remote sensing

MDPI

Article

A New Method for Estimating Soil Fertility Using Extreme Gradient Boosting and a Backpropagation Neural Network

Yiping Peng [1,†], Zhenhua Liu [1,2,†], Chenjie Lin [1], Yueming Hu [2,3,4,*], Li Zhao [1,4], Runyan Zou [1,4], Ya Wen [1] and Xiaoyun Mao [1]

[1] College of Natural Resources and Environment, South China Agricultural University, Guangzhou 510642, China; pengyp@stu.scau.edu.cn (Y.P.); zhenhua@scau.edu.cn (Z.L.); linchenjie@stu.scau.edu.cn (C.L.); zhaoli@stu.scau.edu.cn (L.Z.); zry804@stu.scau.edu.cn (R.Z.); wenya26@scau.edu.cn (Y.W.); xymao@scau.edu.cn (X.M.)
[2] Key Laboratory of Construction Land Transformation, Ministry of Land and Resources, South China Agricultural University, Guangzhou 510642, China
[3] College of Tropical Crops, Hainan University, Haikou 570228, China
[4] South China Academy of Natural Resources Science and Technology, Guangzhou 510642, China
* Correspondence: ymhu@scau.edu.cn; Tel.: +86-020-852-88307
† These authors contributed equally to this work.

Abstract: Soil fertility affects crop yield and quality. A quick, accurate evaluation of soil fertility is crucial for agricultural production. Few satellite image-based evaluation studies have quantified soil fertility during the crop growth period. Therefore, this study proposes a new approach to the quantitative evaluation of soil fertility. Firstly, the optimal crop spectral variables were selected using the integration of an extreme gradient boosting (XGBoost) algorithm with variance inflation factor (VIF). Then, based on the optimal crop spectral variables where the red-edge indices were introduced for the first time, the estimation models were developed using the backpropagation neural network (BPNN) algorithm to assess soil fertility. The model was finally adopted to map the soil fertility using Sentinel-2 imagery. This study was performed in the Conghua District of Guangzhou, Guangdong Province, China. The results of our research are as follows: (1) five crop spectral variables (inverted red-edge chlorophyll index (IRECI), chlorophyll vegetation index (CVI), normalized green-red difference index (NGRDI), red-edge position (REP), and triangular greenness index (TGI)) were the optimal variables. (2) The BPNN model established with optimal variables provided reliable estimates of soil fertility, with the determination coefficient (R^2) of 0.66 and a root mean square error (RMSE) of 0.17. A nonlinear relation was found between soil fertility and the optimal crop spectral variables. (3) The BPNN model provides the potential for soil fertility mapping using Sentinel-2 images, with an R^2 of 0.62 and an RMSE of 0.09 for the measured and estimated results. This study suggests that the proposed method is suitable for the estimation of soil fertility in paddy fields.

Keywords: soil fertility estimation; crop spectral variables; red-edge; XGBoost; BPNN

Citation: Peng, Y.; Liu, Z.; Lin, C.; Hu, Y.; Zhao, L.; Zou, R.; Wen, Y.; Mao, X. A New Method for Estimating Soil Fertility Using Extreme Gradient Boosting and a Backpropagation Neural Network. *Remote Sens.* **2022**, *14*, 3311. https://doi.org/10.3390/rs14143311

Academic Editors: Huajun Tang, Wenbin Wu and Wenjuan Li

Received: 13 June 2022
Accepted: 7 July 2022
Published: 9 July 2022

1. Introduction

Soil fertility refers to the capacity for soil to offer different nutrients for crop growth, significantly affecting crop yield [1,2]. Scientific and reasonable evaluation of soil fertility can provide a reference for land-use planning and fertilizer prescriptions, guiding agricultural production [3]. Soil fertility is typically evaluated using soil sampling and laboratory chemical analysis to determine the chemical properties (soil pH, soil organic matter (SOM), available phosphorus (AP), available potassium (AK), and total nitrogen (TN)). It proves to be time-consuming and expensive for deriving spatially explicit estimates across a large study area [4,5]. In contrast, remote sensing techniques can be used to evaluate

soil fertility using spectral variables and obtain spatially explicit estimates of soil fertility relatively quickly.

Current research methods for soil fertility estimation in arable land using remote sensing can be divided into soil spectrum-based and crop spectral index-based methods. The soil spectrum-based method evaluates the correlation between soil spectral indices and soil fertility. Rossel et al. [6] evaluated the soil fertility of sugarcane using a decision tree algorithm based on visible/near-infrared (vis-NIR) soil spectra and terrain attributes, which showed that the method could improve the efficiency of soil fertility evaluation. Munnaf et al. [7] developed a method for assessing soil fertility indices based on online vis-NIR soil spectroscopy. The results demonstrated that the method can be effective in assessing soil fertility. Yang et al. [8] evaluated soil fertility of paddy fields in southern China using vis–NIR soil spectral indices and partial least-squares regression. According to the results, vis–NIR spectroscopy improved the efficiency of estimating soil fertility. Wang et al. [9] explored the association of a comprehensive soil fertility index with soil spectral curves for agricultural soils in different states under optimal observation conditions. The results indicated that soil spectral indices were suitable for estimating soil fertility. However, arable land in southern China has few bare soil areas; therefore, it is difficult to obtain soil fertility using soil spectral indicators.

The crop spectral index-based method evaluates the correlation between spectral vegetation indices and soil fertility. Zeeshan et al. [10] found that a higher normalized difference vegetation index (NDVI) value corresponds to better soil fertility. Wang et al. [11] set thresholds for the NDVI based on cotton growth and performed density partitioning to obtain information on different levels of soil fertility in cotton fields. The research results showed that the NDVI could be used to estimate soil fertility. Duan et al. [12] used the mean and the coefficient of variation (CV) of the NDVI of arable land for three consecutive years to determine the magnitude and stability of soil fertility, respectively. A larger mean value and a smaller CV indicated higher soil fertility of the arable land. In these studies, soil fertility was estimated qualitatively. Moreover, only vis–NIR spectral indices were used to estimate soil fertility, and other spectral indices variables (e.g., red-edge) were not considered; thus, the accuracy of soil fertility estimation was not very high.

The contribution of this work was to design a novel method for the quantitative estimation of soil fertility in paddy fields using crop spectral variables. In this method, the machine learning algorithms (extreme gradient boosting (XGBoost) and backpropagation neural network (BPNN)) were used for the inaugural time to quantitatively assess soil fertility. For the first time, the red-edge index contributes to the soil fertility evaluation model, a method which provides the potential to quantitatively estimate soil fertility at both the soil sample level and regional scale. The study was performed in the Conghua District of Guangzhou, Guangdong Province, China using Sentinel-2 data.

2. Materials and Methods

In this study, a novel method for quantitative estimation of soil fertility was developed. The method combines the XGBoost algorithm for variable selection with the BPNN algorithm for soil fertility estimation. Figure 1 is the flow chart for the method. The methodological framework involves data collection, data pre-processing, determination of optimal crop spectral indicators, construction of the soil fertility estimation model, and the spatial mapping of soil fertility.

2.1. Study Area

The study area is located in the Conghua District of Guangzhou, Guangdong Province, China (113°17′E–114°04′E, 23°22′N–23°56′N). According to the statistical yearbook of the Conghua District, the annual average temperature was 22.3 °C, the annual rainfall was 1297.5 mm, and the annual sunshine hours were 1976.8 h in 2021. The main soil types in the Conghua District are red loam, yellow loam, lateritic red soil, and rice soil. The total area of arable land in Conghua is 205 km^2; most of it is located in the northwestern area of

the Conghua District. Arable land planted with rice was chosen because rice accounts for the highest proportion of cropland in the study area.

