed using a wooden shovel. Soil samples were collected using a five-point sampling method and then placed in brown glass jars. A total of 117 soil samples were collected in October 2021 from a soil depth of 0 to 20 cm (Figure 1). The soil samples were air dried, ground, and sieved to determine the content of the indicators; these included SOC, pH, and electrical conductivity (EC). The SOC content was determined using the potassium dichromate volumetric method.

2.3. Environmental Covariates

Twelve environmental covariates (Table 1) were selected to predict the SOC, including geographic coordinates (X and Y coordinates), topographic factors (elevation, surface curvature, profile curvature, slope, and aspect), remote sensing factors (NIR, SWIR.1, and SWIR.2), and soil physicochemical factors (EC and pH). The pH and electrical conductivity (EC) were determined in the laboratory by the electrode method. The spatial distribution of soil EC and pH in the study area was obtained through OK interpolation. Topographic factors are the most widely used environmental factors in the prediction of soil properties. In this study, a Landsat 8 digital elevation model (DEM) with a 30 m resolution was downloaded from the geospatial data cloud platform to obtain various topographic factors (https://www.gscloud.cn/ (accessed on 25 February 2022)). Following this, and having considered previous research [33,35–37], five topographic factors (elevation, surface curvature, profile curvature, slope, and aspect) were obtained after processing with ArcGIS 10.6 software. Landsat 8 remote sensing data from the Computer Network Centre of the Chinese

Academy of Sciences' geospatial data cloud platform were used to derive remote sensing variables. After radiometric calibration and atmospheric correction using ENVI5.3 software, the short infrared wave 1 (SWIR.1), short infrared wave 2 (SWIR.2), and near-infrared band (NIR) were extracted because the SOC was sensitive to these bands. Soil pH and EC were selected as environmental covariates. Figure 2 clearly shows the auxiliary variables in space. For better predictions, the variance inflation factor (VIF) for the environmental covariates was calculated before being embedded into the models. When the VIF > 10, the variables were excluded before being substituted into the models.

Table 1. Sources of environment variables.

Extracted Parameters	Sources	Reference	Spatial Resolution
EC	Laboratory measurement	[33,38]	30 m
pH		[36,39]	
Slope	Geospatial data cloud platform to obtain various topographic factors (https://www.gscloud.cn/ (accessed on 25 February 2022)).	[33,35–37]	30 m
SWIR.2		[7,40]	
Y		[41]	
SWIR.1		[7,40]	
NIR		[7,39]	
X		[41]	
Profile curvature		[33]	
Surface curvature		[33]	
Aspect		[33]	
Elevation		[33,35,36]	

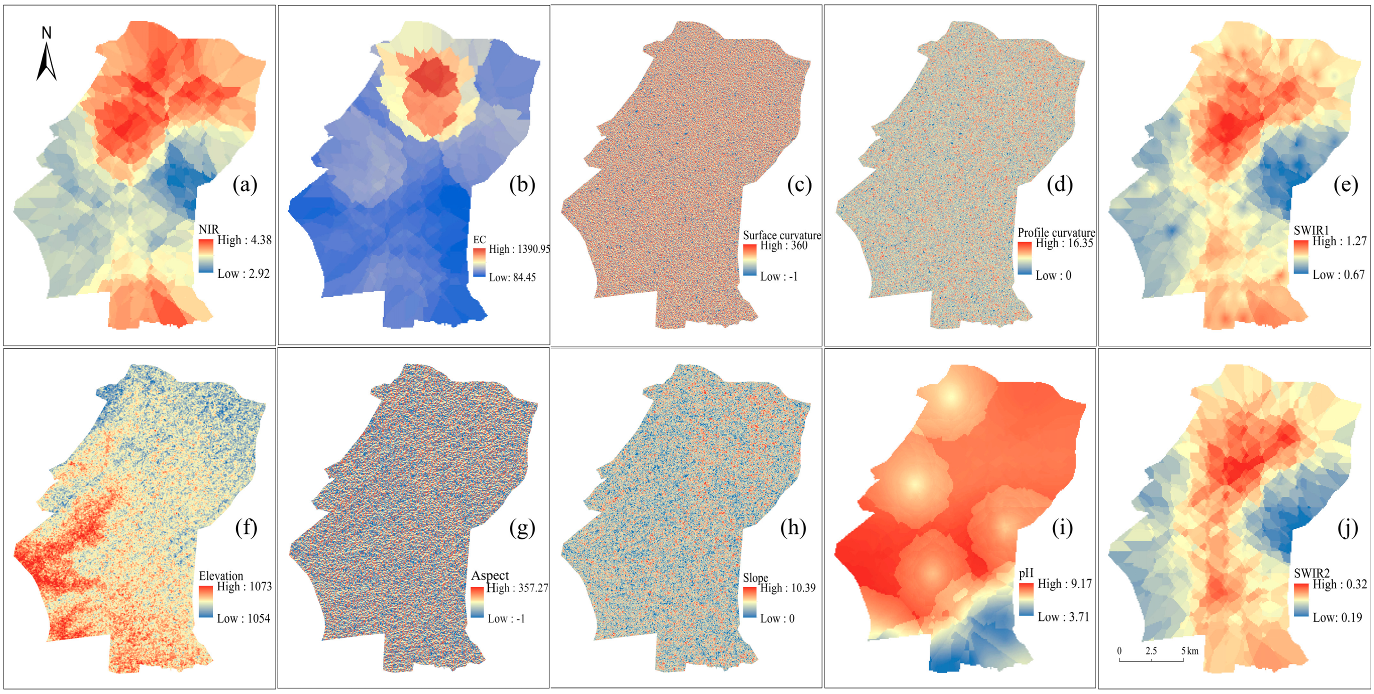

Figure 2. Spatial distribution of environmental factors in the study area of (**a**) NIR: near-infrared; (**b**) EC: soil conductivity; (**c**) surface curvature; (**d**) profile curvature; (**e**) SWIR.1: short infrared wave 1; (**f**) elevation; (**g**) aspect; (**h**) slope; (**i**) pH; and (**j**) SWIR 2: short infrared wave 2.

2.4. Theory and Algorithms

2.4.1. Variance Inflation Factor

The *VIF* is an indicator used to assess the degree of multicollinearity in a multiple linear regression model. It is the ratio of the variance in explanatory variables in the presence of multicollinearity to the variance in explanatory variables in the absence of multicollinearity [42]. The formula is calculated as follows:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{1}$$

where R_i^2 denotes the coefficient of determination x_i between the ith variable in the independent variable matrix and the variables outside it. It is the result of an ordinary least-squares (OLS) regression in which x_i is the dependent variable and the remaining variables are the independent variables [43].

