sustainability

MDPI

Article

Modeling the Spatial Distribution of Soil Nitrogen Content at Smallholder Maize Farms Using Machine Learning Regression and Sentinel-2 Data

Zinhle Mashaba-Munghemezulu [1,2,*], George Johannes Chirima [1,2] and Cilence Munghemezulu [2]

[1] Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria 0028, South Africa; ChirimaJ@arc.agric.za
[2] Geoinformation Science Division, Agricultural Research Council, Natural Resources and Engineering, Pretoria 0001, South Africa; MunghemezuluC@arc.agric.za
* Correspondence: MashabaZ@arc.agric.za

Abstract: Nitrogen is one of the key nutrients that indicate soil quality and an important component for plant development. Accurate knowledge and management of soil nitrogen is crucial for food security in rural communities, especially for smallholder maize farms. However, less research has been done on generating digital soil nitrogen maps for these farmers. This study examines the utility of Sentinel-2 satellite data and environmental variables to map soil nitrogen at smallholder maize farms. Three machine learning algorithms—random forest (RF), gradient boosting (GB), and extreme gradient boosting (XG) were investigated for this purpose. The findings indicate that the RF (R^2 = 0.90, RMSE = 0.0076%) model performs slightly better than the GB (R^2 = 0.88, RMSE = 0.0083%) and XG (R^2 = 0.89, RMSE = 0.0077%) models. Furthermore, the variable importance measure showed that the Sentinel-2 bands, particularly the red and red-edge bands, have a superior performance in comparison to the environmental variables and soil indices. The digital maps generated in this study show the high capability of Sentinel-2 satellite data to generate accurate nitrogen content maps with the application of machine learning. The developed framework can be implemented to map the spatial pattern of soil nitrogen. This will also contribute to soil fertility interventions and nitrogen fertilization management to improve food security in rural communities. This application contributes to Sustainable Development Goal number 2.

Keywords: satellite data; random forest; gradient boosting; extreme gradient boosting; soil fertility; digital mapping

Citation: Mashaba-Munghemezulu, Z.; Chirima, G.J.; Munghemezulu, C. Modeling the Spatial Distribution of Soil Nitrogen Content at Smallholder Maize Farms Using Machine Learning Regression and Sentinel-2 Data. *Sustainability* **2021**, *13*, 11591. https://doi.org/10.3390/su132111591

Academic Editor: Emanuele Radicetti

Received: 19 August 2021
Accepted: 12 October 2021
Published: 20 October 2021

1. Introduction

Improving soil nutrient management at smallholder maize (*Zea mays* L.) farms is imperative for ensuring food security in developing countries. Smallholder maize farms are crucial for the livelihoods of rural communities in Africa who depend on agriculture for food security and their local economic activities. Amongst the most important nutrients is nitrogen; not only is it a component of the chlorophyll molecule but is also essential for maize growth, quality, and yield [1–3]. The soil is one of the most important nitrogen reservoirs in terrestrial ecosystems [4]. Developing frameworks to map the spatial variability of soil nitrogen is necessary for the local government, farmers, and stakeholders to identify nitrogen excesses or deficiencies. Such information will guide soil fertility interventions at smallholder farms. In the long term, improved soil nitrogen content management will enhance maize productivity [5,6]. This application is particularly important for resource limited smallholder maize farms such as those in developing countries, for example South Africa, which have reported sub-optimal yields, infertile land, and land degradation in previous studies [7,8].

Several soil databases and sources are available that archive soil nutrient information for South Africa. Examples of these include the Africa Soil Information Service (AfSIS),

which archives soil nutrient maps at a 250 m spatial resolution for Africa (http://africasoils.net/, accessed on 9 August 2021). The Harmonized World Soil Database (HWSD) nutrient map, which has a spatial resolution of 1 km [9], is another example. Other products such as the SOTER-based soil parameter estimates (SOTWIS) product for Southern Africa have a 1:2 M (million) scale resolution [10]. The soil Atlas of Africa dataset for soil groups has a 1:3 M scale resolution [11]. Although these products are available, they have a coarse spatial resolution to guide soil nutrient management efforts at smallholder farms, which are typically 0.5–2 ha in size. These types of farms are often fragmented and heterogeneous in most parts of the world including South Africa, which necessitates the use of improved resolution data for digital soil mapping [12].

The Sentinel-2 mission has sensor capabilities with a potential to estimate soil nutrients at smallholder farms. This satellite has an improved spatial resolution of 10–60 m, a wide swath of 290 m, and a frequent revisit cycle of 5–10 days [13]. Additionally, the Sentinel-2 data are compatible with Landsat-8 and Satellite Pour l'Observation de la Terre (SPOT) data [14]. The difference between Sentinel-2 and other medium resolution sensors such as Landsat-8 is the presence of the red-edge band region in Sentinel-2. The red-edge region lies between the red and near infrared portions of the electromagnetic spectrum and is distinguished by a sharp increase in vegetation reflectance [15]. This current study relies on soil and vegetation indices derived from strategic locations of the electromagnetic spectrum to estimate the soil nitrogen content for smallholder maize farms.

Different techniques have been applied for digital soil mapping. The commonly used models are multiple linear regression [16], principal component analysis regression [17], generalized additive model [18], and kriging [19]. Recently, machine learning algorithms (support vector machines, decision trees, random forests, artificial neural networks) have been widely used in remote sensing studies [20–23]. These algorithms are beneficial because they can learn from limited data and reduce errors through an adaptive learning process [24,25]. However, studies using these techniques for soil nitrogen mapping at smallholder maize farms are lacking [19]. Machine learning algorithms are not universally applicable in different environments. This necessitates the evaluation of different machine learning algorithms for applicability in our own context to understand the distribution of soil nitrogen content at the locality.

This paper uses the random forest (RF) algorithm, gradient boosting algorithm (GB), and extreme gradient boosting (XG) machine learning algorithm in a regression format. These algorithms were used because they can deal with noisy, high-dimensional, and non-linear data [26,27]. The algorithms are applied to Sentinel-2 imagery to predict the spatial patterns of soil nitrogen content at selected smallholder maize farms in Makhuduthamaga district, South Africa. The study addresses the following specific research questions: (1) What is the relationship between soil nitrogen content and different predictor variables? (2) How effective are the selected machine learning algorithms in predicting soil nitrogen content? (3) Which predictor variables are fundamental for modelling soil nitrogen content? Finally, (4) What is the spatial distribution pattern of soil nitrogen at smallholder maize farms?

2. Materials and Methods

The overview of the methodological approach used in this study is summarized in Figure 1. The Sentinel-2 imageries were pre-processed to correct for atmospheric effects, and band indices were calculated. Ancillary data describing the environmental variables and some of the Sentinel-2 bands were resampled to 10 m. Nine experiments with different data configurations were conducted using the Sentinel-2 bands, spectral indices, and environmental variables. Three machine learning regression algorithms—RF, GB, and XG—were then applied in each experiment using 70% of the nitrogen content measurements for training the model. The remaining 30% of the data was used for model evaluation with commonly used statistical metrics. Variable importance for the predictors

was determined from the scores derived by the three machine learning regression models. Finally, the spatial pattern of soil nitrogen at the smallholder maize farms was mapped.

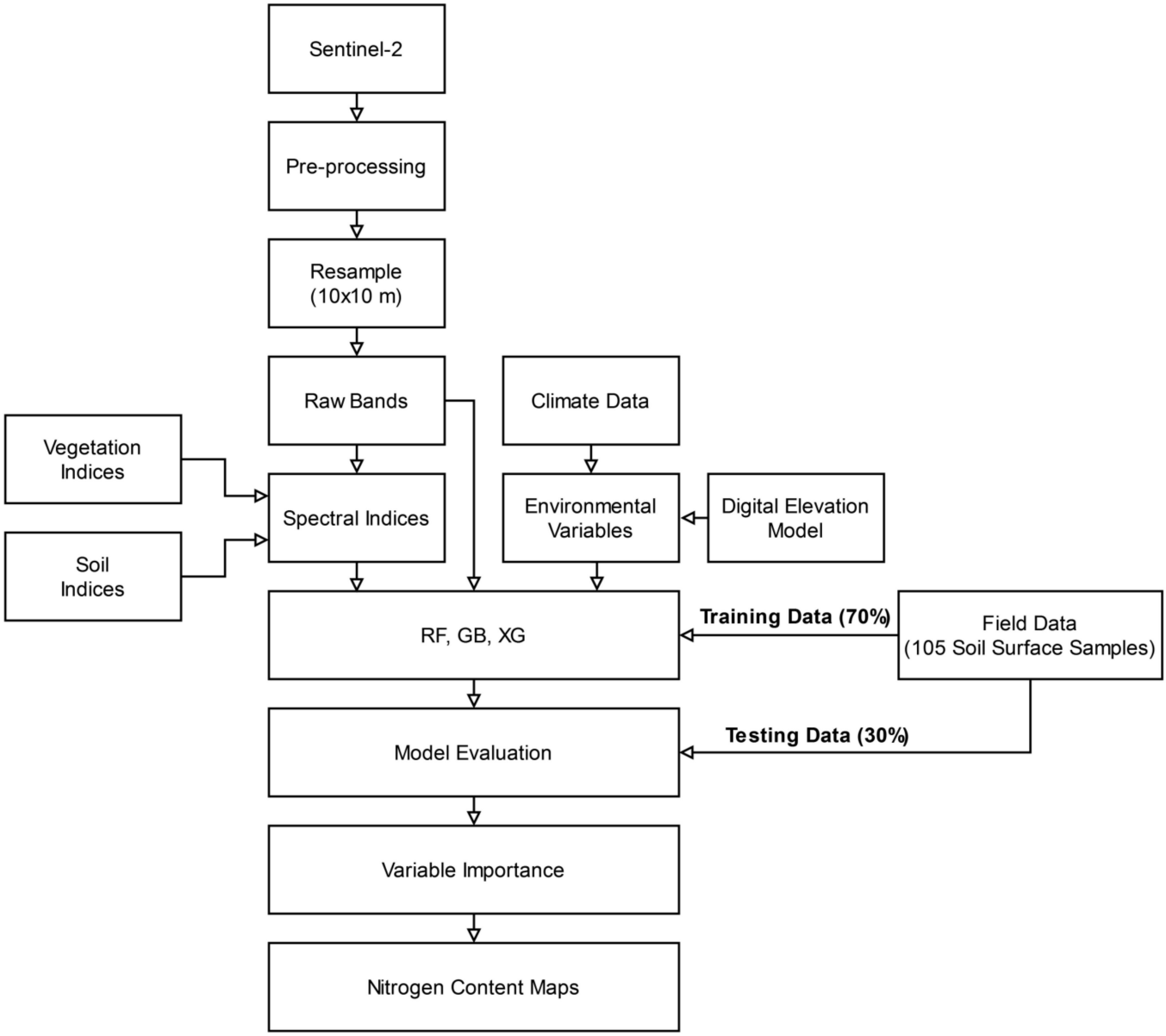

Figure 1. The proposed methodological framework for mapping soil nitrogen content at smallholder maize farms.