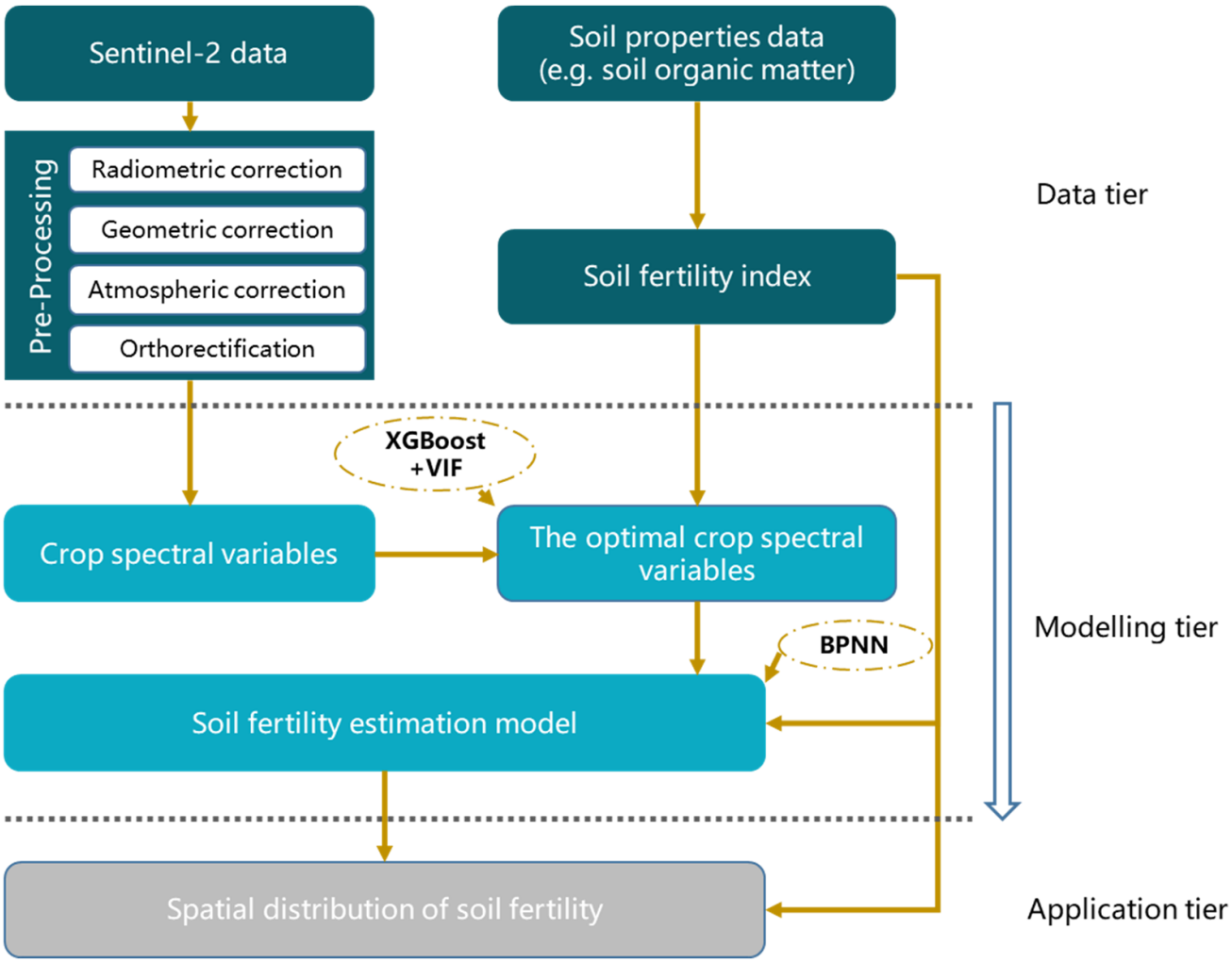

Figure 1. A flow chart displaying soil fertility estimation based on crop spectral variables.

2.2. Data Sources and Preprocessing

2.2.1. Soil Samples

A total of 150 sample points were obtained in the rice-growing area of Conghua District in September 2017 using stratified random sampling and considering different land units, soil types, land-use patterns, and agricultural facility construction levels. The 150 points fell into three groups: 90 samples (yellow plots of Figure 2) were introduced for model training, 30 samples (black plots of Figure 2) for evaluating model accuracy, and 30 samples (red plots of Figure 2) for validating mapping accuracy. At each sample point, five soil sub-samples of the topsoil (0–20 cm) were collected in an X shape, mixed, and used as the soil sample of this point. Basic information, such as the crop type, was recorded during sample collection, and the latitude and longitude of the sample points were recorded by a Global Positioning System (GPS) receiver. The samples underwent air drying at room temperature and were milled and screened via a 100-mesh sieve (0.15 mm) to observe soil properties. We used the regulation of classification of paddy soil fertility and fertilizer technology (DB43/T 2087-2021) and other information [5] to obtain the soil fertility index (*SFI*) of the 150 sample points using Equation (1):

$$SFI = \sum W_i \times N_i \quad (1)$$

where W_i and N_i represent the weight coefficient and membership degree of the ith indicator (soil pH, SOM, AP, AK, and TN). We followed the measurement procedures and methods described in Lu [13]. Table 1 shows the descriptive statistics for soil properties. W_i was obtained from correlation analysis between the indicators, and N_i was obtained from an affiliation function of the ith indicator. The boxplot of the SFI of the sample points is presented in Figure 3.