If the variables x_i and other variables have no covariance, then $R_i^2 = 0$ and $VIF_i = 1$. However, this is simply an ideal condition. Covariance between numerous independent variables is almost always present. VIF_i increases as the linearity between the variables increases. Therefore, the higher the VIF, the closer the multicollinearity between the variables. When the *VIF* is greater than 10, multicollinearity exists [44].

2.4.2. Ordinary Kriging Model

The OK method is an element of geostatistics and is based on structural analysis and the semi-variance function theory. The method provides an optimal, continuous regionalized variable evaluation of known sampling point data based on a linear combination of values between sampling points [45].

$$Z \times (x_0) = \sum_{i=1}^{n} \lambda_i (Zx_i) \tag{2}$$

In equation: $Z \times (x_0)$ denotes the value of the point waiting for valuation x_0, $Z(x_i)$ denotes the ith valid observation $(i = 1, 2, \ldots, n)$, and λ is the weight generated by the semi-variance function and $\sum \lambda = 1$.

2.4.3. Random Forest Model

The RF model is an algorithm combined with a bagging algorithm and a machine tilting method. To boost model performance, the RF model trains several CART decision trees. In the training phase, it employs bootstrap sampling to obtain various sub-training datasets from the input training dataset to train several decision trees in succession. Notably, the RF model assesses variable importance using the variable importance metric, which is the overall reduction in nodal impurity of the split variable as determined with a regression averaging the residual sum of squares over all tree branches [46]. The error values in this study leveled off and were minimized when trees = 1000, as shown in Figure 3. In the training phase of the RF model, the regression variables (influences) and points of regression were evaluated using the mean square error function [47]. The RF classification result is obtained by voting on the output of each classification decision tree; however, in regression prediction [48], the projected value is the average of all regression tree outputs, and the expression is as follows:

$$\overline{h}(x) = (\frac{1}{k}) \sum_{i=1}^{k} h(X; \theta_i) \tag{3}$$

where $\overline{h}(x)$ is the predicted value, θ_i is an independently distributed random vector that can determine the growth of the decision tree, X is the input matrix, $h(X; \theta_i)$ is the output of the ith regression tree, and k is the number of regression trees.

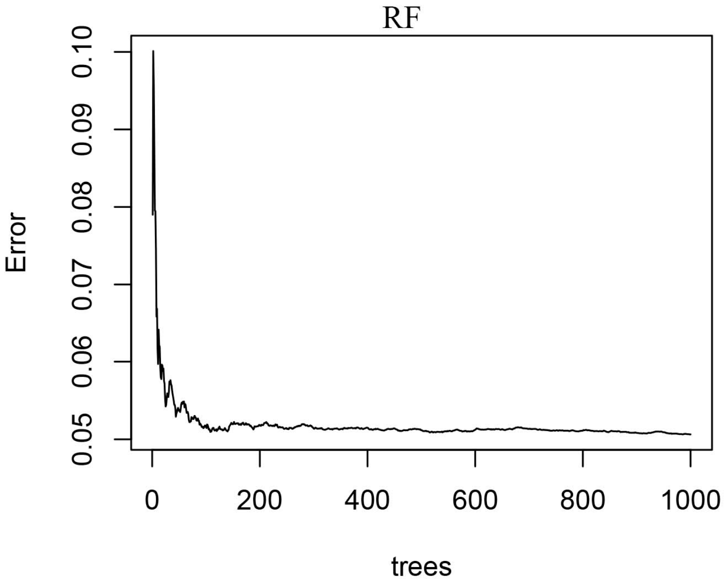

Figure 3. Error values when the trees ranged from 0 to 1000.

2.4.4. Geographically Weighted Regression Models

GWR displayed the regionally varied correlations between the dependent and independent variables, as well as the set of parameter estimates for each site [26].

$$Y_{GWR}(s_0) = \beta(s_0) + \sum_{i=1}^{k} \beta(s_0)X(s_0) \tag{4}$$

where $Y_{GWR}(s_0)$ is the estimated value of the dependent variable Y at point s_0, $X_k(s_0)$ is the measured value of the kth explanatory variable at point s_0, $\beta_i(s_0)$ is the local estimation, and k is the number of independent variables.

2.4.5. Back-Propagation Neural Network Model

As a multi-layer feed-forward neural network, the BPNN is trained according to the error reverse propagation algorithm and uses the gradient search technique to minimize the error of the network's actual and expected output values [49]. In this paper, the BPNN model was used for SOC prediction, and a 3-layer BPNN was constructed, i.e., with an input, middle, and output layer (Figure 4).

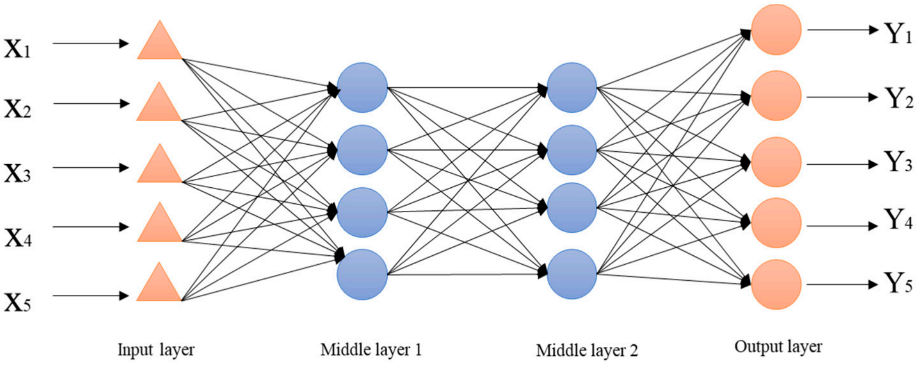

Figure 4. Structure of a 3-layer BPNN.

2.4.6. Precision Evaluation

The 82 randomly generated points used for fitting were involved in all analysis processes, and the remaining 35 validation points were used to verify the spatial results.

Both the root mean square error (RMSE) and mean relative estimation error (MRE) among the predicted and test data are commonly used to assess the accuracy of the model [50]. The MRE and RMSE of the estimates and measurements at the validation points were used to evaluate the precision of the different methods in predicting the SOC content of the study area. The lower values of the RMSE and MRE demonstrated that the model was more accurate [51]:

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{6}$$

where $\hat{y}_i$ is the estimated value of the SOC at i, y_i is the measured value of SOC at i, n is the number of validation points; in this paper, $n = 35$. When MRE and $RMSE$ are close to 0, the interpolation accuracy is higher.