2.1. Study Area

Soil nitrogen samples were collected from the smallholder maize farms of Makhudu-thamaga district located in the northern part of South Africa (Figure 2). This district has a low elevation (799–1047 m) in the northwestern part and a higher elevation (1295–1791 m) in the central and southern parts. The topography is undulating with rock habitats such as rock outcrops, rocky ridges, and rocky refugia [28]. This district was selected because most of the rural population are smallholder maize farmers; they farm mainly for subsistence and partially for selling in local markets. Smallholder maize production is predominant in the southern part of the district [29]. The farmers add manure to their fields in November. Maize is planted during December and January. The growing period is between February and May. Harvesting takes place in June and no maize is present in the smallholder farms during July–November. The smallholder farms in the district are rain-fed. The annual rainfall is 536 mm with an average annual temperature of 7 °C in winter and 35 °C in

summer according to the Agricultural Research Council stations located in Nchabeleng, Ga-Rantho, and Leeuwkraal areas.

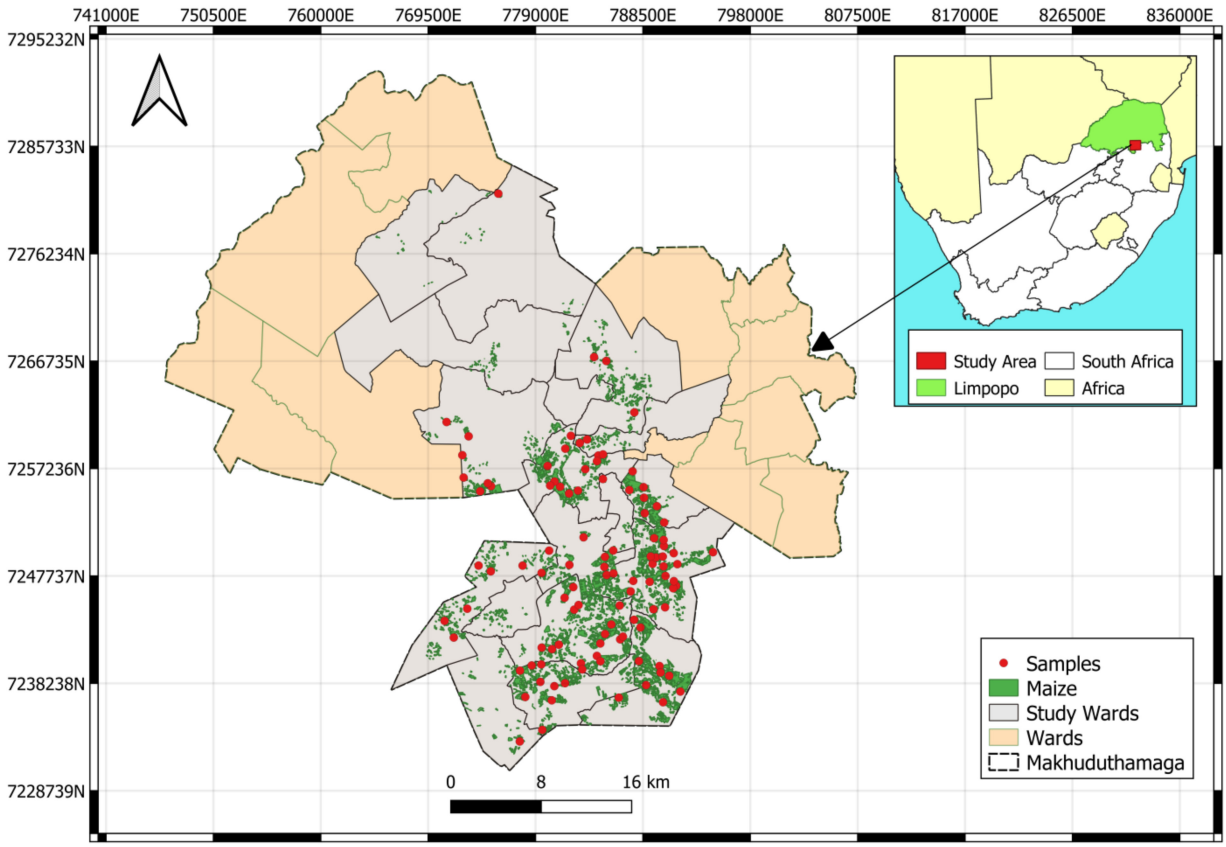

Figure 2. The location of the study wards and smallholder maize farms that are considered for soil nitrogen data collection in Makhuduthamaga district, South Africa.

2.2. Field Data Collection and Laboratory Analysis

A total of 105 soil surface samples were collected from the topsoil layer (0–20 cm) at the smallholder maize sample farms during 14–17 May 2019 corresponding to a period of low rainfall. The positions for each sample were captured with a handheld Global Positioning System (GPS). The samples were then processed at the Agricultural Research Council Analytical Laboratory where they were air-dried at room temperature (25 °C), crushed, and passed through a 2 mm sieve to remove coarse soil materials such as gravel or plant roots. The soil total nitrogen content was then determined through analytical processing with the Kjeldahl digestion method. The soil properties are summarized in Table 1 according to the dominant soil type at the top (haplic acrisols) and least dominant soil at the bottom (lithic leptosols). These were extracted from the Harmonized World Soil Database [9].

Table 1. Soil attributes for the dominant soil types in smallholder farms.

Soil Type	Topsoil Sand Fraction (%)	Topsoil Silt Fraction (%)	Topsoil Clay Fraction (%)	Topsoil Texture	pH (H_2O)	Bulk Density (kg/dm^3)	Organic Carbon (% Weight)
Haplic Acrisols	57	19	24	Sand clay loam	5.1	1.4	0.8
Ferric Luvisols	65	18	17	Sandy loam	6.4	1.5	0.6
Lithic Leptosols	43	29	28	Clay loam	7.5	1.3	0.4

2.3. Sentinel-2 Data Acquisition and Pre-Processing

We used Sentinel-2 MSI level-1C (L1C) data acquired from the Copernicus Open Access Hub. The image for 17 May 2019 was used in this study. This image covered the field sampling date and was appropriate considering that the image was cloud free. The L1C product images consist of top-of-atmosphere (TOA) reflectance after radiometric correction and geometric corrections (ortho-rectification and spatial registration) with a sub-pixel accuracy (https://sentinel.esa.int, accessed on 10 August 2021). Sentinel-2 MSI has 13 bands, which have different spatial resolutions. This study made used of 10 bands (visible, near-infrared, red-edge, and shortwave infrared) as summarized in Table 2 and excluded the bands that are related to water and atmosphere elements. The Sentinel-2 TOA images were pre-processed with Sen2Cor plugin in Sentinel Application Platform (SNAP) to convert them to bottom-of-atmosphere reflectance (BOA), and the 20 m bands were resampled to a 10 m spatial resolution.

Table 2. Sentinel-2 multi-spectral bands used in this study (https://www.usgs.gov, accessed on 10 August 2021).

Variable	Description		
Raw Bands	**Central Wavelength (nm)**	**Bandwidth (nm)**	**Spatial Resolution (m)**
B2–Blue	490	65	10
B3–Green	560	35	10
B4–Red	665	30	10
B5–RE1	705	15	20
B6–RE2	740	15	20
B7–RE3	783	20	20
B8–NIR	842	115	10
B8a–RE4	865	20	20
B11–SWIR1	1610	90	20
B12–SWIR2	2190	180	20

Note: Red Edge (RE), Near Infrared (NIR), Short Wave Infrared (SWIR).

2.4. Spectral Indices

Spectral indices were generated from the Sentinel-2 bands. The vegetation indices that are included in the current study were selected by fitting the RF, XG, and GB machine learning regression models. Vegetation indices that optimized the coefficient of determination (R^2) in relation to the nitrogen content for each model were retained. This procedure was done because similar studies have reported a diverse range of vegetation indices [19,30,31]. The vegetation indices evaluated based on the RE were the following: Normalized Difference Vegetation Index RE 1, 2, and 3 narrow (NDVIRE1n, NDVIRE2n, NDVIRE3n), Normalized Difference Vegetation Index RE 1 (NDRE1), Normalized Difference Vegetation Index RE 1 modified (NDRE1m), Modified Simple Ratio RE (MSRRE), Chlorophyll Index RE (CLRE), and Normalized Difference Vegetation Index RE (NDVIRE). Other indices based on the NIR, SWIR1, SWIR2, and visible parts of the electromagnetic spectrum were also evaluated. These indices included the Plant Senescence Reflectance Index (PSRI), Enhanced Vegetation Index (EVI), and the Green Normalized Difference Vegetation Index (GNDVI). Additionally, the Difference Vegetation Index (DVI), Normalized Difference Water Index (NDWI), Renormalized Difference Vegetation Index (RDVI), Normalized Difference Vegetation Index (NDVI), Optimized Soil Adjusted Vegetation Index (OSAVI), Soil Adjusted Vegetation Index (SAVI), and Triangular Vegetation Index (TVI) were also evaluated. The final spectral indices used in this study are summarized in Table 3.

Table 3. The collection of spectral indices considered in this study.

Vegetation Indices	Equation	Source	Property
PSRI	$\frac{(Red-Green)}{RE2}$	[32]	Senescence-induced reflectance changes
NDVIRE1n	$\frac{(RE4-RE1)}{(RE4+RE1)}$	[33]	Sparse biomass
NDVIRE2n	$\frac{(RE4-RE2)}{(RE4+RE2)}$	[33]	Sparse biomass
NDVIRE3n	$\frac{(RE4-RE3)}{(RE4+RE3)}$	[33]	Sparse biomass
MSRRE	$\frac{(NIR/RE1)-1}{\sqrt{(NIR/RE1)+1}}$	[34]	Correction for leaf specular reflection
EVI	$2.5 \times \frac{(NIR-Red)}{(NIR+6\times Red-7.5\times Blue)+1}$	[35]	Chlorophyll sensitive
GNDVI	$\frac{(NIR-Green)}{(NIR+Green)}$	[36]	Chlorophyll sensitive
Soil Indices	**Equation**	**Source**	**Property**
BI	$\left(\frac{(Red^2+Green^2+Blue^2)}{3}\right)^{0.5}$	[31,37]	Average reflectance magnitude
CI	$\frac{(Red-Green)}{(Red+Green)}$	[31,37]	Soil Colour
HI	$\frac{(2\times Red-Green-Blue)}{(Green-Blue)}$	[31,37]	Primary Colours
RI	$\frac{Red^2}{(Blue\times Green^3)}$	[38]	Hematite content
SI	$\frac{(Red-Blue)}{(Red+Blue)}$	[31,37]	Spectral slope

Note: Brightness Index (BI), Coloration Index (CI), Hue Index (HI), Redness Index (RI), Saturation Index (SI).