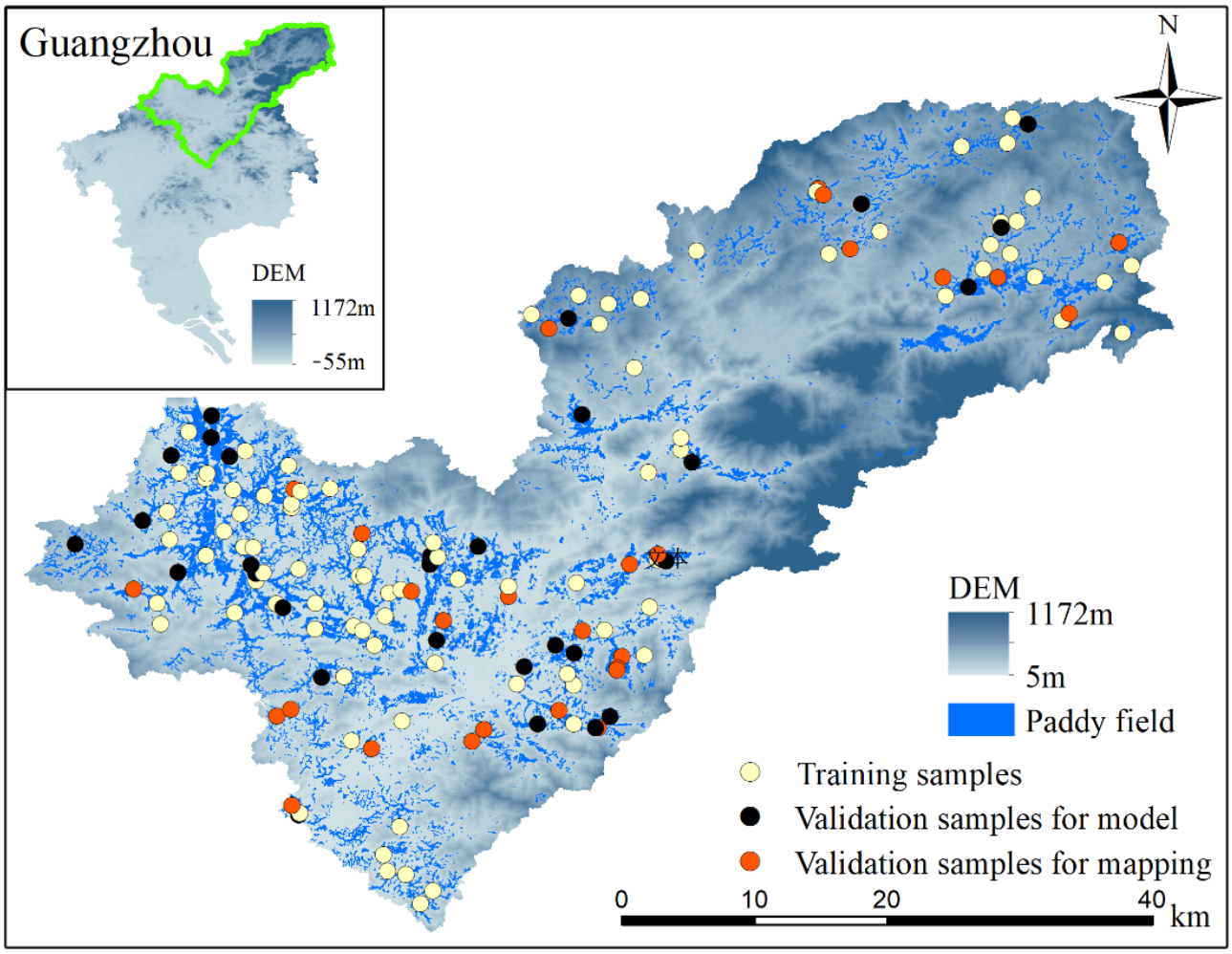

Figure 2. The study area and the spatial distribution of 150 soil sample plots (the sample plots used for training are yellow, those for validating the model are black, and those for validating the mapping results are red).

Table 1. Descriptive statistics of the soil properties.

Soil Properties	Min	Max	Mean	SD	Skewness	Kurtosis	CV (%)
pH	4.90	8.20	5.84	0.44	1.23	4.95	7.52
SOM	6.42	68.90	23.72	8.88	1.10	3.43	37.43
TN	0.37	2.14	0.86	0.42	1.22	0.83	47.95
AP	6.80	140.8	43.89	24.64	1.14	1.35	56.15
AK	2.00	235.00	74.30	49.19	1.06	0.54	66.21

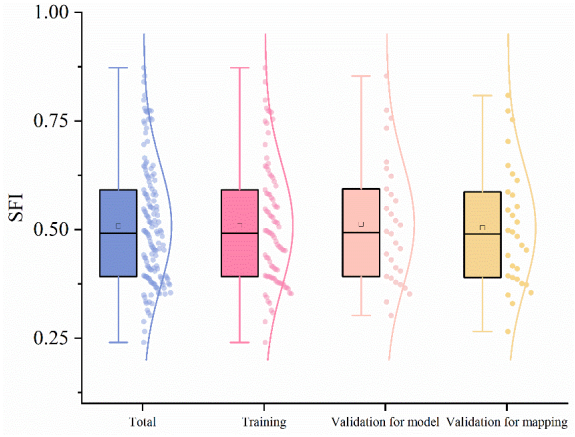

Figure 3. A boxplot of the soil fertility index of the sample points.

2.2.2. Satellite Image Data and Preprocessing

According to the rice key growth stages in the study area, 6 Sentinel-2 images (Figure 4) ranging from 17 September 2017 to 11 November 2017 were acquired. The band information (13 bands) and the spatial resolutions are listed in Table 2. The images were acquired from the Google Earth Engine platform. They had been preprocessed, including radiometric, geometric, and atmospheric corrections and orthorectification.

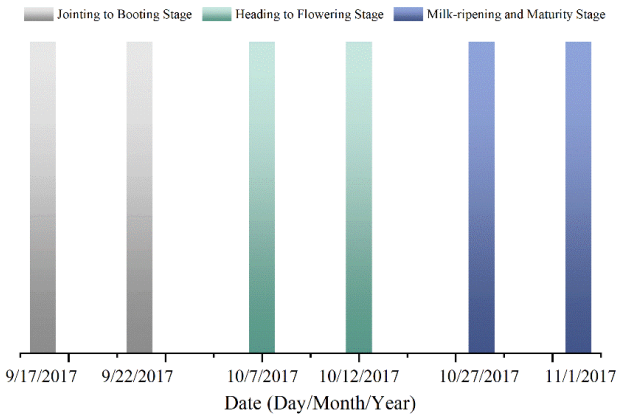

Figure 4. Dates of the Sentinel-2 images and rice growth stages.

Table 2. Band information of the Sentinel-2 images.

Band	Description	CW (nm)	SR (m)	Band	Description	CW (nm)	SR (m)
B1	Coastal aerosol	443	60	B8	NIR-1	842	10
B2	Blue	490	10	B8A	NIR-2	865	20
B3	Green	560	10	B9	Water vapor	945	60
B4	Red	665	10	B10	SMIR-Cirrus	1375	60
B5	Red edge-1	705	20	B11	SMIR-1	1610	20
B6	Red edge-2	740	20	B12	SMIR-2	2190	20
B7	Red edge-3	783	20				

Note: Central wavelength, CW; Spatial resolution, SR.

2.3. Methods

2.3.1. Determination of Optimal Crop Spectral Variables for Estimating SFI

(1) Acquisition of crop spectral variables

The characteristics related to crop growth and the fertility of arable land, which is closely related to vegetation indices [14]. Twenty-seven crop spectral variables were calculated on the Google Earth Engine Platform (https://earthengine.google.com/ (accessed on 12 April 2022)) using the Sentinel-2 images. The details of the crop spectral variables are listed in Table 3.

Table 3. Crop spectral variables for estimating SFI.

