AI

MDPI

Article

Leveraging Spectral Neighborhood Information for Corn Yield Prediction with Spatial-Lagged Machine Learning Modeling: Can Neighborhood Information Outperform Vegetation Indices?

Efrain Noa-Yarasca [1,*], Javier M. Osorio Leyton [1], Chad B. Hajda [2], Kabindra Adhikari [2] and Douglas R. Smith [2]

[1] Texas A&M AgriLife Research, Blackland Research and Extension Center, Temple, TX 76502, USA; javier.osorio@ag.tamu.edu
[2] Grassland Soil and Water Research Laboratory, United States Department of Agriculture–Agriculture Research Service, Temple, TX 76502, USA; chad.hajda@usda.gov (C.B.H.); kabindra.adhikari@usda.gov (K.A.); douglas.r.smith@usda.gov (D.R.S.)
* Correspondence: efrain.noa-yarasca@ag.tamu.edu

Abstract: Accurate and reliable crop yield prediction is essential for optimizing agricultural management, resource allocation, and decision-making, while also supporting farmers and stakeholders in adapting to climate change and increasing global demand. This study introduces an innovative approach to crop yield prediction by incorporating spatially lagged spectral data (SLSD) through the spatial-lagged machine learning (SLML) model, an enhanced version of the spatial lag X (SLX) model. The research aims to show that SLSD improves prediction compared to traditional vegetation index (VI)-based methods. Conducted on a 19-hectare cornfield at the ARS Grassland, Soil, and Water Research Laboratory during the 2023 growing season, this study used five-band multispectral image data and 8581 yield measurements ranging from 1.69 to 15.86 Mg/Ha. Four predictor sets were evaluated: Set 1 (spectral bands), Set 2 (spectral bands + neighborhood data), Set 3 (spectral bands + VIs), and Set 4 (spectral bands + top VIs + neighborhood data). These were evaluated using the SLX model and four decision-tree-based SLML models (RF, XGB, ET, GBR), with performance assessed using R^2 and RMSE. Results showed that incorporating spatial neighborhood data (Set 2) outperformed VI-based approaches (Set 3), emphasizing the importance of spatial context. SLML models, particularly XGB, RF, and ET, performed best with 4–8 neighbors, while excessive neighbors slightly reduced accuracy. In Set 3, VIs improved predictions, but a smaller subset (10–15 indices) was sufficient for optimal yield prediction. Set 4 showed slight gains over Sets 2 and 3, with XGB and RF achieving the highest R^2 values. Key predictors included spatially lagged spectral bands (e.g., Green_lag, NIR_lag, RedEdge_lag) and VIs (e.g., CREI, GCI, NCPI, ARI, CCCI), highlighting the value of integrating neighborhood data for improved corn yield prediction. This study underscores the importance of spatial context in corn yield prediction and lays the foundation for future research across diverse agricultural settings, focusing on optimizing neighborhood size, integrating spatial and spectral data, and refining spatial dependencies through localized search algorithms.

Keywords: corn yield prediction; spatial-lagged machine learning model; spectral neighborhood data; vegetation indices (VIs); spatial autocorrelation

check for updates

Academic Editors: Gianni D'Angelo and Arslan Munir

Received: 13 December 2024
Revised: 7 March 2025
Accepted: 10 March 2025
Published: 13 March 2025

Citation: Noa-Yarasca, E.; Osorio Leyton, J.M.; Hajda, C.B.; Adhikari, K.; Smith, D.R. Leveraging Spectral Neighborhood Information for Corn Yield Prediction with Spatial-Lagged Machine Learning Modeling: Can Neighborhood Information Outperform Vegetation Indices? *AI* **2025**, *6*, 58. https://doi.org/10.3390/ai6030058

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As the primary grain crop in the United States, corn production significantly impacts both local economies and global food supply chains [1]. Accurate yield estimation is therefore essential for efficient agricultural management, decision-making, resource allocation, crop insurance, and policy planning [1]. Moreover, reliable yield forecasts are critical for guiding decisions on harvesting, marketing, and resource management [2]. With growing pressures of climate change and rising demand, reliable forecasting models are becoming increasingly essential for guiding farmers and stakeholders in crop management and resource allocation [3]. Enhanced prediction capabilities can further optimize planting practices, improve pest and nutrient management, and increase agricultural productivity [3]. Thus, advancing yield prediction methods is not only a scientific challenge but a critical need to support the production agriculture sector in adapting to environmental changes.

The integration of machine learning (ML) into agricultural practices has transformed traditional yield prediction methodologies. Unlike classical statistical models, which often rely on linear assumptions, ML algorithms can capture complex, non-linear relationships within high-dimensional data through automated transformations and large datasets [4,5]. Studies using these algorithms have used extensive datasets, including agronomic variables, remote sensing spectral data (satellite and UAV imagery), and environmental factors, to model crop yield responses more effectively [6]. In this context, crop yield prediction with remote sensing data typically correlates spectral information from specific spatial coordinates with corresponding yield values.

Previous studies have explored various spectral band transformations into vegetation indices (VIs), ranging from simple to complex mathematical formulations [7,8]. However, this approach, which often relies on VIs derived from reflectance data, has limitations, as it primarily considers localized information and overlooks neighborhood spatial patterns that may influence yield outcomes [9,10]. In response, spatial regression analysis has increasingly highlighted the importance of spatial autocorrelation, where the dependent variable at (x, y) is influenced not only by direct predictors at that point in (x, y) but also by neighboring values ($x \pm i$, $y \pm i$). Here, spatial metrics like Moran's I index provide a quantitative measure of autocorrelation, enhancing our understanding of spatial dependencies among target variable values. This insight highlights the value of incorporating neighborhood information into predictive models, illustrating the critical role of spatial regression in capturing the interplay among spatial data points [11].

Spatial regression models have been widely used across various fields, including mixed forest production [12], urbanization's impact on air quality [13], and milk production [14]. They have also been applied to predict flood frequencies [15], assess agricultural impacts on property values [16], and analyze soil carbon stocks [17]. These diverse applications highlight the versatility of spatial regression in addressing complex, location-dependent phenomena. In agriculture, spatial regression analysis has notably enhanced crop yield predictions by integrating spatial data with agronomic variables. For instance, spatially structured data have improved model performance in assessing weather and climate impacts on yields [18], while accounting for spatial dependencies has yielded more accurate crop yield estimates [19]. Furthermore, spatially varying coefficients have enabled better predictions of temperature effects on yields [20]. Recent innovations include using UAV spectral reflectance to analyze the spatial structure of wheat grain yield and protein content [21], as well as applying Bayesian spatial regression models with satellite imagery to predict soybean yields and evaluate computational efficiency [22]. These studies highlight the growing importance of spatial regression in agricultural modeling.

The typical spatial regression approaches include the spatial autoregressive (SAR), spatial error (SEM), and spatial lag of X (SLX) models, each treating spatial interactions

differently [23,24]. The SAR model captures spatial dependence in the dependent variable, assuming that outcomes in one location are influenced by outcomes in neighboring locations. The SEM, on the other hand, accounts for spatial dependence in the error terms, attributing spatial autocorrelation to unobserved factors that vary spatially. The SLX model includes spatially lagged predictors, allowing evaluation of how neighboring explanatory variables influence the dependent variable. These models are often paired with spatial weight matrices to define spatial relationships. While SAR and SEM models are effective, they are less practical for predictive tasks where the dependent variable is unknown, such as in predictions involving new or future data. In contrast, the SLX model, which does not require spatially lagged dependent variables, is better suited for prediction, making it ideal for forecasting at new locations [24]. By explicitly incorporating spatial dependencies, these models improve prediction accuracy and offer valuable insights across fields like agriculture, environmental science, and urban planning.

Despite the growing interest in incorporating spatial information into agricultural models, there remains a significant gap in research focused on the integration of neighboring spectral data within machine learning frameworks for crop yield prediction. Existing studies predominantly concentrate on optimizing VIs [10,25], with little attention given to spatially integrated methods that leverage the relationships between neighboring spectral values. This study aims to address this gap by introducing a novel approach that integrates spatial dependencies directly into machine learning models. Specifically, we propose the spatially lagged machine learning (SLML) approach, which combines the power of spatially lagged predictors with decision-tree-based algorithms to enhance the predictive accuracy of corn yield models.

The SLML approach involves incorporating spatially lagged data—specifically, the spectral values at neighboring coordinates—into machine learning models such as random forest (RF), extreme gradient boosting (XGB), extra trees regressor (ET), and gradient boosting regression (GBR). These decision-tree-based ensemble models were chosen for their ability to capture complex, non-linear relationships within high-dimensional data and their strong performance with moderate-sized datasets, which are common in agricultural studies [4,26]. In addition, decision trees are highly interpretable, offering direct insights into feature importance, which is crucial for evaluating how spatially lagged spectral data improves yield predictions. This interpretability is a key advantage over more complex models, such as neural networks or support vector machines, which require additional techniques for interpretation and often demand large datasets—typically in the tens of thousands to millions of data points—to perform effectively due to their complex architectures [27–29].

The spatial lag X (SLX) model and the proposed SLML model both incorporate spatial dependencies, but they differ significantly in their ability to capture complex spatial patterns. The SLX model, a linear approach, explicitly includes spatially lagged independent variables but remains limited by its linear assumptions, which may overlook intricate spatial interactions common in agricultural systems. In contrast, SLML models use machine learning algorithms to capture non-linear relationships and model complex spatial dependencies that the linear SLX model cannot detect. This non-linearity allows SLML models to identify and process spatial patterns more effectively, which is especially crucial in agricultural settings, where spatial interactions can be highly complex and variable. The flexibility of SLML models in handling these complex spatial relationships makes them particularly valuable for applications like crop yield prediction, where the spatial structure of the data plays a critical role in model accuracy.

By including neighborhood spectral data, this approach aims to capture broader spatial interactions that are often overlooked in traditional models relying solely on point-specific

data. The research hypothesizes that these extended spatial relationships will improve predictive performance by better accounting for the spatial structure and autocorrelation inherent in agricultural landscapes. To assess the effectiveness of the SLML approach, this study compares its performance against conventional machine learning models that do not include spatially lagged predictors, as well as models including raw spectral data + VIs, and SLX models. By evaluating these approaches, this study seeks to determine whether incorporating spatial neighborhood information significantly outperforms standard modeling techniques in corn yield prediction. Ultimately, this research aims to contribute to the field of agricultural prediction by demonstrating the potential benefits of integrating spatially lagged data within machine learning frameworks, offering new avenues for more accurate and reliable yield forecasts.

2. Methodology

2.1. Study Area

The research was conducted at the Texas A&M AgriLife Blackland Research and Extension Center in Temple, Bell County, Texas (Figure 1). The study site encompassed a 19.03 ha rainfed field, collaboratively managed by Texas A&M AgriLife's Blackland Research and Extension Center and the USDA-ARS Grassland, Soil, and Water Research Laboratory in Temple, Texas (31.059444° N, 97.345833° W; 192 m elevation). Corn was sown on 28 February 2023 and fertilized with 448 kg/ha of a 32.5N–16.2P–0K blend, which was broadcasted prior to planting. No-till practices were employed, with a planting density of 4.8 plants/m^2. The corn was harvested on 14 August 2023. The topography features a relatively flat landscape, with an 8 m difference in elevation between its highest and lowest points. The predominant soil types are Houston Black clay (fine, thermic, smectitic, Udic Haplusterts) and Austin clay (fine-silty, thermic, carbonatic, Udorthentic Haplustolls), both characteristic of the Blackland Prairies region [30]. These soils are known for their high fertility and strong moisture retention capacity, though they are also prone to compaction. During the growing season, the field received a total of 389.4 mm of precipitation.

