AgriEngineering

Article

Environmental Covariates for Sampling Optimization and Pest Prediction in Soybean Crops

Cenneya Lopes Martins [1,*], Maiara Pusch [1], Wesley Augusto Conde Godoy [2] and Lucas Rios do Amaral [1]

[1] School of Agricultural Engineering, University of Campinas—FEAGRI/UNICAMP, Campinas 13083-875, São Paulo, Brazil; maiarapuschh@hotmail.com (M.P.); lucas.amaral@feagri.unicamp.br (L.R.d.A.)
[2] Luiz de Queiroz College of Agriculture, University of São Paulo—ESALQ/USP, Piracicaba 13418-900, São Paulo, Brazil; wacgodoy@usp.br
* Correspondence: cenneya.martins@gmail.com

Abstract: Insect pest infestations can vary due to spatial differences in microclimates and food availability within agroecosystems. Covariates can reflect these environmental conditions. This study tested whether using environmental covariates in two-phase sample optimization improved the spatial predictions for soybean insect pests. During the 2021–2022 crop season, insect pest samples were collected at 50 georeferenced points in a commercial soybean field in Brazil, alongside data on environmental covariates such as vegetation indices, soil properties, terrain topography, and distances from riparian areas. Three covariates were selected using correlation and principal component analysis (PCA). In the 2022–2023 crop season, sample designs were optimized using the iterative algorithm optimization of sample configurations using spatial simulated annealing (SPSANN) using the selected covariates, resulting in two optimized designs that were compared to a regular grid. Data from the three sampling designs comprising 50 points were evaluated using geostatistical methods, regression analysis (pest abundance), and classification (pest presence or absence) via the random forest algorithm. The data showed no spatial dependence, making using geostatistical interpolators inappropriate. However, a multi-objective optimized sampling design, tailored to refine configurations for identifying and estimating variograms and spatial trends essential for spatial interpolation, produced the most accurate predictions. Therefore, a two-phase sample optimization with prior in situ selection of environmental covariates improves pest predictions in agricultural systems, contributing to more efficient and sustainable agricultural management.

Keywords: ecology; precision agriculture; remote sensing; site-specific management; integrated pest management

check for updates

Academic Editor: Fatwa Ramdani

Received: 13 December 2024
Revised: 9 January 2025
Accepted: 14 January 2025
Published: 18 January 2025

Citation: Martins, C.L.; Pusch, M.; Godoy, W.A.C.; Amaral, L.R.d. Environmental Covariates for Sampling Optimization and Pest Prediction in Soybean Crops. *AgriEngineering* **2025**, *7*, 21. https://doi.org/10.3390/agriengineering7010021

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

The absence of efficient pest monitoring methods limits the implementation of site-specific management practices in agriculture. Visual scouting is the most common procedure for sampling low-mobility insects such as caterpillars and stink bug nymphs in annual crops. However, visual scouting is costly and challenging for large areas. Due to the limitations of scouting methods, pest control is traditionally based on the average infestation levels or prophylactic insecticide use throughout the area. This increases the production costs, reduces the quality of agricultural products, and causes environmental contamination.

Integrated pest management (IPM) and precision agriculture (PA) offer alternatives to traditional pest monitoring and prophylactic insecticide use. IPM, based on ecological principles, is a set of management strategies that propose monitoring the pest population levels to prevent them from reaching limits that could economically damage agricultural production [1,2]. In PA, using tools for the monitoring, mapping, and site-specific management (SSM) of various factors influencing agricultural production, such as soil conditions, pest populations, and irrigation needs, promotes resource efficiency, contributing to agroecosystem sustainability [3–5]. In the context of PA, research has explored sample optimization and digital soil mapping. These strategies use auxiliary variables (covariates) to guide soil sampling and mapping, improving the cost-effectiveness and accuracy of spatial estimates [6,7]. For example, topography, climatic factors, vegetation indices, and the soil parent material serve as covariates for mapping the soil chemical and physical properties [7,8].

Similar to soil, pest distribution can exhibit significant spatial and temporal variability within agricultural areas [9]. As ectothermic organisms, the body temperature of insects is regulated by the environmental temperature [10,11], making temperature one of the most influential variables in their development and life cycle, and has profound ecological, physiological, and molecular effects on these organisms [12]. Microclimate variables not only influence the distribution, abundance, and diversity of insects, but also directly impact primary production and the availability of their food resources [13,14]. Consequently, insects inhabit climatic zones promoting rapid development, high reproduction rates, and low mortality [15]. Thus, understanding the ecology of insects in agroecosystems is crucial for effective monitoring and management.

Research into the spatial distribution of insects in response to microclimate variability often faces significant challenges due to difficulties in obtaining high-resolution microclimate data and the limitations of meteorological satellite data, which usually lack the spatial resolution required for crop-level studies [16,17]. However, the temperature variability can vary due to environmental covariates. For example, temperature varies due to vegetation heterogeneity, soil [18], wind, cloud cover [16], and terrain topography [19]. Additionally, riparian areas (i.e., regions adjacent to water resources such as rivers and lakes) are zones of high biodiversity. They influence soil and air moisture, impact local temperature, and affect the abundance of beneficial and non-beneficial insects [20–22]. This study, therefore, posits that these environmental covariates in agricultural areas can be proxies for temperature variability and food availability for insects, enabling inferences regarding insect pest infestations. Understanding how these covariates relate to insect abundance or presence can also support the development of optimized sampling strategies and spatial predictions for insect populations.

A single-phase optimized sampling plan that assumes correlations between covariates based on knowledge acquired in other areas may compromise the selection of covariates and the sample optimization results [23]. To enhance reliability, a two-phase sampling optimization can be employed. The first phase characterizes the relationship between the covariates and the target variable, while the second phase optimizes the sampling plan using the selected covariates [7]. Thus, this study tested whether using environmental covariates in two-phase sample optimization improved the spatial predictions for soybean insect pests.

2. Materials and Methods

This study presented sampling optimization and spatial prediction approaches for soybean insect pests using environmental covariates (Figure 1). Optimized sampling was divided into two phases. In the first phase, during the 2021–2022 crop season, the relationships between environmental covariates and insects were explored, which allowed the most relevant covariates to be selected for the next phase. In the second phase, during the 2022–2023 crop season, the selected covariates were used in a sampling optimization

algorithm to define the best locations to sample insect pests. This process resulted in two optimized sampling designs that were then compared to a regular grid. The three sampling designs were assessed using geostatistical methods, regression analyses, and classification through the random forest algorithm to predict pest spatial distributions. The prediction performances were evaluated based on modeling metrics, validation with external data, and the quality of the resulting spatial maps.

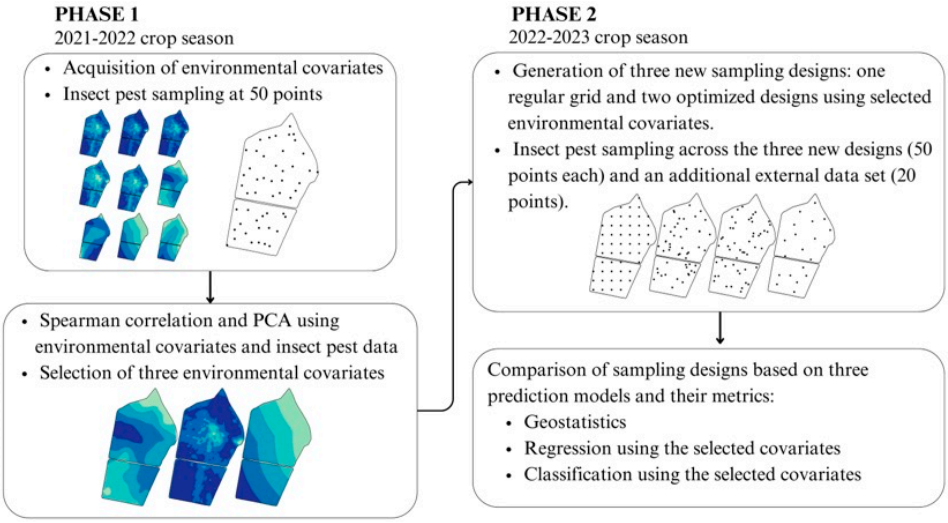

Figure 1. Diagram of the two-phase sample optimization research using environmental covariates.

2.1. Experimental Area

Research was conducted in two soybean cultivation plots totaling 26 hectares, managed under crop rotation with sorghum or oats in the fall, located in the Cosmópolis municipality (22°41′56″ S and 47°10′32″ W), São Paulo state, Brazil (Figure 2). According to Köppen, the area's climate is classified as tropical with hot summers (Cfa), with an average annual temperature of approximately 20 °C and an average annual precipitation of 1600 mm [24]. Its terrain is gently undulated. The soil is predominantly classified as clayey red latosol.

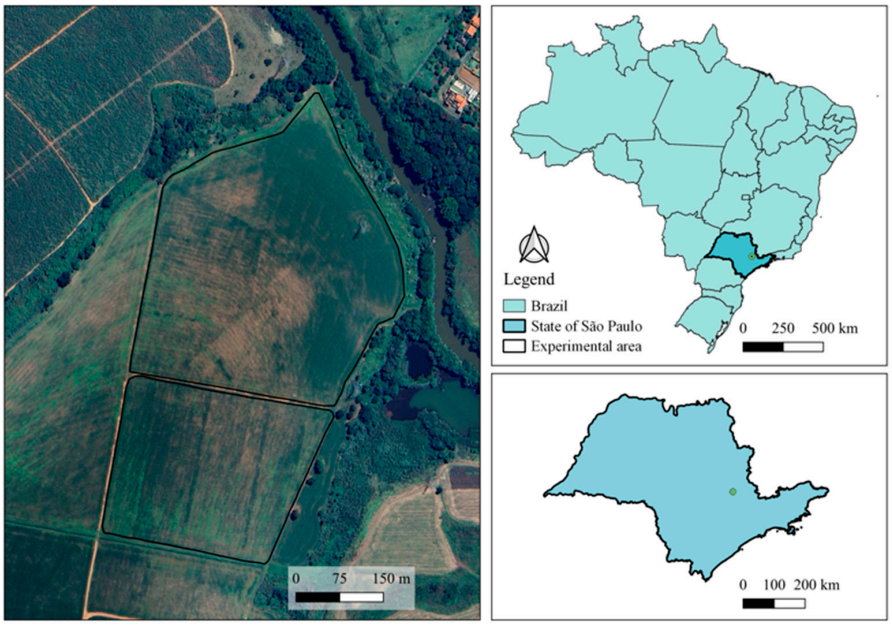

Figure 2. Location of the experimental area showing the two fields. Cartographic base: IBGE, 2023. Basemap: Google Satellite.