Vegetation Index	Formulation in Sentinel-2	References	Vegetation Index	Formulation in Sentinel-2	References
NDVI	$(B8 - B4)/(B8 + B4)$	Haboudane et al. [15]	MCARI	$((B5 - B4) - 0.2 \times (B5 - B3)) \times (B5/B4)$	Daughtry et al. [16]
MTCI	$(B6 - B5)/(B5 - B4)$	Dash et al. [17]	MCARI1	$1.2 \times (2.5 \times (B8 - B4) - 1.3 \times (B8 - B3))$	Haboudane et al. [15]
MGRVI	$((B3)^2 - (B4)^2)/((B3)^2 + (B4)^2)$	Bendig et al. [18]	MCARI2	$1.5 \times (2.5 \times (B8 - B4) - 1.3 \times (B8 - B3)/((2.0 \times B8 + 1)^2) - (6.0 \times B8 - 5 \times ((B4)^{0.5})) - 0.5)^{0.5}$	
REP	$705 + 35 \times ((((B7 + B4)/2) - B5)/(B6 - B5))$	Frampton et al. [19]	MTVI1	$1.2 \times (1.2 \times (B8 - B3) - 2.5 \times (B4 - B3))$	
IRECI	$(B7 - B4)/(B5/B6)$		MTVI2	$1.5 \times (1.2 \times (B8 - B3) - 2.5 \times (B4 - B3)/((2.0 \times B8 + 1)^2) - (6.0 \times B8 - 5 \times ((B4)^{0.5})) - 0.5)^{0.5}$	
RVI	$B8/B4$	Birth et al. [20]	NDREI	$(B8 - B5)/(B8 + B5)$	Gitelson et al. [21]
EVI	$2.5 \times ((B8 - B4)/(B8 + 6 \times B4 - 7.5 \times B2 + 1))$	Huete [22]	NGRDI	$(B3 - B4)/(B3 + B4)$	Tucker [23]
DVI	$B8 - B4$	Jordan [24]	NIRv	$((B8 - B4)/(B8 + B4)) \times B8$	Badgley et al. [25]
SAVI	$(B8 - B4) \times 1.5/(B8 + B4 + 0.5)$	Huete [26]	OSAVI	$(B8 - B4)/(B8 + B4 + 0.16)$	Rondeaux et al. [27]
MASVI	$0.5 \times (2 \times B8 + 1 - ((2 \times B8 + 1)^2 - 8 \times (B8 - B4))^{0.5})$	Qi et al. [28]	SELI	$(B8A - B5)/(B8A + B5)$	Pasqualotto et al. [29]
CIG	$(B8/B3) - 1$	Anatoly et al. [30]	TCARI	$3 \times ((B5 - B4) - 0.2 \times (B5 - B3) \times (B5/B4))$	Haboudane et al. [31]
CIRE	$(B8/B5) - 1$		TCI	$1.2 \times (B5 - B3) - 1.5 \times (B4 - B3) \times (B5/B4)^{0.5}$	
CVI	$(B8 \times B4)/((B3)^2)$	Meng et al. [32]	TGI	$-0.5 \times (190 \times (B4 - B3) - 120 \times (B4 - B2))$	Hunt et al. [33]
TVI	$0.5 \times (120 \times (B8 - B3) - 200 \times (B4 - B3))$	Broge et al. [34]			

(2) Determining the optimal crop spectral variables

Feature selection is a key step in regression analysis to improve prediction accuracy and reduce redundant indicators. Compared to other screening algorithms (e.g., random forest, deep learning), related research [35,36] showed that the extreme gradient boosting (XGBoost) algorithm offers characteristics such as interpretability, computationally efficient, and being less prone to over-fitting under small sample size condition. Because of the relatively small sample size in this study (n = 150), the XGBoost algorithm was employed for selecting optimal crop spectral variables to estimate SFI. The XGBoost is a machine learning algorithm based on a decision-tree ensemble and gradient boosting framework. It gives importance scores for each feature (FI) in each iteration of the training process, so as to indicate the importance of each feature to the training of the model. The FI is directly used as a basis for feature selection [37]. The specific screening steps are as follows [38]:

(1) A classification model is built on the basis of all the features.
(2) Based on the information from the generated model process, the FI is obtained and ranked in descending order. FI is calculated as follows [38]:

$$IG(T, F) = H(T) - H(T|F) = -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{F} p(F) * \sum_{i=1}^{J} p(i|F) \log_2 p(i|F), \quad (2)$$

where $H(T)$ and $H(T|F)$ denote the entropy of parent and child nodes on the basis of the F-feature segmentation, separately; p_i represents the score of the labeled samples at the node.
(3) A subset of features is generated by selecting a number of features with the highest FI values.
(4) Classification experiments were performed on the subset of features to examine their classification ability.
(5) Repeat steps (3) and (4) until all features have been selected.
(6) Check the classification for all subsets and choose the optimal subset of features (namely, the subset having relatively high area under the curve values and fewer features).

2.3.2. Model Construction

Two algorithms (multiple linear regression (MLR) and a backpropagation neural network (BPNN)) were employed for determining the association of optimal crop spectral variables with the SFI, and accuracy assessments were carried out. We present a brief summary of each algorithm.

(1) Multiple linear regression model

MLR refers to a linear regression model that uses multiple explanatory variables to depict the linear correlation between independent and dependent variables. The model can describe the degree of influence of a variable on the soil properties. MLR has been widely used for predicting soil properties [39–41]. In this study, MLR was performed in SPSS software. The definition of MLR is as follows [40]:

$$Y = \sum_{i=1}^{n} \beta_i \check{X}_i + a \quad (3)$$

where $\check{X}_i$ denotes the ith optimal crop spectral variable, β_i represents the regression coefficient of the ith variable, and a represents the intercept.

(2) Backpropagation Neural Network model

The BPNN model is a multilayer feed-forward network that uses fine-tuning of the weights according to the error rate of the former epoch. It comprises an input layer, an output layer, and several hidden layers (Figure 5). The model uses a gradient descent algorithm and backpropagation algorithm to iteratively adjust the weights and biases of the network. The training ends when the predicted value is as close as possible to the

actual value. The learning process comprises forward propagation of the input signal and backward propagation of the error [42–44].

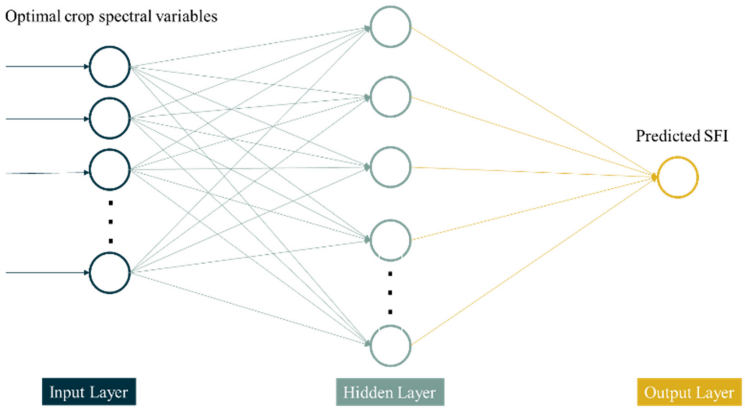

Figure 5. The structure of the backpropagation neural network.

(1) Forward propagation

As for neural networks, forward propagation requires the calculation of both neuron input and output values. The output value (o_j) was written as:

$$O_j = f_i \sum(\omega_{ji} O_i + \theta_j) \quad (4)$$

where o_k is the output layer information (each of the SFI); f_i means the transfer function of the hidden layer to the output layer, where the *Purelin* function is chosen as f_i by the current research [42]; ω_{kj} suggests the weight of the hidden layer to the output layer; θ_j represents the threshold value in the output layer.

When the output value (o_k) of the hidden layer was transferred to the output layer, the o_k was written as:

$$O_k = f_j \sum(\omega_{kj} O_j + \theta_k) \quad (5)$$

where o_i means the input layer information (crop spectral variables); o_j refers to the hidden layer information; ω_{kj} denotes the weight of the input layer to the hidden layer; f_j signifies the transfer function of the input layer to the hidden layer, where the *trainlm* function is selected by the current research [42]; θ_k suggests the threshold value in the hidden layer.

(2) Error back propagation

In cases where the predicted value is different significantly from the measured value, the difference can be transferred to the error in the backpropagation process. The backpropagation process utilizes the Levenberg–Marquardt algorithm for modifying connection weights from the output layer to the input layer to decrease the mean squared error (*MSE*).

$$MSE = \frac{1}{N} \sum (O - O_k)^2 \quad (6)$$

where o and O_k represent the measured and predicted SFI, respectively; N means the number of training samples.

In this study, the number of neuron nodes of the hidden layer (H) is decided using the empirical formula [43]:

$$H = 2n + 1 \quad (7)$$

where n refers to the number of input units.