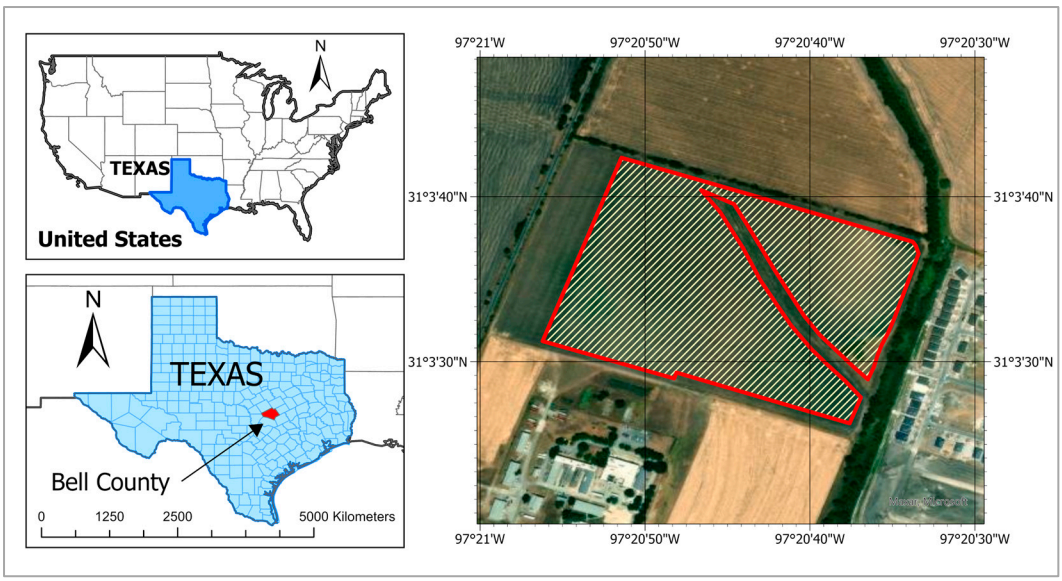


Figure 1. Study area location.

Harvesting was conducted using a John Deere 9510 combine (Deere and Company, Moline, IL, USA) fitted with an AgLeader corn yield monitor (AgLeader Technology Inc., Ames, IA, USA) to record yield data. The yield-monitor recorded data for six-row harvests (16.9 m^2 per section) and logged GPS coordinates for the centroid of each harvested area

using an integrated GPS recorder. To ensure measurement accuracy, pre-harvest and in-field calibrations were conducted for load size, weighing, and GPS data. The raw yield data were then processed using Yield Editor software Version 2.0.7 to correct for factors such as speed and pass delays, overlaps, moisture variability, and outliers [31]. The refined yield data, paired with their corresponding GPS coordinates, were used as ground truth for model training and validation.

2.2. Data Collection and Processing

The corn imagery for this study was collected on 22 May, corresponding to Day 83 and the tasseling/silking (VT/R1) stage. A WingtraOne GenII fixed-wing UAV (Wingtra, Zurich, Switzerland) equipped with a MicaSense RedEdge-P multispectral sensor (AgEagle Aerial Systems, Seattle, WA, USA) was used for image acquisition. The sensor captured five spectral bands: red (R), green (G), blue (B), red-edge (RE), and near-infrared (NIR) (Figure 2a), each with a resolution of 1456 × 1088 pixels (1.6 MP), enabling detailed monitoring of crop health. The UAV flew at an altitude of 60 m with 75% overlap for both forward and sideways directions, ensuring high-quality data for analysis. The flight was conducted autonomously using pre-defined mission routes, with the flight operations being controlled through the WingtraHub 1.0 software (Wingtra, Zurich, Switzerland).

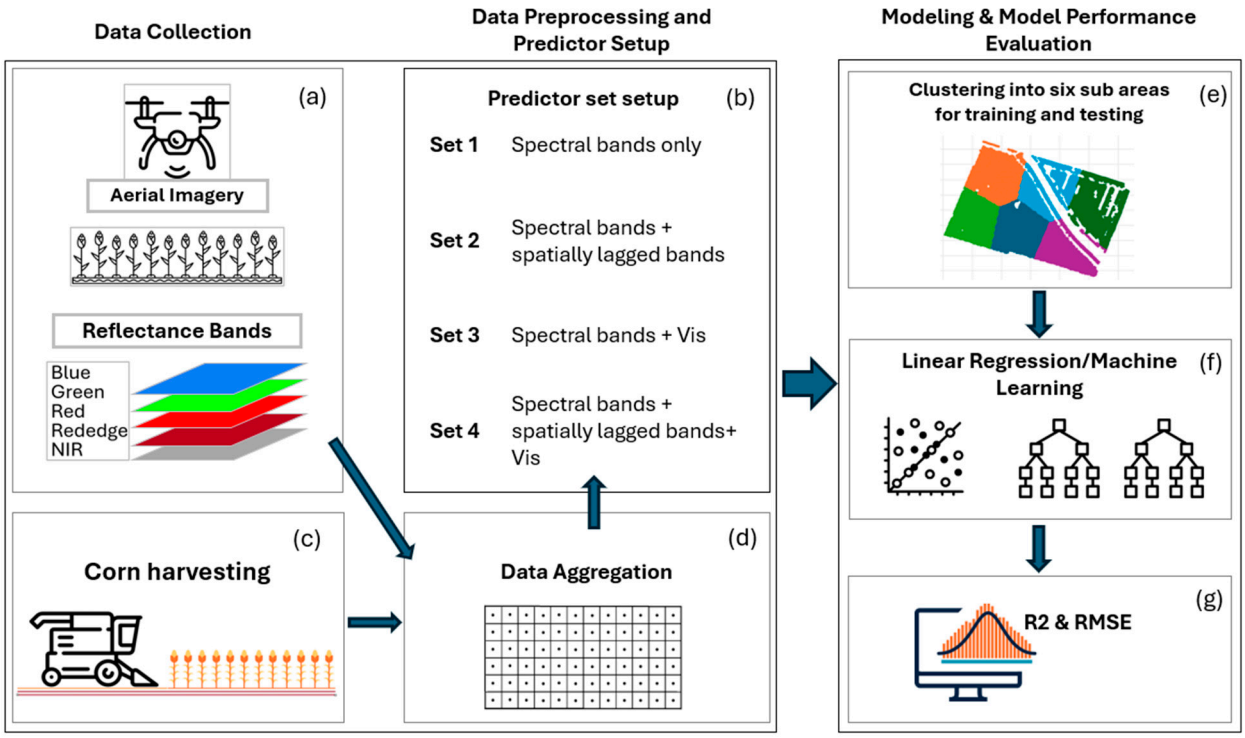


Figure 2. Graphical abstract of the methodology: (**a**) Crop imagery collection, (**b**) Predictor set setup, (**c**) Corn yield data collection, (**d**) Spectral and corn yield data aggregation, (**e**) Spatial clustering of data into six subareas for training and testing, (**f**) Modeling, and (**g**) Model performance evaluation.

For accurate georeferencing, the UAV utilized its built-in GPS system for navigation, nadir image acquisition, and recording coordinates of individual images. The images were processed and orthomosaiced using Pix4Dmapper v4.8.4 (Prilly, Switzerland), resulting in multispectral rasters with 6 cm resolution for each spectral band.

2.2.1. Corn Yield Data Collection Processing

Corn yield data were collected using a John Deere 9510 harvester equipped with an AgLeader sensor (Figure 2c), aggregating yield values across six rows and a length

of approximately 3.6 m, covering an area of about 21.6 square meters per measurement. These yield values were mapped in ArcGIS, where influence polygons were generated for each measurement. Irregular polygons and those at the field edges were excluded to improve data accuracy, resulting in 8581 yield values, each precisely linked to its corresponding influence polygon as a shapefile (Figure 2d). Corn yield varied between 1.69 and 15.86 Mg/ha (Figure 3), with a mean yield of 10.19 Mg/ha, which exceeded the Texas state average of 6.62 Mg/ha for the 2022/2023 growing season [32,33]. The interquartile range (Q1: 9.46, Q3: 10.95) was 1.49 Mg/Ha, covering 50% of the values. A coefficient of variation of 12.11% indicates consistent yields with moderate variability, enabling reliable estimation.

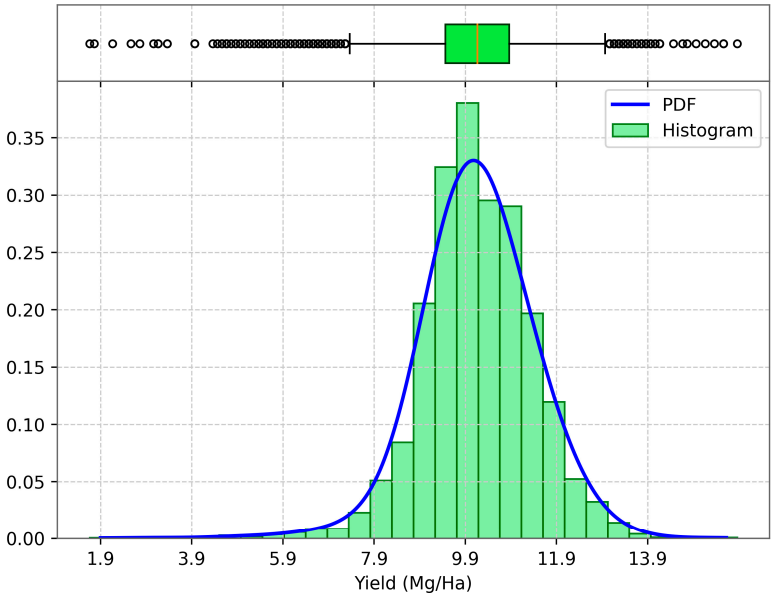


Figure 3. Frequency distribution of corn yield.

The box plot displays Q1 (25th percentile), the median (Q2), and Q3 (75th percentile), illustrating data distribution. Whiskers extend beyond the interquartile range (IQR = Q3 − Q1), with lower and upper limits set at Q1 − 1.5 × IQR and Q3 + 1.5 × IQR, respectively. Data points beyond these limits were not excluded or classified as outliers, as variability in agricultural yield often reflects real environmental factors such as soil heterogeneity, water availability (particularly in rain-fed systems), and topography. Removing such values could lead to information loss and bias. Potential biases from events like pest outbreaks were not explicitly identified or processed, as the yield data showed no significant clustering or unusual patterns. Additionally, the machine learning models used—RF, XGB, ET, and GBR—are robust to extreme values due to their non-parametric and ensemble structures, mitigating the influence of outliers more effectively than linear regression. Data normalization was not required, as decision-tree-based ML models split data based on feature thresholds rather than computing distances or gradients, making them inherently robust to varying data magnitudes.

2.2.2. Imagery Processing

To correlate UAV-derived image data with corn yield data, multispectral images were segmented and aggregated using the corn yield polygon shapefile, producing 8581 polygons with spectral band information. Pixels representing bare ground were excluded, retaining only those corresponding to corn vegetation. Using the zonal statistics function in ArcGIS, mean reflectance values for each band (B, G, R, RE, and NIR) were calculated within each polygon. This process generated a dataset of average reflectance values aligned with corresponding corn yield values, enabling a consistent analysis of the

relationship between spectral reflectance and crop yield. To ensure representativeness, the Shapiro–Wilk test showed that over 98% of the spectral values in the 8581 polygons had p-values > 0.05, indicating normality. This aligns with expectations, as only vegetation pixels were included.

