

Article

Spatial Prediction in Agricultural Landscapes of Soil Nutrients Using NIR Spectroscopy and Advanced Regression Models for Precision Agriculture in Rice Paddies, Tomato Fields, and Montado Ecosystems

Maria Navalho ¹, Ana Rita F. Coelho ^{1,2,*} , Ana Coelho Marques ^{1,2} , Inês Ferreira ³ , José Rafael Silva ^{4,5} , Cristina Houghton ³ and Manuela Simões ^{1,2,*} 

¹ NOVA School of Sciences and Technology, Earth Sciences Department, Campus de Caparica, 2829-516 Caparica, Portugal; mr.navalho@campus.fct.unl.pt (M.N.); amc.marques@fct.unl.pt (A.C.M.)

² GeoBioTec Research Center, NOVA University of Lisbon, 2829-516 Caparica, Portugal

³ cE3c—Centre for Ecology, Evolution and Environmental Changes & Change—Global Change and Sustainability Institute, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Bloco C2, 1749-016 Lisboa, Portugal; iiferreira@fc.ul.pt (I.F.); cmhoughton@fc.ul.pt (C.H.)

⁴ MED—Mediterranean Institute for Agriculture, Environment and Development, CHANGE—Global Change and Sustainability Institute, Universidade de Évora, Pólo da Mitra, Ap. 94, 7006-554 Évora, Portugal; jmsilva@uevora.pt

⁵ Agroinsider Lda, 7005-841 Évora, Portugal

* Correspondence: arf.coelho@fct.unl.pt (A.R.F.C.); mmsr@fct.unl.pt (M.S.)

Abstract

Soil fertility is a critical factor for sustainable agricultural production, yet traditional soil analysis methods are often time-consuming and costly. As such, the aim of our research was to evaluate the potential of near-infrared (NIR) spectroscopy as a rapid, cost-effective, and non-destructive tool to assess soil fertility properties in agricultural soils. The soil samples were collected from the Quinta da Foz located in Benavente (Portugal) across agricultural regions including rice paddies, tomato fields, and Montado ecosystems. The sampling locations were guided by the Normalized Difference Vegetation Index (NDVI) to capture spatial variability, and soil analyses included near-infrared (NIR) spectral measurements and laboratory-based chemical determinations of soil fertility parameters (pH, electrical conductivity, total carbon and nitrogen, organic matter, macro- and micronutrients) as well as multiple soil carbon fractions. Two predictive modeling approaches (Random Forest (RF) and Partial Least Squares Regression (PLSR)) were developed to estimate soil chemical properties from spectral data. The RF models consistently outperformed PLSR, achieving high accuracy ($R^2 = 0.85$) for nutrients such as Mg, Fe, Ni, Ca, and Na and organic matter. A moderate predictive performance (R^2 between 0.70 and 0.80) was observed for different elements, namely, K and Mn. On the other hand, P, S, Zn, electrical conductivity, pH, total N, and various carbon fractions were poorly predicted. The spatial interpolation of predicted values enabled the generation of soil fertility maps that informed site-specific nutrient management. The results indicate that NIR spectroscopy combined with robust modeling offers a promising approach for rapid spatial assessment of selected soil nutrients, supporting precision agriculture and sustainable land management.

Keywords: NIR spectroscopy; spectral modeling; soil fertility; soil carbon



Academic Editor: Spyros Fountas

Received: 15 December 2025

Revised: 10 February 2026

Accepted: 15 February 2026

Published: 20 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The continuous growth in demand for food has increased pressure on agricultural systems, requiring greater efficiency in resource use and a reduction in environmental impact [1]. Projections indicate that production will have to increase substantially in the coming decades, reinforcing the need for technologies that support more precise soil and input management [2]. Precision agriculture responds to this challenge by integrating spatial and temporal variability into decision-making, allowing interventions tailored to the specific conditions of each plot [3]. However, traditional soil characterization, although rigorous, is time-consuming and expensive when applied to large volumes of samples [4]. This creates a critical operational bottleneck: the high cost and latency of laboratory data often prevent producers from implementing the very site-specific interventions that precision agriculture demands. Consequently, there is a fundamental need for a “high throughput” soil sensing solution that can bridge the gap between complex field variability and real-time management.

The primary gap in current research lies in the limited transferability of predictive models across diverse agroecosystems. Near-infrared (NIR) spectroscopy has emerged as an effective alternative due to its speed, non-destructive nature, and low cost per analysis. Several studies demonstrate its ability to predict chemical properties, especially carbon and nitrogen, which exhibit well-defined spectral responses [5,6]. However, nutrients such as phosphorus and potassium require specific calibrations due to the technique’s sensitivity to local soil characteristics [7]. Most of the existing literature focuses on single-crop systems, leaving a significant void in our understanding of how these models perform when applied to a landscape mosaic—such as the coexistence of flooded rice paddies, intensive tomato fields, and silvo-pastoral Montado systems. Furthermore, the ability to translate these spectral predictions into functional, high-resolution fertility maps remains an underutilized area of research.

Soil organic carbon is particularly relevant due to its influence on fertility and the global carbon cycle [8]. Classical laboratory methods, such as those proposed by Walkley or Nelson, are accurate but demanding, making faster methodologies attractive for continuous monitoring, especially given the increasing demands of carbon markets [9]. This study addresses the core problem of providing reliable, low-cost soil data by testing the hypothesis that NIR spectroscopy, when combined with advanced machine learning, can accurately characterize multiple nutrient fractions simultaneously across distinct soil types.

The work presented follows a clear, structured plan designed to validate this approach from sampling to spatial application. First, we utilize a remote-sensing-guided sampling design based on NDVI to ensure that the variability of the three agroecosystems is fully captured. Second, we establish a robust calibration database by pairing NIR spectral data with standardized chemical analyses. Third, we develop and compare Random Forest (RF) and Partial Least Squares Regression (PLSR) models to identify the most accurate predictive algorithms. Finally, the predicted values are used to generate interpolated spatial maps, moving the research from theoretical modeling to a practical decision-support tool.

The importance and usability of this work are found in its potential to transform soil management from a reactive, infrequent task into a proactive, digital workflow. By providing an operational and accessible alternative to conventional analyses, this methodology supports fertilization management and the identification of spatial variability through maps generated from the models. The targeted audience for this research includes agronomists and precision farming specialists seeking more efficient diagnostic tools, agricultural producers aiming to optimize fertilizer costs, and environmental scientists focused on monitoring soil carbon sequestration in heterogeneous landscapes. Ultimately, this work provides

a scalable framework for sustainable soil management that is both economically viable and scientifically robust.

2. Materials and Methods

2.1. Study Area

The study was conducted at Quinta da Foz, located in Benavente, Santarém district, Portugal (GPS coordinates: 38°58' N, 8°50' W) (Figure 1). According to 2023 data, the area of 650 ha is characterized by the coexistence of three contrasting agroecosystems: rice paddies (541.6 ha), tomato fields (80.06 ha), and Montado (21.91 ha) (Figure 2). Specifically, rice paddies are seasonally inundated, producing prolonged saturation and reducing (anaerobic) soil conditions that favor fine-textured, waterlogged soils; tomato fields are intensively tilled, irrigated and fertilized, typically on well-drained, lighter-textured soils; and the Montado is a low-intensity agro-silvo-pastoral system with tree cover, limited soil disturbance, and relatively higher surface organic matter.

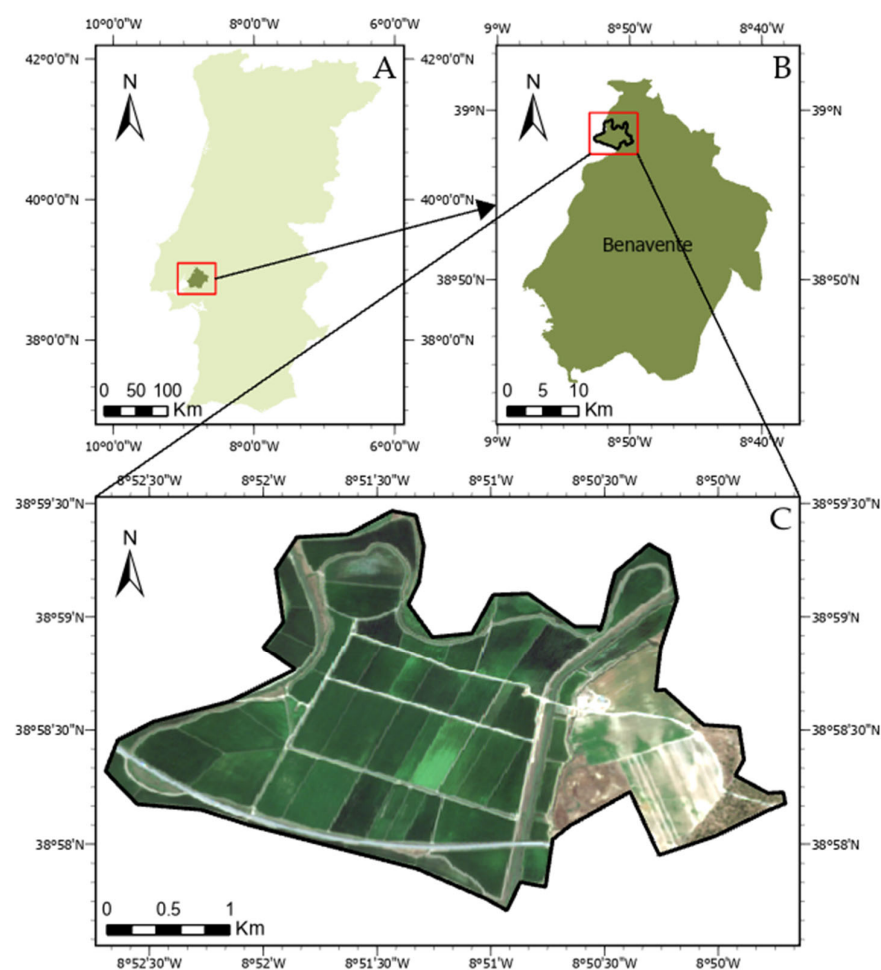


Figure 1. Geographical location at the national scale (Portugal) of the Benavente municipality (A), indication of the location of the field at Quinta da Foz within Benavente municipality limits (B), and detailed view of the study parcel boundary (C). Maps were produced in ArcGIS Pro (version 3.2.2), Esri World Imagery was used as the basemap, and all spatial data were projected in ETRS89/PT-TM06.