2.2. Crop Management

In the 2021–2022 crop season, the cultivar used was M5917 IPRO, with an indeterminate growth habit, sown on 23 November 2021. In the 2022–2023 crop season, the cultivars NS5933 IPRO and Coliseu 631X65 RSF I2X, both also with indeterminate growth habits, were sown on 5 November 2022. Crop management followed traditional farm protocols. During the 2021–2022 crop season, insecticides were applied across the entire area upon pest detection through visual scouting. In the 2022–2023 crop season, fungicides, herbicides, and insecticides were applied 15 days after plant emergence, with treatments repeated at 15-day intervals. In both crop seasons, a broad-spectrum insecticide composed of pyrethroid and carbamate was used, which is recommended for controlling *Anticarsia gemmatalis* (Hübner, 1818) (Lepidoptera: Noctuidae), *Chrysodeixis includens* (Walker, 1857) (Lepidoptera: Noctuidae), *Nezara viridula* (Linnaeus, 1758) (Hemiptera: Pentatomidae), *Piezodorus guildinii* (Westwood, 1837) (Hemiptera: Pentatomidae), *Euschistus heros* (Fabricius, 1798) (Hemiptera: Pentatomidae), and *Bemisia tabaci* (Gennadius, 1887) (Hemiptera: Aleyrodidae), race B, in soybean fields, with a recommended limit of five applications per crop cycle. In the second crop season, an organophosphate insecticide was also used that is recommended for controlling *Anticarsia gemmatalis* (Hübner, 1818) (Lepidoptera: Noctuidae), *Caliothrips phaseoli* (Hood, 1912) (Thysanoptera: Thripidae), *Epinotia aporema* (Walsingham, 1914) (Lepidoptera: Tortricidae), *Frankliniella rodeos* (Moulton, 1936) (Thysanoptera: Thripidae), *Frankliniella schultzei* (Trybom, 1910) (Thysanoptera: Thripidae), *Hedylepta indicata* (Fabricius, 1794) (Lepidoptera: Crambidae), *Trichoplusia ni* (Hübner, 1803) (Lepidoptera: Noctuidae), *Piezodorus guildinii* (Westwood, 1837) (Hemiptera: Pentatomidae), *Nezara viridula* (Linnaeus, 1758) (Hemiptera: Pentatomidae), and *Euschistus heros* (Fabricius, 1798) (Hemiptera: Pentatomidae) in soybean fields, with a recommended maximum of two applications per crop cycle. Despite the recommended application limits of five or two times per soybean cycle, insecticides were applied at 15-day intervals.

2.3. Phase 1

2.3.1. Obtaining Environmental Covariates

To investigate the relationships between the environmental covariates and insect spatial variability, data were collected on nine covariates that could reflect the environmental conditions within the area. These covariates included five vegetation indices (VIs), the soil clay content, terrain slope, and distances from rivers and riparian forest.

(a) Vegetation indices

Vegetation influences microclimates and the availability of food for herbivorous insects. Thus, vegetation architecture often influences the intensity of pest insect infestation [25]. Considering that vegetation indices provide information on vegetation variability, five vegetation indices (VIs) were tested in this research (Table 1 and Figure 3A–E). Among the VIs, there were two widely used in agricultural studies (EVI and NDVI—Figure 3A,B), two that use the red-edge band in their calculations and therefore tend to correlate more with the leaf chlorophyll content (NDRE and SFDVI—Figure 3C,D), and one that presented a simple mathematical combination of two spectral bands, which is easy to calculate and interpret (DVI—Figure 3E). For this purpose, 11 cloudless PlanetScope images were obtained throughout the crop cycle (January 15, 18, 21, 22, 24, 25; February 8, 9, 12; and March 2, 3), captured using the SuperDove sensor with a spatial resolution of 3 m. Six were acquired close to the sampling dates, with each image used to calculate the VIs for assessments related to each phenological stage. For comprehensive data evaluation throughout the cycle, the mean, median, sum, and standard deviation of the VIs from the 11 images were determined, and a single image for each vegetation index was obtained to represent the entire crop season.

Table 1. Vegetation indices: name, formula, and references.

Vegetation Index	Name	Formula	References
EVI	Enhanced vegetation index	$2\frac{NIR-R}{NIR+C1\ R-C2\ B+1}$	[26]
NDVI	Normalized difference vegetation index	$\frac{NIR-R}{NIR+R}$	[27]
NDRE	Red-edge normalized difference vegetation index	$\frac{NIR-RE}{NIR+RE}$	[28]
SFDVI	Spectral feature depth vegetation index	$\frac{\frac{NIR+G}{2}}{\frac{R+RE}{2}}$	[29]
DVI	Difference vegetation index	$NIR - R$	[30]

Spectral bands: NIR = near infrared; R = red; RE = red edge; G = green, and B = blue. EVI: C1 and C2 = aerosol influence adjustment coefficients (C1 = 6; C2 = 7.5).

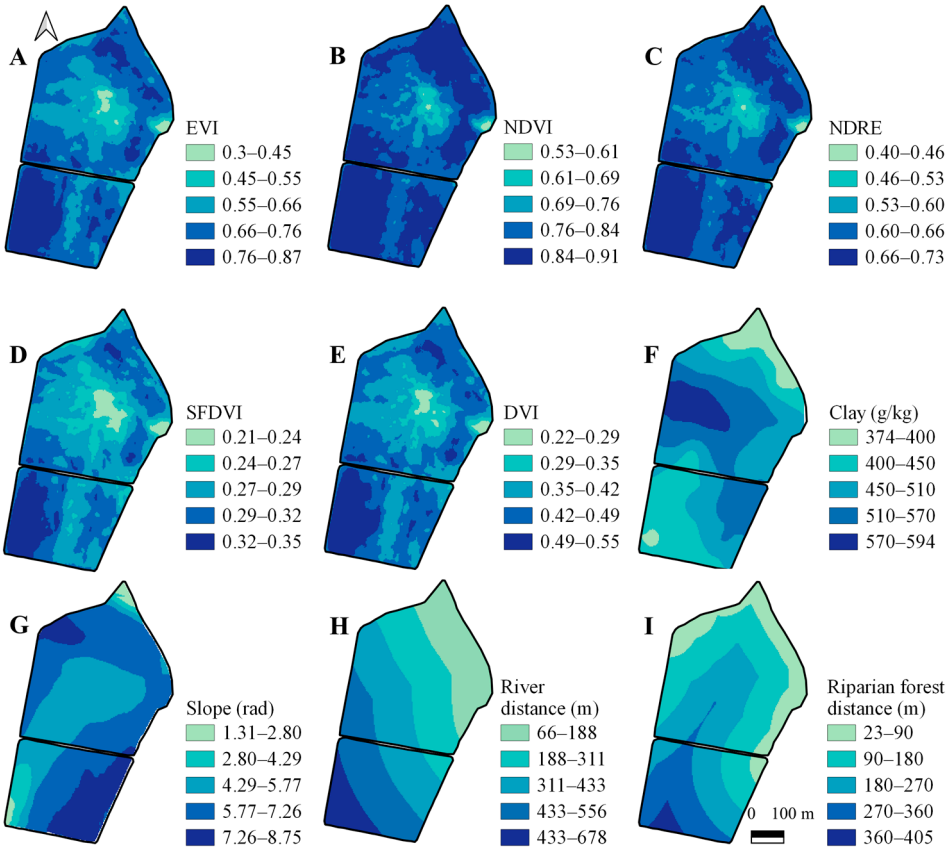

Figure 3. Environmental covariates. Vegetation indices (cycle image): EVI, NDVI, NDRE, SFDVI, and DVI ((**A**–**E**), respectively); soil clay content (**F**); slope (**G**); river distance (**H**), and riparian forest distance (**I**).

(b) Soil and Terrain topography

The soil clay content and terrain slope data (Figure 3F,G) were included as covariates, as they can be related to soil moisture [31] and influence vegetation development and the microclimate under the canopy.

Soil samples were collected from 67 georeferenced sampling points, based on an optimized MSSD (minimizing mean squared shortest distance) sampling design previously developed by Pusch et al. (2023) [23]. The soil samples were sent to a commercial laboratory for physical property analysis. The samples were dried at 40 °C and sieved through a 2 mm mesh to obtain fine air-dried soil for texture determination. The particles were separated using the chemical dispersion method, in which sodium hexametaphosphate was added

to the sample to break up the particle aggregates and isolate the individual fractions, according to the Brazilian Soil Classification System (SiBICS), described by Santos et al. (2018) [32]. The soil was then categorized into textural fractions: sand (>0.053 mm, in g/kg), clay (<0.002 mm, in g/kg), and silt (values between clay and sand). The soil clay content distribution map was generated using ordinary kriging interpolation. Among the 67 sampling points, 28 points were located within the evaluated plots. The variogram was modeled by testing spherical, exponential, and Gaussian models using the SmartMap plugin [33] in QGIS software, version 3.28.13. The spherical model was selected to achieve the best cross-validation performance, with a root mean square error (RMSE) of 52.18 and a coefficient of determination (R^2) of 0.68.

Terrain slope was calculated by densely collecting elevation data across the entire area using a global navigation satellite system (GNSS) receiver with differential correction mounted on the harvester. These data were interpolated using ordinary kriging following the same protocol used for the clay content. However, a 10-fold cross-validation was implemented to validate the interpolation model due to the high data density (626 points per hectare). From the elevation map, the terrain slope was calculated in radians (rad) using the RSAGA package [34] in R software, version 4.3.2.

(c) Riparian areas

Riparian areas significantly impact insect habitats. Proximity to rivers influences the soil and air moisture levels. In addition to influencing moisture, riparian forests affect wind speed, contributing to unique microclimate conditions nearby [35]. These areas also serve as biodiversity sources for beneficial and non-beneficial insects, potentially influencing pest infestations in the surrounding areas [22,36]. To determine the distances from the river and riparian forest, a raster layer with a spatial resolution of 3 m was created to cover the entire experimental area. A vector dataset was also generated to represent points along each feature (river or riparian forest). The distance for each pixel in the raster was then calculated based on its proximity to the nearest point in the vector dataset (Figure 3H,I).

2.3.2. Insect Pest Sampling

In the 2021–2022 crop season, insect pest infestations were monitored at 50 georeferenced points distributed within the area (Figure 4A). Of these, 28 points had the same location as the data used to generate the soil clay content map obtained from an optimized soil sampling project. This allowed for pest sampling at the same points with the available soil data, providing better local information than only using the interpolated soil data. To complete the 50 sampling points, 22 points were randomly generated.

Samplings were performed during the soybean phenological stages R2 (50 days after sowing–DAS), R3 (57 DAS), R4 (64 DAS), R5 (77 and 84 DAS), and R7 (99 DAS), following the phenological scale proposed by Fehr and Caviness (1977) [37]. Pest sampling used the beat cloth method, in which a cloth measuring 1 × 1 m was placed between two soybean rows, and plants from one row were shaken onto the beat cloth. Insects on the cloth were identified at the species or genus level and then counted. Each sample consisted of three sub-samples collected within a 3 m radius from the sampling point, mirroring the pixel size in the satellite images used in the study (PlanetScope).

For each sampling date, the abundance of each insect species and the total pests were obtained for each sampling point. In addition, data on the abundances' mean, median, sum, and standard deviation were determined for each sampling point throughout the crop cycle.