2.5. Environmental Variables

Different datasets in Table 4 were used to describe the environmental variables needed to estimate nitrogen content. These included the slope, elevation, aspect, catchment area, topographic wetness index (TWI), precipitation, and temperature. The ASTER digital elevation model (DEM) with a 30 m spatial resolution was used to extract the terrain variables. This product was used because it is freely available and was closer to the 10 m spatial resolution of Sentinel-2 data. The ASTER DEM tiles were mosaicked and resampled to a 10 m resolution using a bilinear interpolation in the R software. The DEM, slope, aspect, catchment area and TWI were subsequently derived. The JAXA Earth Observation Research Center precipitation and Landsat land surface temperature (LST) covering 7 years from 2013 to 2019 were used. This period was selected based on the continuity of the Landsat LST collection. These images were also resampled to a 10 m resolution. The environmental variables have shown to be valuable in previous studies for modeling nitrogen content [3,30].

Table 4. The list of selected environmental variables used in this study.

Environmental Variables	Units	Source	Property
Slope (SLP)	Degrees	[39]	Rise or fall of the land surface
Elevation (EL)	Meters	[39]	Distance above sea level
Aspect (ASP)	Degrees	[39]	Direction of terrain
Catchment area (CA)	Square Meters	[39]	Flow accumulation
TWI	-	[40]	Soil moisture
Precipitation (RAIN)	Millimeter/hour	[41]	Rainfall
LST	Kelvin	[42]	Temperature

2.6. Machine Learning Regression Models

2.6.1. Random Forest Regression

Random Forest is a bagging ensemble learning method [43]. This algorithm can be applied to both classification and regression problems. The principle of RF regression is to predict a continuous response variable using a bootstrapping method based on the classification and regression trees. Decision tree models are fitted to the data. Every tree is trained using different bootstrap samples from the training data, referred to as

in-bag samples. The final model is generated by averaging the individual tree outputs [43]. Samples that are not used in the bootstrap are referred to as the out-of-bag samples; these can be used for model evaluation and variable importance [44]. The RF algorithm is applied in this study because of its superior performance capabilities. RF can handle high dimensional data, requires relatively few tuning parameters, and processes non-linear data without overestimation [45]. The tuning parameters necessary to train the RF model (number of trees and features) were determined using Gridsearch method in Python; further details can be obtained in Lerman [46]. Variable importance for the RF algorithm was determined using the built-in Python variable importance measure for RF; readers are referred to Dangeti [47] for further details on this procedure.

2.6.2. Gradient Boosting Regression

Gradient boosting is an ensemble-based decision tree machine learning method developed by Friedman [48]. This method can be adapted for both regression and classification problems. The purpose of gradient boosting is to improve the performance of weak learners to achieve over random guessing [49]. At each iteration, a new regression tree is trained to improve the loss function determined by the steepest gradient. This procedure reduces the model residuals along the gradient direction. The results of the individual regression trees are combined to give the final result [48]. The gradient boosting algorithm is applied in the present study because it can handle unbalanced data and it is robust to outliers [50]. The parameters needed for gradient boosting are the number of trees, number of features for the best split, maximum depth, learning rate, and the minimum number of samples required at a leaf node. These were optimized using the Gridsearch method. Variable importance for the GB algorithm was determined using the built-in Python variable importance measure for GB; readers are referred to Dangeti [47] for further details on this procedure.

2.6.3. Extreme Gradient Boosting Regression

The Extreme Gradient Boosting algorithm is part of the classification and regression ensemble gradient boosting machine algorithms. This model can be applied for both classification and regression problems [51]. The XG uses additive training strategies: the first learning phase is fitted to the entire input dataset and the second phase is fitted to the residuals. This procedure enhances the performance of weak supervised learning. The fitting process is done repeatedly until the stopping criteria are achieved [51]. The XG algorithm was applied because it overcomes problems with overfitting and has an optimized performance [52]. This algorithm requires a rigorous number of regularization parameters; these were determined using Gridsearch. Variable importance for the XG algorithm was determined using the built-in Python variable importance measure for XG; readers are referred to Dangeti [47] for further details on this procedure.

2.6.4. Experiments

We investigated the effect of different feature variables for modeling nitrogen content in smallholder maize farms. The data were split into 70% training and 30% testing. Three models RF, GB, and XG with different combinations of variables summarized in Table 5 were implemented. The experiments consisted of: (1) raw bands, (2) raw bands + vegetation indices, (3) raw bands + soil indices, (4) raw bands + environmental variables, (5) raw bands + vegetation indices + soil indices + environmental variables, (6) raw bands + vegetation indices + soil indices, (7) raw bands + vegetation indices + environmental variables, (8) raw bands + soil indices + environmental variables, and (9) raw bands + environmental variables + soil indices.

Table 5. The different data configurations for the nine machine learning regression experiments.

Experiment	Number of Variables	Data Configuration
1	10	Raw bands
2	17	Raw bands and vegetation indices
3	15	Raw bands and soil indices
4	17	Raw bands and environmental variables
5	29	Raw bands, vegetation indices, soil indices, and environmental variables
6	22	Raw bands, vegetation indices, and soil indices
7	24	Raw bands, vegetation indices, and environmental variables
8	22	Raw bands, soil indices, and environmental variables
9	19	Raw bands, environmental variables, and soil indices

2.7. Model Evaluation

The predictive performances of the RF, GB, and XG models were evaluated using validation indices. These included the fraction of predictions within a factor of two (*FAC2*), mean absolute error (*MAE*), mean bias error (*MBE*), root mean square error (*RMSE*), Pearson correlation (*r*), R^2, and cross validation (*CV*) as shown in Equations (1)–(7):

$$FAC2 : 0.5 \leq \frac{P_i}{O_i} \leq 2.0 \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \tag{2}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{4}$$

$$r = \frac{1}{(n-1)}\sum_{i=1}^{n}(\frac{P_i - \overline{P}}{\sigma_p})(\frac{O_i - \overline{O}}{\sigma_o}) \tag{5}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(P_i - \overline{O}_i\right)^2}{\sum_{i=1}^{n}\left(P_i - \overline{O}_i\right)^2} \tag{6}$$

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}R_i \tag{7}$$

where n represents the number of sample points, P_i represents the predicted soil nitrogen content, O_i represents the observed soil nitrogen content in site i, and σ represents the standard deviation. The reader is directed to Carslaw and Ropkins [53] for further information on these model evaluation matrices. The Taylor diagram was derived using the Openair package in R software [53].

3. Results

3.1. Statistical Analaysis for Soil Nitrogen Content Measurements

Different vegetation indices (Figure 3) described in Section 2.4 were evaluated to retain indices that perform optimally for soil nitrogen content estimation. The RF, XG, and GB models were used to relate the vegetation indices to soil nitrogen. The PSRI, NDVIRE1n, EVI, NDVIRE2n, NDVIRE3n, GNDVI, and MSRRE were retained for further analysis. These vegetation indices were strongly related to the soil nitrogen content with an R^2 of 0.62 to 0.81. The soil nitrogen content measurements collected at the smallholder maize farms are characterized in Table 6. The nitrogen content was low for the farms, ranging from 0.014–0.088%. The mean is lower than the standard deviation, which shows that the data are clustered closely around the mean. The mean is greater than the median, indicating a positively skewed distribution similar to the skewness value of 1.42 [54]. The

nitrogen content measurements were related to each of the variables in the regression experiments through a correlation matrix (Table 5). The MSRRE, NDVIRE1-3n, EVI, LST, and TWI had positive relationships with the soil nitrogen content. The remaining variables had a negative relationship with soil nitrogen. The PSRI, NDVIRE1-3n, EVI, CI, BI, SI, RI, and B4-B12 were strongly related to the soil nitrogen content. However, the SLP, CA, ASP, DEM, TWI, LST, and RAIN had a weak relationship with soil nitrogen. Moderate relationships where observed for the HI, B3, and soil nitrogen. Multicollinearity was identified between the vegetation indices, soil indices, and raw bands. These variables were highly linearly related.

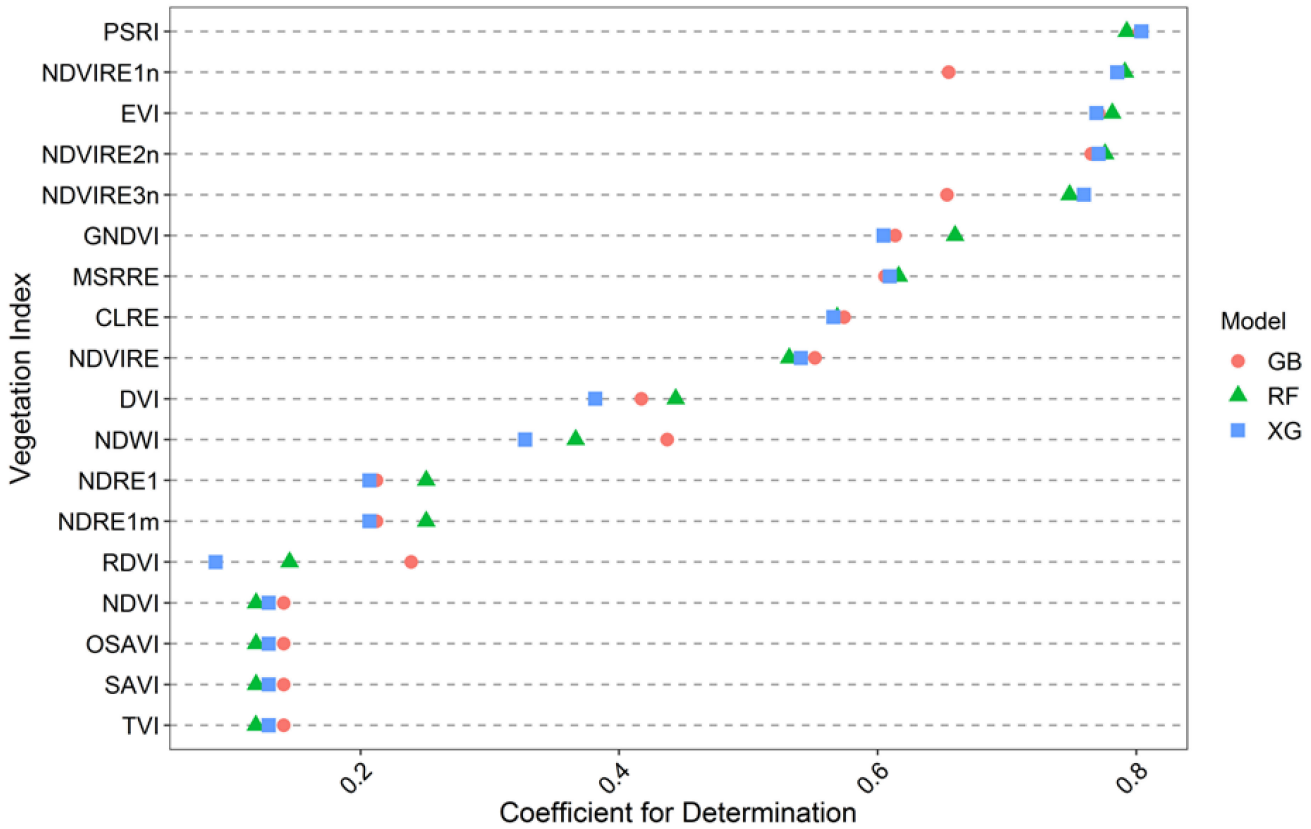

Figure 3. Vegetation indices evaluated for mapping soil nitrogen content.