2.3. Spatial Autocorrelation Evaluation

Spatial autocorrelation of corn yield data was assessed using Moran's I, a widely used spatial statistic for evaluating autocorrelation [34]. This method involved cross-product statistics between the target variable and its spatial lag, with the variable expressed as deviations from its mean ($z_i = x_i - \bar{x}$). Moran's I statistic is calculated as:

$$I = \frac{n}{S_o} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

where w_{ij} represents spatial weights in a matrix computed with $w_{ij} = 1/d_{ij}^2$, n is the total features, and $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ is the sum of all weights. Global Moran's index values range between -1 and 1, where a value of 1 signifies a high degree of positive spatial autocorrelation, -1 indicates strong negative autocorrelation, and 0 reflects a random spatial distribution. Perfect autocorrelation is theoretically unattainable, but an index above 0.3 or below -0.3 is considered strong autocorrelation [35].

The spatial autocorrelation of corn yield was computed for four neighborhood levels to evaluate the trend with increasing neighbor inclusion. These levels involved the immediate surrounding 8 cells and progressively larger neighborhoods encompassing 24, 48, and 80 cells around each target cell. Figure 4 illustrates this structure, with the target cell at the center and neighborhood levels expanding outward. The first level comprises the eight adjacent cells, while subsequent levels incorporate additional surrounding layers, reaching up to 80 neighboring cells in the fourth level.

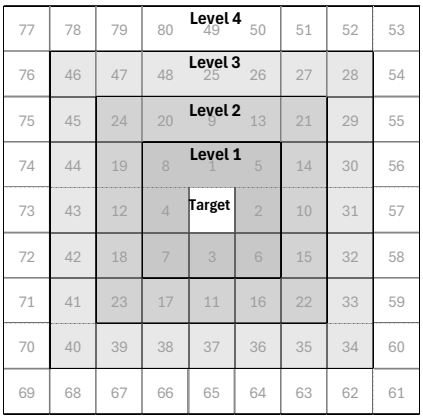


Figure 4. Neighborhood levels for spatial autocorrelation analysis, showing the target cell and expansions to 8, 24, 48, and 80 neighboring cells, each shaded differently for clarity.

2.4. Setting up Predictor Sets for Modeling

The predictors were grouped into four sets, some with subsets (Table 1). Set 1 consisted of the five spectral bands (R, G, B, RE, NIR) as the baseline. Set 2 expanded on Set 1 by adding neighboring information in four variations: Set 2A (4 neighbors), Set 2B (8 neighbors), Set 2C (20 neighbors), and Set 2D (24 neighbors). Figure 5 shows how neighborhood data were considered around the dependent or target corn yield value. Set 3 included the five spectral bands plus eight variations of the most commonly used vegetation indices (VIs) (Table 2). VIs were grouped into subsets through a pre-computed process (pre-process) using a backward stepwise approach. The backward stepwise approach is

a method used in regression analysis to improve model performance by systematically removing the least significant predictors. The process begins with a full model and then iteratively eliminates the least important variables based on feature importance criteria. After each removal, the model is re-fitted, and the process is repeated until only the most significant predictors remain.

Table 1. Overview of predictor sets and subsets used in corn yield regression modeling.

Set	Description	Sub-Set	Details
Set 1 (S-1)	Spectral bands only (Baseline)	S-1	Blue (B), green (G), red (R), red-edge (RE), near infrared (NIR)
Set 2 (S-2)	Spectral bands + spatially lagged bands	S-2A	S-1 + 4 neighbors
		S-2B	S-1 + 8 neighbors
		S-2C	S-1 + 20 neighbors
		S-2D	S-1 + 24 neighbors
Set 3 (S-3)	Spectral bands + vegetation indices	S-3A	S-1 + CREI
		S-3B	S-3A + GCI
		S-3C	S-3B + NPCI
		S-3D	S-3C + ARI, CCCI
		S-3E	S-3D + EVI, MCARI, MCCI, NDRE, NG
		S-3F	S-3E + BGI, NGRDI, PPR, PSRI, TVI
		S-3G	S-3F + GNDVI, MTVI2, NDVI, B-NDVI, TCI
		S-3H	S-3G + MSAVI, RDVI, SAVI, TrVI, TSAVI
Set 4 (S-4)	Spectral bands + spatially lagged bands + VIs	S-4	S-2B + 20 VIs

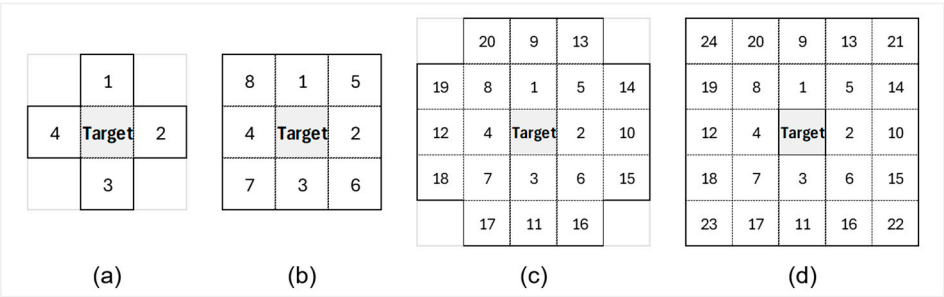


Figure 5. Spatial neighborhood configurations in Set 2 predictors for assessing lagged spectral band values in corn yield regression modeling: (**a**) four neighbors, (**b**) eight neighbors, (**c**) twenty neighbors, and (**d**) twenty-four neighbors.

In this study, the process began with a full set of 25 VIs, and models were initially run with all variables. The five least important VIs (MSAVI, RDVI, SAVI, TrVI, TSAVI) were removed based on feature importance, leaving 20 VIs. This process was repeated, with the five least important VIs removed at each iteration. In the last three iterations, fewer than five predictors were removed to better control model performance, as previous studies have shown that prediction accuracy increases rapidly with the addition of initial variables but diminishes asymptotically as more predictors are added [36]. Throughout the process, the most important VIs for the SLML regression models were retained. For example, in Set 3, the most important VI was CREI (S-3A), followed by GCI (S-3B). Set 4 combined the five spectral bands, the 20 optimal VIs, and data from eight neighboring cells (Figure 2b).

Table 2. Vegetation indices included in the subsets of Set 3 predictors for corn yield regression modeling.

Vegetation Index	Name	Equation	Ref.
CREI	Chlorophyll red-edge index	$NIR/RE - 1$	[37]
GCI	Green chlorophyll index	$NIR/G - 1$	[37]
NPCI	Normalized pigment chlorophyll index	$(R - B)/(R + B)$	[38]
ARI	Anthocyanin reflectance index	$1/G - 1/R$	[39]
CCCI	Canopy chlorophyll content index	$\left(\frac{NIR-RE}{NIR+RE}\right) \Big/ \left(\frac{NIR-R}{NIR+R}\right)$	[40]
EVI	Enhanced vegetation index	$\frac{2.5(NIR-R)}{NIR+6R-7.5B+1}$	[41]
MCARI	Modified chlorophyll absorption in reflectance index (red)	$((RE - R) - 0.2(RE - G))\left(\frac{RE}{R}\right)$	[42]
MCCI	Modified chlorophyll content index	$\left(\frac{RE-R}{RE+R}\right) - \left(\frac{RE-G}{RE+G}\right)$	
NDRE	Normalized difference red-edge index	$(NIR - RE)/(NIR + RE)$	[40]
NG	Normalized green index	$G/(NIR + R + G)$	[43]
BGI	Blue green pigment index	B/G	[44]
NGRDI	Normalized green red difference index	$\frac{G-R}{G+R}$	[45]
PPR	Plant pigment ratio	$(G - B)/(G + B)$	[46]
PSRI	Plant senescence reflectance index	$(R - B)/RE$	[47]
TVI	Triangular vegetation index	$0.5(120(NIR - G) - 200(R - G))$	[48]
GNDVI	Green normalized difference vegetation index	$\frac{NIR-G}{NIR+G}$	[49]
MTVI2	Modified triangular vegetation index (TVI) 2	$\frac{1.5[1.2(NIR-G)-2.5(R-G)]}{\sqrt{(2NIR+1)^2-(6NIR-5\sqrt{R})-0.5}}$	[50]
NDVI	Normalized difference vegetation index	$(NIR - R)/(NIR + R)$	[51]
B-NDVI	Blue normalized difference vegetation index	$\frac{NIR-B}{NIR+B}$	[52]
TCI	Triangular chlorophyl index	$1.2(RE - G) - 1.5(R - G)\sqrt{RE/R}$	[50]
MSAVI	Modified soil-adjusted vegetation index	$\frac{2NIR+1-\sqrt{(2NIR+1)^2-8(NIR-R)}}{2}$	[53]
RDVI	Renormalized difference vegetation index	$(NIR - R)/\sqrt{NIR + R}$	[54]
SAVI	Soil-adjusted vegetation index	$(1 + 0.5)\left(\frac{NIR-R}{NIR+R+0.5}\right)$	[41]
TrVI	Transformed vegetation index	$\sqrt{\frac{NIR-R}{NIR+R} + 0.5}$	[51]
TSAVI	Transformed soil-adjusted vegetation index	$\frac{0.5(NIR-0.5R-0.5)}{0.5NIR+R+0.75}$	[55]

2.5. Spatial Regression Modeling

Spatially lagged models are advanced extensions of linear regression designed to account for spatial dependencies and interactions among observations in geographical data [23,24]. These models incorporate spatial lag terms to capture the influence of neighboring locations on the dependent variable, improving both the accuracy and interpretability of the analysis. Spatial regression models, such as spatial autoregressive (SAR), spatial error (SEM), and spatial lag of X (SLX), handle spatial interactions differently: SAR captures spatial dependence in the dependent variable, SEM addresses spatial dependence in the error terms, and SLX includes spatially lagged predictors. While SAR and SEM are

useful, they are less practical for predicting the dependent variable when it is unknown. In contrast, the SLX model, which does not require spatially lagged dependent variables, is better suited for forecasting.

2.5.1. Spatial Lag X Model (SLX)

The SLX model incorporates spatially lagged independent variables to examine the impact of neighboring observations on each observation's outcome [24]. The model is defined as:

$$y_i = X_i\beta + \sum_{j=1}^{n} W_{ij}X_j\theta + \epsilon_i$$

where y_i is the dependent variable at location i, X_i represents the value of the independent variables at location i, W_{ij} represents the spatial weight between locations i and j, X_j is the value of the independent variable at neighboring location j, β is the vector of coefficients ($n \times 1$), while θ is the coefficient vector of size $k \times 1$ associated with the spatially lagged independent variables. The error term ϵ is a vector of size $n \times 1$, assumed to follow independent and identically distributed (i.i.d.) errors with a mean of zero and constant variance, n is the number of observations (or spatial units) in the dataset, and k is the number of independent variables in the model. W_{ij} determines which observations actually contribute to the sum. If $W_{ij} = 0$ for a give j, the observation j does not influence i (i.e., it is not a "neighbor"). In practice, for immediate or fixed distance, W_{ij} would be nonzero only for those locations within the specified neighborhood or distance.