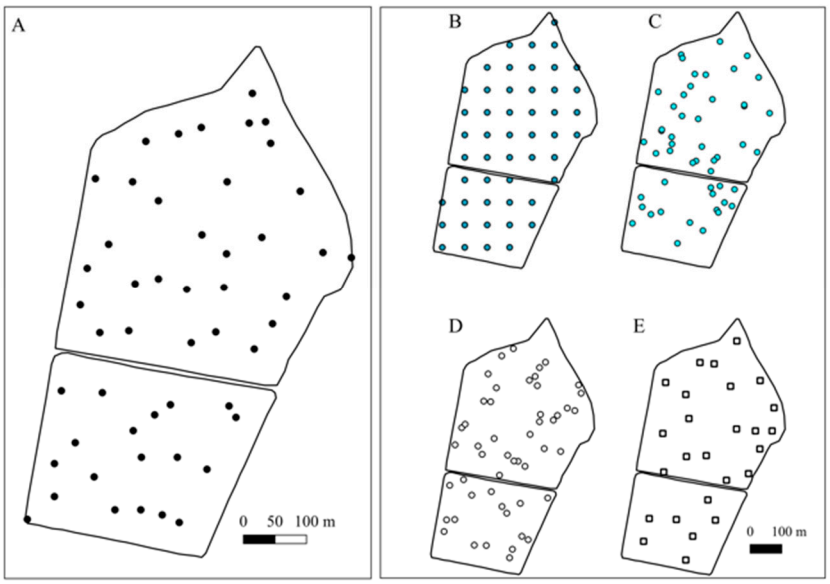

Figure 4. Sampling designs in Phase 1 (**A**) and Phase 2 (**B**–**E**). (**A**) Phase 1: 28 optimized MSSD sampling points combined with 22 random points (50 points). Phase 2: Regular grid (**B**), optimized CORR design (**C**), optimized SPAN design (**D**), each with 50 points, and external dataset (20 points).

2.3.3. Data Analyses

The insect pest sampling data were subjected to the Shapiro–Wilk normality test. Due to the non-normal distribution, a Spearman correlation analysis was performed using the PerformanceAnalytics package [38] in R software, version 4.3.2. Correlations between the species abundance, total pest abundance, and covariates were evaluated individually at each phenological stage and throughout the crop cycle. For the crop cycle, correlations of covariates were analyzed with the mean, median, sum, and standard deviation of each pest species and the total pests present at each sampling point. After comparing the correlations across different pest statistics and covariates, further analyses focused on the median values of pest species and total pests and the mean vegetation indices from 11 images taken throughout the soybean cycle (Figure 3A–E).

Aside from correlation analysis, a principal component analysis (PCA) was performed to select covariates for the sampling optimization algorithm in the second phase. In the PCA, data from the covariates (mean of VIs) and the median of total pests throughout the soybean cycle were used, employing the dudi.pca function from the ade4 package [39] in R software, version 4.3.2. First, principal components explaining at least 70% of the total data variance were identified. Subsequently, the contributions of each original variable to these principal components were evaluated based on their loadings. Following this analysis, the soil clay content, NDVI, and river distance were selected for the next research phase.

2.4. Phase 2

2.4.1. Sampling Designs

In the second phase, the 2022–2023 crop season, three sampling designs, each containing 50 georeferenced points, were generated: a regular grid with two samples/ha (71 × 71 m), and two optimized sampling designs. Optimization employed the iterative algorithm optimization of sample configurations using spatial simulated annealing (SPSANN) [40] in R software. SPSANN is a variation of the spatial simulated annealing (SSA) algorithm used to find optimal solutions for sampling optimization problems [41]. During the algorithm's execution, sampling points are adjusted through random perturbations to explore potential solutions for optimization. After each perturbation, the objective function is recalculated to assess the "energy" of the new sampling configuration. The objective

function minimizes this energy, refining the configuration to achieve an optimal distribution aligned with the sampling design objectives. Simulated annealing (SA) employs a cooling schedule to facilitate this, which prevents the algorithm from becoming trapped in local optima. The cooling process across iterations gradually restricts the movement of points, thereby guiding the algorithm toward a more precise solution. Schedule parameters are fine-tuned through trial and error, with the initial temperature enabling broad exploration of the search space. At each iteration, points are adjusted, and the objective function is evaluated to determine whether the new configuration should be accepted. The algorithm continues iterating until achieving the configuration with the lowest energy and no further changes in the distribution of the sampling points, thereby reaching the best configuration.

As interactions between environmental covariates influence insect pest locations, one of the sampling optimization criteria used the CORR objective function to replicate the bivariate correlation between covariates [40]. Points were allocated based on correlations between the population (covariates) and the samples (from the population) in a specific row and column of the correlation matrix with dimension p (number of covariates). The second optimized sampling design used the multi-objective function optimization of sampling locations for variogram calculations (SPAN), which optimizes the sampling configuration to identify and estimate variograms and spatial trends for spatial interpolation [40]. SPAN optimization criteria include the (a) distribution of variables (DIST); (b) correlation between variables (CORR); (c) minimizing the mean squared shortest distance (MSSD); and (d) optimizing sampling for variogram identification and estimation (PPL). Each optimization criterion in this sampling design received equal weighting. In addition, insect pest data were collected at 20 external points on each sampling date for validation of the results. These points were randomly generated.

2.4.2. Vegetation Index

Since the vegetation index is the only covariate that changes between crop seasons, the selected VI (NDVI) was calculated for the second season for pest prediction. However, due to heavy cloud cover, it was not possible to obtain cloudless images for the third sampling date. To compensate, three images were used: two from January (26 and 27) and one from February 25, while the sampling took place on February 8 and 9 (R5–94 and 95 DAS). The VI was calculated for each image and then averaged to represent the sampling date.

2.4.3. Insect Pest Sampling

Insect pest monitoring was conducted during the phenological stages R2 (59 and 60 DAS), R5 (80 and 81 DAS), and R5 (94 and 95 DAS). The 150 points from the three sampling designs and 20 external data points were collected over two consecutive days, with each plot sampled on a separate day (Figure 4B–E). As in the previous crop season, the beat cloth method was used for samplings. Due to the low pest incidence, only the data from the third sampling were utilized for analyses in the second phase of the research.

2.4.4. Data Analyses

Three methods, geostatistics, regression, and classification, were evaluated using the three sampling designs to predict the total pests and the most abundant species (*Euschistus heros*) in the second crop season. In geostatistics, the spatial dependence was initially assessed using Moran's index, and its statistical significance was assessed using the SmartMap plugin in QGIS software, version 3.28.13 [33]. Variograms and their parameters were analyzed in the spherical, exponential, and Gaussian models. The best-fitted models were selected based on cross-validation metrics, R^2, and RMSE.

The effectiveness of environmental covariates in the prediction stage was assessed using the covariates selected in Phase 1 for the predictions with regression and classification

models using the random forest (RF) algorithm. RF is a machine learning approach that combines predictions from multiple individual decision trees to identify linear and nonlinear patterns in data, providing a comprehensive analysis of relevant variables in the mapping. Regression and classification RF analyses were conducted using the mlr package [42] in R software, version 4.3.2, with hyperparameter optimization.

The regression analysis used covariates to predict the abundance of total pests and the most abundant pest species. Data were randomly split for modeling, with 80% used for model training and 20% for testing. During hyperparameter optimization, the candidate range for the number of trees (ntree) was set from 100 to 1000; the number of random variables (mtry) used at each tree node split ranged from 1 to 10; and the minimum number of samples required in a terminal node (nodesize) ranged from 2 to 5. A 5-fold cross-validation, in which the dataset was divided into five parts, was applied during hyperparameter optimization to assess the performance of the different parameter sets. After optimization, the RF model was trained on the complete training data with the best-adjusted hyperparameters. RMSE and R^2 were the metrics used to evaluate the prediction accuracy. The models were used to predict the insect pest distribution maps, which were externally validated using observed data from 20 independent points. Predicted values at these locations were compared with the observed data using R^2 and RMSE.

The classification analysis used the covariates selected in Phase 1 to predict classes associated with insect pest absence (0) and presence (1). For effective modeling, a substantial number of observations in the two classes (0 e 1) is essential, even before applying the class balance techniques. In this case, the data on the total pests were unsuitable for classification due to the scarcity of the absence class. Consequently, the classification was focused on predicting the presence of the most abundant species during the growing season (*Euschistus heros*). Initially, class balancing was implemented using the synthetic minority over-sampling technique (SMOTE) to reduce the model's bias toward the majority class. This technique creates synthetic examples of the minority class by generating weighted combinations of neighboring samples [43]. Subsequently, these data were randomly divided into training (70%) and test (30%) sets. To select the optimal hyperparameters, we conducted optimization using the random search method (1000 iterations). Hyperparameters included the number of trees, tested in the range of 100 to 3000 (ntree), the number of variables selected at each split, ranging from 1 to 14 (mtry), and the minimum number of samples required in a terminal node, ranging from 2 to 30 (nodesize). Hyperparameter evaluation was performed using 10-fold cross-validation on the training set, with the highest accuracy considered optimal. After optimization, the RF model was trained on the full training dataset using the best-tuned hyperparameters. Subsequently, confusion matrices were generated using the test data to evaluate the models. The models generated insect pest distribution maps, which were externally validated using observed data from 20 independent points. The predicted values at these locations were compared with observed data using the confusion matrices and their performance metrics.

The following metrics were calculated from the confusion matrices: accuracy, precision, specificity, recall (sensitivity), and F1 score. Accuracy represents the percentage of correct predictions (true positives and true negatives) relative to the total number of predictions made by the model (Equation (1)). Precision represents the proportion of true positive predictions relative to the total number of positive predictions (true positives and false positives) (Equation (2)). Specificity measures the proportion of true negatives relative to the total number of actual negative cases (Equation (3)). Recall quantifies the percentage of correct positive predictions that the model makes relative to the total number of positive

cases (Equation (4)). Finally, the F1 score combines precision and sensitivity into a single score, indicating the balance between these metrics (Equation (5)).

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FP} + \text{FN} + \text{TP})} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (4)$$

$$\text{F1 Score} = 2\frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

where TN—true negative; TP—true positive; FN—false negative; and FP—false positive.

3. Results

Both crop seasons showed the presence of *Spodoptera* spp. (Lepidoptera: Noctuidae) and looper caterpillars, potentially including species such as *Rachiplusia nu* (Guenée, 1852) (Lepidoptera: Noctuidae), *Trichoplusia ni* (Hübner, 1800–1803) (Lepidoptera: Noctuidae), and *Chrysodeixis includens* (Walker, 1857) (Lepidoptera: Noctuidae). Due to the difficulty of visually differentiating these species, this study used the popular name "looper caterpillar" to refer to any species in this group. Stink bugs identified included *Euschistus heros* (Fabricius, 1798) (Hemiptera: Pentatomidae), *Nezara viridula* (Linnaeus, 1758) (Hemiptera: Pentatomidae), *Dichelops* spp. (Hemiptera: Pentatomidae), and *Arvelius albopunctatus* (De Geer, 1773) (Hemiptera: Pentatomidae). Additionally, beetle species observed were *Lagria villosa* (Fabricius, 1781) (Coleoptera: Tenebrionidae), *Blapstinus punctulatus* (Solier, 1849) (Coleoptera: Tenebrionidae), *Aracanthus murei* (Marshall, 1958) (Coleoptera: Curculionidae), *Diabrotica speciosa* (Germar, 1824) (Coleoptera: Chrysomelidae), *Cerotoma arcuata* (Olivier, 1791) (Coleoptera: Chrysomelidae), and *Colaspis* sp. (Coleoptera: Chrysomelidae). The latter three species, which are defoliators, were evaluated together and are referred to here by their family name, Chrysomelidae. Overall, the most abundant species in Phase 1 was the looper caterpillar, whereas *Euschistus heros* predominated in Phase 2, showing not only adult insects, but also a high incidence of nymphs.

3.1. Phase 1

3.1.1. Correlation Between Insect Pests and Environmental Covariates

(a) Phenological stages

The insect pest species infestations showed varied correlations with the environmental covariates across phenological stages. The most abundant species throughout the cycle, looper caterpillar and *Spodoptera* spp., presented greater correlation with the covariates, especially with river and riparian forest distances (Table 2). When assessing the total pests at the sampling points, these correlations also varied. At R5 (77 DAS), when sampling took place two days after insecticide application, only the *E. heros* species showed correlation, and only with distance from riparian forest. Conversely, at 84 DAS, we again observed correlations between some species and covariates.