Figure 2. Geographic representation of the division of the field of Quinta da Foz into rice paddy, tomato fields, and Montado. Map produced in ArcGIS Pro (version 3.2.2), Esri World Imagery was used as the basemap, and all spatial data were projected in ETRS89/PT-TM06.

The three agroecosystems (identified in Figure 2), lead to different land uses (rice, tomato, and Montado) which differ markedly in soil type, management intensity, and hydrological regime.

2.2. Soil Sampling Design

Soil sampling was carried out according to the spatial variability of vegetation vigor obtained from the Normalized Difference Vegetation Index (NDVI), calculated from Sentinel-2 satellite images available on the SmartAg platform (AgroInsider). The NDVIs from 2023 were used to delineate homogeneous zones within each agricultural area and to select representative sampling sites.

A total of 200 soil samples were collected (Figure 3)—100 from the rice paddies, 50 from the tomato fields, and 50 from the Montado area—and the sampling points were distributed proportionally to the surface area and NDVI variability. In the Montado area, samples were taken both beneath tree canopies and in open areas to capture within-site heterogeneity. As such, soil sampling was performed between July and October 2023 and at each point, the surface layer (0–10 cm) was collected using a metallic cylinder after removing plant residues and litter, with approximately 1 kg of soil collected per sample for analyses. Each sample was georeferenced using Google Earth and properly labeled.

2.3. Soil Fertility Analysis

Soil samples collected in the different zones (rice paddies, tomato fields, and Montado) were analyzed at a chemical laboratory (A2 Análises Químicas located in Benavente, Portugal), which was a certified facility following ISO protocols for soil analysis [10]. Samples were air-dried at 40 °C for 15 h, sieved through a 2 mm mesh (ISO 11464:2006), and homogenized.

Soil pH was measured potentiometrically (ISO 10390:2005) in aqueous, CaCl_2 (0.01 M), and $\text{Ca}(\text{OH})_2$ buffer suspensions, while electrical conductivity (EC) was determined by conductimetry according to ISO 11265:1994, using the same aqueous suspension and values were expressed in $\mu\text{S}\cdot\text{cm}^{-1}$ standardized to room temperature.

The macro- and microelement concentrations, including those of phosphorus (P_2O_5), potassium (K_2O), calcium (CaO), magnesium (MgO), sulphur (S), boron (B), iron (Fe), total and available manganese (Mn and MnAI), copper (Cu), zinc (Zn), molybdenum (Mo),

cobalt (Co), nickel (Ni), and sodium (Na), were quantified by inductively coupled plasma optical emission spectrometry (ICP-OES) (ISO 22036:2008). For extraction, two grams of soil were treated with the Mehlich-3 and cation-exchange-capacity (CEC) solutions, and the supernatant was subsequently diluted (1:9) for instrumental analysis. The results were expressed in mg.kg^{-1} of soil. For P_2O_5 (P), K_2O (K), CaO (Ca), MgO (Mg), and MnAl, values refer to plant-available fractions, while for the remaining elements, they correspond to total concentrations [10].



Figure 3. Geographic location of the soil sampling points within the study parcel. Each point represents an individual sampling site. Map produced in ArcGIS Pro (version 3.2.2), Esri World Imagery was used as the basemap, and all spatial data were projected in ETRS89/PT-TM06.

Total carbon and nitrogen concentrations were determined using an elemental analyzer based on the dynamic combustion of finely ground soil samples, followed by separation of gaseous products by gas chromatography. The analyzer separates the gases into two distinct peaks—elemental nitrogen (N_2) and carbon dioxide (CO_2)—which are detected by thermal conductivity. The C:N ratio was subsequently calculated from the obtained carbon and nitrogen contents and expressed as a percentage of each element, corresponding to the mass ratio between carbon and nitrogen.

Soil organic matter (OM) was estimated by multiplying the total carbon content, determined by the elemental analyzer, by the Van Bemmelen factor (1.724), assuming that 58% of the total organic matter corresponds to organic carbon, while total nitrogen was directly obtained from the same elemental analysis.

2.4. Soil Carbon Fraction Analysis

Soil carbon was quantified through three complementary analytical methods (Figure 4) designed to capture distinct organic and inorganic carbon pools: total carbon (through elemental analysis—described in Section 2.3), organic carbon determined by loss-on-ignition (LOI), and dissolved organic carbon determined by TOC analysis.

Additionally, to understand analysis variation, the organic carbon was also estimated by measuring the mass loss of the soil after combustion, carried out according to the loss-on-ignition (LOI) technique. Approximately 1 g of fresh soil was weighed into pre-cleaned porcelain crucibles (COBOS CB-Junior, precision ± 0.0005 g) and samples were oven-dried at 45°C for 24 h until reaching a constant weight and then ignited in a muffle furnace (Nabertherm L3/11/06) at 500°C for 16 h. After combustion, the crucibles were allowed to cool to 100°C inside the furnace, transferred to a desiccator, and reweighed. Organic matter content (%) was calculated from the mass difference before and after ignition, using

triplicate measurements per sample to ensure analytical reproducibility. After that, organic matter content was also carried out according to the following equation [11]:

$$\text{Organic matter (\%)} = \frac{\text{Dry sample weight (g)} - \text{After burned sample weight (g)}}{\text{Dry sample weight (g)}} \times 100$$

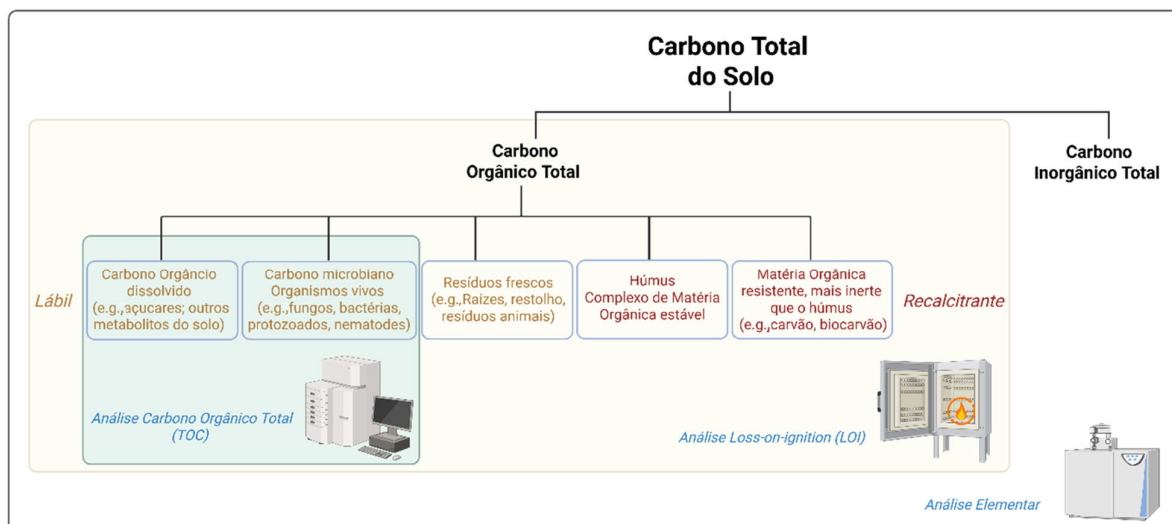


Figure 4. Systematic framework of soil carbon fractions and the corresponding analyses used for their determination.

The microbial and dissolved organic carbon were determined through extraction and analysis of soil carbon fractions using a TOC analyzer. About 5 g of fresh soil was weighed into acid-cleaned containers (cleaned with HCl 10%) and properly labeled. Non-fumigated samples were directly transferred into 50 mL Falcon tubes, while fumigated samples were placed in 50 mL glass beakers labeled with a graphite pencil. The fumigation was performed inside a laboratory fume hood using a glass desiccator containing a moistened paper sheet at the base to maintain humidity, and within the desiccator, the beakers with soil samples were placed alongside a separate beaker containing approximately 50 mL of chloroform and glass beads, which prevented overflow and cross-contamination. A thin layer of vacuum grease was applied to the junction between the lid and base to ensure proper sealing. The desiccator was connected to a vacuum pump, which was activated until the chloroform began to boil and after about 30 s of boiling, the stopcock was closed, and the desiccator was covered with aluminum foil to block light and air entry. After 24 h of fumigation, the aluminum cover was removed, the valve was opened to allow air to enter, and the lid was lifted. The beaker with chloroform and the humid paper were removed, and the desiccator was re-sealed and connected again to the vacuum pump for several minutes to eliminate residual chloroform vapor; after that, the soil samples were then transferred into 50 mL Falcon tubes and labeled.