2.5.2. Machine Learning Models

Random Forest

Random forest is an ensemble learning technique that generates a series of decision trees to improve the accuracy of predictions. Each tree uses a bootstrap sample D_t from the original dataset D (where $D_t \subseteq D$ and $|D_t| = |D|$). At each split, a random subset of features Θ is considered, and the best split θ^* is selected to maximize the split criterion G [56]:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}} G(D_t, \theta)$$

For regression analysis, G measures the reduction in variance:

$$G(D_t, \theta) = \mathrm{Var}(D_t) - \left(\frac{|D_{t,L}|}{|D_t|}\mathrm{Var}(D_{t,L}) + \frac{|D_{t,R}|}{|D_t|}\mathrm{Var}(D_{t,R}) \right)$$

where $D_{t,L}$ and $D_{t,R}$ are the subsets after the split. The final prediction is the average of all trees: $\hat{y}_{RF} = \frac{1}{T}\sum_{t=1}^{T} f_t^{RF}(x)$, where T is the total number of trees, and $f_t^{RF}(x)$ is the prediction from the t^{th} tree. This ensemble approach reduces overfitting and enhances generalization by averaging the outputs of multiple independently trained trees.

Extreme Gradient Boosting (XGB)

XGB constructs an ensemble of decision trees in a sequential manner, with each subsequent tree aiming to correct the errors made by the previous ones by optimizing a regularized objective function [57]:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

where l is the loss function (e.g., mean squared error), y_i is the true label, $\hat{y}_i^{(t-1)}$ is the prediction at iteration $t-1$, and $f_t(x_i)$ is the new tree's prediction. The regularization term $\Omega(f_t)$ penalizes the model complexity to avoid overfitting and is defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where T is the number of leaves in the tree, w_j are the leaf weights, γ is the penalty for adding a new leaf, and λ is the regularization parameter. The prediction is updated as: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$. XGB's efficient handling of missing data, parallelized tree construction, and robust regularization make it highly effective for preventing overfitting.

Extremely Randomized Tree Regression (ET)

ET constructs multiple decision trees by introducing randomness in both data sampling and feature splits to enhance model generalization and reduce overfitting [58]. Unlike traditional decision trees, ET uses the entire dataset D for each tree and split nodes using random thresholds. Specifically, at each split, a random subset of features Θ_{random} is considered, and a random split point θ_{random} is chosen from each feature's range ($\theta_{random}\Theta_{random}$). The split criterion G evaluates the chosen split $G(D, \theta_{random})$. For regression analysis, G measures the reduction in variance:

$$G(D, \theta_{random}) = Var(D) - \left(\frac{|D_L|}{|D|} Var(D_L) + \frac{|D_R|}{|D|} Var(D_R) \right)$$

where D_L and D_R are the subsets after the split. The final prediction is the average of all tree predictions: $\hat{y}_{ET} = \frac{1}{T}\sum_{t=1}^{T} f_t^{ET}(x)$, where T is the total number of trees, and $f_t^{ET}(x)$ is the prediction from the t^{th} tree. By averaging the predictions from multiple randomized trees, this approach achieves lower variance and improved generalization compared to individual decision trees.

Gradient Boosting Regressor (GBR)

GBR combines multiple weak decision-tree learners to produce a strong predictive model. It operates iteratively, where each new tree corrects the residual errors of the previous ones by fitting to the negative gradient of the loss function [59,60]. The procedure starts with an initial estimation, typically represented by the mean value of the target variable, which is expressed as

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \gamma)$$

where $L(y_i, \gamma)$ is the loss function (e.g., squared error), and y_i represents the observed values (true value). At each iteration m, a decision tree $h_m(x)$ is trained on the residuals ($r_i^{(m-1)} = y_i - F_{m-1}(x_i)$), and the model is updated as

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

where η is the learning rate controlling the step size. After M iterations, the final prediction is $\hat{y} = F_M(x)$. By progressively reducing errors, GBR achieves high predictive accuracy, with regularization techniques such as limiting tree depth and applying shrinkage (η) to prevent overfitting [61].

2.6. Model Training and Testing

For model training and testing, the corn field area was partitioned into six subareas. Five subareas were used for training while the remaining subarea served as the test data. Like the leave-one-out cross-validation approach, each subarea was designated as the testing set at least once. Therefore, six combinations of training/testing sets were conducted.

2.6.1. Computation Tools

All computations, including the calculation of Moran's I and the development of SLX and SLML models, were performed using Python (version 3.8). The Python environment facilitated comprehensive data analysis and modeling through libraries such as NumPy, pandas, and scikit-learn [62–65]. The code for modeling and calculations is available in the repository at https://github.com/noayarae/Spatial_Lagged_ML, accessed on 11 December 2024.

The models were trained and validated on a Lenovo workstation running Windows 11 Enterprise (23H2, build 22631.4751), equipped with an AMD Ryzen Threadripper PRO 5975WX processor (32 cores, 64 threads, 3.6 GHz), 128GB RAM, and an NVIDIA GeForce RTX 4090 GPU (24GB VRAM, CUDA 12.0, Driver Version 528.24).

2.6.2. Hyperparameter Tunning

To ensure the robustness and reliability of the machine learning models, hyperparameters were optimized using a 6-fold cross-validation approach combined with grid search. In this process, the dataset was partitioned into six subsets, with each subset serving as a validation set while the remaining five were used for training. This iterative procedure was repeated six times, allowing every data point to be used for both training and validation, thereby reducing the risk of overfitting and ensuring that the models' performance was not overly sensitive to specific data divisions. By systematically evaluating all combinations of specified hyperparameter values, this approach identified the optimal configurations that maximized predictive performance while enhancing model generalization. The final hyperparameter settings derived from this process are detailed in Table 3.

Table 3. Tuned hyperparameters, ranges, and values for machine learning models in corn yield regression.

Model	Hyperparameter	Range or List of Values	Pace Step	Tunned Value
RF	n_estimators	100–900	100	500
	max_depth	1–40	5	11
	max_features	1–20	5	11
XGB	n_estimators	[100, 200, 300, 400]		200
	max_depth	1–12	3	7
	learning_rate	0–2	0.05	0.05
	subsample	0.4–0.8	0.1	0.7
	gamma	0.1–0.4	0.1	0.3
	colsample_bytree	[0.7, 0.8, 0.9]		0.9
ET	n_estimators	100–900	100	300
	max_depth	1–40	5	21
	max_features	1–21	4	21
DT	n_estimators	100–900	100	300
	max_depth	1–15	3	12
	max_features	1–21	4	21
GBR	n_estimators	100–900	100	200
	max_depth	1–13	4	5
	learning_rate	0–2	0.05	0.1
	subsample	0.4–0.8	0.1	0.7

2.7. Model Performance Evaluation

Model performance is assessed using two key metrics: root mean square error (RMSE) and coefficient of determination (R^2) (Figure 2g). RMSE indicates the average magnitude of prediction errors, reflecting model accuracy.

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

where y_i are the observed values, $\hat{y}_i$ are the predicted values, and n is the number of observations. A lower RMSE indicates better model performance.

The R^2 value represents the proportion of variation in the dependent variable that can be explained by the independent variables:

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

where $\bar{y}$ is the mean of the observed values. An R^2 value closer to 1 indicates a better fit.

3. Results

3.1. Spatial Autocorrelation

The corn yield autocorrelation results show significant positive spatial autocorrelation, with Moran's I values decreasing as the neighborhood size increases (Figure 6). For the immediate neighborhood (8 neighbors), Moran's I was 0.48, dropping to 0.34 at 24 neighbors, 0.31 at 48 neighbors, and 0.28 at 80 neighbors. These results indicate that the predicted corn yield values are strongly correlated within local areas. As the neighborhood radius expands, the spatial dependency diminishes, suggesting that the influence of spatial autocorrelation is most pronounced within the closest cells and gradually reduces over larger distances. The observed spatial autocorrelation supports the consideration of spatial lag effects in predictive modeling, as high Moran's I values in close neighborhoods highlight the potential benefit of integrating nearby yield data to enhance model accuracy.

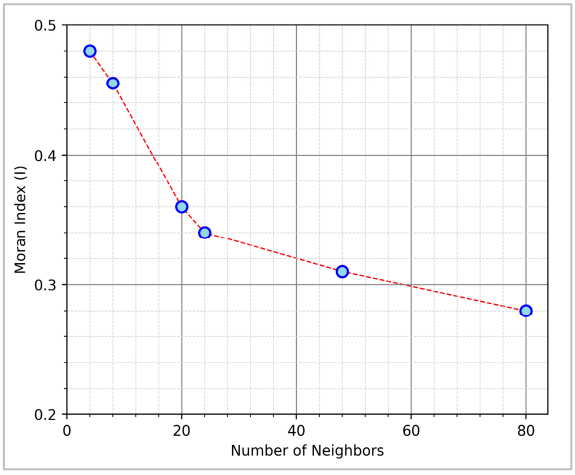


Figure 6. Variation in corn yield autocorrelation (Moran's I) with the number of neighbors, showing a significant drop after 8 neighbors. Blue dots indicate Moran's I, with a dashed trend line.

3.2. Model Performance With and Without Neighbor Information

Figure 7 summarizes the performance of SLX and SLML models (R^2 and RMSE) for predictor sets S-1 (baseline) and S-2 (with neighbor information) (The markers and lines are color-modelled as follows: blue for LR, green for RF, red for XGB, cyan for ET,

and magenta for GBR). Incorporating neighborhood data consistently improved model performance compared to the baseline. For example, SLX showed a notable R^2 increase from 0.19 (S-1) to 0.48 with 20 neighbors (S-2C), demonstrating the utility of spatial data in enhancing even simpler models. However, its performance plateaued with the inclusion of additional neighbors.

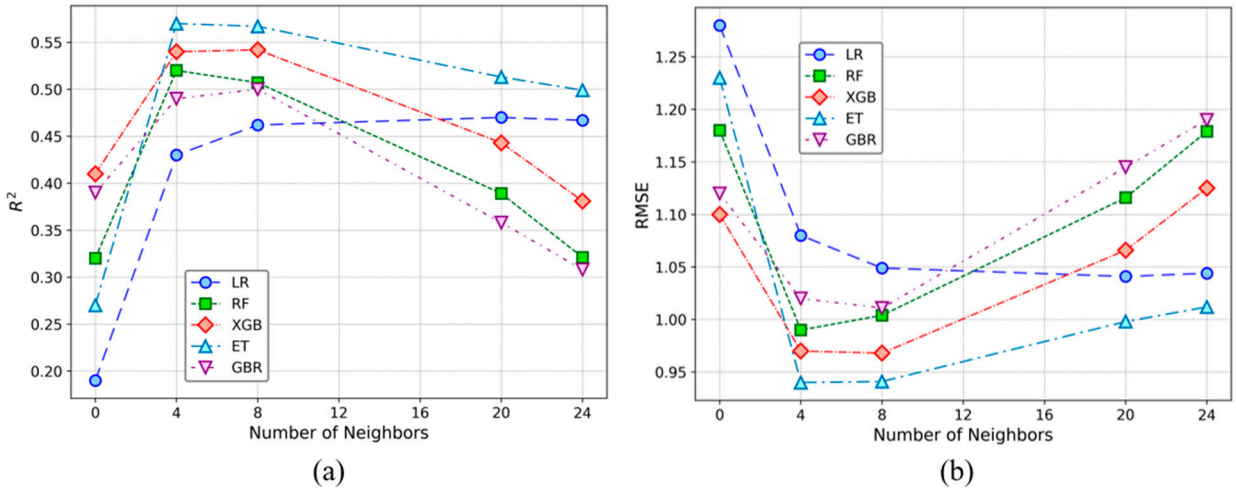


Figure 7. Performance metrics of regression models for corn yield prediction: (**a**) R^2 and (**b**) RMSE for Set S-1 (baseline) and Set S-2 (incorporating neighborhood information).