2.5. Data Preprocessing

The OK, GWR, BPNN, and RF models were selected to predict the spatial distribution of soil organic carbon, based on the applicability of different spatial distribution prediction models in the study area. The 117 datasets described in Section 2.2 were randomly divided into a 70% fitting set and 30% validation set. The SOC content was calculated using Microsoft Excel 2021 software, and the statistical analysis of the relevant data was completed with SPSS 26.0 software. The multisource environmental factors were obtained from the Chinese Geospatial Data Cloud, and the data were processed with ENVI 5.3 and ArcGIS 10.6. The OK method was performed with ArcGIS 10.6, the RF model with R4.2.0, the BPNN model with MATLAB 2022, and the GWR model with GWR4 software.

3. Results

3.1. Descriptive Statistics

The results showed that the SOC content ranged from 1.176 to 53.134 g/kg (Table 2). The coefficient of variation for the SOC content was 0.544, which indicated that the spatial variability was strong. According to the Second National Soil Census Nutrient Classification Standards [52], the average organic carbon content in the study area was at a medium level (10~20 g/kg). Comparing the SOC content in other arid and semiarid regions (Shaanxi 8.2 g/kg [53] and Gansu 11.17 g/kg [54]), the SOC content in the study area was high.

Table 2. Descriptive statistical characteristics of SOC content.

	Soil Sample Sites	No. of Samples	Max(g/kg)	Min(g/kg)	Mean(g/kg)	Standard Deviation	Coefficient of Variation	Skewness	Kurtosis
Group 1	Fitted points.1	82	53.134	1.178	12.155	7.419	0.610	3.447	15.704
	Validation points.1	35	18.815	3.710	11.105	3.288	0.296	−0.023	0.158
Group 2	Fitted points.2	82	36.728	1.178	11.391	4.575	0.402	1.913	10.756
	Validation points.2	35	53.134	3.710	12.896	9.383	0.728	3.340	12.052
	Total sample points	117	53.134	1.178	11.841	6.440	0.544	3.643	17.154

To ensure the validity of the experimental results, the dataset was randomly divided into 70% fitted points and 30% validation points, randomly selected twice and divided into two groups; the group with a low degree of variability was selected as the final result. The mean value of the SOC content was 12.155 g/kg at the fitted site and 11.105 g/kg at the validation site in Group 1. The mean value of the SOC content was 11.391 g/kg at the fitted site and 12.896 g/kg at the validation site in Group 2. The degree of variability for Group 1 was better than that of Group 2, so Group 1 data were chosen for a spatial prediction analysis.

3.2. Relationship between SOC Content and Multisource Environmental Factors

3.2.1. Importance Ranking of the Influencing Factors

Environmental variables such as soil physical, chemical, and topographic factors determine SOC distribution on regional and continuous scales [55,56]. The significance and ordering of the effects of different environmental factors on SOC levels in the study area greatly varied. As shown in Figure 5, EC contributed the most to the accuracy of SOC spatial distribution prediction, having an importance of 38.65%, followed by SWIR.2 at 24.01%, then NIR at 14.85%, slope at 13.31%, and SWIR.1 at 13.08%; the remainder had an importance of less than 10%. Among them, the remote sensing factor accounted for the largest share of the overall importance at 44%, the soil physicochemical factor accounted for 27%, the topographic factor for 21%, and the geographical coordinates for 8%.

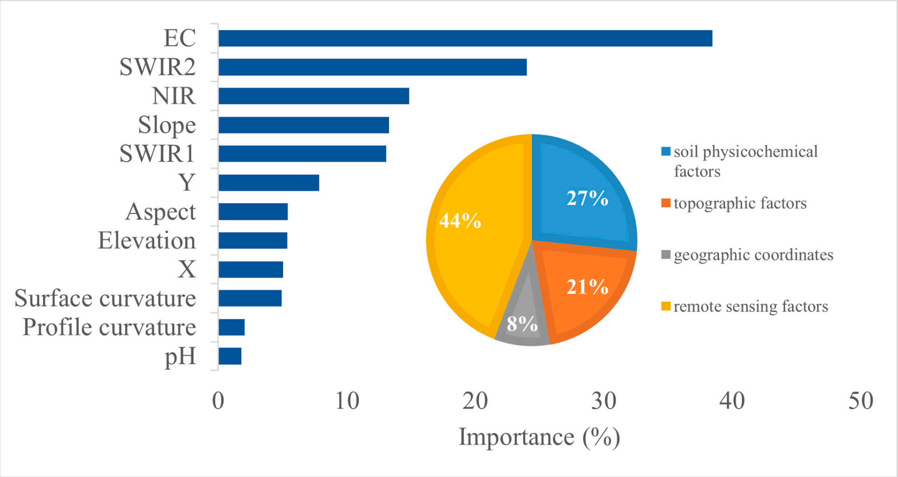

Figure 5. Importance ranking of the influencing factors (X: longitude, Y: latitude, SWIR.1: short infrared wave 1, NIR: near-infrared, SWIR 2: short infrared wave 2, and EC: soil conductivity).

3.2.2. Interactive Correlation of Impact Factors

The results of the interaction correlation analysis (Figure 6) showed that the SOC had different correlations with the six auxiliary variables, X, EC, SWIR.1, SWIR.2, DEM, and slope, at different lag distances and in different directions. There was an interactive negative correlation between SOC and X when the lag distance was four, with an interactive correlation coefficient of −0.184. In addition, EC, SWIR.1, SWIR.2, DEM, and slope interacted with the SOC at lag distances of 7, 0, 4, 3, and 6. Among them, X, SWIR.1, and DEM showed an interactive negative correlation with SOC, while EC, SWIR.2 and NIR, pH, Y, aspect, surface curvature, and profile curvature were not interactively correlated with SOC at lag distances from −7 to 7. These illustrated the non-linear relationship between the environmental factors and SOC.

3.3. Spatial Distribution of Organic Carbon Content

The general trend of the spatial distribution of SOC content predicted with the OK, GWK, BPNN, and RF models was basically the same, with all showing a spatial pattern of higher values in the western region and lower values in the eastern region (Figure 7). The predictive SOC content ranged from 5.97 to 30.72 g/kg, 10.15 to 14.18 g/kg, 7.40 to 27.09 g/kg, and 9.54 to 13.79 g/kg for the OK, GWR, BPNN, and RF models, respectively. The results in Figure 8 showed that the BPNN frequencies (frequencies = 60) were the closest statistical ranges (frequencies = 60) to the original SOC data compared with those of the OK, GWR, and RF models, indicating a more accurate range of SOC prediction for the BPNN model. The relative importance of environmental factors (Figure 5) suggested that EC, SWIR.2, and NIR were the main environmental factors affecting SOC content. Combined with the spatial distribution map of environmental factors (Figure 2), the above influencing factors showed higher values in the northwestern region than in the other regions. This was consistent with the spatial distribution of SOC content in the study area.