Table 6. Statistical analysis for the soil nitrogen content samples.

Soil Nitrogen							
(a) Descriptive Statistics							
	Count	**Minimum (%)**	**Maximum (%)**	**Mean (%)**	**Median (%)**	**Standard Deviation**	**Skewness**
Nitrogen	105	0.014	0.088	0.033	0.025	0.019	1.424
(b) Correlation							
Variable	**r**	**Variable**	**r**	**Variable**	**r**	**Variable**	**r**
MSRRE	0.579	CI	−0.713	B6	−0.899	TWI	0.081
PSRI	−0.793	BI	−0.798	B7	−0.894	DEM	−0.292
NDVIRE3n	0.835	SI	−0.804	B8	−0.883	ASP	−0.011
NDVIRE2n	0.840	RI	−0.748	B8A	−0.889	CA	−0.024
NDVIRE1n	0.737	B2	−0.061	B11	−0.883	SLP	−0.154
EVI	0.838	B3	−0.463	B12	−0.870		
GNDVI	−0.757	B4	−0.884	RAIN	−0.268		
HI	−0.591	B5	−0.898	LST	0.117		

3.2. Model Evaluation

The model performance statistics derived from the testing data (n = 32 samples) are summarized in Table 7. The best performing model from all experiments was the RF model for experiment 4. This model had the highest accuracy for soil nitrogen content estimation based on the lowest values for RMSE and MAE (RMSE = 0.0076% and MAE = 0.0054%) and the highest r and R^2 (r = 0.95 and R^2 = 0.90). The predicted soil nitrogen values were smaller than the observed values based on the MBE (MBE = − 0.0013%). Additionally, this model had a FAC2 = 1, indicating a perfect model similar to the FAC2 values for the other experiments. The least optimal performing model overall was the XG model for experiment 6 containing the raw bands, soil indices, and vegetation indices. This model had a high error rate based on the high RMSE and MAE (RMSE = 0.0090% and MAE = 0.0063%) and the lowest r and R^2 (r = 0.9149 and R^2 = 0.8371). Furthermore, this model overestimated the soil nitrogen content based on the MBE (MBE = 0.0004%). The raw bands and environmental variables were sufficient to model soil nitrogen content with the RF (RF4) and GB (GB4) model. However, additional soil indices were needed in XG (XG8) for estimating soil nitrogen more accurately.

Table 7. Model evaluation statistics for the three machine learning models in different experiments.

Model	FAC2	MAE (%)	MBE (%)	RMSE (%)	r	R^2	CV
RF1	0.9688	0.0067	0.0012	0.0086	0.9324	0.8694	0.7563
RF2	0.9688	0.0061	0.0000	0.0086	0.9302	0.8653	0.8079
RF3	0.9688	0.0071	0.0004	0.0092	0.9204	0.8472	0.7891
RF4	1.0000	0.0054	−0.0013	0.0076	0.9486	0.8998	0.6625
RF5	1.0000	0.0066	−0.0007	0.0086	0.9232	0.8523	0.7720
RF6	0.9688	0.0063	−0.0003	0.0089	0.9256	0.8568	0.6604
RF7	1.0000	0.0053	0.0000	0.0080	0.9433	0.8898	0.7104
RF8	1.0000	0.0059	0.0002	0.0083	0.9368	0.8775	0.6885
RF9	1.0000	0.0056	0.0000	0.0082	0.9395	0.8827	0.8645
GB1	0.9688	0.0070	0.0007	0.0092	0.9210	0.8482	0.5325
GB2	1.0000	0.0059	−0.0001	0.0084	0.9348	0.8739	0.6670
GB3	1.0000	0.0068	−0.0003	0.0092	0.9177	0.8423	0.6124
GB4	1.0000	0.0061	0.0001	0.0083	0.9369	0.8778	0.6354
GB5	1.0000	0.0061	0.0000	0.0084	0.9347	0.8737	0.7043
GB6	1.0000	0.0062	−0.0006	0.0087	0.9298	0.8645	0.7942
GB7	1.0000	0.0060	0.0002	0.0084	0.9336	0.8716	0.7734
GB8	0.9688	0.0064	−0.0009	0.0094	0.9172	0.8413	0.7556
GB9	1.0000	0.0058	0.0008	0.0083	0.9315	0.8676	0.7296
XG1	0.9688	0.0062	0.0003	0.0084	0.9311	0.8669	0.5671
XG2	0.9688	0.0057	0.0001	0.0085	0.9257	0.8569	0.8546
XG3	0.9688	0.0065	0.0005	0.0089	0.9227	0.8513	0.5970
XG4	1.0000	0.0062	0.0004	0.0088	0.9221	0.8502	0.5711
XG5	1.0000	0.0059	0.0004	0.0081	0.9352	0.8747	0.6121
XG6	0.9688	0.0063	0.0004	0.0090	0.9149	0.8371	0.6367
XG7	1.0000	0.0061	0.0007	0.0087	0.9234	0.8527	0.6453
XG8	1.0000	0.0054	0.0003	0.0077	0.9434	0.8900	0.5954
XG9	0.9688	0.0058	0.0002	0.0086	0.9300	0.8648	0.5839

Note: Random forest experiment number (RFx), gradient boosting experiment number (GBx), extreme gradient boosting experiment number (XGx) defined in Table 4.

The Taylor diagram in Figure 4 was used to verify the model performance. All models had high correlation coefficients ranging from 0.91 to 0.95 and they plotted close to the observed reference value at the origin. Additionally, they had a similar performance shown by the clustering of points with the same location on the Taylor diagram [55]. However, the RF4 model had a slightly better performance compared to the other models based on the lowest standard deviation and root mean squared (RMS) error. The correlation coefficient was also high for this model, signifying a good fit between the observed and predicted values. The XG8 and GB4 models were the optimal performing models for the XG and GB models. They had a considerably lower standard deviation and RMS values but a high correlation. Additionally, the predicted values from these models were closer to the observed values.

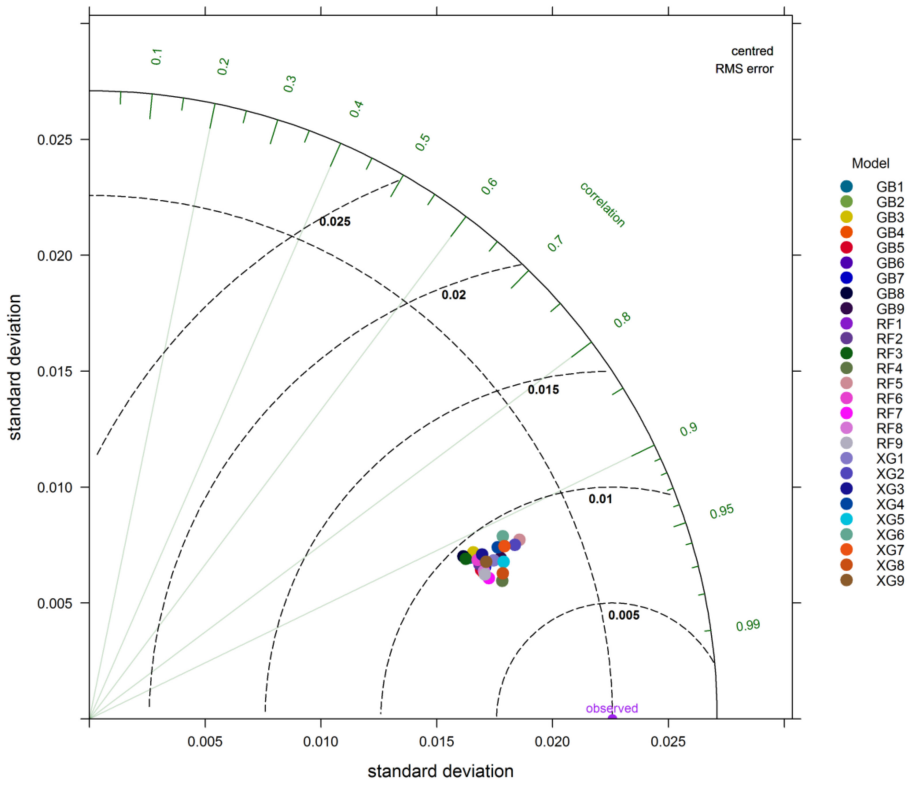

Figure 4. Taylor diagram for the nine experiments applying the three machine learning models.

Scatterplots were constructed for optimal performing RF, GB, and XG models to relate the observed and predicted soil nitrogen content in Figure 5. The data points are close to the diagonal line for all three models, indicating a good agreement between the observed and predicted values. The RF4 model had a slightly better performance R^2 (R^2 = 0.90) than the other models and was statistically significant ($p = 1.6 \times 10^{-16}$) at a 95% confidence interval. The GB and XG models had similar R^2 values (R^2 = 0.88 and R^2 = 0.89). However, GB had a higher p-value of 3.1×10^{-15} in comparison to XG with a p-value of 6.3×10^{-16}. Both models were statistically significant at a 95% confidence interval.

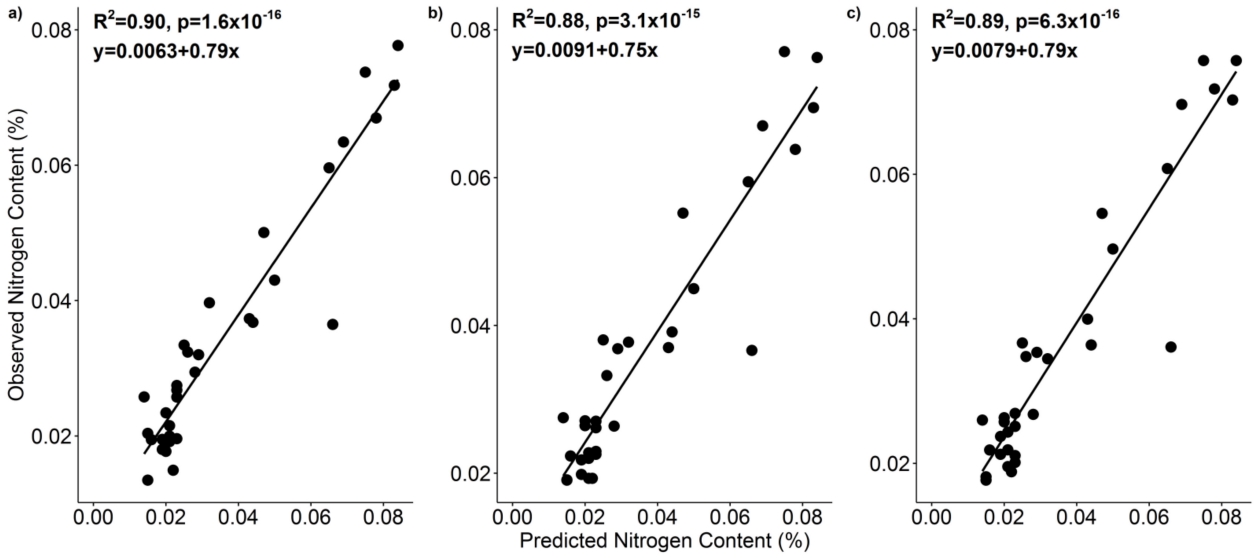

Figure 5. The relationship between observed soil nitrogen and predicted soil nitrogen where (**a**) is RF4, (**b**) is GB4, and (**c**) is XG8.