Table 2. Spearman's correlation between environmental covariates and insect pests during soybean phenological stages R2 to R7.

	EVI	NDVI	NDRE	SFDVI	DVI	Clay	Slope	River_dist	Forest_dist
R2–50 DAS									
Looper	0.32 *	0.33 *	0.34 *	0.29 *	0.30 *	−0.24	−0.09	0.41 *	0.40 *
Spodoptera	0.22	0.23	0.22	0.2	0.20	−0.17	−0.23	0.22	0.19
Euschistus	−0.12	−0.13	−0.17	−0.08	−0.12	0.21	0.07	0.10	−0.08
Dichelops	−0.05	−0.08	−0.05	−0.03	−0.06	0.09	−0.03	−0.25 *	−0.15
Chrysomelidae	−0.03	−0.06	−0.1	0.03	0	0.12	0.01	−0.01	−0.12
Lagria	0.28 *	0.22	0.21	0.31 *	0.29 *	−0.27 *	−0.15	−0.07	−0.13
Aracanthus	−0.11	−0.09	−0.09	−0.14	−0.12	−0.07	−0.06	−0.22	−0.14
Total pests	0.30 *	0.26 *	0.26 *	0.32 *	0.29 *	−0.25 *	−0.14	0	−0.03
R3–57 DAS									
Looper	0.11	0.13	0.1	0.14	0.14	−0.26 *	−0.20	0.47 *	0.50 *
Spodoptera	0.34 *	0.37 *	0.34 *	0.30 *	0.34 *	−0.21	0.01	0.35 *	0.35 *
Nezara	0.17	0.19	0.2	0.10	0.14	−0.28 *	0.07	−0.01	0.02
Dichelops	−0.12	−0.12	−0.11	−0.13	−0.11	0.15	−0.04	−0.14	−0.21
Chrysomelidae	−0.21	−0.21	−0.20	−0.16	−0.20	0.25 *	−0.047	0.05	−0.08
Lagria	0.00	−0.02	−0.02	−0.04	−0.02	−0.02	0.07	0.06	−0.01
Aracanthus	0.04	0.045	0.08	−0.06	0.01	−0.20	0.01	−0.32 *	−0.30 *
Total pests	0.17	0.15	0.14	0.12	0.16	−0.23	−0.07	0.35 *	0.27 *
R4–64 DAS									
Looper	0.16	0.16	0.13	0.23	0.2	−0.21	−0.063	0.26 *	0.11
Spodoptera	0.38 *	0.35 *	0.35 *	0.40 *	0.38 *	−0.12	0.1	0.43 *	0.29 *
Euschistus	0.18	0.19	0.17	0.16	0.2	−0.04	−0.18	0.09	0.12
Nezara	0.23	0.24	0.22	0.24	0.23	−0.13	0.06	0.19	0.23
Dichelops	−0.14	−0.13	−0.13	−0.22	−0.16	0.18	−0.06	−0.12	0.06
Chrysomelidae	−0.23	−0.22	−0.23	−0.23	−0.21	0.22	−0.21	−0.05	−0.01
Lagria	0.06	0.08	0.06	0.07	0.07	0.14	0.014	−0.016	−0.07
Arvelius	0.11	0.11	0.09	0.08	0.09	−0.33 *	−0.17	−0.34 *	−0.31 *
Total pests	0.19	0.2	0.17	0.22	0.22	−0.07	−0.04	0.17	0.09
R5–77 DAS									
Looper	0.22	0.11	0.22	0.22	0.22	−0.12	0.11	−0.06	−0.05
Spodoptera	0.18	0.16	0.17	0.2	0.2	−0.02	0.13	−0.02	−0.12
Euschistus	0.08	0.03	0.06	0.1	0.09	0.16	0.16	−0.15	−0.34 *
Dichelops	−0.10	−0.13	−0.09	−0.10	−0.10	0.04	0.12	0.02	0.02
Chrysomelidae	−0.07	−0.06	−0.11	−0.05	−0.08	−0.08	−0.014	0.23	0.19
Lagria	−0.07	−0.09	−0.16	−0.08	−0.08	0.06	−0.12	0.06	0.09
Blapstinus	0.12	0.06	0.07	0.13	0.12	−0.10	−0.11	−0.14	−0.16
Total pests	0.05	−0.04	−0.01	0.07	0.06	−0.17	−0.019	−0.20	−0.24 *
R5–84 DAS									
Looper	0.08	0.13	−0.04	0.12	0.09	0.13	0.31 *	0.48 *	0.37 *
Spodoptera	−0.06	−0.17	−0.05	0	−0.04	0	−0.08	0.03	−0.20
Euschistus	0.04	0.04	0.05	0.03	0.06	0.08	0.03	0.07	−0.08
Dichelops	−0.01	−0.15	0	0.02	0.03	−0.15	0.09	−0.21	−0.18
Chrysomelidae	0.06	0.18	0.11	0.02	0.03	0.08	−0.05	−0.08	−0.03
Lagria	−0.01	−0.06	−0.08	−0.02	0.01	0.11	0.22	−0.12	−0.17
Blapstinus	0.29 *	0.2	0.25 *	0.31 *	0.30 *	−0.21	−0.10	0.14	0.21
Total pests	0.2	0.12	0.12	0.27 *	0.24 *	−0.06	0.20	0.37 *	0.18

Table 2. *Cont.*

	EVI	NDVI	NDRE	SFDVI	DVI	Clay	Slope	River_dist	Forest_dist
					R7–99 DAS				
Looper	−0.09	−0.04	0	−0.09	−0.06	0.23	0.20	0.04	0.02
Spodoptera	0.16	0.19	0.22	0.10	0.15	0.06	0.11	0.15	−0.07
Euschistus	0.22	0.26 *	0.31 *	0.16	0.24	−0.15	−0.15	0.43 *	0.44 *
Nezara	0.15	0.12	0.11	0.27 *	0.17	−0.23	−0.40 *	0.35 *	0.31 *
Dichelops	−0.14	−0.12	−0.13	−0.15	−0.14	0.13	0.01	−0.03	−0.14
Chrysomelidae	0.16	0.15	0.18	0.09	0.15	0.05	−0.01	0.13	0.25
Lagria	0.15	0.15	0.10	0.18	0.17	0.01	0.09	0.08	0.15
Blapstinus	0.03	0.02	0.05	0.11	0.05	−0.38 *	0.14	−0.06	0.02
Total pests	0.19	0.24	0.28 *	−0.03	0.19	0	0.14	0.22	0.04

Looper = Looper caterpillar. * Significant at the 0.05 level.

(b) Crop cycle

Regarding infestations at sampling points throughout the crop cycle (median per sampling point of each species and total species), a correlation between covariates and pest species infestations was found (Table 3). Looper caterpillar showed the highest correlation with the covariates, particularly with river and riparian forest distances, similar to observations made at each phenological stage.

Table 3. Spearman's correlation between environmental covariates and the median number of species per point and the total pest median in the soybean cycle.

	EVI	NDVI	NDRE	SFDVI	DVI	Clay	Slope	River_dist	Forest_dist
Looper	0.27	0.28 *	0.27	0.29 *	0.28 *	−0.26 *	−0.11	0.50 *	0.43 *
Spodoptera	0.26	0.22	0.21	0.27	0.26	−0.05	−0.02	0.17	−0.00
Euschistus	−0.20	−0.25	−0.24	−0.20	−0.21	0.28 *	−0.06	0.01	−0.08
Nezara	0.22	0.23	0.23	0.22	0.23	−0.13	0.25 *	0.19	0.23
Aracanthus	−0.15	−0.07	−0.08	−0.20	−0.17	−0.08	−0.15	−0.32 *	−0.24
Lagria	−0.04	−0.05	−0.04	−0.04	−0.03	0.14	−0.05	−0.12	−0.13
Blapstinus	0.23	0.17	0.20	0.22	0.24	−0.22	0.11	−0.22	−0.12
Total pests	0.32 *	0.38 *	0.35 *	0.31 *	0.32 *	−0.34 *	−0.07	0.34 *	0.29 *

Looper = Looper caterpillar. * Significant at the 0.05 level.

3.1.2. Selection of Environmental Covariates

Two principal components explained 75% of the variance in the data (Figure 5). In the first principal component (PC1), clay content and vegetation indices (VIs) contributed the most significantly. Distances from the river and riparian forest were the primary contributors to the second principal component (PC2). Based on these results, the clay content was selected for the next research stage. Given the similar contributions of the vegetation indices in PC1, NDVI was included due to its broad applicability and highest correlation with the total number of pests (Table 3). Distance from the river was chosen as one of the most important variables in PC2 because of its stronger correlation with pest infestations compared to the distance from the riparian forest (Table 3). Regarding covariates, there was an inverse relationship between the VIs and soil clay content. The slope was inversely proportional to the distance from the river and riparian forest.

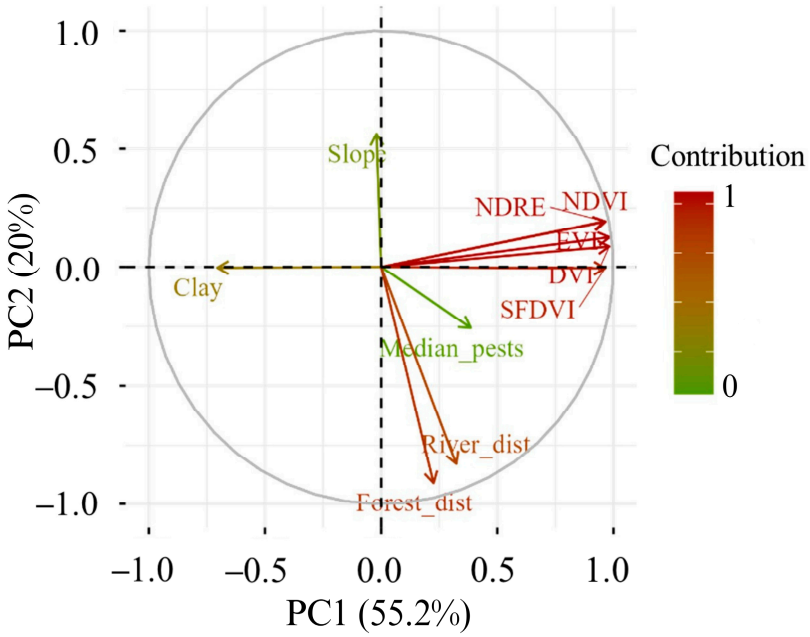

Figure 5. PCA results using the median of pests and the mean of the VIs in the soybean cycle.

3.2. Phase 2

3.2.1. Geostatistical Modeling

The three different sampling designs failed to capture spatial dependence in the data corresponding to the total pests or in the data of the most abundant species during the 2022–2023 crop season, *E. heros* (Table 4). Thus, the data proved inadequate for geostatistical interpolation. Despite this limitation, SPAN optimization showed the lowest errors measured by cross-validation during variogram modeling compared with the other designs.

Table 4. Spatial dependence measured by Moran's index (MI) and its significance (*p*); variogram parameters: a (range), c0 (nugget effect), c1 (sill), and c0 + c1 (partial sill); and cross-validation metrics: R^2 (coefficient of determination), RMSE (root mean square error), and best-fitted model.