For both fumigated and non-fumigated samples, 0.5 M K₂SO₄ solution was added up to a total volume of 50 mL, and the tubes were shaken manually and homogenized using a vortex mixer when necessary, followed by horizontal shaking for 30 min. Samples were then centrifuged at 3500 rpm for 20 min, after prior balancing by weight. The supernatant was filtered under vacuum, and the filtrate was collected into new, labeled Falcon tubes. The filtered extracts were analyzed in a TOC analyzer to determine total organic carbon (TOC), inorganic carbon, and total carbon in fumigated samples. The microbial carbon was calculated as the difference between TOC concentrations measured in fumigated and non-fumigated extracts, based on the assumption that chloroform fumigation lyses microbial

cells and releases intracellular carbon, which is then detected as additional organic carbon in the extract. All analyses were carried out in triplicate according to the following equation:

$$\text{Microbial carbon (mg} \cdot \text{kg}^{-1}) = \text{TOC after fumigation} - \text{TOC}$$

The combined use of LOI and TOC measurements provided an integrated view of soil carbon dynamics, enabling differentiation between stable and labile carbon forms. LOI represented the total organic carbon pool associated with bulk organic matter, while TOC analysis provided detailed quantification of dissolved, inorganic, and microbial carbon fractions.

2.5. NIR Spectroscopy

Near-infrared spectral measurements were conducted using an Ocean Insight NIRQuest+2.5 reflectance spectrometer (wavelength range: 894–2497 nm). This instrument records diffuse reflectance spectra at approximately 1 nm resolution, enabling detection of absorption features associated with C–H, N–H, and O–H molecular bonds commonly present in organic matter and mineral components of soils. As such, approximately 20 g of soil was placed in 9 cm diameter Petri dishes, ensuring that each dish was completely filled and the surface evenly leveled to avoid spectral distortion. Three replicate dishes were prepared per sample and for each dish, three consecutive readings were acquired, resulting in a total of 27 spectra per soil sample. The spectrometer's optical fiber probe was positioned vertically at a constant distance from the sample surface and ambient light, and vibration was minimized during all measurements.

Instrument calibration was performed regularly using a certified reference standard to maintain reflectance accuracy and recalibration was conducted approximately every 30 min or whenever a visible shift in the baseline of the spectral curve was observed. The laboratory environment was kept free from drafts, mechanical vibration, or variable illumination that could interfere with the sensor's readings.

2.6. Data Processing and Modeling

All spectral and chemical data processing, as well as model development, were carried out using Python 3.11, with the libraries *pandas*, *numpy*, *scikit-learn*, *matplotlib*, and *joblib*. The aim was to establish predictive relationships between near-infrared (NIR) spectral data and measured soil chemical properties, using both linear and non-linear regression algorithms. A systematic diagram of the applied modeling process is presented in Figure 5.

2.6.1. Data Organization and Pre-Processing

Spectral data obtained from the NIRQuest+2.5 sensor were first averaged across the 27 replicate measurements per soil sample to produce a single representative reflectance curve. The original spectra contained 512 wavelength bands ranging from 894 to 2498 nm and noisy regions corresponding to water absorption and low signal (1350–1430 nm and >2000 nm) were removed; a moving-average smoothing filter (window size = 5) was applied to reduce high-frequency noise, and the resulting dataset retained 298 effective wavelengths per sample. Each reflectance curve was normalized to a range between 0 and 1 before model calibration and the final spectral dataset was merged with the chemical dataset through a common sample ID code, resulting in a comprehensive data matrix containing 25 soil variables as response factors and 298 NIR predictors.

The 25 response variables used as supervised model targets were pH; electrical conductivity (CE); P, K, Ca, Mg, S, B, Fe, Mn, MnAI, Cu, Zn, Co, Ni, and Na ($\text{mg} \cdot \text{kg}^{-1}$); and the soil carbon/nitrogen fractions MO_LOI, TOC, IC, F_TC, C_microb, MO, Ratio_C/N, Carbon_T, and Nitrogen_T. These chemical measurements were the laboratory-labeled targets for the

calibration set. The predictors were the 298 retained NIR wavelength bands (post-trimming and smoothing). Sample metadata (sample ID, land-use class, and geographic coordinates) were attached to every record.

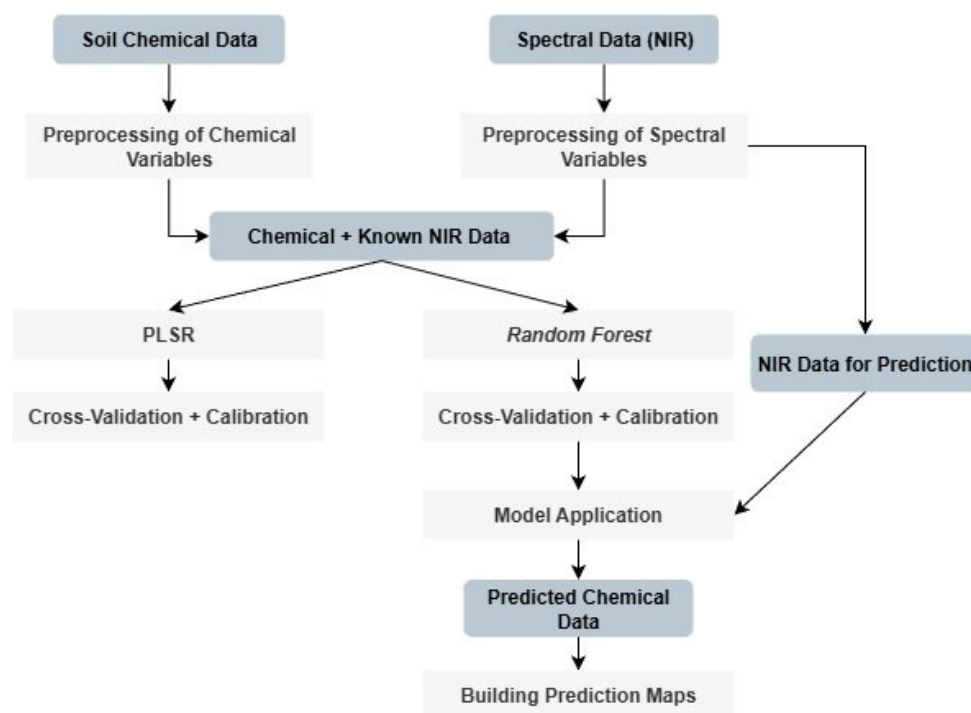


Figure 5. Systematic diagram of the applied modeling process.

The chemical data were cleaned prior to modeling and variables with all values below detection limits (i.e., molybdenum (Mo)) were excluded, while for the remaining elements, values below detection thresholds were replaced by the detection limit value. Outliers exceeding ± 3 standard deviations within each land-use class were removed and new variables were computed from analytical results, such as total carbon ($\text{Carbon_Total} = \text{Ratio_C_N} \times \text{Nitrogen_Total}$) and microbial carbon (C_microb), derived from TOC measurements.

2.6.2. Model Development, Validation and Prediction

Two algorithms were selected to predict soil chemical and carbon properties from near-infrared (NIR) spectral data—*Random Forest* (RF) and *Partial Least Squares Regression* (PLSR)—both being widely applied in soil spectroscopy due to their ability to handle high-dimensional and correlated predictor variables [12,13], and considering that adjacent spectral bands are often highly collinear, models must be capable of capturing both linear and non-linear relationships while remaining robust to noise. The PLSR reduces data dimensionality by extracting latent variables, being particularly effective when the relationship between spectra and target variables is linear. In contrast, RF, an ensemble-based algorithm, identifies complex non-linear interactions among wavelengths and can model intricate soil–spectra relationships with high stability.

To ensure the reliability and generalization of the models, a five-fold cross-validation (CV) procedure was implemented for both algorithms [14], and the CV is a resampling technique that partitions the dataset into five equally sized subsets. In each iteration, four subsets (80% of the data) are used to train the model, while the remaining subset (20%) serves as the test set and this process is repeated five times, each time using a different subset for validation, ensuring that all samples are used for testing once. The results from

all iterations are combined to compute global performance metrics. Moreover, this method is particularly appropriate for small to medium datasets, as it maximizes data utilization while minimizing bias and overfitting [15,16].

During each iteration, the following metrics were computed: the *Coefficient of Determination* (R^2_{cv}) and the *Root Mean Square Error* (RMSE_{cv}). These metrics were used to tune hyperparameters and identify the most efficient model configuration and once optimized, the final model was retrained on the entire dataset to capture all available spectral variation, and its calibration metrics (R^2_{cal} and RMSE_{cal}) were calculated to evaluate fit quality. Both R^2 and RMSE are standard indicators of model performance in soil spectroscopy [14]. The R^2 expresses the proportion of total variance in the measured data explained by the model: values above 0.9 are considered excellent, while values above 0.7 indicate good to very good agreement. However, an excessively high calibration R^2 without corresponding validation accuracy can indicate overfitting. Therefore, maintaining a high R^2_{cv} is essential, as it reflects the model's capacity to generalize to unseen data. RMSE quantifies the average prediction error in the same units as the measured property, where lower values indicate higher precision. An additional performance metric, the *Ratio of Performance to Deviation* (RPD), was also computed to assess prediction quality and reliability [17]. RPD values lower than 2 indicate insufficient predictive ability, 2–2.5 indicate approximate prediction, 2.5–3 indicate good prediction, and values above 3 indicate excellent quantitative performance.