2.3.3. Accuracy Metrics

The coefficient of determination (R^2), concordance correlation coefficient (CCC), root mean square error (RMSE), and the ratio of performance to interquartile range (RPIQ) were

employed for assessing the performance of the SFI estimation models using the training and validation set. The metrics are expressed as follows [45,46]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (8)$$

$$CCC = \frac{2\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})\left(\hat{y}_i - \hat{y}\right)}{\frac{1}{n}\sum_{I=1}^{n}(y_i - \overline{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \hat{y})^2 + (\overline{y} - \hat{y})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (10)$$

$$RPIQ = \frac{IQ}{RMSE} \quad (11)$$

where y_i is the measured SFI and $\hat{y}_i$ is the estimated SFI of the *i*th sample point, n represents the number of samples, and $\overline{y}$ and $\hat{y}$ denote the average value of observations and estimations, respectively. IQ signifies the interquartile range (IQ = Q3 − Q1) of the observed values. Q1 and Q3 denote the first and third quartile, respectively.

3. Results

3.1. Optimal Crop Spectral Variables for Estimating SFI

The spatial distribution of the crop spectral variables in the study area was obtained using the Google Earth Engine platform and the calculation formula (Table 3). The results are shown in Figure 6.

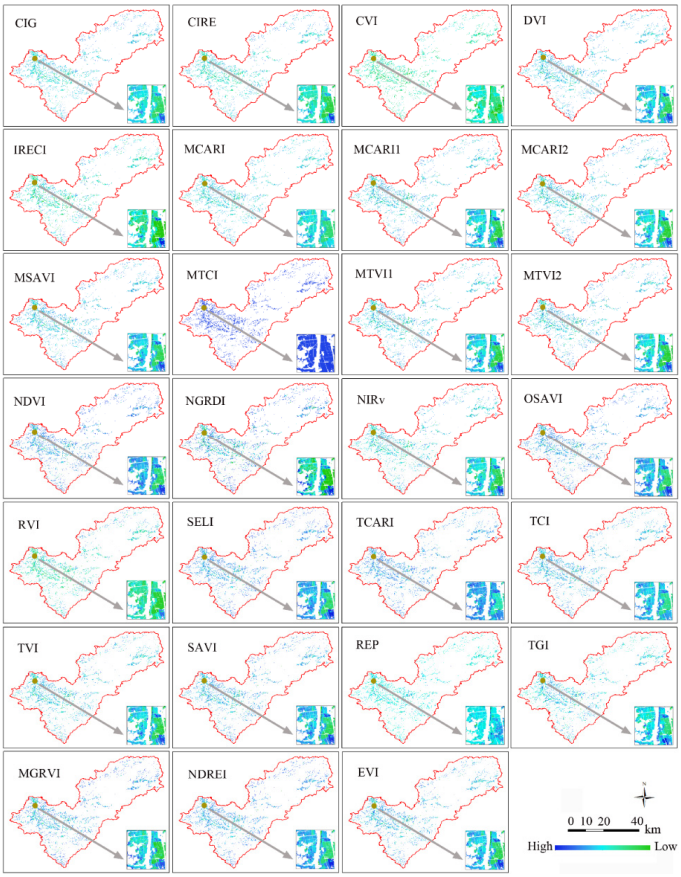

Figure 6. Spatial distribution of the crop spectral variables.

Numerous experiments showed the prediction error in the XGBoost algorithm stabilized with the shrinkage feature weight (eta), the maximum depth (max_depth), and the number of iterations (nround) being 0.4, 10, and 150, separately. The results showed that 6 crop spectral variables (inverted red-edge chlorophyll index (IRECI), chlorophyll vegetation index (CVI), normalized green-red difference index (NGRDI), red-edge position (REP), triangular greenness index (TGI), and optimized soil-adjusted vegetation index (OSAVI)) derived optimal initial results (Figure 7). Subsequently, the variance inflation factor (VIF) was employed for eliminating the collinearity between these characteristic variables with the screening criteria VIF < 10 [47]. Finally, five crop spectral variables (IRECI, CVI, NGRDI, REP, and TGI) were determined as the optimal crop spectral variables for estimating the SFI.

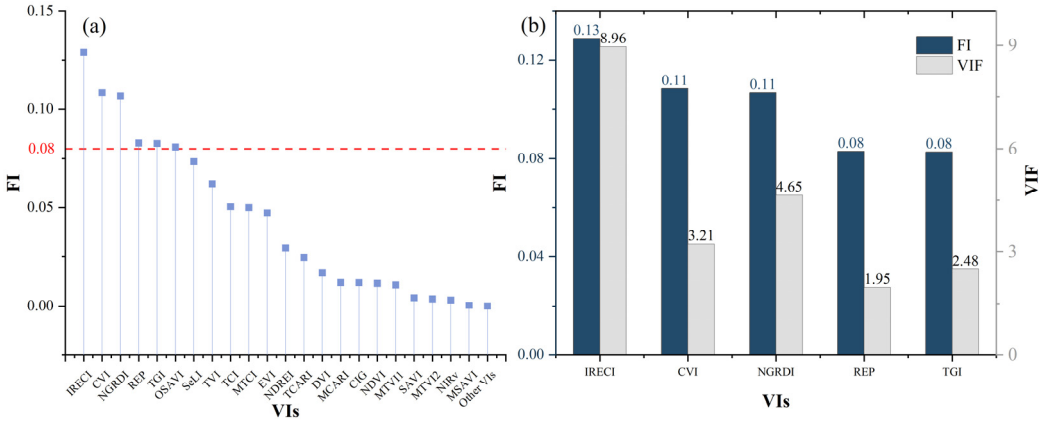

Figure 7. (**a**) The FI of the variables; (**b**) the FI and VIF of the optimal crop spectral variables.

3.2. Model Construction and Accuracy Evaluation

The MLR and BPNN models were adopted for determining the association of optimal crop spectral variables (IRECI, CVI, NGRDI, REP, and TGI) and the SFI using 90 training samples. For the BPNN algorithm, referred to relevant literature [48–50] and through numerous experiments, the number of neuron nodes of the hidden layer was eventually set to 11, the number of iterations was set to 5000, and the learning rate and learning objective were set to 0.01. Figure 8 indicates that the BPNN model provides more accurate estimates than the MLR because the values in the scatter plot approach the 1:1 line. Table 4 presents the accuracy assessment findings of SFI based on 30 validation samples. The RMSE values of the BPNN are smaller than those of the MLR, whereas the R^2 values are larger, indicating that the BPNN model is optimal. Thus, the BPNN model is used to estimate the SFI based on the five spectral variables selected by XGBoost.

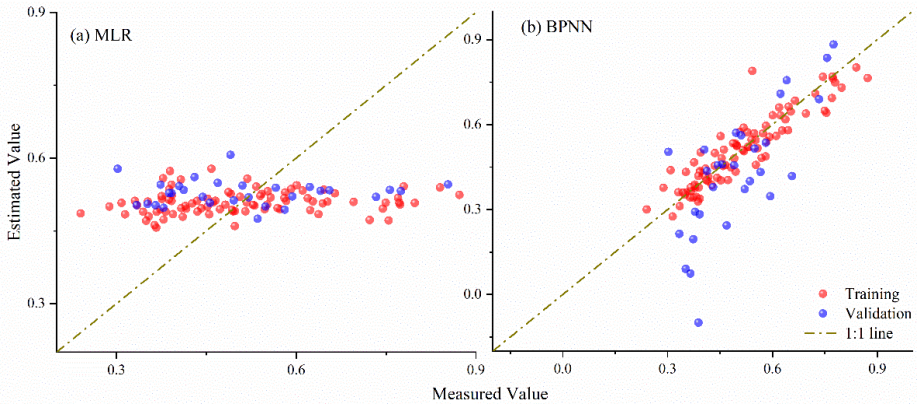

Figure 8. The scatter plots of measured and estimated values: (**a**) MLR and (**b**) BPNN.