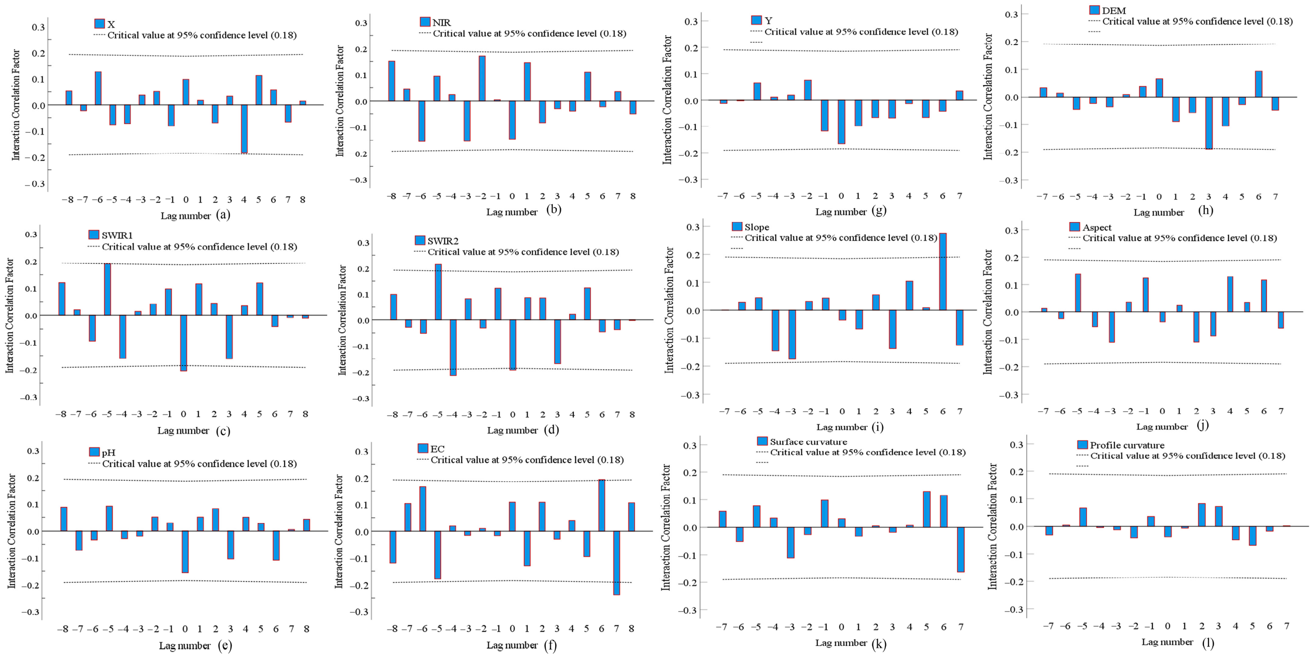

Figure 6. (**a**–**l**) Interactive correlation analysis between SOC content and factors.

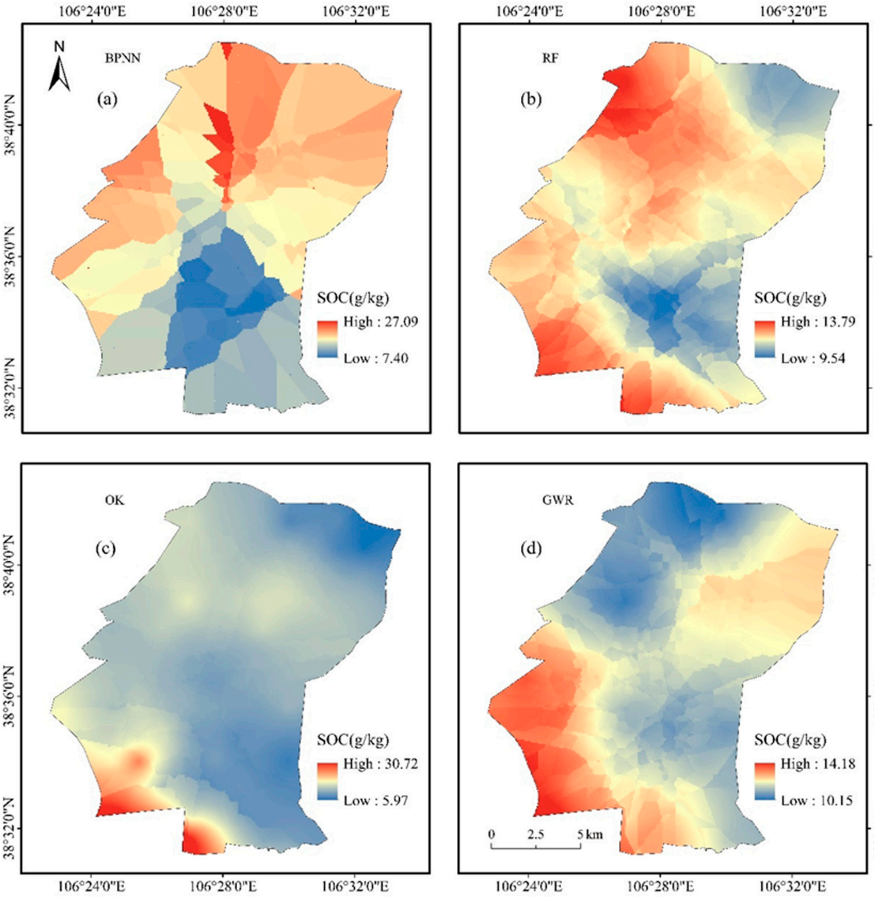

Figure 7. Simulation of the spatial distribution of SOC content with BPNN (**a**), RF (**b**), OK (**c**), and GWR (**d**) models.