For PLSR, the optimal number of latent components was determined iteratively during cross-validation by balancing model simplicity with predictive accuracy; using too few components may lead to underfitting, failing to capture relevant spectral–soil relationships, whereas too many components may result in overfitting, limiting the model's ability to generalize beyond the training data. The RF does not rely on latent components; instead, it builds multiple decision trees using subsets of predictors and samples, aggregating their results to produce stable predictions while minimizing variance.

After training and validation, both RF and PLSR models were compared for all soil properties and each optimized model was saved individually using the *joblib* library, enabling reuse without retraining. Predictions were then generated for new soil samples analyzed exclusively by NIR spectroscopy, extending the model's applicability beyond the calibration set. The predicted chemical variables were compiled into a single output file containing the sample identifiers, corresponding land-use type, and geographic coordinates, to support subsequent spatial interpolation and mapping.

2.7. Spatial Mapping of Predicted Values

Considering the predicted soil chemical and carbon values generated by the *Random Forest* and PLSR models, spatial interpolation was performed using ArcGIS Pro 3.2.2 software. The Inverse Distance Weighted (IDW) interpolation method was selected to generate continuous raster maps representing the predicted spatial distribution of each soil property, and this method assumes that the influence of a sampled point decreases with distance, meaning that the soil at a given location is more similar to nearby samples than to distant ones. This deterministic interpolation technique creates a smoothed surface that captures local variability while preserving the structure of the measured data.

All interpolations were carried out using the GCS_WGS_1984 coordinate reference system. Separate prediction maps were produced for each of the three land-use zones—rice paddies, tomato fields, and Montado ecosystems—and for the soil properties that achieved the best predictive performance (i.e., Mg, Fe, Ni, Na, K, and organic matter). The raster layers were symbolized using five color classes to represent gradients of nutrient concen-

tration and soil fertility, facilitating visual interpretation of spatial variability across each agricultural system.

3. Results

3.1. Soil Chemical Characterization

Considering Table 1, it is possible to observe a clear difference in soil chemical properties among the three land-use zones (Montado, rice fields, and tomato fields). As such, the tomato soils displayed the lowest pH values (average of 5.7) while tomato soils were slightly more alkaline (average of 7.4). The electrical conductivity (CE) was low across all the areas analyzed, ranging from 0.1 to 0.3 dS.m⁻¹, indicating an overall non-saline field conditions.

Table 1. Mean values and standard deviation (SD) of soil nutrient concentrations, pH, and electrical conductivity (CE) measured in the three land-use zones.

Soil Property	Montado	Rice	Tomato
pH	5.7 ± 0.6	6.3 ± 0.7	7.4 ± 0.5
CE (mS.cm ⁻¹)	0.2 ± 0.0	0.3 ± 0.2	0.1 ± 0.0
P (mg.kg ⁻¹)	141.5 ± 107.7	38.5 ± 24.6	336.1 ± 174.4
K (mg.kg ⁻¹)	59.5 ± 36.9	351.7 ± 87.1	58.4 ± 28.8
Ca (mg.kg ⁻¹)	522.0 ± 344.3	3412.0 ± 795.1	767.2 ± 315.5
Mg (mg.kg ⁻¹)	60.0 ± 48.8	1186.5 ± 131.1	87.3 ± 36.7
S (mg.kg ⁻¹)	7.6 ± 5.2	125.5 ± 94.2	5.7 ± 1.6
B (mg.kg ⁻¹)	0.3 ± 0.1	1.3 ± 0.5	0.3 ± 0.0
Fe (mg.kg ⁻¹)	135.7 ± 37.2	838.8 ± 108.1	220.7 ± 91.5
Mn (mg.kg ⁻¹)	83.9 ± 45.7	215.3 ± 59.0	28.9 ± 14.2
MnAl (mg.kg ⁻¹)	329.5 ± 172.6	813.2 ± 221.8	96.4 ± 50.6
Cu (mg.kg ⁻¹)	0.6 ± 1.1	0.9 ± 0.2	3.6 ± 2.0
Zn (mg.kg ⁻¹)	7.6 ± 9.9	2.4 ± 0.6	6.0 ± 5.5
Co (mg.kg ⁻¹)	0.4 ± 0.1	2.3 ± 0.6	0.4 ± 0.2
Ni (mg.kg ⁻¹)	0.2 ± 0.0	2.5 ± 0.4	0.3 ± 0.1
Na (mg.kg ⁻¹)	6.0 ± 3.2	197.9 ± 58.5	13.3 ± 16.4

Regarding the P content, the tomato fields (336 mg.kg⁻¹) showed the highest and rice fields the lowest values (38.5 mg.kg⁻¹). Moreover, a similar pattern was observed for K, with exceptionally high values in rice soils (351.7 mg.kg⁻¹) compared to those in the Montado and tomato fields, which presented similar concentrations (59.5 and 58.4 mg.kg⁻¹, respectively). On the other hand, Ca and Mg exhibited the strongest differences between land uses, with rice soils containing higher Ca (3412 mg.kg⁻¹) and Mg (1186.5 mg.kg⁻¹) concentrations compared to the Montado and tomato soils (with Ca content < 800 mg.kg⁻¹ and Mg content < 350 mg.kg⁻¹). Also, S followed the same pattern observed for Ca and Mg, with rice fields presenting the highest mean values (125.5 mg.kg⁻¹). Additionally, the microelement distributions further showed the distinct soil environments, with Fe showing the highest content in rice soils (838 mg.kg⁻¹), followed by tomato (220 mg.kg⁻¹) and Montado (135.7 mg.kg⁻¹). Manganese and Al were also higher in the rice soils, while tomato soils had the lowest Mn levels (28.9 mg.kg⁻¹). For instance, Zn exhibited a contrasting trend, with higher values obtained in the Montado (7.6 mg.kg⁻¹) and tomato fields (6.0 mg.kg⁻¹) compared to the rice fields (2.4 mg.kg⁻¹).

Copper, Co, and Ni showed moderate variations across the three different lands uses, but Ni was considerably higher in rice fields (2.5 mg.kg⁻¹). Additionally, the Na concentrations increased from the Montado (6.0 mg.kg⁻¹) to tomato fields (13.3 mg.kg⁻¹)

and peaked in the rice fields (197.9 mg.kg^{-1}), consistent with the flooded conditions and potential salinity inputs here.

Overall, considering the results obtained, it is possible to verify substantial chemical differentiation between the different land-use types studied, especially regarding Ca, Mg, F, and Na, which reflect the distinct management practices, hydrological regimes, and even the general conditions of each production system.

According to Table 2, it is possible to verify a clear difference in soil carbon pools and total nitrogen among the three land uses studied. In fact, the strongest contrasts can be observed in the organic matter fraction measured by loss-on-ignition (MO_LOI), and considering the rice soils, they exhibited the highest values (7.9%), reflecting the accumulation of organic residues under flooded conditions, while the Montado showed moderate content (2.3%) and tomato fields the lowest (1.2%). The total organic carbon (TOC) showed a different pattern, with Montado presenting the highest value (6.3%) followed by the tomato and rice fields. Thus, the inorganic carbon (IC) was generally low across all three zones but slightly higher in the rice fields (1.1%). Moreover, the fractions of total carbon associated with F_TC showed comparable mean values across the three land uses. Microbial carbon (C_microb) did not exhibit contrasts between systems, showing a range between 7.9 and 8.3% considering the three land uses. The other technique used for measuring the organic matter content (MO) showed higher value than MO_LOI, with the tomato field also showing the lowest content (1.7%). Regarding the C:N ratio, this was higher in the Montado (12.6), which could reflect more stable organic inputs, followed by the rice (10.2) and tomato fields (9.3). On the other hand, the total carbon (Carbon_T) was slightly higher in the Montado (2.3%) compared to the rice fields (2.1%), and the total nitrogen (Nitrogen_T) presented uniformly low values across the three land-use zones, ranging between 0.1 and 0.2%.

Table 2. Mean values and standard deviation (SD) of soil carbon fractions and total nitrogen measured in the three land-use zones.

Soil Property	Montado	Rice	Tomato
MO_LOI (%)	2.3 ± 1.1	7.9 ± 1.1	1.2 ± 0.4
TOC (mg.L^{-1})	6.3 ± 2.2	3.4 ± 2.4	4.5 ± 2.1
IC (mg.L^{-1})	0.5 ± 0.1	1.1 ± 1.8	0.6 ± 0.3
F_TC (mg.L^{-1})	14.6 ± 9.7	12.5 ± 3.9	13.1 ± 16.8
C_microb (mg.L^{-1})	7.9 ± 9.6	8.3 ± 4.3	8.2 ± 16.2
MO (%)	4.0 ± 1.7	3.6 ± 0.8	1.7 ± 0.7
Ratio_C_N (%)	12.6 ± 1.9	10.2 ± 1.0	9.3 ± 2.4
Carbon_T (%)	2.3 ± 1.0	2.1 ± 0.4	1.0 ± 0.3
Nitrogen_T (%)	0.2 ± 0.1	0.2 ± 0.0	0.1 ± 0.0

3.2. Spatial and Temporal Analysis of NDVI

In Figure 6, on the left side, the rice-growing area is dominated by hydromorphic alluvial soils with poor drainage and high clay content. These soils receive periodic sediment inputs, resulting in good fertility and high agricultural suitability, despite the presence of moderately saline and young alluvial soils prone to waterlogging.

On the right is represented the area corresponding to the Montado, with the predomination of psammitic regosols (incipient soils), with a sandy, loose texture, poorly differentiated horizons, and, sometimes, only a thin surface horizon with low accumulation of organic matter. They are essentially suitable for forestry, although they can be used in intensive irrigation agriculture. The tomato field zone, where planosols are observed, is characterized by the presence of light surface horizons, over more compact and

poorly permeable clayey layers, and non-humic litholic soils, poorly developed and with coarse sandstones.

