Multispectral Sensors and Machine Learning as Modern Tools for Nutrient Content Prediction in Soil

Rafael Felippe Ratke [1], Paulo Roberto Nunes Viana [1], Larissa Pereira Ribeiro Teodoro [1], Fábio Henrique Rojo Baio [1], Paulo Eduardo Teodoro [1], Dthenifer Cordeiro Santana [1], Carlos Eduardo da Silva Santos [2], Alan Mario Zuffo [3] and Jorge González Aguilera [4,*]

1   Department of Agronomic, Federal University of Mato Grosso do Sul, Rodovia MS-306, km 105, Zona Rural, Chapadão do Sul 79560-000, MS, Brazil; rafael.ratke@ufms.br (R.F.R.); paulo.viana@ufms.br (P.R.N.V.); larissa_ribeiro@ufms.br (L.P.R.T.); fabiobaio@ufms.br (F.H.R.B.); paulo.teodoro@ufms.br (P.E.T.); dthenifer.santana@unesp.br (D.C.S.)

2   Federal Institute of Education, Science and Technology of the Tocantins, Quadra Ae 310 Sul, Av. NS 10, S/N-Plano Diretor Sul, Palmas 77021-090, TO, Brazil; carlosedu@ifto.edu.br

3   Department of Agronomic, State University of Maranhão, Praça Gonçalves Dias, s/n, Centro, Balsas 65800-000, MA, Brazil; alanzuffo@professor.uema.br

4   Department of Crop Science, State University of Mato Grosso do Sul, Cassilândia 79540-000, MS, Brazil

*   Correspondence: jorge.aguilera@uems.br

Abstract: The combination of multispectral data and machine learning provides effective and flexible monitoring of the soil nutrient content, which consequently positively impacts plant productivity and food security, and ultimately promotes sustainable agricultural development overall. The aim of this study was to investigate the associations between spectral variables and soil physicochemical attributes, as well as to predict these attributes using spectral variables as inputs in machine learning models. One thousand soil samples were selected from agricultural areas 0–20 cm deep and collected from Northeast Mato Grosso do Sul state of Brazil. A total of 20 g of the dried and homogenized soil sample was added to the Petri dish to perform spectral measurements. Reflectance spectra were obtained by CROP CIRCLE ACS-470 using three spectral bands: green (532–550 nm), red (670–700 nm), and red-edge (730–760 nm). The models were developed with the aid of the Weka environment to predict the soil chemical attributes via the obtained dataset. The models tested were linear regression, random forest (RF), reptree M5P, multilayer preference neural network, and decision tree algorithms, with the correlation coefficient (r) and mean absolute error (MAE) used as accuracy parameters. According to our findings, sulfur exhibited a correlation greater than 0.6 and a reduced mean absolute error, with better performance for the M5P and RF algorithms. On the other hand, the macronutrients S, Ca, Mg, and K presented modest r values (approximately 0.3), indicating a moderate correlation with actual observations, which are not recommended for use in soil analysis. This soil analysis technique requires more refined correlation models for accurate prediction.

Keywords: machine learning; soil analysis; reflectance spectra

1. Introduction

Ferralsols constitute 32.9% of Brazil's population [1]. Ferralsols are characterized by low levels of plant nutrients such as Ca, Mg, and P and high levels of Al [2]. Soybean is most commonly grown in the Ferralsols of Brazil [3].

Modern agriculture requires automation for analysis and management processes. Agricultural automation is well known for the ability to diagnose and observe pests and diseases [4,5]. Automation in soil analysis processes is complex and requires further study.

The assessment of soil properties, including a series of chemical processes, aims to determine the ability of the soil to provide specific nutrients necessary for the plant pathway cycle. The most common approach adopted for this purpose consists of extracting

a chemical solution using extractants that simulate the absorption of nutrients by plants. However, the use of this traditional method of soil analysis raises environmental challenges, notably owing to the inadequate disposal of waste generated during the assessment of soil properties, including a range of chemical processes, which aims to determine the ability of the soil to provide specific nutrients essential for the plant growth cycle [6].