	Sampling Designs	MI	*p*	*a*	c0	c1	c0 + c1	RMSE	R^2	Model
Total pests	Regular	0.61	0.07	181.52	7.35	5.89	13.24	3.25	0.14	Gaussian
	CORR	0.29	0.37	163.19	7.67	0.54	8.21	3.02	0.00	Spherical
	SPAN	0.49	0.01	384.34	6.41	2.81	9.22	2.82	0.13	Gaussian
Euschistus	Regular	0.58	0.15	280.33	6.18	4.78	10.96	3.20	0.01	Spherical
	CORR	0.44	0.04	257.31	2.03	3.84	5.87	2.46	0.04	Spherical
	SPAN	0.43	0.05	384.34	3.65	1.20	4.84	2.07	0.06	Gaussian

3.2.2. Random Forest Regression

Predictions made using RF regression for total pests showed the best model fits (Figure 6A–C) compared with the predictions specifically for *E. heros* (Figure 6D–F). In the first scenario, the SPAN sampling design presented lower error and higher R^2 values, followed by the regular and CORR designs (Figure 6A–C). Regarding the prediction for *E. heros*, the regular and CORR designs showed high errors and R^2 values close to zero. Despite the lower error and higher R^2, prediction with the SPAN design was not satisfactory in generating an infestation map for this species (Figure 6D–F).

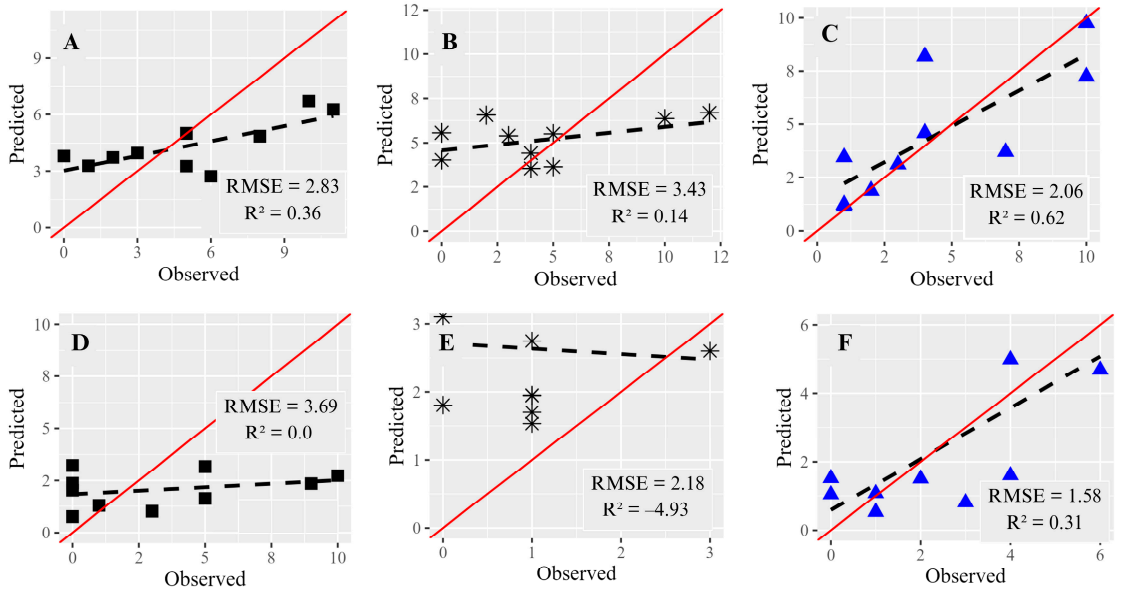

Figure 6. Scatter plots and metrics of RF regression modeling using environmental covariates in predicting the total pests using the regular (squares), CORR (stars), and SPAN (triangles) sampling designs ((**A**–**C**), respectively), and the prediction of *E. heros* in the same sampling designs (**D**–**F**). The red line represents the 1:1 ideal relationship (observed = predicted), while the dashed line indicates the regression line of the model predictions.

The validation results using external data (Figure 7A–C) were consistent with the modeling results for the total pest predictions (Figure 6A–C). The SPAN sampling design demonstrated the lowest errors and the highest R^2 values, outperforming both the regular and CORR designs.

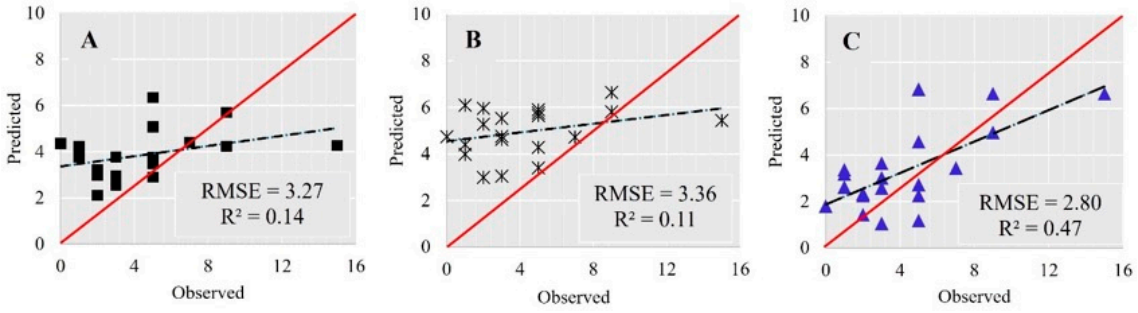

Figure 7. Scatter plots and external validation metrics for the total pest predictions using the regular (squares), CORR (stars), and SPAN (triangles) sampling designs ((**A**–**C**), respectively). The red line represents the 1:1 ideal relationship (observed = predicted), while the dashed line indicates the regression line of the model predictions.

3.2.3. Random Forest Classification

In classifying the presence and absence of *E. heros*, the SPAN sampling design showed fewer errors, achieving superior performance metrics (Tables 5 and 6). The optimized designs (SPAN and CORR) made more errors in predicting the presence of pests where they were absent (FP) than in predicting absence where pests were present (FN). The regular grid showed an equal number of FP and FN, which resulted in more errors overall, and consequently the worst metrics.

Table 5. Confusion matrix. RF classification modeling for predicting the presence–absence of *E. heros*, using the regular, CORR, and SPAN sampling designs. FN—false negative, FP—false positive, TN—true negative, and TP—true positive.

		Regular		**CORR**		**SPAN**	
		Predicted					
		Absence (0)	**Presence (1)**	**Absence (0)**	**Presence (1)**	**Absence (0)**	**Presence (1)**
Observed	Absence (0)	5 (TN)	4 (FP)	4 (TN)	4 (FP)	7 (TN)	2 (FP)
	Presence (1)	4 (FN)	8 (TP)	2 (FN)	7 (TP)	1 (FN)	11 (TP)
	Errors %	38.1		35.3		14.3	

Table 6. Metrics from the confusion matrix. RF classification modeling for predicting the presence–absence of *E. heros* using the regular, CORR, and SPAN sampling designs.

	Accuracy	**Precision**	**Specificity**	**Recall**	**F1 Score**
Regular	0.62	0.56	0.67	0.56	0.56
CORR	0.65	0.67	0.78	0.50	0.57
SPAN	0.86	0.88	0.92	0.78	0.82

In the external validation results (Tables 7 and 8), the findings differed from those of the modeling phase (Tables 5 and 6). In this case, the regular design demonstrated the fewest errors and the best metrics for classifying the presence and absence of *E. heros*. Although the CORR design achieved an accuracy comparable to the SPAN design, it showed the weakest performance across all evaluation metrics. The SPAN design exhibited more FN, while the CORR design had more FP.

Table 7. Confusion matrix. External validation metrics for *E. heros* presence–absence classification using the regular, CORR, and SPAN sampling designs. FN—false negative, FP—false positive, TN—true negative, and TP—true positive.

		Regular		**CORR**		**SPAN**	
		Predicted					
		Absence (0)	**Presence (1)**	**Absence (0)**	**Presence (1)**	**Absence (0)**	**Presence (1)**
Observed	Absence (0)	5 (TN)	2 (FP)	1 (TN)	6 (FP)	6 (TN)	1 (FP)
	Presence (1)	6 (FN)	7 (TP)	4 (FN)	9 (TP)	9 (FN)	4 (TP)
	Errors %	40		50		50	

Table 8. Metrics from the confusion matrix. External validation metrics for *E. heros* presence–absence classification using the regular, CORR, and SPAN sampling designs.

	Accuracy	**Precision**	**Specificity**	**Recall**	**F1 Score**
Regular	0.60	0.45	0.54	0.71	0.55
CORR	0.50	0.20	0.69	0.14	0.16
SPAN	0.50	0.40	0.31	0.86	0.54

4. Discussion

Among the main species found by the samplings, the looper caterpillar stood out, which is economically significant for soybeans, tobacco, alfalfa, and sunflower crops in South America due to its defoliating nature during the juvenile stage [44]. From the same family, the *Spodoptera* genus also produces defoliator caterpillars, which are highly relevant in corn, soybean, and cotton crops. In Brazil, species within this genus include *S.*

albula, *S. cosmioides*, *S. eridania*, and *S. frugiperda*. During the soybean reproductive phase, some of these species can cause direct damage to soybean pods in addition to defoliation. Species within this genus have shown resistance to *Bt* transgenics [45,46], one of the leading technologies used for caterpillar control in major crops, complicating their management.

The species *E. heros*, *N. viridula*, and *Dichelops* spp. are the most common among the stink bug species belonging to the family Pentatomidae recorded in soybean cultivation areas in Brazil. In their juvenile (nymph) and adult stages, they feed by inserting stylets into plants and grains, thus reducing the yield and quality. As for the Coleoptera beetles found in the samplings, the main species were *C. arcuata* and *D. speciosa*, defoliators of the family Chrysomelidae. Both are polyphagous and *D. speciosa*, and in addition to feeding on leaves in their adult stage, they attack the root system in its immature phase.

It is important to note that the insecticide applications were more intensive during the second crop season, resulting in a lower incidence of pests than in first crop season. Although looper caterpillar and *E. heros* were the most abundant species in Phase 1 and Phase 2, respectively, the selection of variables included all species. Despite insects generally exhibiting similar ecological behaviors, this does not invalidate our assessment of sampling optimizations in the second crop season.

4.1. Phase 1

Environmental Covariates and Insect Pest Infestations

Species behavior and abundance significantly influence the correlation between environmental covariates and insect pest populations in the area. Among the insect species found in this research, looper caterpillar and *Spodoptera* spp. have lower mobility. When they find favorable environments, larvae with low mobility adjust their behavior for thermoregulation, thereby minimizing their mortality rate, accelerating their development [47]. In conditions of high temperature and low relative humidity, instead of seeking other places for shelter, Noctuidae larvae move to the lower, more shaded, and cooler parts of plants [48]. Conversely, they move to the upper parts of plants under low temperatures and high humidity. Hence, we might initially assume that these species would correlate more with environmental covariates because they move less horizontally (across the area) and more vertically (on the plant). However, at different phenological stages, other species were more correlated with the covariates (Table 2). Thus, the greater abundance of Noctuidae caterpillars at different stages favored the perception of their correlations with covariates compared to other species.

Among the covariates, distances from the river and riparian forest were the most correlated with pest infestations (Table 2). The influence of distance from riparian forests and rivers on insect abundance is expected because insects can use riparian zones and vegetation for reproduction, foraging, or movement corridors [20,21]. As a source of plant species biodiversity, riparian areas provide food for many animal species including pest arthropods and their natural predators. Additionally, refuge suitability for ectotherms may depend on its proximity to water and the organism's ability to tolerate water loss [49]. Habitats near water sources, even in monoculture areas, may exhibit higher insect abundance, which can then disperse toward the crop's interior [22]. On the other hand, pest insect enemies in natural areas near crops can reduce the presence of these pests [50]. This complex interplay between natural habitats and cultivated fields highlights the importance of integrated pest management strategies that consider the ecological dynamics of these environments.