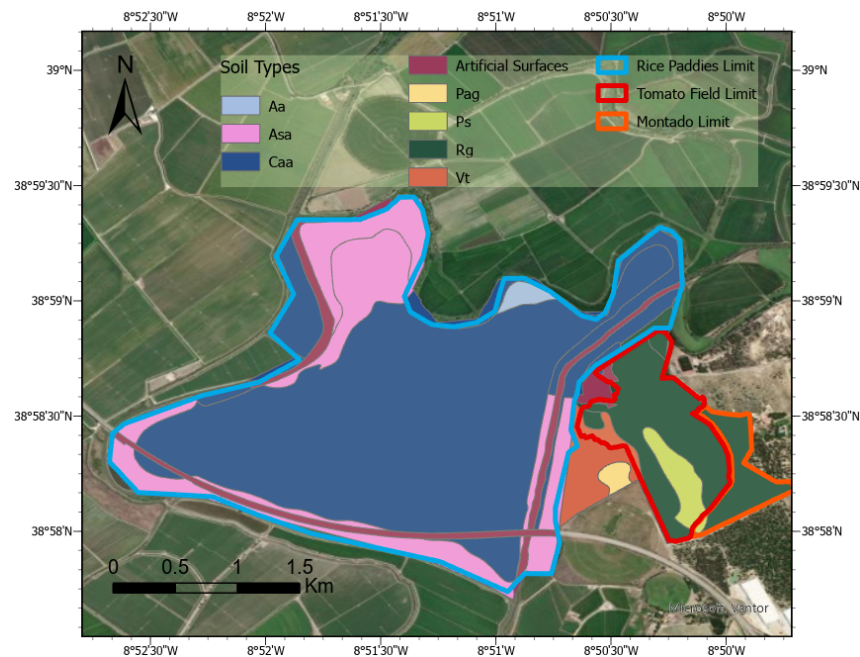


Figure 6. Distribution and soil types in field at Quinta da Foz (Source: Agrogos). Legend: Aa—incipient soils, modern alluvial soils, non-calcareous, heavy texture; Asa—halomorphonic soils, moderately saline, heavy-textured alluvial soils; Caa—heavy-textured hydromorphic soils (without eluvial horizon) and para-alluvial/para-colluvial material; Pag—slightly unsaturated clay-enriched soils; Ps—hydromorphic soils with eluvial horizon and planosols of arenites or argillaceous conglomerates or clays; Rg—incipient soils, psammitic regosols, normal, not moist; Vt—lithic, non-humic, slightly unsaturated soils on coarse sandstone. (Source: Soil type legend adapted from the Direção-Geral de Agricultura e Desenvolvimento Rural [18].)

The spatial distribution of the NDVI values at the time of sampling is shown in Figure 7. In the Montado area, satellite imagery from 27 August 2023 was used for point selection. This date corresponded to the end of the dry summer period, when NDVI values reached their annual minimum. Under these conditions, the vegetation vigor mainly reflects soil characteristics rather than active plant growth, allowing for better differentiation of soil-driven variability.

In the tomato fields, the Sentinel-2 satellite images revealed pronounced heterogeneity during the growing season, and on 17 August 2023, part of the field showed almost no vegetative cover, while other sections displayed relatively high NDVI values (Figure 7A). This contrast indicated strong variability in soil properties and crop growth. To adequately capture this variability, the tomato field was subdivided into three subzones, each representing distinct NDVI patterns, and the sampling dates were adjusted according to the last observed NDVI decline in each subzone: late July for Subzone 1, and 1 September for Subzones 2 and 3.

In the rice paddies, vegetation growth began in June and extended until October, following the typical flooded paddy cycle. Soil sampling was scheduled for the end of September, immediately after the peak of the growth stage, when the NDVI values stabilized before harvest and this timing ensured that the collected samples corresponded to the phase of maximum plant–soil interaction, capturing spatial differences in soil fertility that directly influence crop vigor.

Overall, the NDVI-based approach successfully delineated areas of contrasting vegetation performance within each land-use system, ensuring that the soil samples represented the full spectrum of spatial variability present across the landscape.

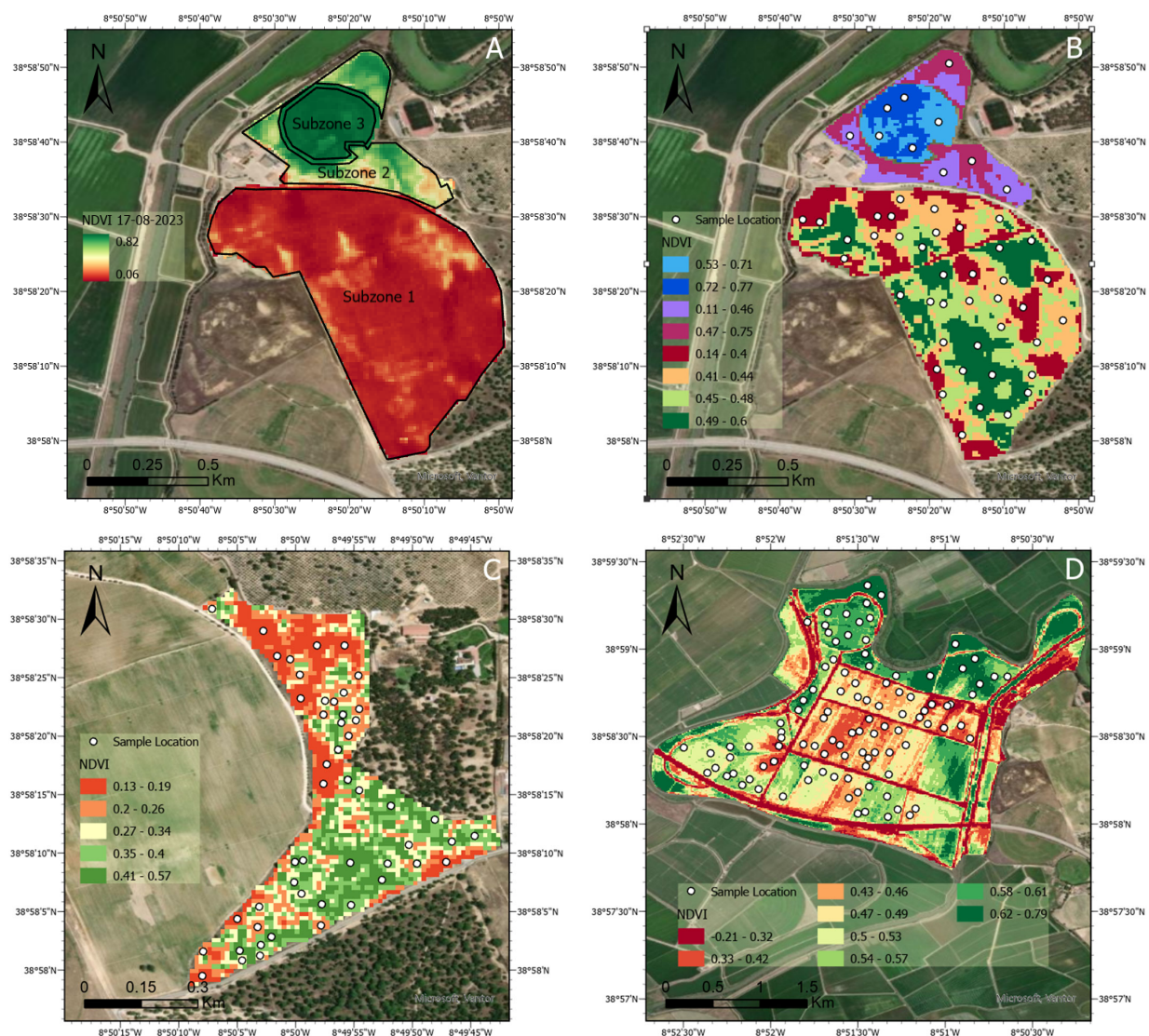


Figure 7. Determination of sampling points according to NDVI classes, showing the subdivision of the tomato field area into three subzones (A); the location of the samples in the tomato field (B); the location of the samples in the Montado (C); and the location of the samples in the rice field (D). Maps were produced in ArcGIS Pro (version 3.2.2), Esri World Imagery was used as the basemap, and all spatial data were projected in ETRS89/PT-TM06.

3.3. Model Dataset, Statistical Performance, and Predictive Evaluation

The datasets were used as the basis for developing and evaluating the Random Forest (RF) and Partial Least Squares Regression (PLSR) models used to predict soil fertility properties. As such, the statistical performance of the models was analyzed based on the application of both algorithms, enabling an assessment of the predictive capacity of NIR spectroscopy for estimating key soil properties. Overall, the RF model consistently outperformed PLSR, demonstrating greater robustness and accuracy, and was therefore selected to generate predictions for those samples analyzed only spectrally. The statistical performance results for both models are summarized in Tables 3 and 4 and the models achieved variable performance across soil properties, with RF generally outperforming PLSR.

Table 3. Statistical results for the Random Forest model (RF). Legend: RMSE—Root Mean Square Error; R^2 —Coefficient of Determination; RPD—Ratio of Performance to Deviation; Cal—calibration result; CV—cross-validation result; NC—number of components.