3.3. Variable Importance

The importance of the predictor variables was determined for the most robust RF, GB, and XG models. All three models in Figure 6 varied in terms of predictor importance. The most important predictors for RF were B7, B5, B6, and B4. These were derived from experiment 4. The GB model ranked B4, B6, B5, and B12 highly from experiment 4. The B4 band was important in the XG model followed by CI and B5 in experiment 8. The RF model had a more even distribution of predictor importance in comparison to GB and XG where there is a greater contrast between the important (highest 4) and least important predictors (after the highest 4 predictors).

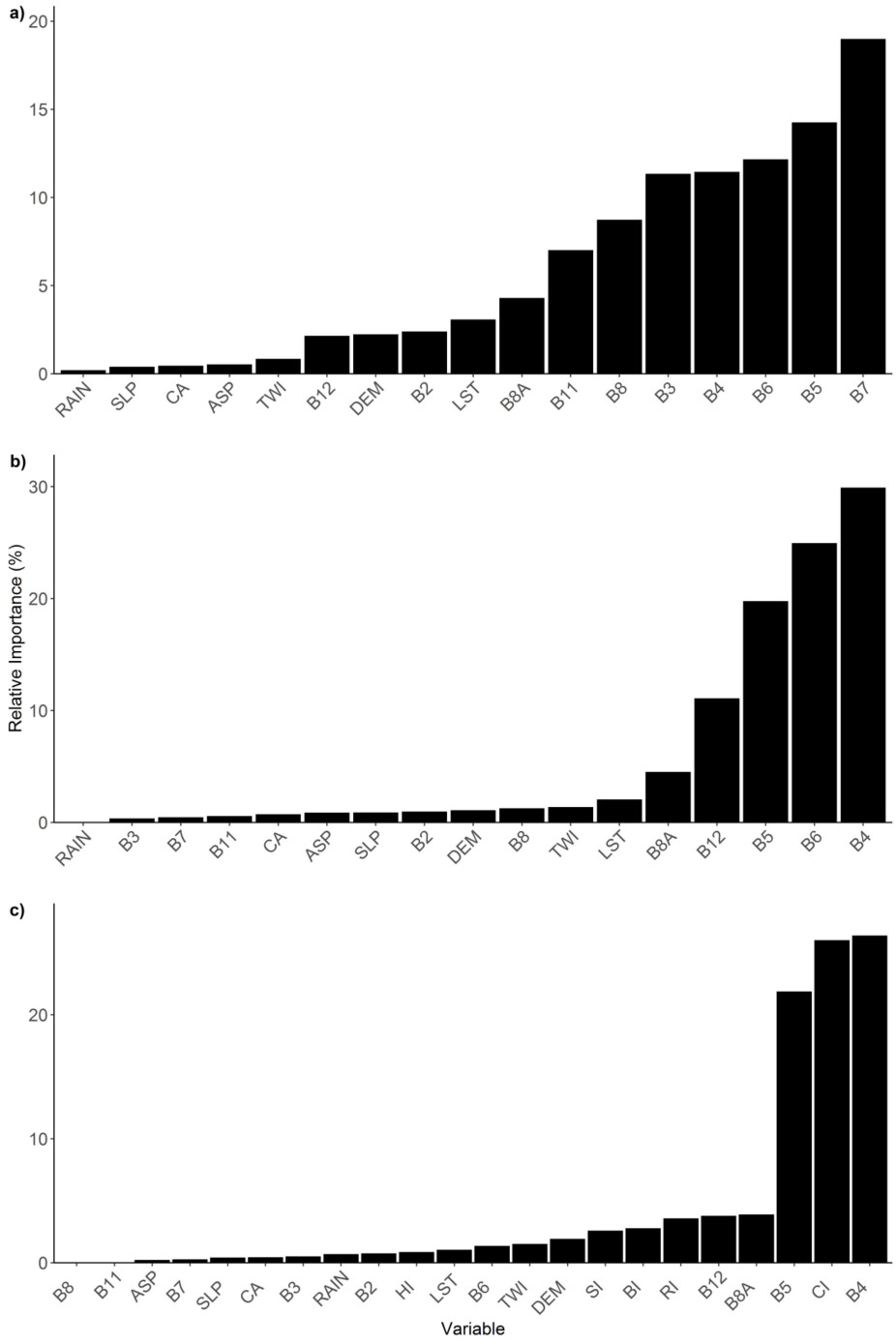

Figure 6. The ranking of variables for predicting soil nitrogen content with (**a**) RF4, (**b**) GB4, and (**c**) XG8 algorithms.

3.4. Mapping Soil Nitrogen Content for Smallholder Maize Farms

The spatial distribution of soil nitrogen was mapped in Figures 7–9. There were differences in the spatial distribution of nitrogen for the smallholder maize farms. The smallholder farms in the central and southeastern part of the study area had a lower nitrogen content. However, the farms in the southern part of the study area had a higher nitrogen content. The maps generated by the RF and XG algorithms were similar, but GB overestimated the nitrogen content.

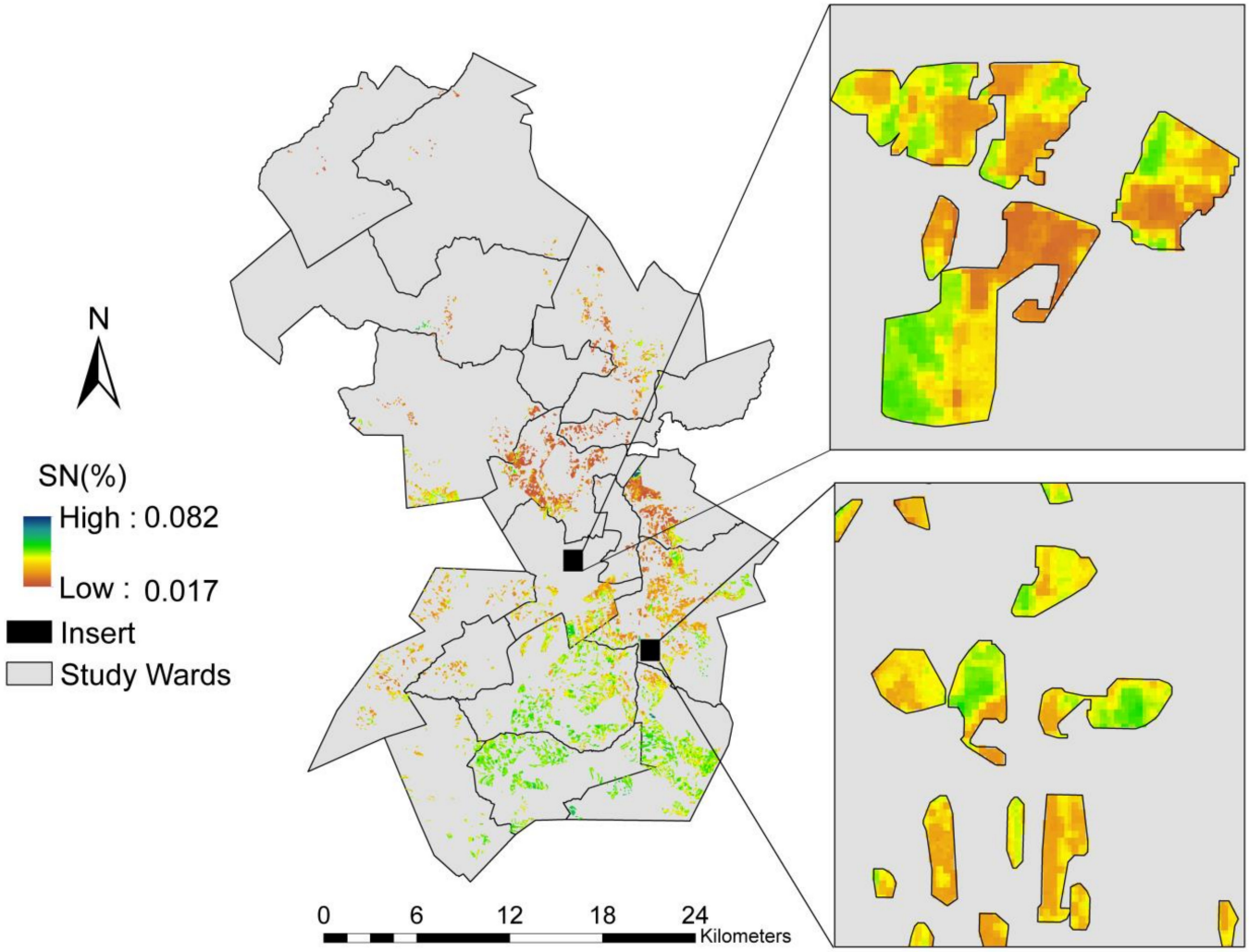

Figure 7. The spatial distribution of soil nitrogen mapped with the random forest model for experiment 4.