The vegetation indices showed positive correlations with the defoliating looper caterpillar and *Spodoptera* spp. (R2 to R4) as well as the defoliating beetle *Lagria villosa* (R2) (Table 2). In agricultural environments, however, such a correlation is expected to be

negative (i.e., a higher pest population in areas with less vegetation). This is because remote sensing (RS) techniques allow for the identification of some physiological and morphological changes associated with pest damage to plants [51]. However, the RS detection of plant damage depends on the pest population size and damage intensity, as the pest feeding time must be sufficient to alter turgor, biomass, and photosynthetic pigments, resulting in changes to the electromagnetic radiation reflected by the plants [52,53]. In this research, the positive correlations between infestations of these pest insects and vegetation biomass can be explained by the plant's greater food availability in areas with denser vegetation, which alters the microclimates under the canopy, favoring insect population growth. Insect infestations in agricultural areas follow survival and reproduction criteria similar to population ecology in natural environments and other study scales [10]. In precision agriculture, however, VI-guided insect sampling is targeted to areas with lower VI values. The results of this research indicate, however, that such directing is not always appropriate, as this relation is not always negative. Low VI values can be associated with other factors affecting vegetation development such as plant nutrients [54], soil parameters like moisture [55], and others. In such a context, these covariates should be considered together with the VIs to increase sampling and pest prediction effectiveness.

Ecology, particularly at a microscale, demonstrates that insect population dynamics are closely linked to specific local environmental conditions such as microclimate, soil composition, and vegetation structure [56,57]. In addition to environmental factors, availability of food resources, and population growth (birth rate, death rate, presence of natural predators, and others) [58], management practices in agricultural environments will influence pest infestations and their natural enemies. All of these intrinsic and complex variables hinder drawing generalizable conclusions about the relations between environmental variables and pest infestations based solely on individual evaluations of each phenological stage. In this regard, considering data from phenological stages together with the entire cycle can provide complementary inferences. Analyzing the correlations of the median of total pests throughout the crop cycle (Table 3) and the PCA results (Figure 5) revealed a consistent pattern of pest–covariate relations. Consequently, insect pest infestations tended to be more pronounced in places with higher VIs values, lower soil clay content, and were closer to riparian areas. Thus, the correlation and interaction between selected covariates influence insect pest distribution.

4.2. Phase 2

The optimized SPAN project stood out in most scenarios compared to the regular sampling and CORR projects. At the tested sampling density, however, none of them could capture the spatial dependence between the samples (Table 4). Spatial dependence is a fundamental requirement for effectively applying kriging and other geostatistical techniques [59]. In this context, machine learning techniques (regression and classification) for insect pest prediction using environmental covariates were more appropriate.

4.2.1. Random Forest Regression

In the regression analysis (Figures 6A–F and 7A–C), the superior metrics of the SPAN design were also reflected in the predicted maps of the total pest abundance generated by the three sampling designs (Figure 8D–F). The SPAN produced more detailed and coherent spatial patterns, while the CORR design showed a pattern that indicated probable estimation errors. Unlike the CORR design, which considers only the correlation between variables to determine the sampling points, SPAN is a multi-objective function with four criteria including the variable distribution, correlation between variables, minimization of the shortest mean squared distance, and sampling optimization for identifying and

estimating variograms [40]. In other words, it considers the correlation between variables, their distribution, and the distance between sampling points. Although the covariates allow for inferences about the environmental conditions and insect locations, points that are too close together can introduce bias and redundancy into the estimates. In contrast, points that are too far apart may not adequately capture the spatial variability.

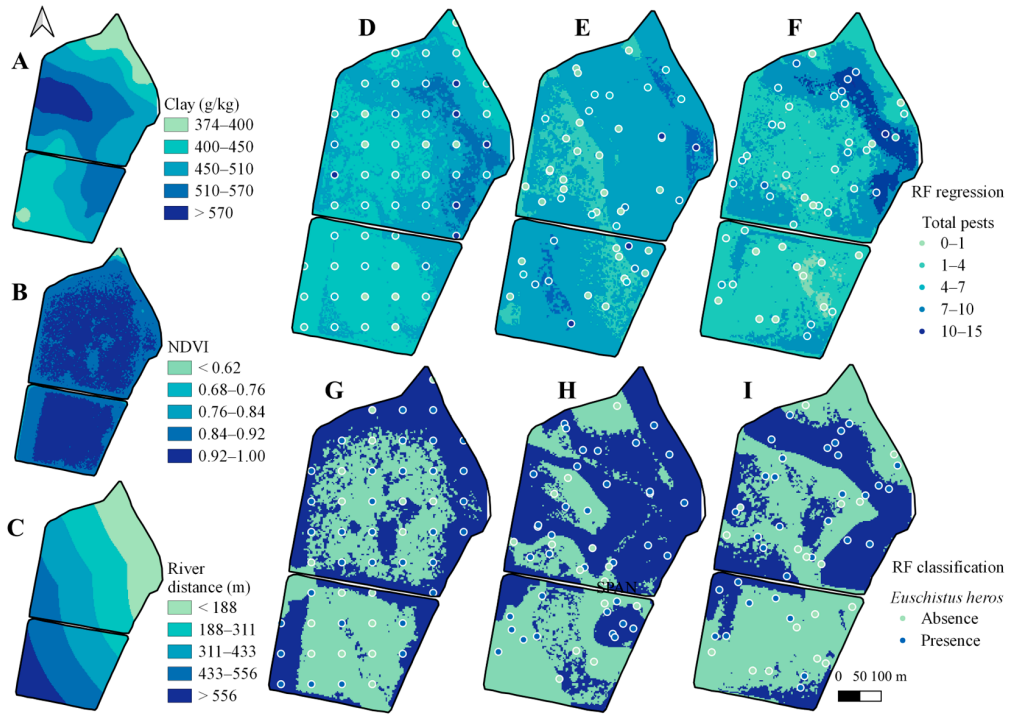

Figure 8. Environmental covariates, pest sampling points, and prediction maps. (**A**–**C**) Environmental covariates selected in phase 1: soil clay content, NDVI, and distance from river. (**D**–**F**) Predicted maps using the RF regression algorithm, with environmental covariates as predictors of total pests abundance in the regular, CORR, and SPAN sampling designs, respectively. (**G**–**I**) Predicted maps using the RF classifier algorithm, with environmental covariates as predictors of the presence and absence of *E. heros* in the regular, CORR, and SPAN sampling designs, respectively.

Regarding the distance between points in each sample design, the regular design presented the highest value of maximum (796 m) and minimum (71 m) distances between points, the CORR design had the smallest maximum (653 m) and minimum (3 m) distances, and the SPAN design showed maximum and minimum distances of 661 m and 13 m, respectively. The SPAN sampling plan had distances between points that allowed for a better capture of the spatial variability patterns of pest insects, resulting in more robust models with a greater ability to generalize for un-sampled locations. Although it was created to benefit geostatistical models, the SPAN design is also efficient for non-geostatistical models like RF regression.

4.2.2. Random Forest Classifier

Regarding the classification, the three sampling designs evaluated showed good metrics for classifying the presence and absence of *E. heros* in the modeling and external validation phases (Tables 5–8). The regular design exhibited more errors, with an equal number of FP and FN during the modeling phase. Still, it showed the best results in external validation, achieving the lowest error rate. An FN occurs when the model incorrectly classifies locations where the pest is present as absent. In terms of management, this would imply failing to apply insecticide in areas where control is necessary. The CORR design showed intermediate performance during the modeling phase, and despite showing fewer

FNs in external validation, it presented a higher number of FPs in both the modeling and validation phases. While the SPAN design demonstrated superior classification metrics during the modeling phase, its performance in external validation was hindered by a higher prevalence of FNs. In managing pests, weeds, and other phytosanitary organisms, the ideal is to minimize both classification errors. However, in terms of control, FNs are generally more detrimental than FPs because they can allow for the persistence of infestation sources in the area [60]. On the other hand, FPs can lead to unnecessary insecticide applications, increasing costs and causing economic and environmental impacts.

Regarding the predicted maps, despite the CORR and regular designs showing good metrics overall, the predicted maps from these models exhibited distribution patterns indicating prediction errors (Figure 8G,H). The regular design showed excessive alignment with the spatial structure of the vegetation index, which was not observed in the best regression models (Figure 8A–I). The tendency of the prediction model to reproduce only the provided data (i.e., the covariate) indicates noise, and that the model failed to capture the patterns of the predicted variable [61]. Conversely, despite external validation results indicating limitations for the SPAN design, this design demonstrated spatial distributions more similar to the regression model in the same sampling design, which presented the best metrics and predictions (Figure 8F,I). This suggests that SPAN, among the three evaluated models, is more effective in capturing the spatial complexity of pest distribution. These observations emphasize the importance of relying on classification metrics and carefully analyzing the predicted maps' spatial and ecological coherence when evaluating sampling designs.

4.3. Environmental Covariates in Pest Sampling and Prediction

The effectiveness of using environmental covariates for insect pest sampling and prediction relies on some key factors: the proper selection of covariates, the sampling method employed, the density and frequency of sampling during the crop season, the assessment and treatment of outliers, and the choice of prediction model.

In this research, distance from the river was one of the most relevant covariates, given its location near the crop in our experimental site, which is not always the case. Thus, evaluating the quality of covariate selection is essential, as each agricultural system is unique and presents a different distribution of environmental covariates, which may affect insect pest infestations differently. Therefore, to optimize the sampling and prediction of insect pests, covariate selection should be conducted for each agricultural area. Pusch et al. (2023) also noted the need to select covariates for each agricultural area in a study on optimized sampling and the prediction of soil variables [23]. Additionally, irrelevant and collinear variables should be avoided since they can affect the performance of prediction models [62,63].

The effectiveness of a sampling method in capturing pest distribution patterns is closely linked to the abundance of pests within the area, which is influenced not only by environmental covariates, but also by the management practices applied. A higher pest abundance increases the likelihood that sampling points will capture meaningful spatial patterns, while low pest density can obscure these patterns, leading to less reliable predictions. Given the variety of factors that can affect pest abundance in agroecosystems, it is recommended that analyses use data from different samplings for covariate selection. This assists in capturing the influence of covariates in different scenarios, both in terms of crop development and pest population, thus increasing the reliability of covariate selection.

Adopting a more strategic sampling approach with multi-objective functions that considers the distribution of covariates and sample points in the area, such as the SPAN design, is superior to relying on a regular grid or a design based solely on covariate corre-

lations (CORR). The distance between sampling points and the distribution of covariates are essential for capturing the spatial variability of insect pests because they ensure a more comprehensive representation of the environmental conditions that drive insect pest distribution patterns. This increases the model's sensitivity to local variations, increasing its generalization ability to new scenarios.