Among SLML models, ET achieved the highest R^2 (0.57) with four and eight neighbors, while XGB reached 0.54 with the same neighbor configurations. Random forest (RF) and gradient boosting regression (GBR) performed best with fewer neighbors (four or eight) but showed declining accuracy with larger neighbor sets (S-2C, S-2D), likely due to overfitting or noise. These results highlight the importance of balancing spatial information to optimize model performance, as excessive inclusion of neighbors can diminish accuracy in some models.

The RMSE values, in Figure 7b, further confirm the benefits of incorporating spatially lagged neighbors into predictive modeling. RMSE decreased significantly across all models when moving from the baseline set (S-1) to spatially enhanced sets (S-2). ET had the lowest RMSE (0.94) with four and eight neighbors, while XGB performed best at 0.97 RMSE for the same neighbor counts. SLX showed a steady RMSE decline, reaching 1.03 with 20 neighbors. RF and GBR showed more variable responses, with RMSE improving at fewer neighbors but increasing with excessive ones. Overall, the RMSE results reinforce the R^2 findings, showcasing the value of spatially lagged data in improving corn yield predictions.

Among the spatial-lagged machine learning (SLML) models incorporating neighborhood information, ET achieved the highest predictive performance, followed by XGB, RF, and GBR. This ranking remained consistent across different neighborhood sizes, suggesting that spatially lagged features introduce complexity that models handle differently. ET's randomized split mechanism likely enhances its ability to capture spatial heterogeneity, while XGB benefits more from structured, engineered features such as vegetation indices.

3.3. Model Performance With and Without Vegetation Indices

Figure 8 summarizes the R^2 and RMSE values for SLX and SLML models using predictor sets S-1 (baseline) and S-3 (including VIs) (The markers and lines are color-modelled as follows: blue for LR, green for RF, red for XGB, cyan for ET, and magenta for GBR). These values markedly show the role of VIs in improving model performance. The incremental inclusion of VIs consistently enhanced the coefficient of determination (R^2)

across all SLML models and SLX, with varying levels of improvement depending on the modeling approach. The five top VIs that showed the best performance in the regression models were CREI, GCI, NCPI, ARI, and CCCI. As mentioned above, VIs were selected using a backward stepwise approach.

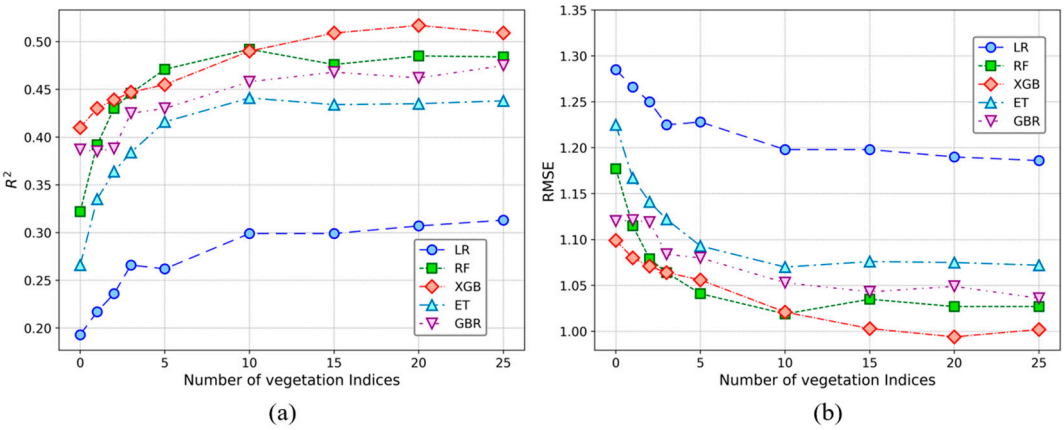


Figure 8. Performance metrics of regression models for corn yield prediction: (**a**) R^2 and (**b**) RMSE for Set S-1 (baseline) and Set S-3 (incorporating vegetation indices).

Among the SLML models, XGB exhibited the highest R^2 of 0.52 when using 20 VIs, demonstrating a substantial improvement over its baseline (R^2 = 0.41). Similarly, RF achieved its peak performance (R^2 = 0.49) with 10 VIs, a marked enhancement compared to its baseline of 0.32. GBR reached a similar R^2 of 0.48 when all 25 VIs were included but showed no meaningful improvements as fewer indices were removed, indicating a lower sensitivity to feature reduction. SLX demonstrated limited predictive power overall, but its R^2 improved steadily from 0.19 at baseline to a maximum of 0.31 with 25 VIs. In contrast, ET showed moderate performance gains, with R^2 plateauing at 0.44 when 10–25 vegetation indices were included.

After a certain number of VIs were included, improvements in R^2 became marginal across all models. For example, XGB and RF showed negligible changes in R^2 beyond 10–20 VIs, with ET and GBR exhibiting similar trends. This plateauing effect highlights the reduced contribution of additional VIs to predictive performance. These results suggest that while VIs provide valuable spectral information, an optimal subset (e.g., 10–15 indices) is sufficient to capture most of the relevant variability in corn yield. Including more indices beyond this threshold adds unnecessary complexity and computational cost without meaningful performance gains. Thus, optimizing the number of indices is essential to balance model complexity and performance.

The RMSE results in Figure 8b corroborate the trends observed in the R^2 values, highlighting the impact of vegetation indices (VIs) on model performance. Across all regression models, including VIs (S-3) consistently reduces RMSE compared to baseline models (Set S-1), which use only spectral bands (B, G, R, RE, NIR). For example, the XGB model shows a notable RMSE reduction from 1.10 in S-1 to 1.00 in S-3 with 25 VIs. However, beyond 15–20 VIs, further RMSE reductions are negligible, indicating diminishing returns in predictive accuracy as less critical indices are added.

When evaluating various vegetation indices in ML models (without neighborhood information), XGB outperformed RF, GBR, and ET, consistently with more than 10 indices. With 10 or fewer indices, performance varied among XGB, RF, and GBR, while ET consistently lagged. This performance could be attributed to how each model processes input features. XGB, RF, and GBR capture complex relationships within structured data, benefiting from multiple indices. XGB excels with boosting and engineered features, but its

performance becomes more variable with fewer indices. ET, relying on randomized splits instead of boosting or feature interactions, struggles to extract meaningful patterns, resulting in lower performance. These results highlight that XGB performs best with vegetation indices, while ET is more effective with spatially lagged spectral data.

3.4. Model Performance Combining Neighborhood Information and Vegetation Indices

Regression models with Set 4 predictors (five bands, 20 VIs, and eight neighbors) slightly outperformed those with sets S-1, S-2, and S-3. Among the SLML models, XGB achieved the highest R^2 value (0.57), followed closely by random forest (RF, R^2 = 0.56) and extra trees (ET, R^2 = 0.55). Gradient boosting regressor (GBR) showed a moderate R^2 of 0.50, while SLX had the lowest performance in this group with an R^2 of 0.46. Furthermore, it can be highlighted that the SLML models showed superior performance over the LR.

The RMSE values support this ranking, reflecting the model's predictive accuracy. XGB showed the lowest RMSE (0.94), indicating superior precision, with RF (0.95) and ET (0.96) performing similarly. GBR had a slightly higher RMSE (1.01), and SLX again showed the least accurate predictions with an RMSE of 1.05. The consistent superiority of XGB, RF, and ET across both metrics underscores the ability of SLML approaches to effectively capture non-linear relationships and complex interactions within the data, including spectral bands, neighborhood information, and vegetation indices.

When neighborhood information was integrated with vegetation indices, XGB outperformed the other models, followed by RF, ET, and GBR. This ranking reflects how models handle Vis and spatially correlated data. XGB and RF effectively utilize vegetation indices while also capturing spatial dependencies, with XGB refining weak predictions through boosting and RF benefiting from ensemble averaging. Although ET performs best when using neighborhood data, it remains less efficient when incorporating both vegetation indices and spatial information due to its reliance on randomized splits. GBR, relying on sequential learning, struggles with spatial autocorrelation, leading to the lowest performance. These findings highlight the varying abilities of ML models to integrate Vis and spatial information for crop yield prediction.

3.5. Feature Importance When Combining All the Predictors

The feature importance analysis for Set 4 predictors showed that spatially lagged spectral bands consistently ranked among the top predictors across various SLML models (Figure 9). In the RF, ET, and GBR models, all spatially lagged features (Green_lag, NIR_lag, RedEdge_lag, Blue_lag, and Red_lag) were within the top 11 features out of 30. In the RF and GBR models, for example, Green_lag and NIR_lag ranked as the top two predictors, surpassing all VIs. In the ET model, these features ranked third and fourth, respectively, highlighting their significance among the large number of VIs evaluated. Meanwhile, in the XGB model, although ranked lower, Green_lag and NIR_lag still stood out for their high importance. These results emphasize the value of spatially lagged features in capturing spatial variability and improving the corn yield prediction.

Regarding vegetation indices, NDRE, CCCI, and CREI consistently emerged as significant contributors across all models, reinforcing their relevance for yield estimation. In contrast, raw spectral bands such as green, red, and rededge ranked lower, which can be attributed to the fact that VIs synthesize and enhance spectral information, making them more relevant for predicting crop yield. Thus, the information from spectral bands is more effectively utilized through VIs, leading to more accurate yield predictions.

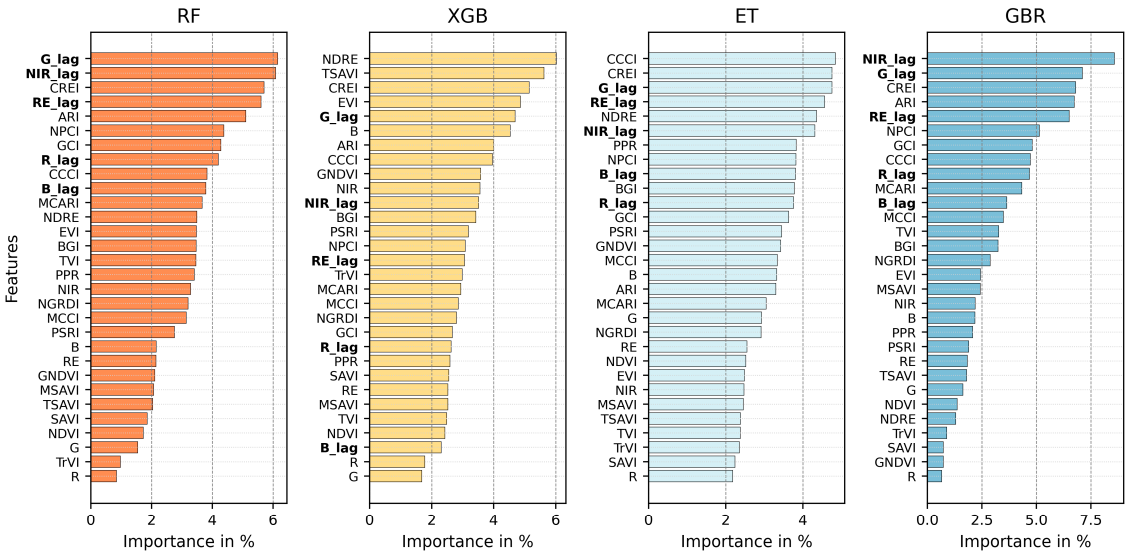


Figure 9. Feature importance of ML modeling in corn yield prediction. The five spatially lagged spectral band features (Green_lag (G-lag), NIR_lag, RedEdge_lag (RE-lag), Blue_lag (B-lag), and Red_lag (R-lag)) are in bold to highlight their significant role among all evaluated predictors.