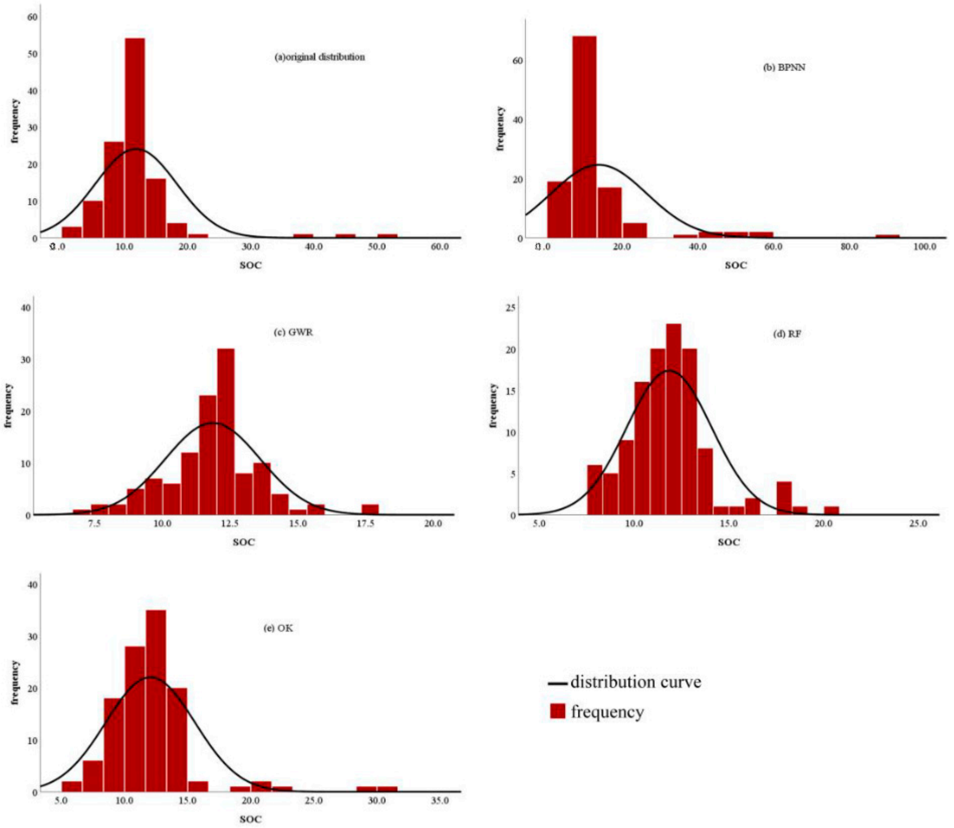

Figure 8. SOC content histogram comparison by original distribution; (**a**), BPNN (**b**), GWR (**c**), RF (**d**), and OK (**e**) models.

3.4. Prediction Precision Analysis

The MRE and RMSE of the values measured and predicted with the OK method were higher than those of the other prediction methods, which indicated its lower accuracy (Table 3). In contrast, the BPNN's MRE and RMSE were the smallest, indicating that the accuracy of the prediction was the best (Table 3). This was consistent with Reda's [57] results which predicted SOC using different machine learning algorithms; the BPNN prediction accuracy was better.

Table 3. MRE and RMSE for different SOC content prediction methods.

Prediction Method	MRE	RMSE
OK	0.198	4.248
RF	0.186	3.320
GWR	0.193	3.595
BPNN	0.066	0.257

The relative accuracy of SOC content prediction in the control group was improved by 1.2%, 13.2%, and 0.5% for the RF, BP, and GWR models, respectively, compared with the OK model. In addition, the RF and BPNN models improved by 0.7% and 12.7%, respectively, with the GWR model as the reference, and the BPNN model improved by 12% with the RF model as the reference. This result indicated that the BPNN model can better capture the complex relationship between SOC and environmental factors, having an overall higher prediction accuracy and better prediction results than those of the OK, RF, or GWR models. Tang et al. [58] predicted wheat yield using a linear regression model and a BP neural network; their results indicated that the BP neural network predicted better than the linear regression model. Yang et al. [59] used the Kriging and BPNN models to predict the

Normalized Difference Vegetation Index (NDVI), demonstrating that the BPNN produced higher estimations than the Kriging model. Reda et al. [57] used Partial Least-Squares Regression (PLS), a BPNN, and Ensemble Learning Modeling (ELM) to predict SOC and total soil nitrogen (STN); their results indicated that the BPNN showed excellent accuracy in the prediction of SOC and STN. In conclusion, the BPNN prediction accuracy has been shown to be superior to other models, as well as being suitable for predicting different variables in different regions.

4. Discussion and Conclusions

4.1. Discussion

The average SOC content in the surface layer of farmland in Helan County was 11.841 g/kg, which was a medium level. The analysis of the SOC content characteristics of the study area showed that there was high variability in the SOC content in the study area. The reason for this was that the SOC content of farmland is closely related to anthropogenic activities. Previous studies have shown that higher carbon inputs usually increase SOC storage [60], and eventually, SOC accumulation improves soil properties, hydraulic conductivity, and agglomeration stability [61]. Wu et al. [62] found that agricultural land has often been ploughed and fertilized for a long time, with a thick plough layer and high SOC content. Ebhin [63] and Ma [64] found that the SOC content in farmland receiving farmyard manure fertilizer or straw, as well as NPK fertilizer, was much greater than that in farmland receiving only NPK chemical fertilizer. Frequently, the animal manure from the farms in the study area is scattered in the surrounding fields, causing an increase in the SOC content of the surrounding farmland. Further investigation showed that some areas of the study fields had been cultivated using fertilizer and animal manure, while others had been cultivated using chemical fertilizer alone, resulting in large variations in SOC content in some parts of the study area.

In this study, the RF, OK, BPNN, and GWR algorithms were selected for SOC content prediction. The prediction accuracy results demonstrated that the BPNN > RF > GWR > OK. The nonlinear relationship between certain environmental auxiliary factors and SOC may have impacted geographical prediction. Among the four algorithms, the BPNN was more suitable for predicting the spatial distribution of SOC in the study area. The researchers found that the BPNN algorithm essentially implemented a mapping function from inputs to outputs, and mathematical theory suggested that the BPNN algorithm had the ability to implement complex nonlinear mapping [65]. In addition, there was less influence from environmental factors in the BPNN model, so its prediction accuracy was improved compared with that of the other models [66]. There are many factors that affect the accuracy of SOC content prediction results in farmland, such as climate change and the spatial resolution of remote sensing images [29,37]. Therefore, more comprehensive investigations into the influence of human factors on SOC are required to improve the accuracy of SOC measurement [67].

4.2. Conclusions

(1) The SOC content ranged from 1.178 to 53.134 g/kg, with an average of 11.841 g/kg, which is a medium level. The SOC greatly varied and was distributed unevenly in the study area, mainly due to the nonuniform application of fertilizer during crop cultivation.

(2) The main environmental factors affecting the spatial distribution of SOC in this study area were EC values and remote sensing factors, with EC values accounting for 38% of the relative importance and remote sensing factors accounting for 44% of all environmental factors.