The current method not only presents risks for the professionals involved, but also has adverse environmental impacts due to the presence of chemical residues, in addition to incurring high costs related to the acquisition of chemical reagents and requiring a considerable time lag. In this context, sustainable and efficient analytical approaches that minimize environmental impacts and optimize the effectiveness of the soil assessment process are needed to guarantee more efficient agricultural practices [7].

Multispectral soil analysis involves the use of various spectroscopic techniques to evaluate soil properties [8]. Different methods have been used for soil analysis, such as microspectrophotometry, X-ray fluorescence spectroscopy, and laser-induced breakdown spectroscopy [9]. These techniques provide information on the composition, purity, and elemental content of soils, helping to discriminate between different types and sources of soils [10]. Data fusion methodologies have been applied to improve classification accuracy, especially when combining information from several spectroscopic analyzers. Multispectral analysis can capture the spectral dimensionality of soils, providing valuable information on the variability of soil elements, despite limitations in the resolution of narrowband absorption.

Soil spectral analysis calibration is crucial for accurately predicting soil properties. Soil parameters such as potassium, phosphorus, and organic matter are already being evaluated by visible-near-infrared (Vis-NIR) spectroscopy, but it requires calibration [11]. Calibration methods involve preprocessing transformations, variable selection techniques, and regression algorithms to increase prediction accuracy [10]. Methods of calibration are employed to calibrate soil spectral data, including preprocessing transformations, variable selection techniques, and regression algorithms [12]. Using spectral libraries and reducing sample processing levels have shown potential for lowering costs and time implications for predicting soil properties such as organic carbon, clay, and pH [13]. Overall, proper calibration methods are essential for leveraging soil spectral analysis to monitor soil properties effectively and contribute to precision agriculture.

The use of indirect analysis through multispectral sensors has enabled expeditious, economically viable, and ecologically sustainable monitoring of elementary soil levels. Moreover, the integration of machine learning (ML) algorithms has proven crucial for obtaining reliable estimates in this context [14]. The synergistic combination of these approaches provides effective and agile monitoring of the soil nutrient content, which is highly relevant for agricultural soil productivity, food security, and the promotion of sustainable agricultural development [15]. The convergence of these techniques not only optimizes the speed and efficiency of monitoring, but also contributes to mitigating the environmental challenges inherent to traditional soil analyses, such as the production of chemical residues.

The central problem of this study is the difficulty in quickly and accurately assessing and monitoring soil physicochemical attributes, which are crucial for the proper and sustainable management of agricultural resources. Traditional soil analysis methods, such as laboratory collection and analysis, are generally time-consuming, expensive, and limited in terms of spatial coverage, which makes large-scale and real-time monitoring difficult. In this context, the use of spectral variables as indirect indicators of soil properties emerges as a promising alternative. However, an important issue is the complexity of the relationship between spectral variables and soil physicochemical attributes, which can vary depending on the type of soil, moisture, and presence of organic matter, among other factors. Therefore, there is a technical challenge in building machine learning models capable of capturing these relationships in a robust and generalizable way so that they can be applied in different scenarios. The objective of the current investigation was to analyze

the associations between the spectral and physicochemical variables of soil in addition to predicting the physicochemical attribute levels of soil via the use of spectral variables as inputs into machine learning models.

2. Materials and Methods

2.1. Sample Collection and Determination of Physicochemical Properties

Soil samples were collected at 0 to 20 cm depth from the municipalities of Cassilândia, Chapadão do Sul, Costa Rica, and Paraíso das Águas (18°46′26″ S 52°37′28″ W, average altitude of 810 m of sea level), with a coverage area of 16,130.84 km^2, located in the State of Mato Grosso do Sul (MS), Brazil. The regional climate is classified as humid tropical, with a rainy season in summer and a dry season in winter, with an average annual rainfall of 1.850 mm, an average annual temperature of 20.5 °C, and a variation of 7.5 °C.

The soil in the region is mostly classified as Rhodic Ferralsol [16]. A total of 33% of the 1000 samples analyzed were characterized as sandy, 25% as sandy loam, 24% as clay loam, and 20% as clay. The soil samples were collected with different augers, i.e., probe-type augers (20 mm diameter) and screw-type augers, at depths of 0–0.20 m. The soil samples were sieved through a 2 mm mesh and air-dried. The elements Ca, Mg, and K were analyzed in the Exata Brasil Laboratory located in Chapadão do Sul-MS.