3.6. Model Performance Comparison Across Predictor Sets

Table 4 summarizes the performance of various corn yield prediction models across the four predictor sets. For Sets 2 (S-2) and Set 3 (S-3), the optimal number of neighbors and vegetation indices, respectively, were considered. The results show that the model performance using the baseline predictors was outperformed by the other sets (S-2, S-3, and S-4), with lower R^2 values compared to the other sets.

Table 4. Best R^2 and RMSE values of regression models for the four predictors sets evaluated in corn yield prediction, with top-performing values in each model highlighted in bold.

Model	Coefficient of Determination (R^2)				Root Mean Square Error (RMSE)			
	S-1	Best S-2	Best S-3	S-4	S-1	Best S-2	Best S-3	S-4
LR	0.19	**0.48**	0.31	0.46	1.28	**1.03**	1.19	1.05
RF	0.32	0.52	0.49	**0.56**	1.18	0.99	1.02	**0.95**
XGB	0.41	0.54	0.52	**0.57**	1.10	0.97	0.99	**0.94**
ET	0.27	**0.57**	0.44	0.55	1.23	**0.94**	1.07	0.96
GBR	0.39	**0.50**	0.48	0.50	1.12	**1.01**	1.04	1.01

Results also show that incorporating spatial neighborhood data (S-2) enhances model performance more significantly than including VIs (S-3). The R^2 values using optimal neighborhood information were always higher than using VIs (S-3). For example, the XGB model achieved an R^2 of 0.54 with eight neighbors (S-2B), outperforming its performance when using both spectral bands and VIs (Set S-3G, R^2 = 0.52 with 20 VIs). Similarly, the ET model showed an R^2 of 0.57 with neighborhood data, compared to 0.44 with VIs. A similar trend was observed in the other models (RF, GBR, and LR) when comparing these approaches. The RMSE values further support this trend. For instance, the XGB model achieved an RMSE of 0.97 with eight neighbors (S-2B), compared to 0.99 when 20 VIs were used (S-3). A similar pattern was observed for RF, with an RMSE of 0.99 when using neighborhood data, compared to 1.02 when incorporating VIs. These results highlight that spatial relationships captured through neighborhood information are more effective for accurate yield prediction than the inclusion of additional spectral indices. These results highlight the greater predictive power of spatial data in corn yield prediction.

The performance of regression models using the S-4 predictor set generally demonstrated comparable or slightly superior results relative to other predictor-sets, particularly with RF and XGB. For example, RF and XGB achieved their highest R^2 values with the S-4 set (0.56 and 0.57, respectively), outperforming their results with S-2 and S-3. In contrast, the ET, GBR, and SLX models showed no significant advantage with S-4 over S-2 but maintained competitive performance. These findings highlight the robustness of the S-4 predictors, especially for ensemble methods like RF and XGB, which effectively utilized the combined spectral and spatial data

4. Discussion

4.1. Spatial Autocorrelation of Corn Yield

The findings confirm a strong positive spatial autocorrelation of corn yield, particularly within the immediate neighborhood, as indicated by Moran's I values. This reinforces the idea that yield variability in agricultural fields is significantly influenced by localized spatial factors. The gradual decrease in autocorrelation with increasing neighborhood size further supports the notion that proximity plays a central role in yield patterns, aligning with the first law of geography [66,67].

The results highlight the importance of incorporating spatial lag effects in yield prediction models. Capturing local spatial dependencies significantly improved model performance, as reflected in the enhanced R^2 and RMSE values. Notably, the spatially lagged predictors consistently ranked among the top features in feature importance analyses, highlighting their predictive value. This finding aligns with previous studies emphasizing the relevance of spatial patterns in agricultural predictions [68]. Additionally, the integration of lagged variables, as suggested by [11], proved essential in improving model accuracy.

However, a limitation of this approach is that spatial autocorrelation weakens as neighborhood size increases, suggesting that incorporating too many distant neighbors may not offer additional predictive benefits [11]. This could be due to the diminishing relevance of spatial dependencies over larger distances.

This finding is particularly relevant for precision agriculture, where optimizing the spatial scale of prediction models is crucial for accurate decision-making and resource allocation. By demonstrating the potential of spatially lagged spectral information, our study establishes a strong foundation for advancing predictive modeling across diverse agricultural settings. Recognizing its significance, we identify this as a key area for future research and plan to investigate it further in upcoming studies.

4.2. Neighborhood Information and Model Performance

Incorporating neighborhood information significantly improved corn yield prediction across all regression techniques. The benefits were most pronounced in spatially lagged machine learning (SLML) models, such as ET and XGB, emphasizing the importance of spatial context in understanding crop yield variability. This finding is consistent with research that underscores the role of spatially structured variables in agricultural modeling [69] and the recognition that agricultural production is inherently spatial, necessitating spatial dependence models for accurate predictions [70].

Despite the positive impact of neighborhood data, R^2 did not scale linearly with the number of neighbors. This suggests that spatial dependencies are primarily localized, with diminishing returns beyond a certain spatial extent. Similar patterns have been observed in other spatial prediction studies [71,72], but are less studied in the context of crop yield prediction. This could be attributed to the variability in agricultural practices and environmental factors, which may limit the benefits of extending spatial dependencies beyond a certain threshold [73,74].

Our results contribute to this gap by demonstrating that an optimal balance exists between model complexity and predictive performance. Specifically, models incorporating four to eight neighboring points provided the highest predictive accuracy, while increasing the number of neighbors led to diminishing returns due to potential overfitting or the inclusion of less relevant spatial information. We recognize that the optimal neighborhood size can vary based on factors such as soil type, climate, management practices, and irrigation, which all affect the relationship between crop yield, spectral data, and spatial dependencies. This highlights the importance of optimizing neighborhood parameters, a challenge that remains underexplored in agricultural research [75,76].

Moreover, while this study used a fixed number of neighbors for neighborhood selection, we acknowledge that a more scientific approach to determining the optimal neighborhood size could improve accuracy. However, our findings provide a solid foundation for further research into systematically identifying both the number and type of neighbors. Future work could explore localized optimal neighbor search algorithms, integrated with spatial-lagged machine learning (SLML), to refine spatial dependency estimates, considering additional spatial features such as soil characteristics.

4.3. Vegetation Indices and Model Performance

The inclusion of vegetation indices (VIs) consistently enhanced regression model performance compared to using only spectral bands, reinforcing the importance of VIs in improving corn yield predictions. SLML models, particularly XGB and RF, showed the most substantial improvements, emphasizing the ability of these indices to capture crop variability effectively. These results align with previous research highlighting the predictive strength of spectral indices in agricultural modeling [77,78]. Particularly in this study, indices like CREI, GCI, NCPI, ARI, and CCCI emerged as the most influential.

A key finding was the plateauing effect when more than 10–15 VIs were included, indicating that only a subset of indices is needed to capture most of the relevant variability. The diminishing returns beyond this threshold suggest that additional indices contribute redundant or weakly relevant information, aligning with feature selection principles in machine learning [79]. These findings highlight the need for balancing model complexity and accuracy [76,80].

While adding VIs reduced prediction errors, the marginal improvements with additional indices also highlighted the challenges of computational inefficiency and potential overfitting. Furthermore, SLX, while showing slight gains, remained limited in its ability to model the complex, non-linear relationships inherent in crop yield predictions. These findings reinforce the importance of optimizing feature selection to achieve robust and efficient models.

4.4. Model Performance When Combining All Predictors

The combination of raw spectral bands, spatially lagged bands, and VIs yielded competitive predictive accuracy. However, improvements over predictor sets that included only spatial neighborhood data (S-2) were relatively modest. This finding suggests that while VIs are valuable, spatially lagged data remain the dominant factor in driving model performance. This aligns with studies emphasizing the role of spatial factors in agricultural modeling, challenging the traditional prioritization of VIs as primary predictors [69,70].

Feature importance analysis further reinforced this perspective, with spatially lagged predictors consistently ranking among the top predictors. This finding contributes to the evolving understanding of crop yield prediction by highlighting the interplay between spatial and spectral data. Unlike previous research, which has not studied the combination of neighborhood spectral values and VIs, our study demonstrates their combined

effects, providing a more comprehensive framework for yield estimation. However, the modest accuracy gains from integrating these predictors suggest potential redundancies or collinearities, which warrant further investigation.

Additionally, raw spectral bands (B, G, R, RE, NIR) ranked lower in feature importance than VIs, reinforcing the value of synthesized indices in capturing meaningful spectral information [78,81]. These findings align with prior studies demonstrating that VIs outperform raw reflectance in biomass estimation and other crop productivity measures [10,25].

The overall performance of decision-tree-based models showed that no single ML algorithm demonstrated clear superiority; performance varied depending on whether neighborhood information, vegetation indices, or both were included. This suggests that model effectiveness in predicting corn yield is context-dependent, influenced by input type (spatially lagged data vs. vegetation indices) and the number of variables used. ET efficiently handles high-dimensional data and captures complex relationships [58], showing strong performance with spatially lagged data by using neighboring spectral features through randomized feature splits. In contrast, XGB excels with structured data, using boosting to refine weak learners [57], though its sensitivity to spatial correlations may limit its effectiveness when spatial autocorrelation is strong. RF, with its ensemble-based approach, balances predictive performance and robustness, making it effective across different input types but less specialized for spatial or structured data. GBR, relying on sequential learning [59], struggles with spatial autocorrelation, reducing its performance when neighborhood information is included. ET performs better with spatially lagged spectral data, while XGB thrives with engineered features like vegetation indices. These findings underscore the importance of context-dependent model selection. Future studies should assess multiple ML approaches based on dataset characteristics, as no single model is universally superior.

4.5. Implications for Future Research

These findings have several methodological and practical implications for agricultural modeling. First, the results highlight the need for context-specific optimization of spatial parameters, as the optimal neighborhood size and the relative importance of spatial and spectral predictors may vary across crops, regions, and data sources. Understanding these variations can help refine adaptive modeling strategies for different agricultural settings.

Second, the study highlights the value of high-resolution UAV data in capturing fine-scale spatial dependencies, suggesting that future research should explore how these findings scale to coarser resolutions or larger areas. It remains unclear whether the same spatial relationships hold at broader scales, necessitating multi-scale validation efforts.

Third, the observed trade-offs between model complexity and predictive performance emphasize the importance of balancing data richness with computational efficiency. While incorporating additional features generally improves accuracy, diminishing returns and the risk of overfitting must be carefully managed. These insights are particularly relevant for operational applications, where computational efficiency is critical for real-time decision support in precision agriculture.

Finally, this study emphasizes the value of neighborhood information in enhancing corn yield prediction, laying the groundwork for future research across diverse crops and agricultural settings with spatial dependencies. Future studies should explore how specific spectral bands (e.g., blue, green, red, NIR, rededge) and their interactions influence prediction accuracy, considering context-dependent factors such as agronomic conditions, crop types, and environmental stressors. Additionally, our fixed neighborhood size may not be optimal for all contexts. Research should focus on developing methods for determining the optimal neighborhood size through localized neighbor search algorithms

integrated with spatial-lagged machine learning (SLML). Incorporating additional spatial features, such as soil characteristics, could further refine these models and improve prediction performance.