(3) Compared with the OK (MRE = 0.198, RMSE = 4.248), GWR (MRE = 0.193, RMSE = 3.595), and RF (MRE = 0.186, RMSE = 3.320) models, the prediction accuracy and results of the BPNN (MRE = 0.066, RMSE = 0.257) model were better, and the simulated spatial distribution map could better represent the actual distribution of the SOC. The results

showed that the BPNN model was more suitable for the prediction of the distribution of SOC in the study area.

(4) In this study, four different models were selected to predict the SOC content in Helan County, and the optimal model suitable for the prediction of SOC content was determined; this provides theoretical support for refined agricultural management. This study was limited to the applicability analysis of the four models described above; however, our method requires enhancement to render it applicable to the spatial distribution of SOC.

Author Contributions: Conceptualization, Y.Z. and Y.W.; methodology, Y.Z. and Y.B.; software, X.L.; validation, R.Z.; formal analysis, X.M. and Y.Z.; investigation, Y.Z.; resources, Y.Z. and X.M.; data curation, Y.Z. and R.Z.; writing—original draft, Y.Z.; writing—review and editing, Y.W. and Y.B.; visualization, Y.W. and Y.B.; supervision, Y.W. and Y.B.; funding acquisition, Y.W. and Y. B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Natural Science Foundation of Ningxia Hui Autonomous Region Project (2023AAC03046, 2023AAC02018) and the National Natural Science Foundation of China (NSFC) (32360321).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to have influenced the work reported in this paper.

Abbreviations

Abridgement	**Full Name**
OK	ordinary Kriging
GWR	geographically weighted regression
RF	random forest
BPNN	back-propagation neural network
SOC	soil organic carbon
ML	machine learning
pH	potential of hydrogen
MLR	multiple linear regression
DEM	digital elevation model
SWIR1	short infrared wave 1
SWIR2	short infrared wave 2
NIR	near-infrared band
EC	electrical conductivity
VIF	variance inflation factor
X	longitude
Y	latitude
RMSE	root mean square error
MRE	mean relative estimation error
NDVI	normalized difference vegetation index
PLS	partial least squares regression
ELM	ensemble learning modeling
STN	total soil nitrogen

References

1. Freier, K.P.; Glaser, B.; Zech, W. Mathematical modeling of soil carbon turnover in natural Podocarpus forest and Eucalyptus plantation in Ethiopia using compound specific δ13C analysis. *Glob. Chang. Biol.* **2010**, *16*, 1487–1502. [CrossRef]
2. Kell, D.B. Large-scale sequestration of atmospheric carbon via plant roots in natural and agricultural ecosystems: Why and how. *Philos. Trans. R. Soc. B Biol. Sci.* **2012**, *367*, 1589–1597. [CrossRef] [PubMed]
3. Hou, G.; Delang, C.O.; Lu, X.; Gao, L. Grouping tree species to estimate afforestation-driven soil organic carbon sequestration. *Plant Soil* **2020**, *455*, 507–518. [CrossRef]
4. Lal, R. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. *Science* **2004**, *304*, 1623–1627. [CrossRef] [PubMed]

5. Rovai, A.S.; Twilley, R.R.; Castañeda-Moya, E.; Riul, P.; Cifuentes-Jara, M.; Manrow-Villalobos, M.; Horta, P.A.; Simonassi, J.C.; Fonseca, A.L.; Pagliosa, P.R. Global controls on carbon storage in mangrove soils. *Nat. Clim. Chang.* **2018**, *8*, 534–538. [CrossRef]
6. Bilen, S.; Turan, V. Enzymatic analyses in soils. In *Practical Handbook on Agricultural Microbiology*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 377–385.
7. Dvorakova, K.; Heiden, U.; Pepers, K.; Staats, G.; van Os, G.; van Wesemael, B. Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties. *Geoderma* **2023**, *429*, 116128. [CrossRef]
8. Mishra, U.; Lal, R.; Liu, D.; Van Meirvenne, M. Predicting the Spatial Variation of the Soil Organic Carbon Pool at a Regional Scale. *Soil Sci. Soc. Am. J.* **2010**, *74*, 906–914. [CrossRef]
9. Fang, X.; Xue, Z.; Li, B.; An, S. Soil organic carbon distribution in relation to land use and its storage in a small watershed of the Loess Plateau, China. *Catena* **2012**, *88*, 6–13. [CrossRef]
10. Guoju, X.; Yanbin, H.; Qiang, Z.; Jing, W.; Ming, L. Impact of cultivation on soil organic carbon and carbon sequestration potential in semiarid regions of China. *Soil Use Manag.* **2020**, *36*, 83–92. [CrossRef]
11. Nguyen, T.T.; Pham, T.D.; Nguyen, C.T.; Delfos, J.; Archibald, R.; Dang, K.B.; Hoang, N.B.; Guo, W.; Ngo, H.H. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Sci. Total Environ.* **2022**, *804*, 150187. [CrossRef]
12. Gholizadeh, A.; Žižala, D.; Saberioon, M.; Borůvka, L. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sens. Environ.* **2018**, *218*, 89–103. [CrossRef]
13. Scaioni, M.; Longoni, L.; Melillo, V.; Papini, M. Remote Sensing for Landslide Investigations: An Overview of Recent Achievements and Perspectives. *Remote Sens.* **2014**, *6*, 9600–9652. [CrossRef]
14. Racek, D.; Thurner, P.W.; Davidson, B.I.; Zhu, X.X.; Kauermann, G. Conflict forecasting using remote sensing data: An application to the Syrian civil war. *Int. J. Forecast.* **2023**. [CrossRef]
15. Chai, H.; Rao, S.; Wang, R.; Liu, J.; Huang, Q.; Mou, X. The Effect of the Geomorphologic Type as Surrogate to the Time Factor on Digital Soil Mapping. *Open J. Soil Sci.* **2015**, *5*, 12. [CrossRef]
16. Abdoli, P.; Khanmirzaei, A.; Hamzeh, S.; Rezaei, S.; Moghimi, S. Use of remote sensing data to predict soil organic carbon in some agricultural soils of Iran. *Remote Sens. Appl. Soc. Environ.* **2023**, *30*, 100969. [CrossRef]
17. Dhiman, G.; Bhattacharya, J.; Roy, S. Soil textures and nutrients estimation using remote sensing data in north india—Punjab region. *Procedia Comput. Sci.* **2023**, *218*, 2041–2048. [CrossRef]
18. Jia, P.; Shang, T.; Zhang, J.; Sun, Y. Inversion of soil pH during the dry and wet seasons in the Yinbei region of Ningxia, China, based on multi-source remote sensing data. *Geoderma Reg.* **2021**, *25*, e00399. [CrossRef]
19. Kalambukattu, J.G.; Kumar, S.; Arya Raj, R. Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. *Environ. Earth Sci.* **2018**, *77*, 203. [CrossRef]
20. Zhao, X.; Zhao, D.; Wang, J.; Triantafilis, J. Soil organic carbon (SOC) prediction in Australian sugarcane fields using Vis–NIR spectroscopy with different model setting approaches. *Geoderma Reg.* **2022**, *30*, e00566. [CrossRef]
21. Hao, J.; Lin, Y.; Ren, G.; Yang, G.; Han, X.; Wang, X.; Ren, C.; Feng, Y. Comprehensive benefit evaluation of conservation tillage based on BP neural network in the Loess Plateau. *Soil Tillage Res.* **2021**, *205*, 104784. [CrossRef]
22. Li, Q.-Q.; Zhang, X.; Wang, C.-Q.; Li, B.; Gao, X.-S.; Yuan, D.-G.; Luo, Y.-L. Spatial prediction of soil nutrient in a hilly area using artificial neural network model combined with kriging. *Arch. Agron. Soil Sci.* **2016**, *62*, 1541–1553. [CrossRef]
23. Halecki, W.; Młyński, D.; Ryczek, M.; Kruk, E.; Radecki-Pawlik, A. Applying an Artificial Neural Network (ANN) to Assess Soil Salinity and Temperature Variability in Agricultural Areas of a Mountain Catchment. *Pol. J. Environ. Stud.* **2017**, *26*, 2545–2554. [CrossRef] [PubMed]
24. Lu, W.; Luo, H.; He, L.; Duan, W.; Tao, Y.; Wang, X.; Li, S. Detection of heavy metals in vegetable soil based on THz spectroscopy. *Comput. Electron. Agric.* **2022**, *197*, 106923. [CrossRef]
25. Wang, D.; Li, X.; Zou, D.; Wu, T.; Xu, H.; Hu, G.; Li, R.; Ding, Y.; Zhao, L.; Li, W.; et al. Modeling soil organic carbon spatial distribution for a complex terrain based on geographically weighted regression in the eastern Qinghai-Tibetan Plateau. *Catena* **2020**, *187*, 104399. [CrossRef]
26. Yuan, Y.; Cave, M.; Xu, H.; Zhang, C. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). *J. Hazard. Mater.* **2020**, *393*, 122377. [CrossRef]
27. Emadi, M.; Taghizadeh-Mehrjardi, R.; Cherati, A.; Danesh, M.; Mosavi, A.; Scholten, T. Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sens.* **2020**, *12*, 2234. [CrossRef]
28. Hu, S.; Xiong, C.; Chen, P.; Schonfeld, P. Examining nonlinearity in population inflow estimation using big data: An empirical comparison of explainable machine learning models. *Transp. Res. Part A Policy Pract.* **2023**, *174*, 103743. [CrossRef]
29. Zeraatpisheh, M.; Ayoubi, S.; Mirbagheri, Z.; Mosaddeghi, M.R.; Xu, M. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. *Geoderma Reg.* **2021**, *27*, e00440. [CrossRef]
30. Mahmoudzadeh, H.; Matinfar, H.R.; Taghizadeh-Mehrjardi, R.; Kerry, R. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. *Geoderma Reg.* **2020**, *21*, e00260. [CrossRef]
31. Takoutsing, B.; Heuvelink, G.B.M. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. *Geoderma* **2022**, *428*, 116192. [CrossRef]