Outliers can affect prediction quality [61], thus carefully evaluating their exclusion is essential to ensure model robustness. If outliers represent extreme values significant to the problem, their removal can distort analysis and lead to incorrect predictions. Pest data are field observations in which high values are not necessarily errors but natural population variations; however, extreme values initially negatively impacted the performance of the tested predictive models. In our study, extreme values were associated with *E. heros* at points of high nymph density. At this cycle stage, these stink bugs have lower mobility than adults and usually occur clustered. Although nymphs also cause crop damage, excluding these points was justifiable since extreme values were relatively rare (only one in the regular grid and one in CORR) and were linked to the species' less mobile stage. However, it is important to note that the decision to exclude such values should be context-dependent, balancing the need for model accuracy with the ecological significance of extreme observations. In this case, excluding outliers helped to refine the predictive models without compromising the integrity of the analysis.

The choice of the best prediction model depends on the sampling method, sampling density, data characteristics, and spatial distribution patterns of pests. For example, when pest data do not have spatial dependence, machine learning models, like regression and classification, tend to be more suitable than geostatistical methods. Machine learning models can capture complex, nonlinear relationships between the environmental covariates and pests without relying on spatial dependence. Moreover, the choice of model should align with the prediction objective: regression models allow for the estimation of insect abundance while classification models assess the presence and absence (binary class) or even infestation levels (multiclass). Each approach can be used for different pest management strategies, optimizing intervention measures in specific locations.

4.4. Integrated Pest Management and Site-Specific Management

Although studies, such as those by Bueno et al. (2011), have demonstrated the efficiency of IPM compared to traditional control methods, its adoption by farmers remains limited due to challenges in achieving effective pest population monitoring [64]. However, combining IPM, which uses ecological concepts, and precision agriculture techniques that utilize covariates allows for more accurate spatial monitoring and insect pest predictions than traditional techniques. Thus, using these approaches, prophylactic insecticide applications can be replaced by site-specific management. For example, predictions using regression techniques can be utilized for variable-rate pesticide applications. However, in Brazil, the current insecticide dose recommendations for agricultural pest control generally do not consider variations in doses based on insect population variability or vegetation coverage. Therefore, incorporating these variabilities into recommendations is a necessary advancement toward more sustainable and precise management.

Integrated pest management guidelines suggest that chemical pesticides should only be applied when pest populations reach economic thresholds [2,64]. Since few insecticides currently allow for variable rate recommendations, classification models offer practical solutions. In this way, predicting the presence or absence of pests or economic thresholds through classification makes it possible to apply a single dose in specific locations, using the on–off function available on many sprayers. This allows pesticides to be applied only in areas where pests are present, or where economic thresholds are reached.

5. Conclusions

Among the analyzed covariates, NDVI (satellite image related to crop vigor variability), soil clay content, and distance from the river provide valuable information for pest infestation prediction. These covariates assist in agricultural pest mapping when included in optimization and prediction models. However, due to the unique characteristics of each agricultural area, covariate selection must be tailored to the specific conditions of each location. In this case, a two-phase sampling optimization ensures that the selected agricultural environment covariates represent the specific area. This approach enhances the effectiveness of pest mapping strategies in agricultural systems, contributing to more efficient and sustainable agricultural management practices.

Multi-objective sampling designs such as SPAN that consider environmental covariates and their distribution as well as the distance between sampling points in the area generate better insect pest maps compared with simpler sampling designs without covariates (e.g., regular grid) or with only one objective function representing the covariates' correlation (e.g., CORR).

Although precision agriculture uses vegetation indices (VIs) to guide the sampling of insect pests, these samples are generally directed toward areas with lower VI values. However, this approach may need to be more accurate. When pest populations are low and their feeding time is insufficient to cause damage, pest locations can be associated with higher VI values, areas with greater food availability, and a favorable microclimate under the plant canopy. Considering that other factors, such as topography, soil characteristics and moisture, can influence crop canopy biomass and microclimate variability, these covariates should be used with VIs to increase the effectiveness of pest sampling and mapping.

Author Contributions: Conceptualization, C.L.M. and M.P.; methodology, C.L.M., M.P. and L.R.d.A.; software C.L.M. and M.P.; validation, C.L.M.; formal analysis, C.L.M.; investigation, C.L.M.; resources, L.R.d.A.; data curation, C.L.M.; writing—original draft preparation, C.L.M.; writing—review and editing, C.L.M., M.P., W.A.C.G. and L.R.d.A.; visualization, C.L.M.; supervision, C.L.M. and L.R.d.A.; project administration, L.R.d.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Council for Scientific and Technological Development (CNPq) through a PhD scholarship awarded to the first author (No. 140060/2021-9).

Data Availability Statement: Martins, Cenneya Lopes; Amaral, Lucas Rios, 2024, "Replication data for: optimized sampling for pest mapping in soybean crop using environmental covariates", https://doi.org/10.25824/redu/CRFFQU, Repositório de Dados de Pesquisa da Unicamp, V1, UNF:6:s0sqcKfL6xs9trS7ndlIwg== [fileUNF].

Acknowledgments: We are very grateful to the owners and managers of São José Farm for their invaluable support and for granting access to their land for our research. We would also like to thank the GITAP members for their assistance with the data collection.

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

References

1. Angon, P.B.; Mondal, S.; Jahan, I.; Datto, M.; Antu, U.B.; Ayshi, F.J.; Islam, S. Integrated Pest Management (IPM) in Agriculture and Its Role in Maintaining Ecological Balance and Biodiversity. *Adv. Agric.* **2023**, *2023*, 1–19. [CrossRef]
2. Batistela, M.J.; Bueno, A.d.F.; Nishikawa, M.A.N.; Bueno, R.C.O.d.F.; Hidalgo, G.; Silva, L.; Corbo, E.; Silva, R.B. Re-evaluation of leaf-lamina consumer thresholds for IPM decisions in short-season soybeans using artificial defoliation. *Crop Prot.* **2012**, *32*, 7–11. [CrossRef]
3. Balafoutis, A.; Beck, B.; Fountas, S.; Vangeyte, J.; Van Der Wal, T.; Soto, I.; Gómez-Barbero, M.; Barnes, A.; Eory, V. Precision agriculture technologies positively contributing to ghg emissions mitigation, farm productivity and economics. *Sustainability* **2017**, *9*, 1339. [CrossRef]

4. Brown, R.M.; Dillon, C.R.; Schieffer, J.; Shockley, J.M. The carbon footprint and economic impact of precision agriculture technology on a corn and soybean farm. *J. Environ. Econ. Policy* **2016**, *5*, 335–348. [CrossRef]
5. Gebbers, R.; Adamchuk, V.I. Precision agriculture and food security. *Science* **2010**, *327*, 828–831. [CrossRef] [PubMed]
6. Hengl, T.; Rossiter, D.G.; Stein, A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Aust. J. Soil Res.* **2003**, *41*, 1403–1422. [CrossRef]
7. Szatmári, G.; László, P.; Takács, K.; Szabó, J.; Bakacsi, Z.; Koós, S.; Pásztor, L. Optimization of second-phase sampling for multivariate soil mapping purposes: Case study from a wine region, Hungary. *Geoderma* **2019**, *352*, 373–384. [CrossRef]
8. Wadoux, A.M.-C.; Brus, D.J.; Heuvelink, G.B. Sampling design optimization for soil mapping with random forest. *Geoderma* **2019**, *355*, 113913. [CrossRef]
9. Pinchao, E.C.; Muñoz, A.C. Mapping the Spatial Distribution of *Conotrachelus psidii* (Coleoptera: Curculionidae): Factors Associated with the Aggregation of Damage. *Neotrop. Entomol.* **2019**, *48*, 678–691. [CrossRef] [PubMed]
10. Haavik, L.J.; Stephen, F.M. Insect Ecology. In *Forest Entomology and Pathology*; Allison, J.D., Paine, T.D., Slippers, B., Wingfield, M.J., Eds.; Springer: Cham, Switzerland, 2023; Volume 1, Chapter 4; pp. 91–114. Available online: https://link.springer.com/book/10.1007/978-3-031-11553-0 (accessed on 9 September 2024).
11. Rebaudo, F.; Faye, E.; Dangles, O. Microclimate data improve predictions of insect abundance models based on calibrated spatiotemporal temperatures. *Front. Physiol.* **2016**, *7*, 139. [CrossRef]
12. Colinet, H.; Sinclair, B.J.; Vernon, P.; Renault, D. Insects in fluctuating thermal environments. *Annu. Rev. Èntomol.* **2015**, *60*, 123–140. [CrossRef] [PubMed]
13. Istifanus, A.P.; Abdelmutalab, A.G.A.; Pirk, C.W.W.; Yusuf, A.A. Predicting the Habitat Suitability and Distribution of Two Species of Mound-Building Termites in Nigeria Using Bioclimatic and Vegetation Variables. *Diversity* **2023**, *15*, 157. [CrossRef]
14. Kamala, M.; Devanand, I. Impact of Climate Change on Insects and their Sustainable Management. In *Sustainable Intensification for Agroecosystem Services and Management*; Springer: Singapore, 2021; pp. 779–815. Available online: https://link.springer.com/chapter/10.1007/978-981-16-3207-5_21 (accessed on 9 September 2024).
15. Fatnassi, H.; Pizzol, J.; Boulard, T.; Poncet, C.; Voisin, S.; Zigler, M. Dependence of thrips infestation on spatial climate distribution in a rose greenhouse crop. *Acta Hortic.* **2012**, *927*, 261–266. [CrossRef]
16. Faye, E.; Rebaudo, F.; Carpio, C.; Herrera, M.; Dangles, O. Does heterogeneity in crop canopy microclimates matter for pests? Evidence from aerial high-resolution thermography. *Agric. Ecosyst. Environ.* **2017**, *246*, 124–133. [CrossRef]
17. Potter, K.A.; Woods, H.A.; Pincebourde, S. Microclimatic challenges in global change biology. *Glob. Change Biol.* **2013**, *19*, 2932–2939. [CrossRef] [PubMed]
18. Duffy, J.P.; Anderson, K.; Fawcett, D.; Curtis, R.J.; Maclean, I.M.D. Drones provide spatial and volumetric data to deliver new insights into microclimate modelling. *Landsc. Ecol.* **2021**, *36*, 685–702. [CrossRef]
19. Suggitt, A.J.; Wilson, R.J.; Isaac, N.J.B.; Beale, C.M.; Auffret, A.G.; August, T.; Bennie, J.J.; Crick, H.Q.P.; Duffield, S.; Fox, R.; et al. Extinction risk from climate change is reduced by microclimatic buffering. *Nat. Clim. Change* **2018**, *8*, 713–717. [CrossRef]
20. Abdullah, N.-A.; Radzi, S.N.F.; Asri, L.-N.; Idris, N.S.; Husin, S.; Sulaiman, A.; Khamis, S.; Sulaiman, N.; Hazmi, I.R. Insect community in riparian zones of Sungai Sepetang, Sungai Rembau and Sungai Chukai of Peninsular Malaysia. *Biodivers. Data J.* **2019**, *7*, e35679. [CrossRef] [PubMed]
21. Brito, T.F.; Phifer, C.C.; Knowlton, J.L.; Fiser, C.M.; Becker, N.M.; Barros, F.C.; Contrera, F.A.L.; Maués, M.M.; Juen, L.; Montag, L.F.A.; et al. Forest reserves and riparian corridors help maintain orchid bee (Hymenoptera: Euglossini) communities in oil palm plantations in Brazil. *Apidologie* **2017**, *48*, 575–587. [CrossRef]
22. Silva, G.S.; Jahnke, S.; Johnson, N. Riparian forest fragments in rice fields under different management: Differences on hymenopteran parasitoids diversity. *Braz. J. Biol.* **2020**, *80*, 122–132. [CrossRef] [PubMed]
23. Pusch, M.; Samuel-Rosa, A.; Magalhães, P.S.G.; Amaral, L.R. Covariates in sample planning optimization for digital soil fertility mapping in agricultural areas. *Geoderma* **2023**, *429*, 116252. [CrossRef]
24. Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; Moraes, G.J.L.; Sparovek, G. Köppen's climate classification map for Brazil. *Meteorol. Z.* **2013**, *22*, 711–728. [CrossRef] [PubMed]
25. Costes, E.; Lauri, P.E.; Simon, S.; Andrieu, B. Plant architecture, its diversity and manipulation in agronomic conditions, in relation with pest and pathogen attacks. *Eur. J. Plant Pathol.* **2013**, *135*, 455–470. [CrossRef]
26. Huete, A.R.; Justice, C.; van Leeuwen, W. Modis Vegetation Index Algorithm Theoretical Basis. *Environ. Sci.* no. Mod 13. 1999. Available online: https://modis.gsfc.nasa.gov/data/atbd/atbd_mod13.pdf (accessed on 15 January 2025).
27. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. *Monitoring Vegetation Systems in the Great Plains with ERTS*; Remote Sensing Center; NASA: Washington, DC, USA, 1973; pp. 309–317. Available online: https://ntrs.nasa.gov/api/citations/19740022614/downloads/19740022614.pdf (accessed on 15 January 2025).
28. Gitelson, A.; Merzlyak, M.N. Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *J. Photochem. Photobiol. B* **1994**, *22*, 247–252. [CrossRef]