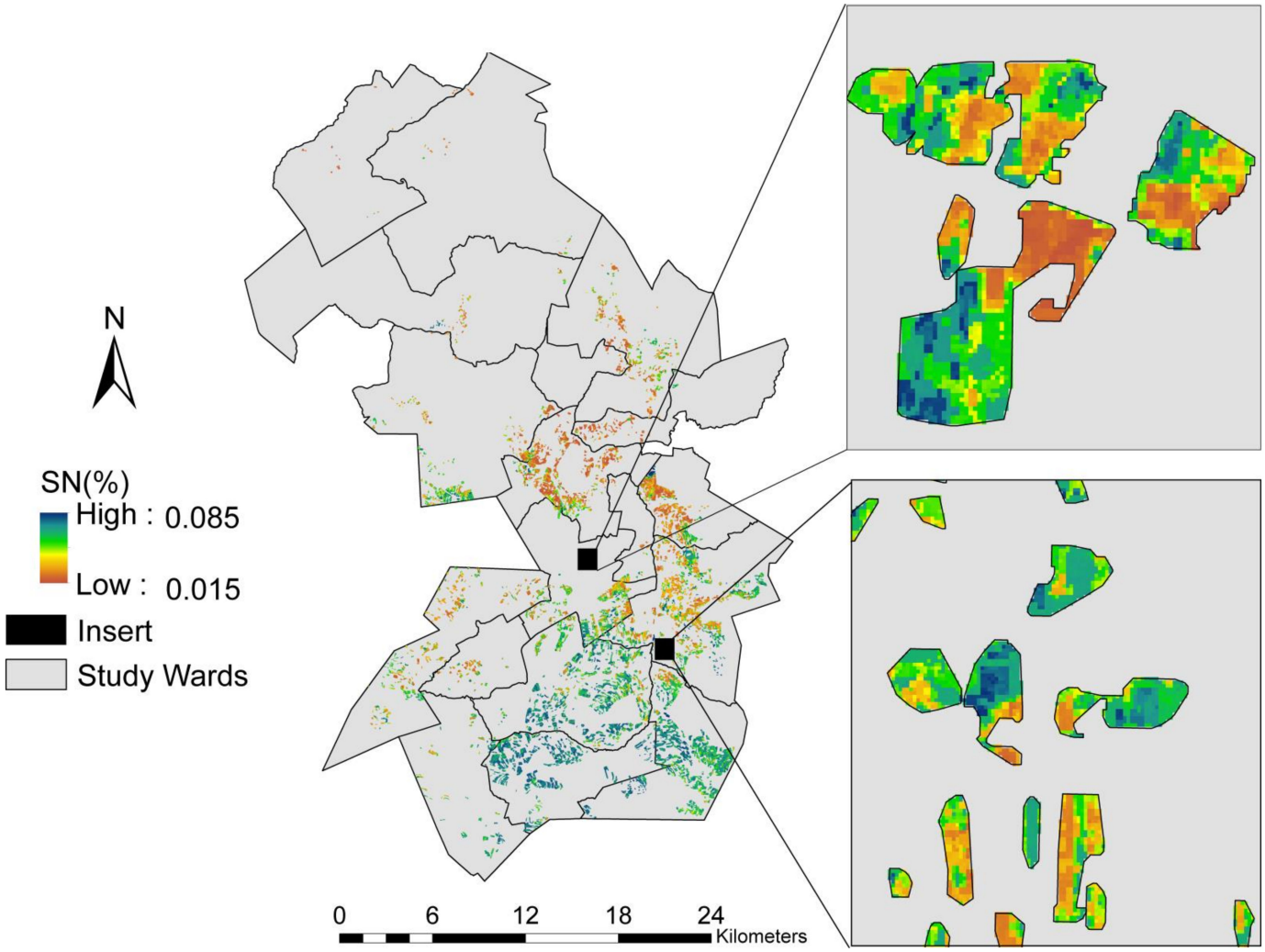

Figure 8. The spatial distribution of soil nitrogen mapped with the gradient boosting model for experiment 4.

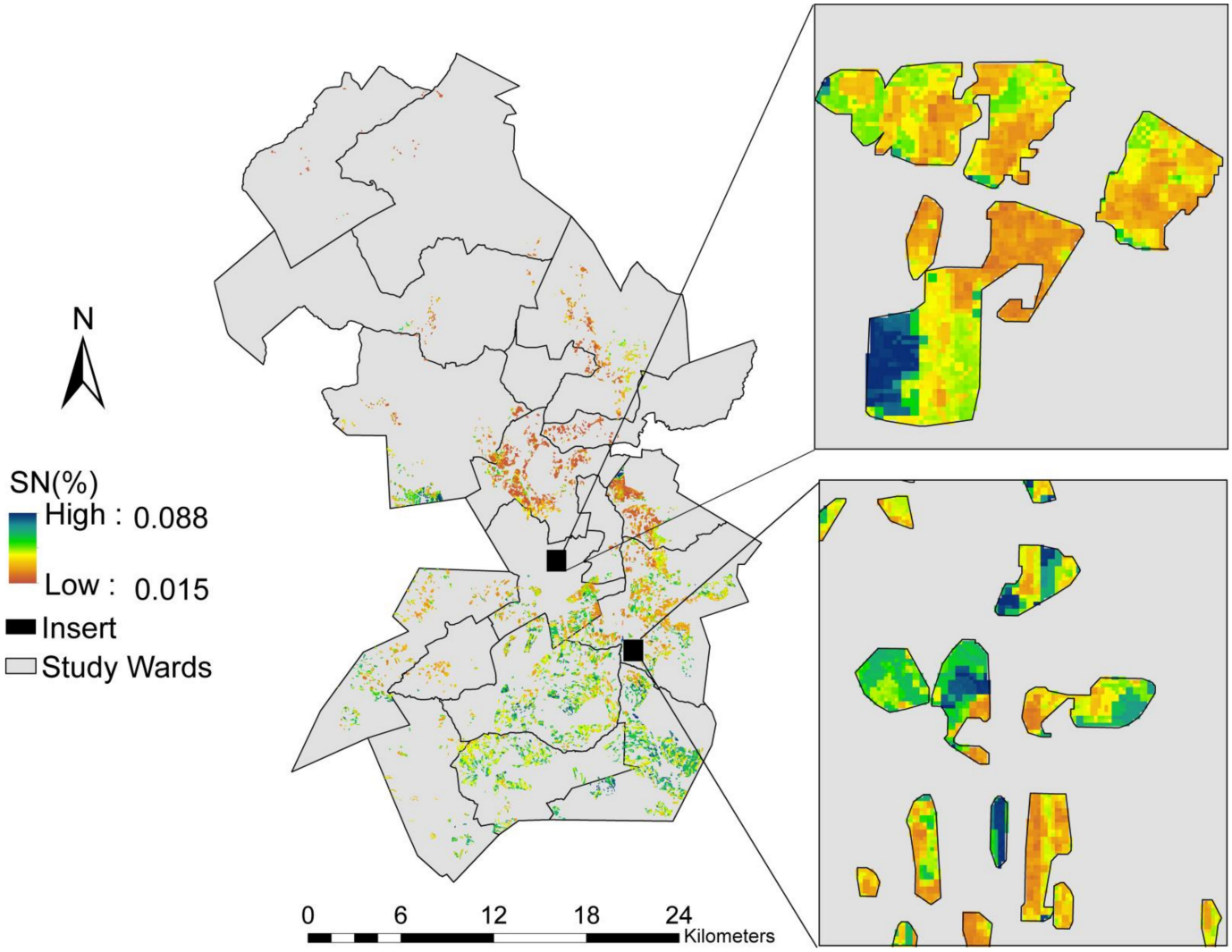

Figure 9. Distribution map of soil nitrogen obtained using the XG model is for experiment 8.

4. Discussion

This study assessed the applicability of Sentinel-2 bands, derived soil and vegetation indices, and environmental data for predicting soil nitrogen in the smallholder maize farms of Makhuduthamaga district. Descriptive statistics were generated for the collected soil nitrogen content samples. Experiments were used to evaluate the performance of RF, GB, and XG machine learning algorithms in a regression format. The variable importance measure for each algorithm was used to determine which predictors had the most influence. The best performing algorithms in each experiment were then used for mapping nitrogen content. The results showed that the Sentinel-2 bands and environmental variables have a superior performance when estimating the soil nitrogen content in comparison to the vegetation indices and soil indices.

Findings from the descriptive statistics indicate that nitrogen content is low (0.014–0.088%) for the smallholder maize farms. This is expected because the smallholder farms within the study area rarely apply nitrogen fertilization and a small proportion of the farmers use cow manure as fertilizer. For example, Nyamangara et al. [56] conducted experiments for three years and found that the combination of cow manure and nitrogen fertilizers in smallholder maize farms in Zimbabwe improved soil nitrogen content and increased maize yield. Furthermore, data exploration in our study revealed that multicollinearity was present when relating soil nitrogen content to the different predictor variables. The presence of multicollinearity implies that the application of multiple linear regression with these

variables to predict the soil nitrogen content would be unreliable [57]. Multicollinearity introduces large variances in the least squares estimators (regression coefficients) and lowers the quality of the resulting parameter estimates, and the variables have a low information content [58]. The main advantage of the machine learning techniques, applied in the present study, is that they are less prone to multicollinearity problems. For example, Jaya et al. [59] found that the artificial neural network model had a lower bias, mean squared error, and minimized residuals in comparison to a multiple linear regression model when multicollinearity was present. Additionally, Farrell et al. [60] observed that multicollinearity removal and correlation removal did not reduce the performance of RF and support vector machine substantially. The robustness of machine learning could be due to the adaptive learning process used by the models that reduces errors [24,25]. For example, RF uses bagging, XG uses additive training strategies, and GB reduces the model residuals along the gradient direction, which minimizes the multicollinearity problem.

Three predictive models were evaluated. Findings show that the RF model performs better than the GB and XG models when estimating soil nitrogen at smallholder maize farms in our study area. These results are similar to other studies that show the high capacity of RF in mapping soil nitrogen content [19,61–65]. Furthermore, the findings suggest that the XG model needs more input variables to model soil nitrogen content in comparison to GB and RF. This can be attributed to the implementation of the models: the XG algorithm is sensitive to outliers because the individual learners are in series format, and RF is not sensitive to outliers because it is a parallel implementation of multiple decision trees [66]. In terms of variability, this study found an R^2 of 0.87–0.90, RMSE of 0.0086–0.0092%, and CV of 0.66–0.81 with RF, which is the most robust model. Our results are similar to López-Calderón et al. [65] that found an R^2 of 0.77 and a mean square error of 0.15% when predicting soil total nitrogen content applying RF for forage maize with UAV imagery. Additionally, Sorenson et al. [62] used field reflectance spectroscopy for estimating soil nitrogen content and reported a cross-validation RMSE of 0.62% and R^2 of 0.78 with RF for reclaimed soils. Furthermore, Deng et al. [64] found a cross validation R^2 = 0.65 and RMSE = 0.43 g kg^{-1} with RF applied on MODIS data when estimating soil nitrogen content for croplands. Contrary to our findings, Xu et al. [19] reported an adjusted R^2 of 0.49 and RMSE of 125.71 mg kg^{-1} with Landsat 8 data applying RF to predict soil nitrogen at smallholder farmlands planting different crops. Jeong et al. [61] observed an R^2 = 0.552 and RMSE = 1.131 mg g^{-1} when applying RF soil nitrogen content estimation in a complex terrain with Landsat TM data. These differences in findings can be influenced by the input variables or other factors such as whether the soil is completely bare or has plant coverage, which can influence the predicted soil nitrogen content. For example, the study by Beguin et al. [67] found that the input predictors affect the predictive capacity of models predicting soil properties. Other studies such as Zhang et al. [63] observed different performance for the digital soil map generated in a vegetated condition (R^2 = 0.67) and completely bare soil condition (R^2 = 0.80) with RF.