KCl solution (1 mol L^{-1}) at a ratio of 1/10 (soil:solution) was used to extract Ca and Mg from the soil. The element potassium (K) was extracted from the Mehlich1 solution (0.05 mol L^{-1} HCl + 0.0125 mol L^{-1} H$_2$SO$_4$) at a ratio of 1/10 (soil:solution). The ammonium acetate solution in a proportion of 10 g of soil to 25 mL of the solution was used to extract S from the soil. Ca, Mg, K, and S contents in the soil extracts were measured via Argon Plasma Optical Atomic Emission Spectrometry (ICP-OES) (Perkin Elmer, Waltham, MA, USA).

Multispectral evaluations were carried out in a 20 g aliquot of each sieved, dried, and homogenized soil sample, which was subsequently added to a Petri dish for spectral measurements (Figure 1). The Petri dish was placed on a flat bench, and the sensor was installed 8 cm from the soil surface. The area of incidence of the spectral beam was 3 cm^2. Two external 50 W halogen lamps were positioned 35 cm from the Petri dish at a zenith angle of 30°, forming a 90° angle to each other following the method described by Franceschini et al. [14].
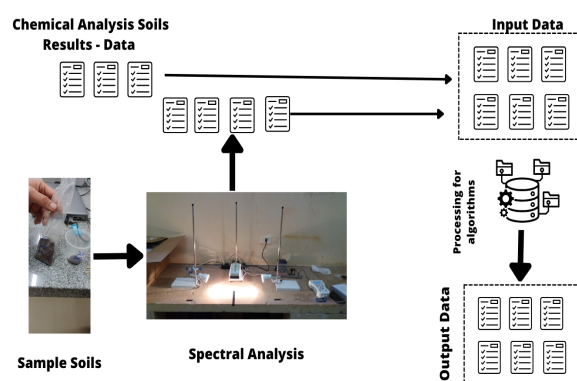
Figure 1. Illustration of spectral analysis and data processing by machine learning. Images are the property of the author.

The reflectance spectra were obtained with a CROP CIRCLE ACS-470 instrument (Holland Scientific, Inc., Lincoln, NE, USA). The six spectral bands used were green (532–550 nm), red (670–700 nm), and red edge (730–760 nm). The sensor was calibrated via FieldCal SC-1. The spectral bands were applied to the surface of the soil samples in 100 replicates for each band. Reflectances were recorded in spreadsheets, and reflectance averages were calculated for each spectral band.

2.2. Data Analysis via Computational Intelligence

The data were subjected to observation and compared via the WEKA (Waikato Environment for Knowledge Analysis) software version 3.9.3(c) 1999–2018, which was accessed by a computer with an AMD Phenom™ IIx4 B97 processor 3.20 GHz, installed memory RAM 4 GB, 32-bit operating system, Windows 7, using cross-validation with 10 folds (K = 10) and 10 repetitions (100 runs) in a spectral analysis of 690 samples with wavelength data obtained as input values and macronutrients as output values to be predicted for Ca, Mg, and K. The data prediction analysis used 370 samples for S.

The models tested were random forest (RF), multilayer perceptron (MLP), decision trees (M5P), REPTree (REPT), and random trees (RTs). All the parameters adopted were set to the default software configuration. The tested models were selected with applicability in other agronomic works according to Refs. [13,14]. MLP is a type of neural network that excels at solving supervised learning problems with multiple inputs. It consists of layers of neurons (or perceptrons) organized into an input layer, one or more hidden layers, and an output layer. Each neuron in a layer is connected to the neurons in the next layer by adjustable weights. In Weka software, the default MLP configuration includes a single hidden layer with a number of neurons defined by the average between the number of attributes and the number of classes, which generally provides a good balance between learning capacity and computational efficiency.

M5P provides more information on mathematical equations and addresses categorical and continuous variables and missing values. This model combines features of decision trees with linear regression, making it particularly useful for continuous and mixed-category data. It builds decision trees where the leaves contain linear equations that facilitate prediction. This allows M5P to handle both categorical and continuous variables and manage missing values, providing a more interpretable model by providing insight into the underlying mathematical relationships between variables.