3.3. Soil Fertility Index Map

The map of the SFI in the Conghua District obtained from the BPNN model is presented in Figure 9. The SFI value is generally concentrated within 0.20–0.60. The soil fertility is lower in the west but higher in the northeast of the study area.

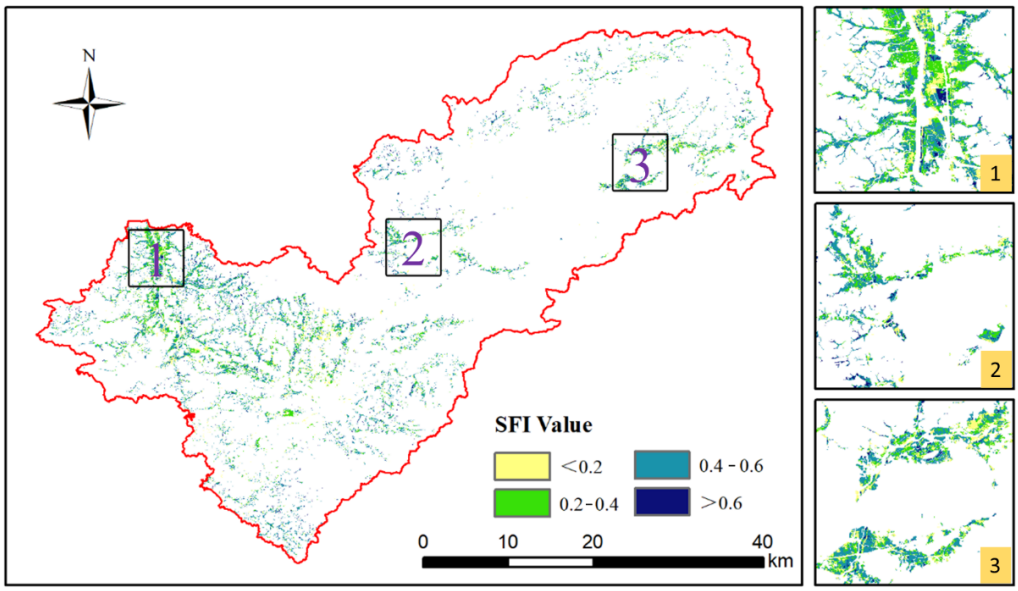

Figure 9. Spatial distribution of the SFI.

Figure 10 shows the measured and estimated SFI for 30 sample plots (red dots in Figure 2). The accuracy metrics (R^2 of 0.62, RMSE of 0.09) show that the proposed model provides potential for mapping the SFI in the Conghua District.

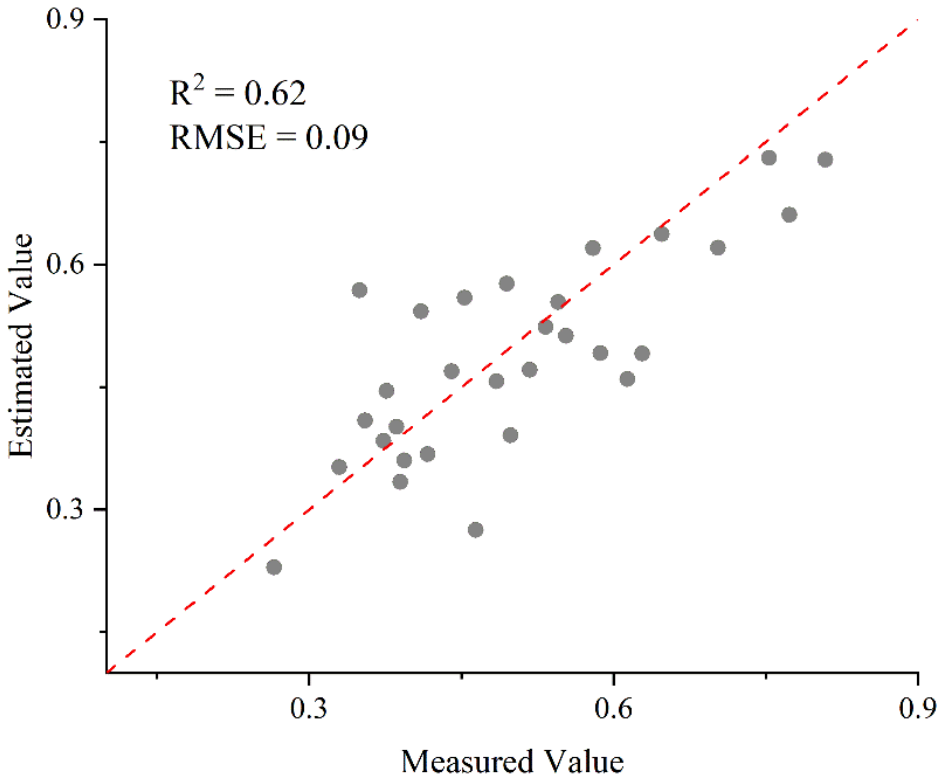

Figure 10. The measured versus estimated SFI values on the basis of the validation set.

Table 4. The accuracy assessment results for the soil fertility index.

Model	Data Set	R^2	RMSE	CCC	RPIQ
MLR	training	0.03	0.26	0.17	0.76
	validation	0.02	0.28	0.02	0.72
BPNN	training	0.84	0.06	0.92	3.60
	validation	0.66	0.17	0.81	1.16

4. Discussion

This study determined the quantitative relationship between crop spectral variables and the SFI. The results indicated that the proposed method has great potential to evaluate soil fertility using crop spectral variables.

4.1. Comparison with Other Similar Studies

Previous studies [6–9] have focused primarily on descriptions of the relationship between soil spectral indicators and soil fertility. Due to the complexity of soil components, the response spectrum of soil fertility may be disturbed, leading to difficulties in identification. In addition, regionally, in southern China there is more vegetation cover and fewer bare soil areas, which also makes it difficult to use soil spectra for soil fertility monitoring. Thus, some scholars used vegetation spectra to assess soil fertility. However, these studies [10–12,51,52] lacked quantitative information on soil fertility and their reliability could not be verified by ground-truth data. In this study, we selected five soil fertility indicators (pH, SOM, TN, AP, and AK) based on previous studies [5,53] and calculated the SFI using a fuzzy approach. The relationship model between crop spectral variables and SFI was established, and quantitative evaluation results of soil fertility were obtained. This quantitative method for SFI estimation provided reasonable accuracy at the sample point level (R^2 = 0.66) and the regional scale (R^2 = 0.62). The BPNN algorithm had higher estimation accuracy than MLR, suggesting the marked nonlinear relationship between crop spectral variables and the SFI.