32. Tan, C.; Yang, J.; Li, M. Temporal-Spatial Variation of Drought Indicated by SPI and SPEI in Ningxia Hui Autonomous Region, China. *Atmosphere* **2015**, *6*, 1399–1421. [CrossRef]
33. Zeraatpisheh, M.; Garosi, Y.; Reza Owliaie, H.; Ayoubi, S.; Taghizadeh-Mehrjardi, R.; Scholten, T.; Xu, M. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *Catena* **2022**, *208*, 105723. [CrossRef]
34. Wang, H.; Yang, Q.; Ma, H.; Liang, J. Chemical compositions evolution of groundwater and its pollution characterization due to agricultural activities in Yinchuan Plain, northwest China. *Environ. Res.* **2021**, *200*, 111449. [CrossRef] [PubMed]
35. Zeraatpisheh, M.; Galford, G.L.; White, A.; Noel, A.; Darby, H.; Adair, E.C. Soil organic carbon stock prediction using multi-spatial resolutions of environmental variables: How well does the prediction match local references? *Catena* **2023**, *229*, 107197. [CrossRef]
36. Wang, Q.; Le Noë, J.; Li, Q.; Lan, T.; Gao, X.; Deng, O.; Li, Y. Incorporating agricultural practices in digital mapping improves prediction of cropland soil organic carbon content: The case of the Tuojiang River Basin. *J. Environ. Manag.* **2023**, *330*, 117203. [CrossRef]
37. Mondal, A.; Khare, D.; Kundu, S.; Mondal, S.; Mukherjee, S.; Mukhopadhyay, A. Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data. *Egypt. J. Remote Sens. Space Sci.* **2017**, *20*, 61–70. [CrossRef]
38. Moloney, J.P.; Malone, B.P.; Karunaratne, S.; Stockmann, U. Leveraging large soil spectral libraries for sensor-agnostic field condition predictions of several agronomically important soil properties. *Geoderma* **2023**, *439*, 116651. [CrossRef]
39. Salazar, O.; Benvenuto, A.; Fajardo, M.; Fuentes, J.P.; Nájera, F.; Celedón, A.; Pfeiffer, M.; Renwick, L.L.R.; Seguel, O.; Tapia, Y.; et al. Evaluation of a miniaturized portable NIR spectrometer for the prediction of soil properties in Mediterranean central Chile. *Geoderma Reg.* **2023**, *34*, e00675. [CrossRef]
40. Zepp, S.; Heiden, U.; Bachmann, M.; Möller, M.; Wiesmeier, M.; van Wesemael, B. Optimized bare soil compositing for soil organic carbon prediction of topsoil croplands in Bavaria using Landsat. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 287–302. [CrossRef]
41. Yang, Z.; Xiaomin, Z.; Xi, G. Prediction of Total Nitrogen Distribution in Surface Soil Based on Multi-source Auxiliary Variables and Random Forest Approach. *Acta Pedol. Sin.* **2021**, *59*, 451–460. [CrossRef]
42. Vu, D.H.; Muttaqi, K.M.; Agalgaonkar, A.P. A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. *Appl. Energy* **2015**, *140*, 385–394. [CrossRef]
43. Cheng, J.; Sun, J.; Yao, K.; Xu, M.; Cao, Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *268*, 120652. [CrossRef] [PubMed]
44. Gómez, R.S.; Sánchez, A.R.; García, C.G.; Pérez, J.G. The VIF and MSE in Raise Regression. *Mathematics* **2020**, *8*, 605. [CrossRef]
45. Maiti, P.; Mitra, D. Complexity reduction of ordinary kriging algorithm for 3D REM design. *Phys. Commun.* **2022**, *55*, 101912. [CrossRef]
46. Li, X.; Geng, T.; Shen, W.; Zhang, J.; Zhou, Y. Quantifying the influencing factors and multi-factor interactions affecting cadmium accumulation in limestone-derived agricultural soil using random forest (RF) approach. *Ecotoxicol. Environ. Saf.* **2021**, *209*, 111773. [CrossRef]
47. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
48. Lindner, C.; Bromiley, P.A.; Ionita, M.C.; Cootes, T.F. Robust and accurate shape model matching using random forest regression-voting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1862–1874. [CrossRef] [PubMed]
49. Wang, L.; Zhou, H.; Yang, J.; Xiong, Y.; She, J.; Chen, W. A decision support system for tobacco cultivation measures based on BPNN and GA. *Comput. Electron. Agric.* **2021**, *181*, 105928. [CrossRef]
50. Ullah, Z.; Khan, M.; Raza Naqvi, S.; Farooq, W.; Yang, H.; Wang, S.; Vo, D.-V.N. A comparative study of machine learning methods for bio-oil yield prediction—A genetic algorithm-based features selection. *Bioresour. Technol.* **2021**, *335*, 125292. [CrossRef]
51. Fang, Y.; Ma, L.; Yao, Z.; Li, W.; You, S. Process optimization of biomass gasification with a Monte Carlo approach and random forest algorithm. *Energy Convers. Manag.* **2022**, *264*, 115734. [CrossRef]
52. Xian-ju, X.; Yong-chun, Z.; Ji-dong, W.; Hui, Z.; Zhong-hou, T.; Ai-jun, Z.; Huan, L.; Qing, L. Soil nutrient status and soil fertility evaluation of farmland in three main sweet potato regions in China. *Soil Fertil. Sci. China* **2021**, *5*, 27–33. [CrossRef]
53. Zhang, F.; Li, S.; Yue, S.; Song, Q. The effect of long-term soil surface mulching on SOC fractions and the carbon management index in a semiarid agroecosystem. *Soil Tillage Res.* **2022**, *216*, 105233. [CrossRef]
54. Dou, X.; Zhang, J.; Zhang, C.; Ma, D.; Chen, L.; Zhou, G.; Li, J.; Duan, Y. Calcium carbonate regulates soil organic carbon accumulation by mediating microbial communities in northern China. *Catena* **2023**, *231*, 107327. [CrossRef]
55. Xu, X.; Shi, Z.; Li, D.; Rey, A.; Ruan, H.; Craine, J.M.; Liang, J.; Zhou, J.; Luo, Y. Soil properties control decomposition of soil organic carbon: Results from data-assimilation analysis. *Geoderma* **2016**, *262*, 235–242. [CrossRef]
56. Wiesmeier, M.; Urbanski, L.; Hobley, E.; Lang, B.; von Lützow, M.; Marin-Spiotta, E.; van Wesemael, B.; Rabot, E.; Ließ, M.; Garcia-Franco, N.; et al. Soil organic carbon storage as a key function of soils—A review of drivers and indicators at various scales. *Geoderma* **2019**, *333*, 149–162. [CrossRef]
57. Reda, R.; Saffaj, T.; Ilham, B.; Saidi, O.; Issam, K.; Brahim, L.; El Hadrami, E.M. A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. *Chemom. Intell. Lab. Syst.* **2019**, *195*, 103873. [CrossRef]
58. Tang, X.; Liu, H.; Feng, D.; Zhang, W.; Chang, J.; Li, L.; Yang, L. Prediction of field winter wheat yield using fewer parameters at middle growth stage by linear regression and the BP neural network method. *Eur. J. Agron.* **2022**, *141*, 126621. [CrossRef]