Variable	RF					
	RMSE Cal	RMSE CV	R^2 Cal	R^2 CV	RPD Cal	RPD CV
Mg	53.85	139.825	0.991	0.939	10.571	4.071
MO_LOI	0.406	1.019	0.984	0.897	7.863	3.131
Fe	39.279	111.127	0.987	0.896	8.81	3.114
Ni	0.156	0.406	0.983	0.887	7.773	2.99
Ca	210.462	534.074	0.979	0.866	6.974	2.748
Na	12.645	37.163	0.983	0.851	7.648	2.602
Co	0.148	0.419	0.981	0.847	7.272	2.569
K	21.596	60.273	0.98	0.842	7.055	2.528
B	0.121	0.314	0.961	0.736	5.077	1.956
Mn	17.856	50.295	0.965	0.725	5.397	1.916
MnAI	68.529	195.966	0.964	0.71	5.333	1.865
Cu	0.449	1.163	0.927	0.511	3.722	1.438
S	22.088	58.129	0.927	0.496	3.727	1.416
P	42.446	122.491	0.931	0.425	3.826	1.326
CE	0.031	0.086	0.916	0.355	3.467	1.252
Zn	1.329	3.43	0.883	0.219	2.934	1.137
pH	0.303	0.815	0.879	0.126	2.889	1.075
Nitrogen_T	0.024	0.063	0.869	0.077	2.779	1.046
TOC	0.721	2.152	0.889	0.011	3.018	1.011
MO	0.52	1.405	0.86	−0.027	2.681	0.992
Carbon_T	0.311	0.826	0.85	−0.055	2.599	0.979
Ratio_C/N	0.812	2.254	0.848	−0.167	2.581	0.93
IC	0.169	0.459	0.837	−0.201	2.489	2.489
C_microb	2.666	8.089	0.867	−0.225	2.757	0.909
F_TC	2.996	9.226	0.852	−0.407	2.611	0.848

Table 4. Statistical results for the PLSR model. Legend: RMSE—Root Mean Square Error; R^2 —Coefficient of Determination; RPD—Ratio of Performance to Deviation; Cal—calibration result; CV—cross-validation result; NC—number of components.

Variable	PLSR						
	RMSE Cal	RMSE CV	R^2 Cal	R^2 CV	RPD Cal	RPD CV	NC
Mg	127.587	147.217	0.949	0.932	4.462	3.867	5
MO_LOI	0.969	1.163	0.907	0.866	3.291	2.742	5
Fe	61.387	118.662	0.968	0.881	5.637	2.916	8
Ni	0.254	0.456	0.956	0.858	4.774	2.667	8
Ca	483.488	581.598	0.89	0.841	3.036	2.524	5
Na	30.305	35.64	0.901	0.863	3.191	2.713	5
Co	0.398	0.475	0.862	0.803	2.705	2.263	5
K	53.284	61.681	0.876	0.834	2.859	2.47	5
B	0.29	0.311	0.776	0.742	2.124	1.977	4
Mn	43.282	46.132	0.796	0.769	2.227	2.089	4
MnAI	165.294	175.986	0.793	0.766	2.211	2.077	4
Cu	0.401	1.179	0.942	0.498	4.173	1.419	12
S	40.726	60.018	0.753	0.463	2.022	1.372	7
P	95.607	114.758	0.65	0.495	1.699	1.415	5
CE	0.079	0.079	0.454	0.448	1.36	1.36	2
Zn	2.672	3.16	0.526	0.337	1.459	1.234	5
pH	0.532	0.72	0.627	0.316	1.646	1.215	6
Nitrogen_T	0.056	0.059	0.263	0.176	1.171	1.107	2
TOC	1.648	1.952	0.42	0.186	1.32	1.114	5
MO	0.991	1.238	0.489	0.202	1.406	1.125	5
Carbon_T	0.671	0.767	0.304	0.09	1.205	1.054	4
Ratio_C/N	1.896	2.076	0.174	0.009	1.106	1.01	4
IC	0.395	0.41	0.111	0.044	1.066	1.028	2
C_microb	7.233	7.459	0.021	−0.042	1.016	0.985	1
F_TC	7.774	8.146	0.001	−0.097	1.006	0.96	1

Considering Tables 3 and 4, both models presented high coefficients of determination (R^2) for several soil parameters. Excellent calibration and cross-validation results (R^2_{cal} and $R^2_{cv} > 0.80$) were obtained for Mg, MO_LOI, Fe, Ni, Ca, Na, Co, and K, with corresponding RPD values above 2.5. For B, total Mn, and available Mn (MnAI), model performance was satisfactory ($0.70 < R^2 < 0.80$; RPD of about 2). The remaining variables, including Cu, S, CE, Zn, pH, total C, total N, TOC, IC, and microbial carbon (C_{microb}), exhibited lower predictive ability, with R^2_{cv} values below 0.70. On the other hand, for the microbial carbon and TOC-derived fractions, RF calibration results were moderately good, but cross-validation values indicated weak prediction reliability.

The scatterplots of predicted versus observed values (Figure 8) show two distinct clusters corresponding to different soil types: the nutrient-rich rice soils and the coarser-textured Montado and tomato soils. The models differentiated these groups but showed reduced precision within each group.

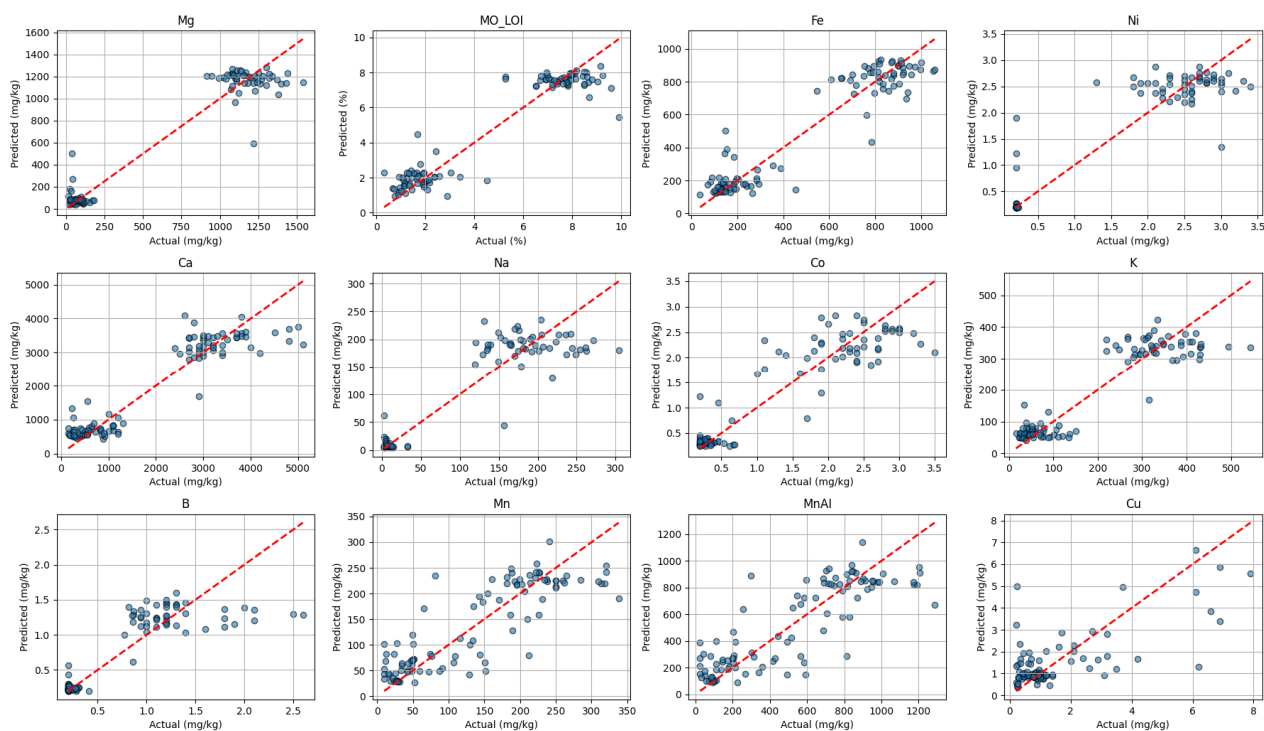


Figure 8. Actual vs. predicted values for the chemical properties in the cross-validation.

Moreover, following the performance comparison, the RF model was selected to generate predicted chemical values for samples analyzed only by NIR. Thus, all resulting predictions were compiled in a single database, including sample identification codes, corresponding land-use categories, and geographic coordinates.

3.4. Prediction Maps

Prediction maps were generated for the five soil properties with the highest calibration performance: magnesium (Mg), iron (Fe), organic matter (MO_LOI), nickel (Ni), and sodium (Na). These are shown in Figure 9 for the Montado, rice paddy, and tomato field areas.