Current studies that used crop spectral indices to estimate soil fertility [10–12,51,52] focused primarily on indices in the vis-NIR spectral range and did not consider other spectral indices (e.g., red-edge bands). References [54,55] showed that red-edge bands provided abundant information not contained in the red, green, and short-wave infrared bands. The bands can be utilized for identifying and monitoring the chlorophyll content, phenological growth status, health status of vegetation, and heavy metal pollution and can also reflect soil fertility [55]. Red-edge indices (e.g., IRECI, NDREI, and REP) were evaluated in this study to evaluate their ability to estimate the SFI. The XGBoost algorithm selected five crop spectral variables (IRECI, CVI, NGRDI, REP, and TGI) as the optimal variables for estimating the SFI. The results showed that five crop spectral variables (IRECI, CVI, NGRDI, REP, and TGI) could explain 66% of the variance in soil fertility.

4.2. Prospects for Future Studies

This study considered only paddy fields. Future research may consider various arable land types (such as dry land and irrigated land) to improve the generalizability of the model. In addition, we evaluated soil fertility using only spectral vegetation indices and did not consider other crop growth indicators (such as gross primary productivity, net primary productivity, or leaf area index). Additional crop growth indicators should be introduced to future research for improving SFI estimation accuracy.

Notably, the BPNN model was applied to SFI estimation, while the large uncertainties of weights and threshold may potentially affect the accuracy of the model. In future research, more parameter optimization algorithms (e.g., whale optimization, particle swarm optimization) should be introduced to promote efficiency and stability in the BPNN model. Finally, some mixed pixels existed in the images, although the spatial resolution was 10

m. It is unclear as to whether the association of spectral variables with SFI holds true for mixed pixels. Thus, further research is required to support our conclusions.

5. Conclusions

The research here proposes a new approach to the quantitative estimation of soil fertility using crop spectral variables during the crop growth period. The method combines the XGBoost algorithm for variable selection with the BPNN algorithm for soil fertility estimation. The five optimal crop spectral variables (IRECI, CVI, NGRDI, REP, and TGI) were screened, which was the first time soil fertility using red-edge indices was assessed. Based on the five optimal crop spectral variables, BPNN algorithm was used to construct the model to realize the quantitative estimation of soil fertility. The research result showed that the proposed method is reliable with the R^2 of 0.62 and RMSE of 0.09 at the regional scale. To the best of our knowledge, this research is the first to provide an efficient solution to quantify soil fertility based on crop spectral variables.

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

Funding: This research was funded by the National Natural Science Foundation of China (No. U1901601), National Key Research and Development Program of China (No. 2020YFD1100204), Natural Science Foundation of Guangdong Province, China (No. 2021A1515011643), and Guangdong Province Agricultural Science and Technology Innovation and Promotion Project (No. 2022KJ102).

Data Availability Statement: Not applicable.

Acknowledgments: We gratefully acknowledge the paper writing assistance of Mingbang Zhu as well as the experimental assistance of Ziqing Xia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stockdale, E.A.; Shepherd, M.A.; Fortune, S.; Cuttle, S.P. Soil fertility in organic farming systems—Fundamentally different? *Soil Use Manag.* **2006**, *18*, 301–308. [CrossRef]
2. Li, X.G.; Liu, X.P.; Liu, X.J. Long-term fertilization effects on crop yield and desalinized soil properties. *Agron. J.* **2020**, *112*, 4321–4331. [CrossRef]
3. Ye, H.C.; Zhang, S.W.; Huang, Y.F.; Huang, Y.F.; Zhou, Z.M.; Shen, C.Y. Application of Rough Set Theory to Determine Weights of Soil Fertility Factor. *Sci. Agric. Sin.* **2014**, *47*, 710–717.
4. Wang, F.; Li, Q.H.; Lin, C.; He, C.M. Characteristics of soil fertility quality and minimum dataset for yellow-mud paddy fields in Fujian Province. *Chin. J. Eco-Agric.* **2018**, *26*, 1855–1865.
5. Huang, J.; Han, T.F.; Shen, Z.; Liu, K.L.; Ma, C.B.; Wang, H.Y.; Qu, X.L.; Yu, Z.K.; Xie, J.H.; Zhang, H.M. Spatiotemporal Variation of Fertility Quality of Chinese Paddy Soil Based on Fuzzy Method in Recent 30 Years. *Acta Pedol. Sin.* **2022**. [CrossRef]
6. Rossel, R.; Rizzo, R.; Demattê, J.A.M.; Behrens, T. Spatial modeling of a soil fertility index using Visible–Near-Infrared spectra and terrain attributes. *Soil Sci. Soc. Am. J.* **2010**, *74*, 1293–1300. [CrossRef]
7. Munnaf, M.A.; Mouazen, A.M. Development of a soil fertility index using on-line Vis-NIR spectroscopy. *Comput. Electron. Agric.* **2021**, *188*, 106341. [CrossRef]
8. Yang, M.H.; Mouazen, A.; Zhao, X.M.; Guo, X. Assessment of a soil fertility index using visible and near-infrared spectroscopy in the rice paddy region of southern China. *Eur. J. Soil Sci.* **2020**, *71*, 615–626. [CrossRef]
9. Wang, L.Z.; Han, Y.; Pan, J. Study on Farmland Soil Fertility Model Based on Multi-Angle Polarized Hyper-Spectrum. *Spectrosc. Spectr. Anal.* **2018**, *38*, 240–245.
10. Zeeshan, M.; Siddique, M.T.; Ali, N.A.; Farooq, M.S. Correlation of Spatial Variability of Soil Macronutrients with Crop Performance by Using Satellite and Remote Sensing Indices for Site Specific Agriculture: Chakwal Region. *Rice Res.* **2017**, *5*, 1000182. [CrossRef]
11. Wang, Q.; Wang, K.R.; Li, S.K.; Xiao, C.H.; Li, J.; Dai, J.G.; Fang, L.F.; Chen, B.; Wang, F.Y. Study on Evaluation Methods for Soil Fertility in Oasis Cotton Field Based on the Nor-malized Difference Vegetation Index (NDVI). *Cotton Sci.* **2013**, *25*, 148–153.