5. Conclusions

This study presents an innovative approach to crop yield prediction by integrating spectral neighborhood data using the spatial-lagged machine learning model (SLML), an enhanced version of the spatial lag X model (SLX). The research aims to demonstrate how integrating spatially lagged spectral information enhances predictive accuracy compared to traditional approaches that primarily rely on vegetation indices (VIs). The study hypothesizes that integrating neighborhood spectral data significantly enhances predictive accuracy, providing novel insights into the role of spatial dependencies in agricultural yield prediction.

This study was conducted on a 19-hectare cornfield at the ARS Grassland, Soil, and Water Research Laboratory during the 2023 growing season. Multispectral imagery was collected using a WintraOne UAV with a MicaSense RedEdge-P sensor, capturing five spectral bands (R, G, B, RE, NIR) at 6 cm resolution during the VT/R1 stage. Yield data covered 21.6 m² per measurement, resulting in 8581 yield values ranging from 1.69 to 15.86 Mg/Ha (M = 10.19 Mg/Ha). SLML models, using decision-tree-based techniques (RF, XGB, ET, and GBR) in addition to SLX, were applied to predict corn yield, incorporating spatially lagged predictors (neighborhood data). Four predictor sets were evaluated: Set 1 (spectral bands—baseline), Set 2 (spectral bands + neighborhood data), Set 3 (spectral bands + VIs), and Set 4 (spectral bands + top VIs + neighborhood data). Model performance was assessed using R^2 and RMSE.

The main finding revealed that the introduced approach (incorporating neighborhood data—Set 2), consistently outperformed the traditional approach using VIs (Set 3) in predicting corn yield. This highlights the superior value of spatial context, where the proximity of neighboring cells provides more relevant information about yield variability than the spectral reflectance captured by VIs. Additionally, this study demonstrated the significant role of spatial autocorrelation in corn yield data, with the strongest correlations observed in immediate neighborhoods (eight neighbors). Incorporating spatial neighborhood data consistently improved model performance, with SLML models such as XGB, RF, and ET performing best with four to eight neighbors. However, excessive inclusion of neighbors led to diminishing returns, underscoring the importance of optimizing neighbor selection.

Evaluation in Set 3 showed that while VIs are effective for assessing crop health and vigor; however, their predictive power for yield plateaus once an optimal set of indices is reached. In this regard, VIs enhanced model performance, particularly with XGB, although a smaller subset (10–15 indices) proved sufficient for optimal yield prediction.

Evaluation in Set 4, combination of spatial and spectral data showed slight performance improvements, with XGB and RF achieving the highest R^2 values, emphasizing the value of integrating both data types for enhanced predictive accuracy. In this case, combining spatial and spectral data, spatially lagged spectral bands (e.g., Green_lag, NIR_lag, RedEdge_lag) were ranked predictors, highlighting the importance of including neighborhood information into the regression model for corn yield prediction. Vegetation indices such as CREI, GCI, NCPI, ARI, and CCCI were also key predictors, competing with spatially lagged data and consistently outperforming raw spectral data alone.

In addition, the performance of decision-tree-based ML models in predicting corn yield varied depending on the inclusion of neighborhood information, vegetation indices, or both. No single model emerged as superior, with effectiveness being context dependent. ET excelled with spatially lagged data, XGB performed best with structured data, RF

balanced performance across data types, and GBR struggled with spatial autocorrelation. These findings highlight the need for context-driven model selection and suggest future studies should evaluate multiple ML approaches based on dataset characteristics.

This study emphasizes the significance of spatial context and neighborhood information in improving corn yield prediction. It highlights the need for optimizing spatial parameters, feature selection, and neighborhood size to enhance model accuracy. Future research should explore the scalability of these methods to larger areas and coarser resolutions, while investigating advanced approaches for integrating spatial and spectral data. Additionally, the influence of specific spectral bands and their interactions, along with agronomic and environmental factors, should be further examined. Incorporating spatial features such as soil characteristics and developing localized search algorithms for optimal neighborhood size could further refine predictive models and improve their robustness across diverse agricultural settings.

Author Contributions: Conceptualization, J.M.O.L. and E.N.-Y.; methodology, J.M.O.L. and E.N.-Y.; software, E.N.-Y.; validation, E.N.-Y.; formal analysis, J.M.O.L. and E.N.-Y.; investigation, E.N.-Y.; resources, J.M.O.L.; field work and data collection, K.A. and C.B.H.; writing—original draft preparation, E.N.-Y.; writing—review and editing, E.N.-Y., J.M.O.L., C.B.H., K.A. and D.R.S.; supervision, J.M.O.L. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported through the U.S. Department of Agriculture's Conservation Effects Assessment Project (CEAP), a multi-agency effort led by the Natural Resources Conservation Service (NRCS) to quantify the effects of voluntary conservation and strengthen data-driven management decisions across the nation's private lands, under Texas A&M AgriLife Cooperative agreement number NR213A750023C012.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code used in this study are available upon request from the corresponding author. Given that the study involves large volumes of aerial images and derived rasters, we are currently seeking a repository that can accommodate high-volume datasets.

Acknowledgments: The authors acknowledge the help of technical and non-technical persons involved in establishment and data collection of this study. This project was supported through the U.S. Department of Agriculture's Conservation Effects Assessment Project (CEAP), a multi-agency effort led by the Natural Resources Conservation Service (NRCS) to quantify the effects of voluntary conservation and strengthen data-driven management decisions across the nation's private lands, under Texas A&M AgriLife Cooperative agreement number NR213A750023C012. We would also like to express our sincere gratitude to ARS Grassland Soil and Water Research Laboratory for their support in conducting the fieldwork and maintaining the crop throughout the study, ensuring the quality and accuracy of the data collected. The authors acknowledge that the geospatial corrections for images using post-processing kinematics (PPK) and the creation of an orthomosaic of multispectral rasters were processed by Sayantan Sarkar at Texas A&M AgriLife Blackland Research and Extension Center, Temple, TX.

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

References

1. Filippi, P.; Jones, E.J.; Wimalathunge, N.S.; Somarathna, P.D.S.N.; Pozza, L.E.; Ugbaje, S.U.; Jephcott, T.G.; Paterson, S.E.; Whelan, B.M.; Bishop, T.F.A. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precis. Agric.* **2019**, *20*, 1015–1029. [CrossRef]
2. Lobell, D.B.; Roberts, M.J.; Schlenker, W.; Braun, N.; Little, B.B.; Rejesus, R.M.; Hammer, G.L. Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest. *Science* **2014**, *344*, 516–519. [CrossRef] [PubMed]

3. Hatfield, J.L.; Prueger, J.H. Temperature extremes: Effect on plant growth and development. *Weather. Clim. Extremes* **2015**, *10*, 4–10. [CrossRef]
4. Zhou, H.; Yang, J.; Lou, W.; Sheng, L.; Li, D.; Hu, H. Improving grain yield prediction through fusion of multi-temporal spectral features and agronomic trait parameters derived from UAV imagery. *Front. Plant Sci.* **2023**, *14*, 1217448. [CrossRef]
5. Noa-Yarasca, E.; Leyton, J.M.O.; Angerer, J. Biomass Time Series Forecasting Using Deep Learning Techniques. Is the Sophisticated Model Superior? In *Biometry and Statistical Computing*; ASA, CSSA, SSSA International Annual Meeting: St. Louis, MO, USA; Available online: https://scisoc.confex.com/scisoc/2023am/meetingapp.cgi/Paper/151648 (accessed on 10 April 2024).
6. Hunt, E.R., Jr.; Hively, W.D.; Fujikawa, S.J.; Linden, D.S.; Daughtry, C.S.T.; McCarty, G.W. Acquisition of NIR-Green-Blue Digital Photographs from Unmanned Aircraft for Crop Monitoring. *Remote Sens.* **2010**, *2*, 290–305. [CrossRef]
7. Cicek, H.; Sunohara, M.; Wilkes, G.; McNairn, H.; Pick, F.; Topp, E.; Lapen, D. Using vegetation indices from satellite remote sensing to assess corn and soybean response to controlled tile drainage. *Agric. Water Manag.* **2010**, *98*, 261–270. [CrossRef]
8. Killeen, P.; Kiringa, I.; Yeap, T.; Branco, P. Corn Grain Yield Prediction Using UAV-Based High Spatiotemporal Resolution Imagery, Machine Learning, and Spatial Cross-Validation. *Remote Sens.* **2024**, *16*, 683. [CrossRef]
9. Moeckel, T.; Dayananda, S.; Nidamanuri, R.R.; Nautiyal, S.; Hanumaiah, N.; Buerkert, A.; Wachendorf, M. Estimation of Vegetable Crop Parameter by Multi-temporal UAV-Borne Images. *Remote Sens.* **2018**, *10*, 805. [CrossRef]
10. Panda, S.S.; Ames, D.P.; Panigrahi, S. Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques. *Remote Sens.* **2010**, *2*, 673–696. [CrossRef]
11. Lesage, J.; Pace, R.K. *Introduction to Spatial Econometrics*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2023.
12. Lou, M.; Zhang, H.; Lei, X.; Li, C.; Zang, H. Spatial Autoregressive Models for Stand Top and Stand Mean Height Relationship in Mixed *Quercus mongolica* Broadleaved Natural Stands of Northeast China. *Forests* **2016**, *7*, 43. [CrossRef]
13. Fang, C.; Liu, H.; Li, G.; Sun, D.; Miao, Z. Estimating the Impact of Urbanization on Air Quality in China Using Spatial Regression Models. *Sustainability* **2015**, *7*, 15570–15592. [CrossRef]
14. Ponciano, P.F.; Scalon, J.D. Análise espacial da produção leiteira usando um modelo autoregressivo condicional. *Semin. Cienc. Agrar.* **2010**, *31*, 487–496. [CrossRef]
15. Ahn, K.-H.; Palmer, R. Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique. *J. Hydrol.* **2016**, *540*, 515–526. [CrossRef]
16. Yoo, J.; Ready, R. The impact of agricultural conservation easement on nearby house prices: Incorporating spatial autocorrelation and spatial heterogeneity. *J. For. Econ.* **2016**, *25*, 78–93. [CrossRef]
17. Guo, L.; Zhao, C.; Zhang, H.; Chen, Y.; Linderman, M.; Zhang, Q.; Liu, Y. Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology. *Geoderma* **2017**, *285*, 280–292. [CrossRef]
18. Auffhammer, M.; Hsiang, S.; Schlenker, W.; Sobel, A. Using Weather Data and Climate Model Output in Economic Analyses of Climate Change. *Rev. Environ. Econ. Policy* **2013**, *7*, 181–198. [CrossRef]
19. Dell, M.; Jones, B.F.; Olken, B.A. What Do We Learn from the Weather? The New Climate-Economy Literature. *J. Econ. Lit.* **2014**, *52*, 740–798. [CrossRef]
20. Schlenker, W.; Roberts, M.J. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15594–15598. [CrossRef]
21. Hawinkel, S.; De Meyer, S.; Maere, S. Spatial Regression Models for Field Trials: A Comparative Study and New Ideas. *Front. Plant Sci.* **2022**, *13*, 858711. [CrossRef]
22. Fischer, R.J.; Rekabdarkolaee, H.M.; Joshi, D.R.; Clay, D.E.; Clay, S.A. Soybean prediction using computationally efficient Bayesian spatial regression models and satellite imagery. *Agron. J.* **2024**, *116*, 2841–2849. [CrossRef]
23. Ward, M.; Gleditsch, K. *Spatial Regression Models*; SAGE Publications: Thousand Oaks, CA, USA, 2008. [CrossRef]
24. Rüttenauer, T. Spatial Regression Models: A Systematic Comparison of Different Model Specifications Using Monte Carlo Experiments. *Sociol. Methods Res.* **2022**, *51*, 728–759. [CrossRef]
25. Rehman, T.H.; Lundy, M.E.; Linquist, B.A. Comparative Sensitivity of Vegetation Indices Measured via Proximal and Aerial Sensors for Assessing N Status and Predicting Grain Yield in Rice Cropping Systems. *Remote Sens.* **2022**, *14*, 2770. [CrossRef]
26. Sarkar, S.; Leyton, J.M.O.; Noa-Yarasca, E.; Adhikari, K.; Hajda, C.B.; Smith, D.R. Integrating Remote Sensing and Soil Features for Enhanced Machine Learning-Based Corn Yield Prediction in the Southern US. *Sensors* **2025**, *25*, 543. [CrossRef] [PubMed]
27. Effrosynidis, D.; Sylaios, G.; Arampatzis, A. The Effect of Training Data Size on Disaster Classification from Twitter. *Information* **2024**, *15*, 393. [CrossRef]
28. Awan, F.M.; Saleem, Y.; Minerva, R.; Crespi, N. A Comparative Analysis of Machine/Deep Learning Models for Parking Space Availability Prediction. *Sensors* **2020**, *20*, 322. [CrossRef]
29. Noa-Yarasca, E.; Leyton, J.M.O.; Angerer, J.P. Deep Learning Model Effectiveness in Forecasting Limited-Size Aboveground Vegetation Biomass Time Series: Kenyan Grasslands Case Study. *Agronomy* **2024**, *14*, 349. [CrossRef]
30. Soil Survey Staff. Keys to Soil Taxonomy 11th Edition. Washington, DC, USA. Available online: https://www.nrcs.usda.gov/sites/default/files/2022-09/Keys-to-Soil-Taxonomy.pdf (accessed on 10 December 2024).