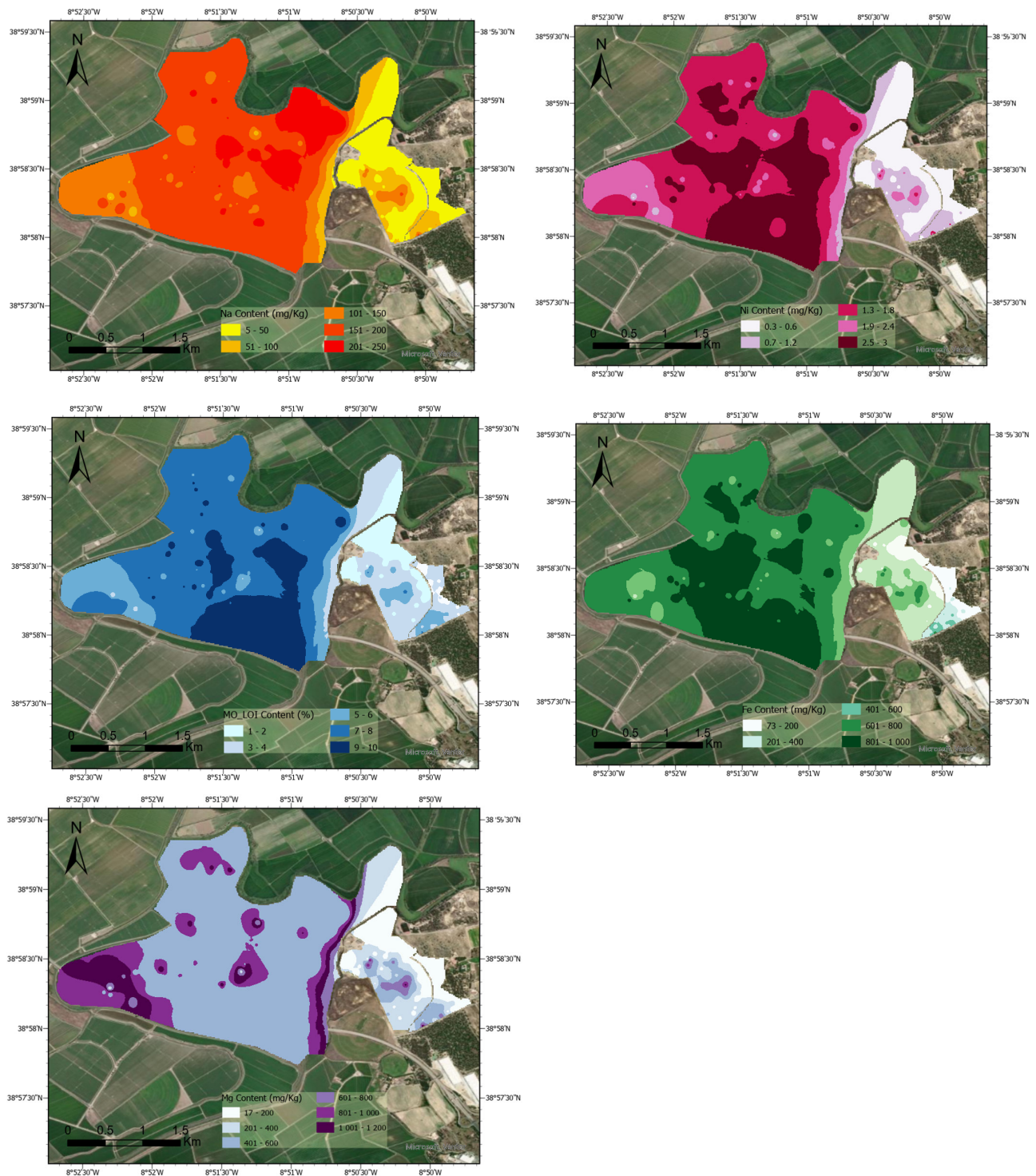


Figure 9. Maps with predicted values of Na, Ni, OM_LOI, Fe, and Mg for the tomato field, rice field, and Montado areas. Each map was created with a class symbology representing the range of predicted values. Maps were produced in ArcGIS Pro (version 3.2.2), Esri World Imagery was used as the basemap, and all spatial data were projected in ETRS89/PT-TM06.

The maps were produced using Geographic Information System (GIS) software, applying the Inverse Distance Weighted (IDW) interpolation method to the predicted values obtained from the Random Forest model. This interpolation generated continuous raster surfaces that represent the spatial variation of each property across the study area. Moreover, the resulting maps clearly display areas of higher and lower concentration for each element, allowing visualization of the spatial distribution of soil fertility within and between

the three land-use systems. For each property, the color gradients indicate the intensity and extent of predicted values, with distinct zones corresponding to different fertility levels. Thus, the maps provide a spatial overview that can be used to identify zones of contrasting soil quality and to delineate potential management zones within each agricultural system. The results illustrate the capacity of the predictive models to capture local variability and translate it into spatial patterns suitable for precision agriculture applications.

4. Discussion

Understanding the spatial variation of soil nutrients is essential for optimizing crop management and implementing precision agriculture strategies. This study made this variation and the potential to use NIR to measure it evident. The use of the sampling selection method proved effective for linking remote sensing indicators with ground-based soil measurements, thereby improving the efficiency and representativeness of the sampling design.

The soil chemical properties varied clearly among the Montado, rice, and tomato fields, reflecting differences in land use, management, and water regimes. The tomato soils were more acidic with high phosphorus, whereas the rice soils were enriched in K, Ca, Mg, S, and micronutrients, and showed elevated Na due to flooding. These distinct chemical profiles provided a challenging test for the spectral models, requiring them to generalize across contrasting pedological environments.

The results show that NIR has high potential for estimating this variation in soil chemical properties, although with variable performance among the parameters.

Organic carbon was predicted with greater accuracy, which is in agreement with studies that highlight the strong spectral response of organic matter [19,20]. In contrast, nutrients such as phosphorus and potassium exhibited a lower predictive capacity, a result also reported in other works due to the weak direct expression of these elements in the spectrum and the influence of local physicochemical characteristics [14,21–23].

This pattern—enabling accurate prediction of organic carbon and elements closely associated with organic or mineral spectral features, and poor prediction of P and some other analytes—is consistent with established findings in soil spectroscopy, where properties with a strong direct spectral expression (e.g., organic matter, certain minerals) are typically more predictable than analytes that are only indirectly expressed in the spectrum, as described by Wang et al. [24] and Santana et al. [25].

Consequently, NIR should be considered a suitable screening or mapping tool for these lower predictive properties only when combined with strong local calibrations or with ancillary predictors (soil texture, carbonate content, farm management history), or when estimating sorption indices rather than absolute extractable concentrations [23].

In our work, RF consistently outperformed PLSR across most properties and was therefore selected to produce the prediction maps. This confirms the suitability of machine learning for handling complex relationships between wavelength and soil properties, as described in previous works, for example, Vestergaard et al. [12], and identifying non-linear patterns compared to linear regression methods like PLSR. While PLSR struggles when the relationship between the target variable and the spectra is not linear (often the case in soils with mixed mineralogy), noisy RF is able to model these intricate interactions.

However, the observed residual variability points to limitations in the representativeness of the sampling, a factor frequently identified as conditioning the precision in soil spectroscopy [20].

The separation into groups observed in the graphs suggests interference from textural or mineralogical composition differences, aspects recognized as determinants in the interactions between radiation and soil. These effects explain the non-linear patterns captured by

the models and reinforce the need for large and heterogeneous calibration sets to improve the generalization of the results.

The presence of distinct clusters for rice soils versus tomato/Montado soils highlights the challenge of “global” calibration; similar difficulties were reported by Viscarra Rossel and Behrens [26,27], who found that splitting models by soil type or texture class often improves predictive performance compared to a single universal model.

For example, this includes soils with intermediate properties, to fill the spectral space between the two groups [28].

In addition, if intended, a model could be used in a different region, or the area could be updated in the future; Purushothaman et al. [29] recommend that approximately 20% of the samples from the new set should be added to the calibration set.

The practical implications for precision agriculture follow directly from these findings, allowing for the rapid monitoring of essential parameters and reducing analytical costs.

The high accuracy obtained for some of the properties indicates that NIR spectroscopy combined with RF can deliver operational, field-scale maps that are suitable to inform site-specific fertilization and management (for example, variable-rate application zones).

However, the moderate performance for some nutrients suggests that NIR-derived estimates should complement—and not replace—laboratory analyses in high-demand contexts.

To improve future investigations into robustness and generalizability, we recommend: (1) larger datasets, expanding the calibration database across multiple soil types and management regimes to capture broader mineralogical and textural variability; (2) testing and preprocessing workflows systematically because optimized preprocessing substantially affects model performance; and (3) testing more advanced non-linear algorithms and hybrid modeling strategies and inclusion of additional predictors (e.g., texture, moisture, geospatial covariates) or testing the integration of other spectral domains, which Malley et al. [30] suggest may offer better specificity for mineral nutrients.

The creation of regional calibration databases could also increase the robustness of the models and expand their operational use [12,31].

5. Conclusions

This study demonstrates the potential of near-infrared (NIR) spectroscopy combined with machine learning techniques to support rapid and cost-effective assessment of soil fertility across contrasting agricultural systems. Thus, by integrating NDVI-guided sampling, laboratory soil analyses, and advanced regression models, it was possible to evaluate the capacity of NIR spectra to predict multiple soil chemical properties and carbon fractions. In this context, the results show that the Random Forest models outperformed Partial Least Squares Regression for most variables, achieving high predictive accuracy for elements such as Mg, Fe, Ni, Na, Ca, K, and organic matter. These properties exhibited strong and consistent relationships with NIR spectral information, allowing reliable spatial prediction and mapping. In contrast, nutrients such as P, S, Zn, pH, total nitrogen, and several carbon fractions showed limited predictive performance, highlighting limitations of NIR spectroscopy for analytes with weak or indirect spectral responses. Additionally, the spatial interpolation of predicted values allowed the generation of detailed soil fertility maps, showing the applicability of the proposed approach for identifying zones of contrasting nutrient availability within rice paddies, tomato fields, and Montado ecosystems. Moreover, these outputs demonstrate clear potential to support precision agriculture practices, contributing to more efficient soil management and reduced analytical costs. As such, future research should focus on expanding sample size, incorporating additional soil properties such as texture, testing alternative modeling approaches, and developing regional spectral libraries to improve model transferability and robustness.