Variable importance was done to determine the most important predictors for estimating soil nitrogen content at smallholder maize farms. The results showed that the Sentinel-2 bands have an advantage when estimating soil nitrogen content. However, environmental variables had a lower ranking, and additional soil indices were necessary in the XG model. These findings are similar to other studies that found that spectral bands are more important than environmental variables [63,68,69]. However, some studies showed contrasting results in which the environmental variables had the highest ranking [30,70]. The differences in findings are attributed to variations in the model input variables in these studies. For example, most of these studies used Landsat optical data for mapping soil nitrogen content, which does not have the RE bands that Sentinel-2 has, which the current study incorporated. Additionally, the presence of maize crops within the smallholder farms in the current study could have contributed to the higher importance of the red-edge bands. These bands are sensitive to variations in chlorophyll content, differences in the leaf structure, and plant biomass [33,35]. The radiation from the red-edge penetrates deeper

into the crop canopy and leaves in comparison to visible light due to lower chlorophyll absorption in the visible region [71]. Xu et al. [19] also found that red-edge spectral bands are important when estimating soil total nitrogen in smallholder farms that have different crops planted. These studies prove that red-edge bands have a high capability to estimate total nitrogen content accurately in smallholder farms that have crop cover. The high importance of the CI and RI amongst the soil indices was expected within the study area because most of the soils are red soils that have a high iron oxide content, possibly related to haematite, to which the RI is sensitive [38]. The most important predictors were LST, DEM, and TWI for the environmental variables. The LST affects the spatial distribution of soil nitrogen through its effect on soil temperature, thereby affecting the process of nitrogen mineralization [72]. The DEM is important because elevation plays a role in microclimate, runoff, evaporation, and transpiration [73]. The TWI is an indicator of soil moisture distribution [40]. Soil moisture conditions, in addition of course to soil nutrients, are determinants of crop vigor and development. The distinction between highly ranked predictors and low-ranking predictors in the GB and XG models shows that further exploration of the influence of the predictors on model performance can be done for both models for model optimization.

The spatial distribution of soil nitrogen was mapped. The resulting spatial maps produced from the three algorithms were similar. This finding proved the high capability of machine learning to estimate soil nitrogen content in smallholder maize farms. The soil nitrogen maps generated in this study can be used as a tool to guide decision making for smallholder farms. Recommendations by crop consultants, extension services, and fertilizer dealers can also benefit from using nitrogen content maps. Government initiatives providing farmers with agricultural inputs can use such maps to determine the soil nitrogen content at the farms and the proportion of fertilizer to use, because different fertilizer quantities affect maize yield differently, as shown by Nyamugara et al. [56]. Improved levels of soil nitrogen content at smallholder farms will increase maize yields, thereby improving food security [1–3]. This application contributes to Sustainable Development Goals (SDG) number 2 (Zero Hunger), target 2.4 and indicator 2.4.1, which are concerned with mitigating factors that affect agricultural production, ensuring sustainable agriculture and increasing the proportion of agricultural area under production [74].

The main limitation of this study is that a small number of farms were visited for field data collection due to the high cost for laboratory processing of samples and fieldwork. This study recommends further exploration of Sentinel-1 and Sentinel-2 data for estimating soil nitrogen in smallholder farms [63,69,70]. Studies focusing on smallholder farms are lacking, especially in an African context, and these farms are important for food security and rural livelihoods [7,8]. Training programs are recommended for the smallholder farms to improve the awareness of farmers on chemical fertilization. For example, nitrogen is essential when the crop is actively growing, but nitrogen application before that time can lead to losses through leaching or subsurface flow [75]. Other more cost-effective alternatives to nitrogen fertilizers such as leguminous trees and shrubs grown with maize are recommended for smallholder farms in resource poor areas. These will provide nitrogen-rich residues that increase soil fertility [76].

5. Conclusions

This study was aimed at assessing Sentinel-2 bands, derived soil and vegetation indices, and environmental variables for predicting soil nitrogen in smallholder maize farms applying machine learning regression. Different predictor variables were related to soil nitrogen content. The red, red-edge, and short-wave infrared bands were strongly related to soil nitrogen with correlations of 0.89–0.90. The machine learning models applied in this study (RF, GB, and XG) were suitable for the data because multicollinearity was present between the predictors, which these models dealt with effectively. Model evaluation results show that machine learning models have a high predictive capacity in estimating soil nitrogen (R^2 = 0.84–0.90 and RMSE = 0.0076–0.0094%) in smallholder

farms. Variable importance revealed that the Sentinel-2 bands, particularly the red and red-edge bands, are fundamental for modeling soil nitrogen in all three models. The soil nitrogen maps generated in this study can be used as a tool to guide decision making for smallholder farms. Recommendations by governments, extension services, and fertilizer dealers can also benefit from using such maps. These maps are useful to establish nitrogen management plans in the smallholder farms, which will increase maize yields, thereby improving food security.

Author Contributions: Z.M.-M. conceptualized and developed the original draft of the manuscript. G.J.C. revised the manuscript, supervised, and provided financial resources for the project. C.M. was involved in data analysis, review and edit of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the Agricultural Research Council, the National Research Foundation (Grant number: SFH170524232697), Spatial Business Intelligence—SIQ, GeoTerraImage and the University of Pretoria.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Sentinel-2 satellite data are freely available from the Copernicus Hub (https://www.sentinel-hub.com/, accessed on 9 August 2021).

Acknowledgments: The authors would like to thank the Agricultural Research Council and University of Pretoria for hosting this research. We would also like to extend our gratitude to the anonymous reviewers for their valuable comments.

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

References

1. Sinclair, T.R.; Muchow, R.C. Effect of Nitrogen Supply on Maize Yield: I. Modeling Physiological Responses. *Agron. J.* **1995**, *87*, 632–641. [CrossRef]
2. Otto, R.; Castro, S.A.Q.; Mariano, E.; Castro, S.G.Q.; Franco, H.C.J.; Trivelin, P.C.O. Nitrogen Use Efficiency for Sugarcane-Biofuel Production: What Is Next? *Bioenerg. Res.* **2016**, *9*, 1272–1289. [CrossRef]
3. Chlingaryan, A.; Sukkarieh, S.; Whelan, B. Machine Learning Approaches for Crop Yield Prediction and Nitrogen Status Estimation in Precision Agriculture: A Review. *Comput. Electron. Agric.* **2018**, *151*, 61–69. [CrossRef]
4. Batjes, N.H. Total Carbon and Nitrogen in the Soils of the World. *Eur. J. Soil Sci.* **1996**, *47*, 151–163. [CrossRef]
5. Lemcoff, J.H.; Loomis, R.S. Nitrogen Influences on Yield Determination in Maize. *Crop Sci.* **1986**, *26*, 1017–1022. [CrossRef]
6. Osterholz, W.R.; Rinot, O.; Liebman, M.; Castellano, M.J. Can Mineralization of Soil Organic Nitrogen Meet Maize Nitrogen Demand? *Plant Soil* **2017**, *415*, 73–84. [CrossRef]
7. Shi, W.; Tao, F. Vulnerability of African Maize Yield to Climate Change and Variability during 1961–2010. *Food Sec.* **2014**, *6*, 471–481. [CrossRef]
8. Fischer, K.; Hajdu, F. Does Raising Maize Yields Lead to Poverty Reduction? A Case Study of the Massive Food Production Programme in South Africa. *Land Use Policy* **2015**, *46*, 304–313. [CrossRef]
9. Jones, P.G.; Thornton, P.K. Representative Soil Profiles for the Harmonized World Soil Database at Different Spatial Resolutions for Agricultural Modelling Applications. *Agric. Syst.* **2015**, *139*, 93–99. [CrossRef]
10. Batjes, N.H. *SOTER-Based Soil Parameter Estimates for Southern Africa;* Report 2004/04; ISRIC—World Soil Information: Wageningen, The Netherlands, 2004.
11. Jones, A.; Breuning-Madsen, H.; Brossard, M.; Dampha, A.; Deckers, J.; Dewitte, O.; Gallali, T.; Hallett, S.; Jones, R.; Kilasara, M.; et al. *Soil Atlas of Africa*; European Commission, Publications Office of the European Union: Luxembourg, Luxembourg, 2013; ISBN 978-92-79-26715-4. [CrossRef]
12. Chivasa, W.; Mutanga, O.; Biradar, C. Application of Remote Sensing in Estimating Maize Grain Yield in Heterogeneous African Agricultural Landscapes: A Review. *Int. J. Remote Sens.* **2017**, *38*, 6816–6845. [CrossRef]
13. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
14. Wang, Q.; Blackburn, G.A.; Onojeghuo, A.O.; Dash, J.; Zhou, L.; Zhang, Y.; Atkinson, P.M. Fusion of Landsat 8 OLI and Sentinel-2 MSI Data. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3885–3899. [CrossRef]
15. Filella, I.; Penuelas, J. The Red Edge Position and Shape as Indicators of Plant Chlorophyll Content, Biomass and Hydric Status. *Int. J. Remote Sens.* **1994**, *15*, 1459–1470. [CrossRef]
16. Shi, T.; Cui, L.; Wang, J.; Fei, T.; Chen, Y.; Wu, G. Comparison of Multivariate Methods for Estimating Soil Total Nitrogen with Visible/near-Infrared Spectroscopy. *Plant Soil* **2013**, *366*, 363–375. [CrossRef]