12. Duan, D.D.; Sun, X.; Liang, S.F.; Sun, J.; Fan, L.L.; Chen, H.; Xia, L.; Zhao, F.; Yang, W.Q.; Yang, P. Spatiotemporal Patterns of Cultivated Land Quality Integrated with Multi-Source Remote Sensing: A Case Study of Guangzhou, China. *Remote Sens.* **2022**, *14*, 1250. [CrossRef]
13. Lu, R.K. *Methods of Soil Agrochemical Analysis*; China Agricultural Science and Technology Press: Beijing, China, 2000.
14. Guan, Y.J.; Zou, Z.L.; Zhang, X.P.; Min, C.W. Research on the inversion model of cultivated land quality based on normalized difference vegetation index. *Chin. J. Soil Sci.* **2018**, *49*, 779–787.
15. Haboudane, D.; Miller, J.R.; Pattey, E.; Zarco-Tejada, P.J.; Strachan, L.B. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sens. Environ.* **2004**, *90*, 337–352. [CrossRef]
16. Daughtry, C.; Walthall, C.L.; Kim, M.S.; Colstoun, E.B.D.; Mcmurtreyll, J.E. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sens. Environ.* **2000**, *74*, 229–239. [CrossRef]
17. Dash, J.; Curran, P.J. MTCI: The meris terrestrial chlorophyll index. *Int. J. Remote Sens.* **2004**, *25*, 151–161. [CrossRef]
18. Bendig, J.L.; Yu, K.; Aasen, H.; Bolten, A.; Bennertz, S.; Broscheit, J.; Gnyp, M.L.; Bareth, G. Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *39*, 79–87. [CrossRef]
19. Frampton, W.J.; Dash, J.; Watmough, G.; Milton, E.J. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS J. Photogramm. Remote Sens.* **2013**, *82*, 83–92. [CrossRef]
20. Birth, G.S.; Mcvey, G.R. Measuring the color of growing turf with a reflectance spectrophotometer. *Agron. J.* **1968**, *60*, 640–643. [CrossRef]
21. Gitelson, A.; Merzlyak, M.N. Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *J. Photochem. Photobiol. B Biol.* **1994**, *22*, 247–252. [CrossRef]
22. Huete, A.R.; Liu, Q.H.; Batchily, K.; Leeuwen, W.V. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sens. Environ.* **1997**, *59*, 440–451. [CrossRef]
23. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
24. Jordan, C.F. Derivation of leaf-area index from quality of light on the forest floor. *Ecology* **1969**, *50*, 663–666. [CrossRef]
25. Badgley, G.; Field, C.B.; Berry, J.A. Canopy near-infrared reflectance and terrestrial photosynthesis. *Sci. Adv.* **2017**, *3*, e1602244. [CrossRef] [PubMed]
26. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
27. Rondeaux, G.; Steven, M.; Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sens. Environ.* **1996**, *55*, 95–107. [CrossRef]
28. Qi, J.G.; Chehbouni, A.R.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]
29. Pasqualotto, N.; Delegido, J.; Wittenberghe, S.V.; Rinaldi, M.; Moreno, J. Multi-crop green LAI estimation with a new simple Sentinel-2 LAI index (SeLI). *Sensors* **2019**, *19*, 904. [CrossRef]
30. Anatoly, A.; Gritz, Y.; Merzlyak, M.N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* **2003**, *160*, 271–282.
31. Haboudane, D.; Tremblay, N.; Miller, J.R.; Vigneault, P. Remote estimation of crop chlorophyll content using spectral indices derived from Hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 423–437. [CrossRef]
32. Meng, J.H.; Xu, J.; You, X.Z. Optimizing soybean harvest date using HJ-1 satellite imagery. *Precis. Agric.* **2015**, *16*, 164–179. [CrossRef]
33. Hunt, E.R.; Doraiswamy, P.C.; Mcmurtrey, J.E.; Daughtry, C.S.T.; Perry, E.M.; Akhmedov, B. A visible band index for remote sensing leaf chlorophyll content at the canopy scale. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *21*, 103–112. [CrossRef]
34. Broge, N.H.; Leblanc, E. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote Sens. Environ.* **2001**, *76*, 156–172. [CrossRef]
35. Cui, H.Y.; Xu, S.; Zhang, L.F.; Welsch, R.E.; Horn, B.K.P. The key techniques and future vision of feature selection in machine learning. *J. Beijing Univ. Posts Telecommun.* **2018**, *41*, 1–12.
36. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
37. Ma, J.; Ding, Y.X.; Cheng, J.C.P.; Jiang, F.F.; Tan, Y.; Gan, V.J.L.; Wan, Z.W. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *J. Clean. Prod.* **2020**, *244*, 118955. [CrossRef]
38. Zhao, L.; Zhou, W.; Peng, Y.P.; Hu, Y.M.; Ma, T.; Xie, Y.K.; Wang, L.Y.; Liu, J.C.; Liu, Z.H. A new AG-AGB estimation model based on MODIS and SRTM data in Qinghai Province, China. *Ecol. Indic.* **2021**, *133*, 108378. [CrossRef]
39. Chagas, C.D.S.; Junior, W.D.C.; Bhering, S.B.; Filho, B.C. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena* **2016**, *139*, 232–240. [CrossRef]
40. Selige, T.; Böhner, J.; Schmidhalter, U. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. *Geoderma* **2006**, *136*, 235–244. [CrossRef]
41. Tavares, T.R.; Mouazen, A.M.; Nunes, L.C.; Santos, F.R.D.; Melquiades, F.L.; Silva, T.R.D.; Krug, F.J.; Molin, J.P. Laser-Induced Breakdown Spectroscopy (LIBS) for tropical soil fertility analysis. *Soil Tillage Res.* **2022**, *216*, 105250. [CrossRef]

42. Peng, Y.P.; Zhao, L.; Hu, Y.M.; Wang, G.X.; Wang, L.; Liu, Z.H. Prediction of Soil Nutrient Contents Using Visible and Near-Infrared Reflectance Spectroscopy. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 437. [CrossRef]
43. Nielsen, R.H. Kolmogorov's mapping neural network existence theorem. In Proceedings of the IEEE 1st International Conference on Neural Networks, San Diego, CA, USA, 21–24 June 1987.
44. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
45. Tziolas, N.; Tsakiridis, N.; Ogen, Y.; Kalopesa, E.; Ben-Dor, E.; Theocharis, J.; Zalidis, G. An integrated methodology using open soil spectral libraries and Earth Observation data for soil organic carbon estimations in support of soil-related SDGs. *Remote Sens. Environ.* **2020**, *244*, 111793. [CrossRef]
46. Chen, S.C.; Xu, H.Y.; Xu, D.Y.; Ji, W.J.; Li, S.; Yang, M.H.; Hu, B.F.; Zhou, Y.; Wang, N.; Arrouays, D.; et al. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. *Geoderma* **2021**, *400*, 115159. [CrossRef]
47. Allouis, T.; Durrieu, S.; Véga, V.; Couteron, P. Stem Volume and Above-Ground Biomass Estimation of Individual Pine Trees from LiDAR Data: Contribution of Full-Waveform Signals. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 924–934. [CrossRef]
48. Dlamini, D.S.; Mishra, A.K.; Mamba, B.B. ANN modeling in Pb(II) removal from water by clay-polymer composites fabricated via the melt-blending. *J. Appl. Polym. Sci.* **2013**, *130*, 3894–3901. [CrossRef]
49. Mouazen, A.M.; Kuang, B.; Baerdemaeker, J.D.; Ramon, H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma* **2010**, *158*, 23–31. [CrossRef]
50. Sheela, K.G.; Deepa, S.N. Review on methods to fix number of hidden neurons in neural networks. *Math. Probl. Eng.* **2013**, *2013*, 425740. [CrossRef]
51. Fang, L.N.; Song, J.P. Cultivated Land Quality Assessment Based on SPOT Multispectral Remote Sensing Image: A Case Study in Jimo City of Shandong Province. *Prog. Geogr.* **2008**, *27*, 71–78.
52. Liu, S.S.; Peng, Y.P.; Xia, Z.Q.; Hu, Y.M.; Wang, G.X.; Zhu, A.X.; Liu, Z.H. The GA-BPNN-Based Evaluation of Cultivated Land Quality in the PSR Framework Using Gaofen-1 Satellite Data. *Sensors* **2019**, *19*, 5127. [CrossRef]
53. Zhou, W.Z.; Dong, B.; Liu, J.J.; Li, Q. Method of comprehensive evaluation on soil fertility on the basis of weight analysis. *J. Irrig. Drain.* **2016**, *35*, 81–86.
54. Liang, J.; Zheng, Z.W.; Xia, S.T.; Zhang, X.T.; Tang, Y.Y. Crop recognition and evaluation using red edge features of GF-6 satellite. *J. Remote Sens.* **2020**, *24*, 1168–1179.
55. Weksler, S.; Rozenstein, O.; Haish, N.; Moshelion, M.; Wallach, R.; Ben-Dor, E. Pepper Plants Leaf Spectral Reflectance Changes as a Result of Root Rot Damage. *Remote Sens.* **2021**, *13*, 980. [CrossRef]