Author Contributions: Conceptualization, M.N., M.S., I.F., C.H. and A.R.F.C.; formal analysis, M.N. and I.F.; investigation, M.N., I.F. and J.R.S.; writing—original draft preparation, M.N., C.H., I.F., A.R.F.C., A.C.M. and M.S.; writing—review and editing, A.R.F.C., M.S., M.N. and C.H.; supervision, M.S. and C.H.; project administration, M.S., J.R.S. and C.H.; funding acquisition M.S., J.R.S. and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundação para a Ciência e Tecnologia (FCT) under project number UID/00329/2025 and by Sistema de Incentivos à Competitividade Empresarial Investigação e Desenvolvimento Empresarial operação n.º 21100; The authors acknowledge the Research Unit Geosciences, Geotechnology and Geoengineering (GeoBioTec), Ref. UID/04035/2025, for its support.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors thank NOVA School of Sciences and Technology, especially the Earth Sciences Department, the Faculty of Sciences of the University of Lisbon, especially to the Centre for Ecology, Evolution and Environmental Changes (cE3c) & Global Change and Sustainability Institute (CHANGE).

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

References

- Jin, T.; Han, X. Robotic Arms in Precision Agriculture: A Comprehensive Review of the Technologies, Applications, Challenges, and Future Prospects. *Comput. Electron. Agric.* **2024**, *221*, 108938. [CrossRef]
- Cravero, A.; Sepúlveda, S. Use and Adaptations of Machine Learning in Big Data—Applications in Real Cases in Agriculture. *Electronics* **2021**, *10*, 552. [CrossRef]
- International Society of Precision Agriculture. Precision Agriculture Definition. 2024. Available online: <https://www.ispag.org/resources/definition> (accessed on 10 December 2025).
- Cozzolino, D.; Morón, A. Potential of Near-Infrared Reflectance Spectroscopy and Chemometrics to Predict Soil Organic Carbon Fractions. *Soil Tillage Res.* **2006**, *85*, 78–85. [CrossRef]
- Morgan, C.L.; Waiser, T.H.; Brown, D.J.; Hallmark, C.T. Simulated In Situ Characterization of Soil Organic and Inorganic Carbon with Visible Near-Infrared Diffuse Reflectance Spectroscopy. *Geoderma* **2009**, *151*, 249–256. [CrossRef]
- Vohland, M.; Besold, J.; Hill, J.; Fründ, H.C. Comparing Different Multivariate Calibration Methods for the Determination of Soil Organic Carbon Pools with Visible to Near Infrared Spectroscopy. *Geoderma* **2011**, *166*, 198–205. [CrossRef]
- Du, C.; Zhou, J. Evaluation of Soil Fertility Using Infrared Spectroscopy—A Review. In *Climate Change, Intercropping, Pest Control and Beneficial Microorganisms*; Lichtfouse, E., Ed.; Springer: Dordrecht, The Netherlands, 2009; Sustainable Agriculture Reviews, Volume 2. [CrossRef]
- Post, W.M.; Izaurralde, R.C.; Mann, L.K.; Bliss, N. Monitoring and Verifying Changes of Organic Carbon in Soil. *Clim. Change* **2001**, *51*, 73–99. [CrossRef]
- Badgley, G.; Chay, F.; Chegwidden, O.S.; Hamman, J.J.; Freeman, J.; Cullenward, D. California’s Forest Carbon Offsets Buffer Pool Is Severely Undercapitalized. *Front. For. Glob. Change* **2022**, *5*, 930426. [CrossRef]
- Neves, A. Técnicas de Análise Laboratorial Para Solos, Matéria Vegetal e Corretivos Orgânicos. Master’s Thesis, Universidade do Minho, Braga, Portugal, 2015.
- Robertson, S. *Direct Estimation of Organic Matter by Loss on Ignition: Methods*; Simon Fraser University, Geography and Soil Science Lab: Burnaby, BC, Canada, 2011; unpublished work.
- Vestergaard, R.J.; Vasava, H.B.; Aspinall, D.; Chen, S.; Gillespie, A.; Adamchuk, V.; Biswas, A. Evaluation of optimized preprocessing and modeling algorithms for prediction of soil properties using VIS-NIR spectroscopy. *Sensors* **2021**, *21*, 6745. [CrossRef]
- Hao, X.; Chen, Z.; Yi, S.; Liu, J. Application of improved stacking ensemble learning in NIR spectral modeling of corn seed germination rate. *Chemom. Intell. Lab. Syst.* **2023**, *243*, 105020. [CrossRef]
- Soriano-Disla, J.M.; Janik, L.J.; Viscarra Rossel, R.A.; Macdonald, L.M.; McLaughlin, M.J. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl. Spectrosc. Rev.* **2013**, *49*, 139–186. [CrossRef]

15. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*, 1st ed.; Sections 5.1.1 and 5.1.4; Springer: New York, NY, USA, 2013.
16. Penn State Eberly College of Science. 10.6—Cross-Validation. In *STAT 501: Regression Methods*; The Pennsylvania State University: University Park, PA, USA, 2024; Available online: <https://online.stat.psu.edu/stat501/lesson/10/10.6> (accessed on 5 February 2026).
17. Zhang, L.; Yang, X.M.; Drury, C.F.; Chantigny, M.; Gregorich, E.; Miller, J.; Bittman, S.; Reynolds, W.; Yang, J. Infrared spectroscopy prediction of organic carbon and total nitrogen in soil and particulate organic matter from diverse Canadian agricultural regions. *Can. J. Soil Sci.* **2017**, *98*, 77–90. [[CrossRef](#)]
18. Direção-Geral de Agricultura e Desenvolvimento Rural (DGADR). Nota Explicativa da Carta dos Solos de Portugal e da Carta de Capacidade de Uso do Solo. Available online: <https://www.dgadr.gov.pt/pt/nota-explicativa> (accessed on 4 February 2026).
19. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* **2010**, *107*, 163–215.
20. Viscarra Rossel, R.A.; Walvoort, D.J.J.; McBratney, A.B.; Janik, L.J.; Skjemstad, J.O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75. [[CrossRef](#)]
21. Chang, C.W.; Laird, D.A.; Mausbach, M.J.; Hurburgh, C.R., Jr. Near-Infrared Reflectance Spectroscopy—Principal Components Regression Analysis of Soil Properties. *Soil Sci. Soc. Am. J.* **2001**, *65*, 480–490. [[CrossRef](#)]
22. Vyavahare, G.D.; Yun, J.-J.; Park, J.-H.; Shim, J.; Kim, S.H.; Kim, K.; Roh, A.; Jang, H.-J.; Jeon, S. Evaluating the Performance of NIR Spectroscopy in Predicting Soil Properties: A Comparative Study. *Appl. Sci.* **2025**, *15*, 13240. [[CrossRef](#)]
23. Eslamifar, M.; Tavakoli, H.; Thiessen, E.; Kock, R.; Correa, J.; Hartung, E. Effective Spectral Pre-Processing Methods Enhance Accuracy of Soil Property Prediction by NIR Spectroscopy. *Discov. Appl. Sci.* **2025**, *7*, 896. [[CrossRef](#)]
24. Wang, Z.; Chen, S.; Lu, R.; Zhang, X.; Ma, Y.; Shi, Z. Non-linear Memory-Based Learning for Predicting Soil Properties Using a Regional Vis-NIR Spectral Library. *Geoderma* **2024**, *441*, 116752. [[CrossRef](#)]
25. de Santana, F.B.; de Souza, A.M.; Poppi, R.J. Visible and Near Infrared Spectroscopy Coupled to Random Forest to Quantify Some Soil Quality Parameters. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2018**, *195*, 182–190. [[CrossRef](#)]
26. Abdi, D.; Tremblay, G.F.; Ziadi, N.; Bélanger, G.; Parent, L.-É. Predicting Soil Phosphorus-Related Properties Using Near-Infrared Reflectance Spectroscopy. *Soil Sci. Soc. Am. J.* **2012**, *76*, 2318–2326. [[CrossRef](#)]
27. Viscarra Rossel, R.A.; Behrens, T. Using Data Mining to Model and Interpret Soil Diffuse Reflectance Spectra. *Geoderma* **2010**, *158*, 46–54. [[CrossRef](#)]
28. Food and Agriculture Organization of the United Nations. *A Primer on Soil Analysis Using Visible and Near-Infrared (Vis-NIR) and Mid-Infrared (MIR) Spectroscopy (Revision)*; FAO: Rome, Italy, 2024. [[CrossRef](#)]
29. Purushothaman, N.K.; Garg, K.K.; Venkataradha, A.; Anantha, K.; Singh, R.; Jat, M.; Das, B.S. Applicability of Calibrated Diffuse Reflectance Spectroscopy Models Across Spatial and Temporal Boundaries. *Geoderma* **2024**, *449*, 117012. [[CrossRef](#)]
30. Malley, D.F.; Martin, P.D.; Ben-Dor, E. Application in Analysis of Soils. In *Near-Infrared Spectroscopy in Agriculture*; Roberts, C.A., Workman, J., Reeves, J.B., Eds.; ASA, CSSA, and SSSA: Madison, WI, USA, 2004; pp. 729–784. [[CrossRef](#)]
31. Canero, F.M.; Rodriguez-Galiano, V.; Aragonés, D. Machine Learning and Feature Selection for Soil Spectroscopy: An Evaluation of Random Forest Wrappers to Predict Soil Organic Matter, Clay, and Carbonates. *Heliyon* **2024**, *10*, e30228. [[CrossRef](#)] [[PubMed](#)]

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.