The random forest (RF) algorithm uses multiple independent decision trees and combines their predictions. It is particularly effective for large-scale problems because of its robustness against overfitting, and facilitates data interpretation by allowing the assessment of variable importance. A random tree (RT) is used to build a decision tree with a random dataset through the division of nodes.

The REPTree algorithm builds decision trees via regression logic in multiple iterations. In each iteration, it evaluates several trees, selecting the best one on the basis of regression error criteria. This model allows a robust pruning approach, where the final tree is simplified to improve the generalizability of the model. The averages of the S, Mg, Ca, and K contents of the actual data and the predicted data of the samples randomly selected via machine learning were contrasted in scatter and line graphs via SigmaPlot 11.0 software.

The accuracy of the prediction models was evaluated by the correlation between the predicted and observed values (r) and the mean absolute error (MAE). The accuracy values of each of the tested models were subjected to analysis of variance to verify the existence of significant differences between the machine learning models. Subsequently, boxplots were generated for r and MAE for each model in the case of macronutrient prediction.

The means of the performance parameters were grouped via the Scott–Knott test at 5% probability. The boxplots and groupings of means were generated via the ggplot2 and ExpDes.pt packages of R software.

For the Pearson correlation coefficient (r), we applied the criterion adapted by Figueiredo Filho and Silva Júnior [15], which is classified into three categories—low, medium, and high—and is considered low when r is approximately 0.10–0.30, moderate when r is between 0.40 and 0.60, and high when r is 0.70–1.

3. Results

The averages of the S, Mg, K, and Ca contents in the analyzed soil show a disparity between the predicted data due to accumulated error (Figure 2). The S content in the soil predicted by machine learning analyses via MLP, M5P, RF, RT, and REPT exhibited

significant dispersion compared with the chemically analyzed mean content (Figure 2A). Conversely, the Mg and K contents predicted by MLP, M5P, RF, and REPT showed low dispersion relative to the mean content, not aligning with the values obtained through chemical analysis (Figure 2B,C). In this context, the use of RT resulted in greater dispersion in the prediction of Mg, K, and Ca contents. However, the mean Ca content predicted by RT was the closest to the real value found in the chemical analyses (Figure 2D).
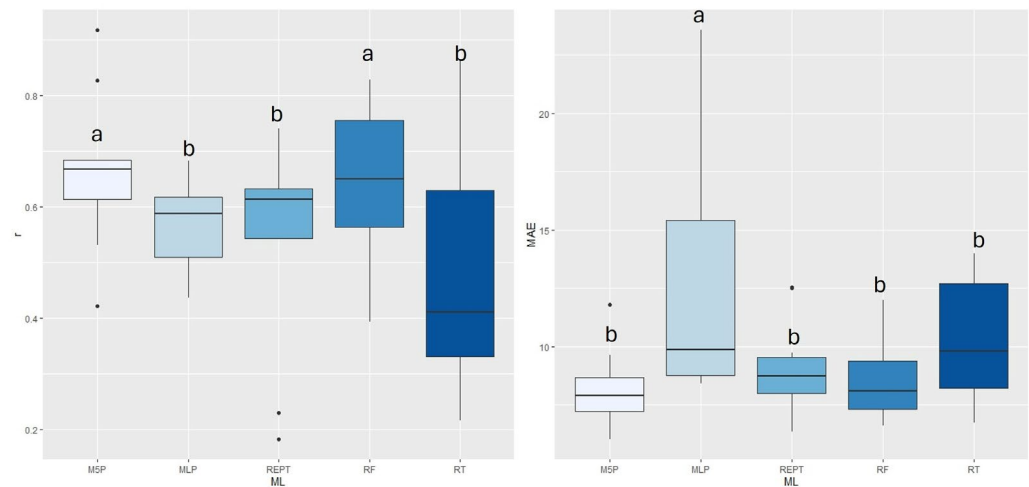

(A)

(B)

(C)