59. Yang, Y.; Zhu, J.; Zhao, C.; Liu, S.; Tong, X. The spatial continuity study of NDVI based on Kriging and BPNN algorithm. *Math. Comput. Model.* **2011**, *54*, 1138–1144. [CrossRef]
60. Virto, I.; Barré, P.; Burlot, A.; Chenu, C. Carbon input differences as the main factor explaining the variability in soil organic C storage in no-tilled compared to inversion tilled agrosystems. *Biogeochemistry* **2012**, *108*, 17–26. [CrossRef]
61. Liu, E.; Yan, C.; Mei, X.; He, W.; Bing, S.H.; Ding, L.; Liu, Q.; Liu, S.; Fan, T. Long-term effect of chemical fertilizer, straw, and manure on soil chemical and biological properties in northwest China. *Geoderma* **2010**, *158*, 173–180. [CrossRef]
62. Wu, Z.; Chen, Y.; Zhu, Y.; Feng, X.; Ou, J.; Li, G.; Tong, Z.; Yan, Q. Mapping Soil Organic Carbon in Floodplain Farmland: Implications of Effective Range of Environmental Variables. *Land* **2023**, *12*, 1198. [CrossRef]
63. Ebhin Masto, R.; Chhonkar, P.K.; Singh, D.; Patra, A.K. Changes in soil biological and biochemical characteristics in a long-term field trial on a sub-tropical inceptisol. *Soil Biol. Biochem.* **2006**, *38*, 1577–1582. [CrossRef]
64. Ma, G.; Cheng, S.; He, W.; Dong, Y.; Qi, S.; Tu, N.; Tao, W. Effects of Organic and Inorganic Fertilizers on Soil Nutrient Conditions in Rice Fields with Varying Soil Fertility. *Land* **2023**, *12*, 1026. [CrossRef]
65. Wang, Q.; Mao, X.; Jiang, X.; Pei, D.; Shao, X. Digital image processing technology under backpropagation neural network and K-Means Clustering algorithm on nitrogen utilization rate of Chinese cabbages. *PLoS ONE* **2021**, *16*, e0248923. [CrossRef]
66. Gao, C.; Yan, X.; Qiao, X.; Wei, K.; Zhang, X.; Yang, S.; Wang, C.; Yang, W.; Feng, M.; Xiao, L.; et al. Multivariate prediction of soil aggregate-associated organic carbon by simulating satellite sensor bands. *Comput. Electron. Agric.* **2023**, *209*, 107859. [CrossRef]
67. Odebiri, O.; Odindi, J.; Mutanga, O. Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102389. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.