17. Yang, J.; Gong, W.; Shi, S.; Du, L.; Sun, J.; Song, S. Estimation of Nitrogen Content Based on Fluorescence Spectrum and Principal Component Analysis in Paddy Rice. *Plant Soil Environ.* **2016**, *62*, 178–183. [CrossRef]
18. de Brogniez, D.; Ballabio, C.; Stevens, A.; Jones, R.J.A.; Montanarella, L.; van Wesemael, B. A Map of the Topsoil Organic Carbon Content of Europe Generated by a Generalized Additive Model: Soil Organic Carbon Content at Pan-European Level. *Eur. J. Soil Sci.* **2015**, *66*, 121–134. [CrossRef]
19. Xu, Y.; Smith, S.E.; Grunwald, S.; Abd-Elrahman, A.; Wani, S.P.; Nair, V.D. Estimating Soil Total Nitrogen in Smallholder Farm Settings Using Remote Sensing Spectral Indices and Regression Kriging. *Catena* **2018**, *163*, 111–122. [CrossRef]
20. Friedl, M.A.; Brodley, C.E. Decision Tree Classification of Land Cover from Remotely Sensed Data. *Remote Sens. Environ.* **1997**, *61*, 399–409. [CrossRef]
21. Chang, D. Estimation of Soil Physical Properties Using Remote Sensing and Artificial Neural Network. *Remote Sen. Environ.* **2000**, *74*, 534–544. [CrossRef]
22. Heumann, B.W. An Object-Based Classification of Mangroves Using a Hybrid Decision Tree—Support Vector Machine Approach. *Remote Sens.* **2011**, *3*, 2440–2460. [CrossRef]
23. Wang, L.; Zhou, X.; Zhu, X.; Dong, Z.; Guo, W. Estimation of Biomass in Wheat Using Random Forest Regression Algorithm and Remote Sensing Data. *Crop J.* **2016**, *4*, 212–219. [CrossRef]
24. Belgiu, M.; Drăguţ, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
25. Cooner, A.; Shao, Y.; Campbell, J. Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake. *Remote Sens.* **2016**, *8*, 868. [CrossRef]
26. Izquierdo-Verdiguier, E.; Gomez-Chova, L.; Bruzzone, L.; Camps-Valls, G. Semisupervised Kernel Feature Extraction for Remote Sensing Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 5567–5578. [CrossRef]
27. Li, X.; Chen, W.; Cheng, X.; Wang, L. A Comparison of Machine Learning Algorithms for Mapping of Complex Surface-Mined and Agricultural Landscapes Using ZiYuan-3 Stereo Satellite Imagery. *Remote Sens.* **2016**, *8*, 514. [CrossRef]
28. Siebert, S.J.; Van Wyk, A.E.; Bredenkamp, G.J.; Siebert, F. Vegetation of the Rock Habitats of the Sekhukhuneland Centre of Plan Endemism, South Africa. *Bothalia* **2003**, *33*, 207–228. [CrossRef]
29. SDM. *Greater Sekhukhune Cross Border District Municipality Integrated Development Plan: 2019/20*; SDM: Groblersdal, South Africa, 2019.
30. Wang, S.; Adhikari, K.; Wang, Q.; Jin, X.; Li, H. Role of Environmental Variables in the Spatial Distribution of Soil Carbon (C), Nitrogen (N), and C:N Ratio from the Northeastern Coastal Agroecosystems in China. *Ecol. Indic.* **2018**, *84*, 263–272. [CrossRef]
31. Mandal, U.K. Spectral color indices based geospatial modeling of soil organic matter in Chitwan district, Nepal. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Prague, Czech Republic, 12–19 July 2016; Volume 41. [CrossRef]
32. Merzlyak, M.N.; Gitelson, A.A.; Chivkunova, O.B.; Rakitin, V.Y. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiol. Plant.* **1999**, *106*, 135–141. [CrossRef]
33. Fernández-Manso, A.; Fernández-Manso, O.; Quintano, C. SENTINEL-2A Red-Edge Spectral Indices Suitability for Discriminating Burn Severity. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *50*, 170–175. [CrossRef]
34. Chen, J.M. Evaluation of Vegetation Indices and a Modified Simple Ratio for Boreal Applications. *Can. J. Remote Sens.* **1996**, *22*, 229–242. [CrossRef]
35. Miura, T.; Huete, A.R.; Yoshioka, H. Evaluation of Sensor Calibration Uncertainties on Vegetation Indices for MODIS. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 1399–1409. [CrossRef]
36. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a Green Channel in Remote Sensing of Global Vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
37. Madeira, J.; Bedidi, A.; Cervelle, B.; Pouget, M.; Flay, N. Visible spectrometric indices of hematite (Hm) and goethite (Gt) content in lateritic soils: The application of a Thematic Mapper (TM) image for soil-mapping in Brasilia, Brazil. *Int. J. Remote Sens.* **1997**, *18*, 2835–2852. [CrossRef]
38. Bullard, J.E. Quantifying Iron Oxide Coatings on Dune Sands Using Spectrometric Measurements: An Example from the Simpson-Strzelecki Desert. *Aust. J. Geophys. Res.* **2002**, *107*, 2125. [CrossRef]
39. Wu, S.; Li, J.; Huang, G.H. A study on DEM-derived primary topographic attributes for hydrologic applications: Sensitivity to elevation data resolution. *Appl. Geogr.* **2008**, *28*, 210–223. [CrossRef]
40. Sörensen, R.; Zinko, U.; Seibert, J. On the calculation of the topographic wetness index: Evaluation of different methods based on field observations. *Hydrol. Earth Syst. Sci.* **2006**, *10*, 101–112. [CrossRef]
41. Kubota, T.; Shige, S.; Hashizume, H.; Aonashi, K.; Takahashi, N.; Seto, S.; Hirose, M.; Takayabu, Y.N.; Ushio, T.; Nakagawa, K.; et al. Global Precipitation Map Using Satellite-Borne Microwave Radiometers by the GSMaP Project: Production and Validation. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 2259–2275. [CrossRef]
42. Ermida, S.L.; Soares, P.; Mantas, V.; Göttsche, F.-M.; Trigo, I.F. Google Earth Engine Open-Source Code for Land Surface Temperature Estimation from the Landsat Series. *Remote Sens.* **2020**, *12*, 1471. [CrossRef]
43. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
44. Pal, M. Random Forest Classifier for Remote Sensing Classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]

45. Hutengs, C.; Vohland, M. Downscaling Land Surface Temperatures at Regional Scales with Random Forest Regression. *Remote Sens. Environ.* **2016**, *178*, 127–141. [CrossRef]
46. Lerman, P.M. Fitting Segmented Regression Models by Grid Search. *J. Appl. Stat.* **1980**, *29*, 77. [CrossRef]
47. Dangeti, P. *Statistics for Machine Learning: Techniques for Exploring Supervised, Unsupervised, and Reinforcement Learning Models with Python and R.*; Packt Publishing: Birmingham, UK, 2017; ISBN 9781788295758.
48. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* **2001**, *29*, 1189–1232. [CrossRef]
49. Zemel, R.S.; Pitassi, T. A gradient-based boosting algorithm for regression problems. *Adv. Neural Inf. Process. Syst.* **2001**, 696–702. [CrossRef]
50. Wei, Z.; Meng, Y.; Zhang, W.; Peng, J.; Meng, L. Downscaling SMAP Soil Moisture Estimation with Gradient Boosting Decision Tree Regression over the Tibetan Plateau. *Remote Sens. Environ.* **2019**, *225*, 30–44. [CrossRef]
51. 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; pp. 785–794.
52. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Wolff, E. Very High Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 607–611. [CrossRef]
53. Carslaw, D.C.; Ropkins, K. Openair—An R Package for Air Quality Data Analysis. *Environ. Model. Softw.* **2012**, *27–28*, 52–61. [CrossRef]
54. Cumming, G.; Calin-Jageman, R. *Introduction to the New Statistics: Estimation, Open Science, and Beyond*; Routledge: New York, NY, USA, 2016.
55. Taylor, K.E. Summarizing Multiple Aspects of Model Performance in a Single Diagram. *J. Geophys. Res.* **2001**, *106*, 7183–7192. [CrossRef]
56. Nyamangara, J.; Mudhara, M.; Giller, K.E. Effectiveness of cattle manure and nitrogen fertilizer application on the agronomic and economic performance of maize. *S. Afr. J. Plant Soil* **2005**, *22*, 59–63. [CrossRef]
57. Mansfield, E.R.; Helms, B.P. Detecting Multicollinearity. *Am. Stat.* **1982**, *36*, 158. [CrossRef]
58. Farrar, D.E.; Glauber, R.R. Multicollinearity in Regression Analysis: The Problem Revisited. *Rev. Econ. Stat.* **1967**, *49*, 92. [CrossRef]
59. Jaya, I.G.N.M.; Ruchjana, B.; Abdullah, A. Comparison of Different Bayesian And Machine Learning Methods in Handling Multicollinearity Problem: A Monte Carlo Simulation Study. *ARPN J. Eng. Appl. Sci.* **2020**, *15*, 1998–2011.
60. Farrell, A.; Wang, G.; Rush, S.A.; Martin, J.A.; Belant, J.L.; Butler, A.B.; Godwin, D. Machine Learning of Large-scale Spatial Distributions of Wild Turkeys with High-dimensional Environmental Data. *Ecol. Evol.* **2019**, *9*, 5938–5949. [CrossRef]
61. Jeong, G.; Oeverdieck, H.; Park, S.J.; Huwe, B.; Ließ, M. Spatial Soil Nutrients Prediction Using Three Supervised Learning Methods for Assessment of Land Potentials in Complex Terrain. *Catena* **2017**, *154*, 73–84. [CrossRef]
62. Sorenson, P.T.; Small, C.; Tappert, M.C.; Quideau, S.A.; Drozdowski, B.; Underwood, A.; Janz, A. Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy. *Can. J. Soil Sci.* **2017**, *97*, 241–248. [CrossRef]
63. Zhang, Y.; Sui, B.; Shen, H.; Ouyang, L. Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors. *Comput. Electron. Agric.* **2019**, *160*, 23–30. [CrossRef]
64. Deng, X.; Ma, W.; Ren, Z.; Zhang, M.; Grieneisen, M.L.; Chen, X.; Fei, X.; Qin, F.; Zhan, Y.; Lv, X. Spatial and Temporal Trends of Soil Total Nitrogen and C/N Ratio for Croplands of East China. *Geoderma* **2020**, *361*, 114035. [CrossRef]
65. López-Calderón, M.J.; Estrada-Ávalos, J.; Rodríguez-Moreno, V.M.; Mauricio-Ruvalcaba, J.E.; Martínez-Sifuentes, A.R.; Delgado-Ramírez, G.; Miguel-Valle, E. Estimation of Total Nitrogen Content in Forage Maize (Zea Mays L.) Using Spectral Indices: Analysis by Random Forest. *Agriculture* **2020**, *10*, 451. [CrossRef]
66. Li, Y.; Li, C.; Li, M.; Liu, Z. Influence of variable selection and forest type on forest aboveground biomass estimation using machine learning algorithms. *Forests* **2019**, *10*, 1073. [CrossRef]
67. Beguin, J.; Fuglstad, G.A.; Mansuy, N.; Paré, D. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. *Geoderma* **2017**, *306*, 195–205. [CrossRef]
68. Forkuor, G.; Hounkpatin, O.K.L.; Welp, G.; Thiel, M. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLoS ONE* **2017**, *12*, e0170478. [CrossRef]
69. Zhou, T.; Geng, Y.; Chen, J.; Sun, C.; Haase, D.; Lausch, A. Mapping of Soil Total Nitrogen Content in the Middle Reaches of the Heihe River Basin in China Using Multi-Source Remote Sensing-Derived Variables. *Remote Sens.* **2019**, *11*, 2934. [CrossRef]
70. Zhou, T.; Geng, Y.; Chen, J.; Pan, J.; Haase, D.; Lausch, A. High-Resolution Digital Mapping of Soil Organic Carbon and Soil Total Nitrogen Using DEM Derivatives, Sentinel-1 and Sentinel-2 Data Based on Machine Learning Algorithms. *Sci. Total Environ.* **2020**, *729*, 138244. [CrossRef] [PubMed]
71. Li, F.; Miao, Y.; Feng, G.; Yuan, F.; Yue, S.; Gao, X.; Liu, Y.; Liu, B.; Ustin, S.L.; Chen, X. Improving Estimation of Summer Maize Nitrogen Status with Red Edge-Based Spectral Vegetation Indices. *Field Crops Res.* **2014**, *157*, 111–123. [CrossRef]
72. Knoepp, J.; Swank, W. Using Soil Temperature and Moisture to Predict Forest Soil Nitrogen Mineralization. *Biol. Fertil. Soils* **2002**, *36*, 177–182. [CrossRef]

73. Baxter, S.J.; Oliver, M.A. The Spatial Prediction of Soil Mineral N and Potentially Available N Using Elevation. *Geoderma* **2005**, *128*, 325–339. [CrossRef]
74. SDG. *Sustainable Development Goals*; United Nations: New York, NY, USA, 2019.
75. Poffenbarger, H.J.; Sawyer, J.E.; Barker, D.W.; Olk, D.C.; Six, J.; Castellano, M.J. Legacy Effects of Long-Term Nitrogen Fertilizer Application on the Fate of Nitrogen Fertilizer Inputs in Continuous Maize. *Agric. Ecosyst. Environ.* **2018**, *265*, 544–555. [CrossRef]
76. FAO. *Save and Grow in Practice: Maize, Rice and Wheat, a Guide to Sustainable Cereal Production*; Food and Agriculture Organization: Rome, Italy, 2016.