29. Baptista, G.M.M. Aplicação do Índice de Vegetação por Profundidade de Feição Espectral (SFDVI—Spectral Feature Depth Vegetation Index) em dados RapidEye. In Proceedings of the Anais XVII Simpósio Brasileiro de Sensoriamento Remoto—SBSR, João Pessoa, PB, Brazil, 25–29 April 2015; INPE: São Paulo, Brazil, 2015; pp. 2277–2284. Available online: http://www.dsr.inpe.br/sbsr2015/files/p0466.pdf (accessed on 15 January 2025).
30. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
31. Murphy, P.N.C.; Ogilvie, J.; Arp, P. Topographic modelling of soil moisture conditions: A comparison and verification of two models. *Eur. J. Soil Sci.* **2009**, *60*, 94–109. [CrossRef]
32. Santos, H.G. *Sistema Brasileiro de Classificação de Solos*; Embrapa: Brasília, Brazil, 2018.
33. Pereira, G.W.; Valente, D.S.M.; de Queiroz, D.M.; de Freitas, A.L. SMART-MAP: Plugin QGIS para Interpolação Utilizando Krigagem Ordinária e Machine Learning. Viçosa—MG. 2023. Available online: https://plugins.qgis.org/plugins/Smart_Map (accessed on 15 January 2025).
34. Brenning, A.; Bangs, D.; Becker, M.; Schratz, P.; Polakowski, F. Package 'RSAGA' Type Package Title SAGA Geoprocessing and Terrain Analysis. 2022. Available online: https://github.com/r-spatial/RSAGA (accessed on 16 January 2025).
35. Briers, R.A.; Gee, J.H.R. Riparian forestry management and adult stream insects. *Hydrol Earth Syst. Sci.* **2004**, *8*, 545–549. [CrossRef]
36. Ramey, T.L.; Richardson, J.S. Terrestrial Invertebrates in the Riparian Zone: Mechanisms Underlying Their Unique Diversity. *Bioscience* **2017**, *67*, 808–819. [CrossRef]
37. Fehr, W.R.; Caviness, C.E. *Stages of Soybean Development*; Special Report 80; Iowa State University: Ames, IA, USA, 1977; pp. 1–12. Available online: https://dr.lib.iastate.edu/entities/publication/58c89bfe-844d-42b6-8b6c-2c6082595ba3 (accessed on 15 January 2025).
38. Peterson, B.G.; Carl, P.; Boudt, K.; Bennett, R.; Ulrich, J.; Zivot, E.; Cornilly, D.; Hung, E.; Lestel, M.; Balkissoon, K.; et al. Package 'PerformanceAnalytics': Econometric Tools for Performance and Risk Analysis. 2020. Available online: https://rdrr.io/cran/PerformanceAnalytics/ (accessed on 7 August 2024).
39. Chessel, D.; Dufour, A.B.; Thioulouse, J. The ade4 Package-I: One-Table Methods. 2004. Available online: https://pbil.univ-lyon1.fr/JTHome/ref/ade4-Rnews.pdf (accessed on 7 August 2024).
40. Samuel-Rosa, A. Package 'spsann': Optimization of Sample Configurations Using Spatial Simulated Anneling. 2017. Available online: https://rdrr.io/github/samuel-rosa/spsann/ (accessed on 7 August 2024).
41. Van Groenigen, J.W.; Siderius, W.; Stein, A. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma* **1998**, *87*, 239–259. [CrossRef]
42. Bischl, B.; Lang, M.; Kotthoff, L.; Schiffner, J.; Richter, J.; Studerus, E.; Casalicchio, G.; Jones, Z.M. Mlr: Machine learning in R. *J. Mach. Learn. Res.* **2016**, *17*, 1–5. Available online: https://www.jmlr.org/papers/volume17/15-066/15-066.pdf (accessed on 7 August 2024).
43. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
44. Specht, A.; Sosa-Gómez, D.R.; Roque-Specht, V.F.; Valduga, E.; Gonzatti, F.; Schuh, S.M.; Carneiro, E. Biotic Potential and Life Tables *of Chrysodeixis includens* (Lepidoptera: Noctuidae), *Rachiplusia nu*, and *Trichoplusia ni* on Soybean and Forage Turnip. *J. Insect Sci.* **2019**, *19*, 1–8. [CrossRef] [PubMed]
45. Bernardi, O.; Sorgatto, R.J.; Barbosa, A.D.; Domingues, F.A.; Dourado, P.M.; Carvalho, R.A.; Martinelli, S.; Head, G.P.; Omoto, C. Low susceptibility of *Spodoptera cosmioides*, *Spodoptera eridania* and *Spodoptera frugiperda* (Lepidoptera: Noctuidae) to genetically-modified soybean expressing Cry1Ac protein. *Crop Prot.* **2014**, *58*, 33–40. [CrossRef]
46. Huang, F. Resistance of the fall armyworm, *Spodoptera frugiperda*, to transgenic *Bacillus thuringiensis* Cry1F corn in the Americas: Lessons and implications for Bt corn IRM in China. *Insect Sci.* **2021**, *28*, 574–589. [CrossRef] [PubMed]
47. Hagstrum, D.W.; Subramanyam, B. Immature insects: Ecological roles of mobility. *Am. Entomol.* **2010**, *56*, 230–241. [CrossRef]
48. Hoy, C.W.; McCulloch, C.E.; Shoemaker, C.A.; Shelton, A.M. Transition Probabilities for *Trichoplusia ni* (Lepidoptera: Noctuidae) Larvae on Cabbage as a Function of Microclimate. *Environ. Entomol.* **1989**, *18*, 187–194. [CrossRef]
49. Cohen, M.P.; Alford, R.A. Factors affecting diurnal shelter use by the cane toad, *Bufo marinus*. *Herpetologica* **1996**, *52*, 172–181. Available online: https://www.researchgate.net/publication/222712043 (accessed on 7 August 2024).
50. Ali, M.P.; Clemente-Orta, G.; Kabir, M.M.M.; Haque, S.S.; Biswas, M.; Landis, D.A. Landscape structure influences natural pest suppression in a rice agroecosystem. *Sci. Rep.* **2023**, *13*, 1–11. [CrossRef]
51. Zhang, J.; Huang, Y.; Pu, R.; Gonzalez-Moreno, P.; Yuan, L.; Wu, K.; Huang, W. Monitoring plant diseases and pests through remote sensing technology: A review. *Comput. Electron. Agric.* **2019**, *165*, 104943. [CrossRef]
52. Barros, P.P.S.; Schutze, I.X.; Iost Filho, F.H.; Yamamoto, P.T.; Fiorio, P.R.; Demattê, J.A.M. Monitoring *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae) Infestation in Soybean by Proximal Sensing. *Insects* **2021**, *12*, 47. [CrossRef] [PubMed]

53. Iost Filho, F.H.; Pazini, J.B.; Medeiros, A.D.; Rosalen, D.L.; Yamamoto, P.T. Assessment of Injury by Four Major Pests in Soybean Plants Using Hyperspectral Proximal Imaging. *Agronomy* **2022**, *12*, 1516. [CrossRef]
54. Maresma, Á.; Ariza, M.; Martínez, E.; Lloveras, J.; Martínez-Casasnovas, J.A. Analysis of vegetation indices to determine nitrogen application and yield prediction in maize (*Zea mays* L.) from a standard UAV service. *Remote Sens.* **2016**, *8*, 973. [CrossRef]
55. Joiner, J.; Yoshida, Y.; Anderson, M.; Holmes, T.; Hain, C.; Reichle, R.; Koster, R.; Middleton, E.; Zeng, F.-W. Global relationships among traditional reflectance vegetation indices (NDVI and NDII), evapotranspiration (ET), and soil moisture variability on weekly timescales. *Remote Sens. Environ.* **2018**, *219*, 339–352. [CrossRef] [PubMed]
56. Checa, M.F.; Rodriguez, J.; Willmott, K.R.; Liger, B. Microclimate variability significantly affects the composition, abundance and phenology of butterfly communities in a highly threatened neotropical dry forest. *Fla. Entomol.* **2014**, *97*, 1–13. [CrossRef]
57. Vives-Ingla, M.; Sala-Garcia, J.; Stefanescu, C.; Casadó-Tortosa, A.; Garcia, M.; Peñuelas, J.; Carnicer, J. Interspecific differences in microhabitat use expose insects to contrasting thermal mortality. *Ecol. Monogr.* **2022**, *93*, e1561. [CrossRef]
58. Chown, S.L.; Terblanche, J.S. Physiological Diversity in Insects: Ecological and Evolutionary Contexts. *Adv. Insect Physiol.* **2006**, *33*, 50–152. [CrossRef]
59. Oliver, M.A.; Webster, R. *Basic Steps in Geostatistics: The Variogram and Kriging*; Springer: Cham, Switzerland, 2015. Available online: https://link.springer.com/book/10.1007/978-3-319-15865-5 (accessed on 22 November 2024).
60. Martins, C.L.; Oliveira, A.L.G.; Cunha, I.A.; Oldoni, H.; Pereira, J.C.; Amaral, L.R. Classification of the Occurrence of Broadleaf Weeds in Narrow-Leaf Crops. *Eng. Agrícola* **2024**, *44*, e20230148. [CrossRef]
61. Wendler, T.; Gröttrup, S. *Data Mining with SPSS Modeler: Theory, Exercises and Solutions*, 2nd ed.; Springer International Publishing: Cham, Switzerland, 2021. [CrossRef]
62. Harrell, F.E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*; Springer: New York, NY, USA, 2001. [CrossRef]
63. Shrestha, N. Detecting Multicollinearity in Regression Analysis. *Am. J. Appl. Math. Stat.* **2020**, *8*, 39–42. [CrossRef]
64. Bueno, A.F.; Batistela, M.J.; Bueno, R.C.O.F.; França-Neto, J.B.; Nishikawa, M.A.N.; Filho, A.L. Effects of integrated pest management, biological control and prophylactic use of insecticides on the management and sustainability of soybean. *Crop Prot.* **2011**, *30*, 937–945. [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.