(D)
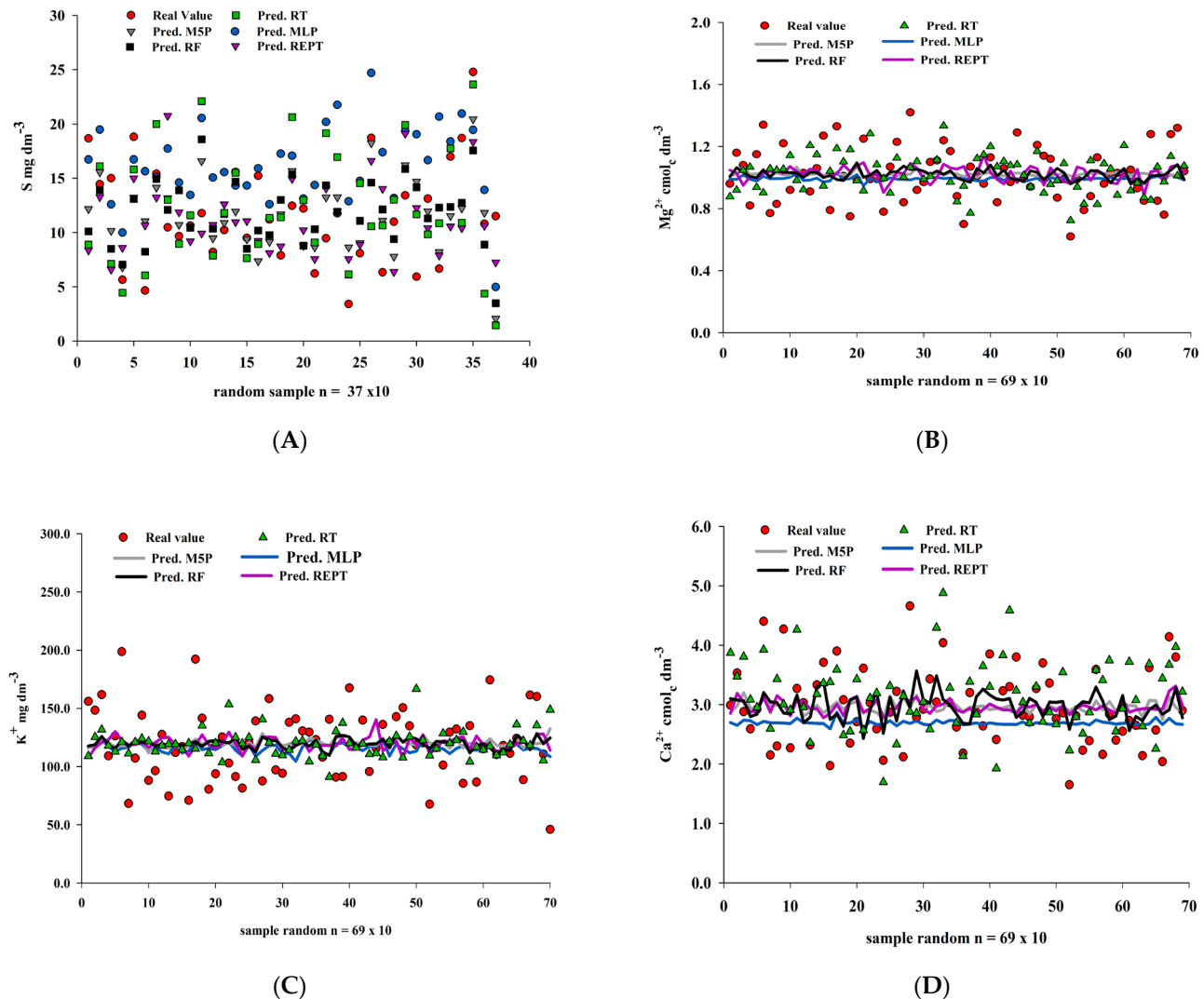
Figure 2. Boxplots of the Pearson correlation coefficient (r, to the left) and mean absolute error (MAE, on the right) for different sulfur-related machine learning models: random forest (RF), multilayer Perceptron (MLP), decision trees (M5P), REPTree (REPT), and random tree (RT). Mean levels of S (**A**), Mg^{+2} (**B**), K^+ (**C**) and Ca^{+2} (**D**) of the soil chemically analyzed and predicted by different algorithms.

With respect to the prediction of the sulfur content, the M5P and RF algorithms outperformed the other algorithms (Figure 3), presenting high r values (higher than 0.6). This value guarantees the high accuracy of these algorithms in estimating the sulfur content on the basis of spectral reflectance. Another factor that contributes to the accuracy of both algorithms is the low MAE, indicating smaller errors in the prediction, ensuring greater precision of these algorithms when predicting S contents.