31. Adhikari, K.; Smith, D.R.; Hajda, C.; Kharel, T.P. Within-field yield stability and gross margin variations across corn fields and implications for precision conservation. *Precis. Agric.* **2023**, *24*, 1401–1416. [CrossRef]
32. FAO. Faostat: Crops and Livestock Products. Food and Agriculture Organization of the United Nations. Available online: https://www.fao.org/faostat/en/#data/QV (accessed on 19 July 2024).
33. USDA-NASS. Quick Stats. United States Department of Agriculture, National Agricultural Statistics Service. Available online: https://quickstats.nass.usda.gov/ (accessed on 19 July 2024).
34. Fu, W.J.; Jiang, P.K.; Zhou, G.M.; Zhao, K.L. Using Moran's I and GIS to study the spatial pattern of forest litter carbon density in a subtropical region of southeastern China. *Biogeosciences* **2014**, *11*, 2401–2409. [CrossRef]
35. O'sullivan, D.; Unwin, D. *Geographic Information Analysis*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2010.
36. Noa-Yarasca, E. A Machine Learning Model of Riparian Vegetation Attenuated Stream Temperatures. Oregon State University, Corvallis, OR, USA. Available online: https://ir.library.oregonstate.edu/downloads/0r967b65c#page=137 (accessed on 10 April 2024).
37. Gitelson, A.A.; Viña, A.; Arkebauer, T.J.; Rundquist, D.C.; Keydan, G.; Leavitt, B. Remote estimation of leaf area index and green leaf biomass in maize canopies. *Geophys. Res. Lett.* **2003**, *30*, 1248. [CrossRef]
38. Peñuelas, J.; Gamon, J.; Fredeen, A.; Merino, J.; Field, C. Reflectance indices associated with physiological changes in nitrogen- and water-limited sunflower leaves. *Remote Sens. Environ.* **1994**, *48*, 135–146. [CrossRef]
39. Gitelson, A.A.; Merzlyak, M.N.; Zur, Y.; Stark, R.; Gritz, U. Non-Destructive and Remote Sensing Techniques for Estimation of Vegetation Status. Papers in Natural Resources, No. 273. 2001. Available online: https://digitalcommons.unl.edu/natrespapers/273/ (accessed on 18 November 2024).
40. Barnes, E.M.; Clarke, T.R.; Richards, S.E.; Colaizzi, P.D.; Haberland, J.; Kostrzewski, M.; Moran, M.S. Coincident Detection of Crop Water Stress, Nitrogen Status and Canopy Density Using Ground-Based Multi-spectral Data. In Proceedings of the 5th International Conference on Precision Agriculture, Bloomington, MN, USA, 16–19 July 2000; Robert, P.C., Rust, R.H., Larson, W.E., Eds.
41. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
42. Daughtry, C.S.T.; Walthall, C.L.; Kim, M.S.; De Colstoun, E.B.; McMurtrey, J.E., III. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sens. Environ.* **2000**, *74*, 229–239. [CrossRef]
43. Marcial-Pablo, M.d.J.; Gonzalez-Sanchez, A.; Jimenez-Jimenez, S.I.; Ontiveros-Capurata, R.E.; Ojeda-Bustamante, W. Estimation of vegetation fraction using RGB and multispectral images from UAV. *Int. J. Remote Sens.* **2019**, *40*, 420–438. [CrossRef]
44. Zarco-Tejada, P.J.; Berjón, A.; López-Lozano, R.; Miller, J.R.; Martín, P.; Cachorro, V.; González, M.R.; De Frutos, A. Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. *Remote Sens. Environ.* **2005**, *99*, 271–287. [CrossRef]
45. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
46. Metternicht, G. Vegetation indices derived from high-resolution airborne videography for precision crop management. *Int. J. Remote Sens.* **2003**, *24*, 2855–2877. [CrossRef]
47. 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]
48. 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]
49. 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]
50. 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]
51. Rouse, J.W., Jr.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring Vegetation Systems in the Great Plains with Erts. 1974. Washington, DC, USA. Available online: https://ui.adsabs.harvard.edu/abs/1974NASSP.351.309R/abstract (accessed on 18 November 2024).
52. Yang, C.; Everitt, J.H.; Bradford, J.M.; Murden, D. Airborne Hyperspectral Imagery and Yield Monitor Data for Mapping Cotton Yield Variability. *Precis. Agric.* **2004**, *5*, 445–461. [CrossRef]
53. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]
54. Zhang, L.; Zhang, H.; Niu, Y.; Han, W. Mapping Maize Water Stress Based on UAV Multispectral Remote Sensing. *Remote Sens.* **2019**, *11*, 605. [CrossRef]

55. Baret, F.; Guyot, G.; Major, D. TSAVI: A Vegetation Index Which Minimizes Soil Brightness Effects On LAI And APAR Estimation. In Proceedings of the 12th Canadian Symposium on Remote Sensing Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 10–14 July 1989; Institute of Electrical and Electronics Engineers (IEEE): Vancouver, BC, Canada, 1989; pp. 1355–1358. [CrossRef]
56. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
57. Chen, T.; Guestrin, C. "XGBoost". In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
58. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
59. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics* **2013**, *7*, 21. [CrossRef]
60. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
61. Zhang, Z.; Zhao, Y.; Canes, A.; Steinberg, D.; Lyashevska, O. Predictive analytics with gradient boosting in clinical medicine. *Ann. Transl. Med.* **2019**, *7*, 152. [CrossRef]
62. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
63. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 51–56.
64. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
65. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
66. Grossfeld, B. Geography and law. *Mich. Law Rev.* **1984**, *82*, 1510–1519. Available online: https://repository.law.umich.edu/mlr/vol82/iss5/25 (accessed on 22 November 2024). [CrossRef]
67. Bai, D.; Ye, L.; Yang, Z.; Wang, G. Impact of climate change on agricultural productivity: A combination of spatial Durbin model and entropy approaches. *Int. J. Clim. Chang. Strat. Manag.* **2024**, *16*, 26–48. [CrossRef]
68. Lichstein, J.W.; Simons, T.R.; Shriner, S.A.; Franzreb, K.E. Spatial Autocorrelation and Autoregressive Models In Ecology. *Ecol. Monogr.* **2003**, *72*, 445–463. [CrossRef]
69. Naher, A.; Almas, L.K.; Guerrero, B.; Shaheen, S. Spatiotemporal Economic Analysis of Corn and Wheat Production in the Texas High Plains. *Water* **2023**, *15*, 3553. [CrossRef]
70. Ayouba, K. Spatial dependence in production frontier models. *J. Prod. Anal.* **2023**, *60*, 21–36. [CrossRef]
71. Huo, X.-N.; Li, H.; Sun, D.-F.; Zhou, L.-D.; Li, B.-G. Combining Geostatistics with Moran's I Analysis for Mapping Soil Heavy Metals in Beijing, China. *Int. J. Environ. Res. Public Health* **2012**, *9*, 995–1017. [CrossRef]
72. Wu, G.; Fan, Y.; Riaz, N. Spatial Analysis of Agriculture Ecological Efficiency and Its Influence on Fiscal Expenditures. *Sustainability* **2022**, *14*, 9994. [CrossRef]
73. Sangoi, L. Understanding plant density effects on maize growth and development: An important issue to maximize grain yield. *Cienc. Rural.* **2001**, *31*, 159–168. [CrossRef]
74. Postma, J.A.; Hecht, V.L.; Hikosaka, K.; Nord, E.A.; Pons, T.L.; Poorter, H. Dividing the pie: A quantitative review on plant density responses. *Plant Cell Environ.* **2020**, *44*, 1072–1094. [CrossRef]
75. Trevisan, R.G.; Bullock, D.S.; Martin, N.F. Spatial variability of crop responses to agronomic inputs in on-farm precision experimentation. *Precis. Agric.* **2021**, *22*, 342–363. [CrossRef]
76. Noa-Yarasca, E.; Babbar-Sebens, M.; Jordan, C.E. Machine Learning Models for Prediction of Shade-Affected Stream Temperatures. *J. Hydrol. Eng.* **2025**, *30*, 04024058. [CrossRef]
77. Shrestha, A.; Bheemanahalli, R.; Adeli, A.; Samiappan, S.; Czarnecki, J.M.P.; McCraine, C.D.; Reddy, K.R.; Moorhead, R. Phenological stage and vegetation index for predicting corn yield under rainfed environments. *Front. Plant Sci.* **2023**, *14*, 1168732. [CrossRef]
78. Pinto, A.A.; Zerbato, C.; Rolim, G.d.S.; Júnior, M.R.B.; da Silva, L.F.V.; de Oliveira, R.P. Corn grain yield forecasting by satellite remote sensing and machine-learning models. *Agron. J.* **2022**, *114*, 2956–2968. [CrossRef]
79. Verma, B.; Prasad, R.; Srivastava, P.K.; Yadav, S.A.; Singh, P.; Singh, R. Investigation of optimal vegetation indices for retrieval of leaf chlorophyll and leaf area index using enhanced learning algorithms. *Comput. Electron. Agric.* **2022**, *192*, 106581. [CrossRef]
80. Radočaj, D.; Šiljeg, A.; Marinović, R.; Jurišić, M. State of Major Vegetation Indices in Precision Agriculture Studies Indexed in Web of Science: A Review. *Agriculture* **2023**, *13*, 707. [CrossRef]
81. Lawrence, R.L.; Ripple, W.J. Comparisons among Vegetation Indices and Bandwise Regression in a Highly Disturbed, Heterogeneous Landscape: Mount St. Helens, Washington. *Remote Sens. Environ.* **1998**, *64*, 91–102. [CrossRef]

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