With respect to the performance of the algorithms in predicting the magnesium (Mg) content, the results revealed that the random forest (RF) algorithm was superior in terms of accuracy (Figure 4). This could be translated into greater consistency between the predictions and observed values, as revealed by the correlation coefficient (r), which

surpassed those of the other algorithms. Additionally, the RF demonstrated a lower mean absolute error (MAE) value, denoting a significantly high precision in its estimates.

Figure 3. Boxplots of the Pearson correlation coefficient (r, to the left) and mean absolute error (MAE, on the right) for different sulfur-related machine learning models: random forest (RF), multilayer Perceptron (MLP), decision trees (M5P), REPTree (REPT), and random tree (RT). Different lowercase letters about the boxplots represent statistical differences at 5% probability by the Scott–Knott test.
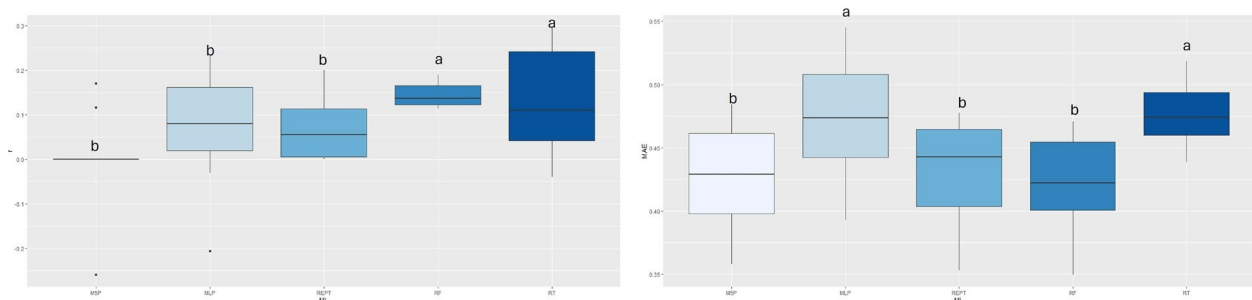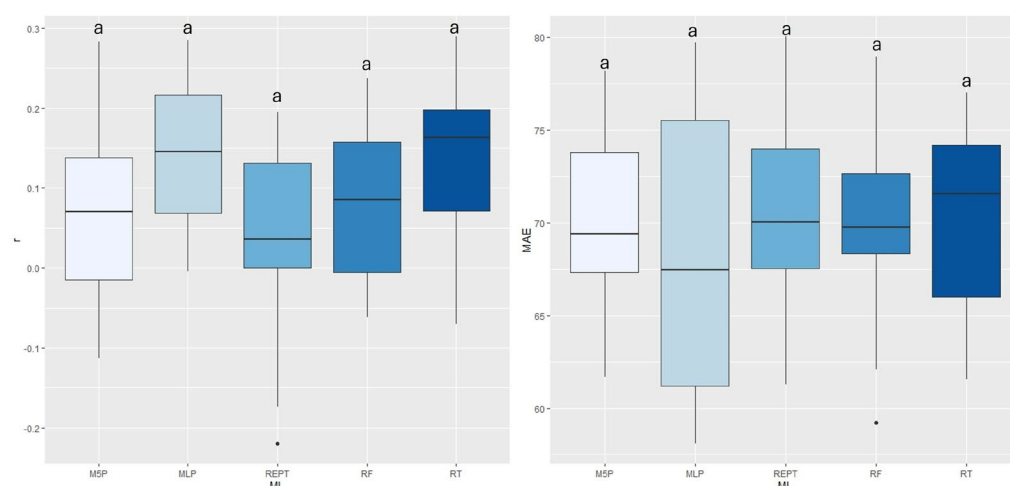

Figure 4. Boxplots of the Pearson correlation coefficient (r, to the left) and mean absolute error (MAE, on the right) for different magnesium-related machine learning models. Different lowercase letters about the boxplots represent statistical differences at 5% probability by the Scott–Knott test.

The RT algorithm exhibited similar behavior to that of the RF when evaluated in relation to the accuracy indicator (r). However, it is important to highlight that RT revealed a significantly high MAE value, which indicates that although it presents relative consistency in predictions, it is not an accurate algorithm for predicting magnesium content.

Statistically, the algorithms had the same behavior for the correlation coefficient (r) in potassium prediction. There was also no significant difference in terms of error. Therefore, using algorithms that maintain better performance for the other elements facilitates processing because, in the case of potassium, all the models have the same performance (Figure 5). The prediction value was approximately 0.3, indicating a moderate value that may be considerably adequate in terms of the variability and dynamicity of P in the soil.

On the other hand, in the analysis of the prediction of the calcium (Ca) content, the M5P algorithm demonstrated superior performance in relation to the other algorithms, as evidenced by correlation coefficient values (r) that approached 0.3 (Figure 6). Furthermore, notably, the M5P algorithm achieved notably lower mean absolute error (MAE) values, approximately 1.50, with the maximum MAE value for predicting this nutrient. These results emphasize the ability of the M5P algorithm to generate accurate estimates, with a relatively low level of error, which is extremely relevant for estimating the calcium content in soil samples via multispectral data.

Figure 5. Boxplots of the Pearson correlation coefficient (r, to the left) and mean absolute error (MAE, on the right) for different potassium-related machine learning models. Different lowercase letters about the boxplots represent statistical differences at 5% probability by the Scott–Knott test.
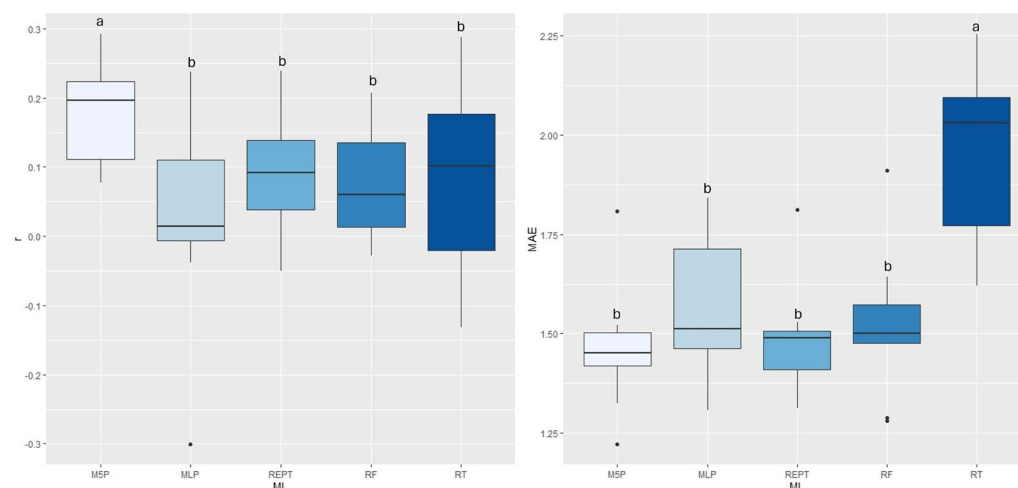
Figure 6. Boxplots of the Pearson correlation coefficient (r, to the left) and mean absolute error (MAE, on the right) for different calcium-related machine learning models. Different lowercase letters about boxplots represent statistical differences at 5% probability by the Scott–Knott test.

In short, the M5P and RF algorithms performed satisfactorily in terms of the correlation coefficient (r). Specifically, these algorithms achieve values remarkably close to the actual values, which is especially evident when the sulfur content is predicted, where the r value is close to 0.8. For the other elements analyzed, the accuracy remained at approximately 0.3, which is moderate and consistently shows that, owing to the dynamism and complexity of these soil elements, the use of multispectral reflectance to determine them is promising. Furthermore, the M5P and RF algorithms yielded lower MAE values, further reinforcing the reliability of the predictions generated by these algorithms. Reducing the error contributes to greater precision in the estimates, improving the accuracy of the predictions obtained via multispectral reflectance.

4. Discussion

The results indicate a significant disparity between the expected concentrations of sulfur (S), magnesium (Mg), potassium (K), and calcium (Ca) and the measurements derived from chemical analysis. The inconsistency of the predicted data is influenced by the quality of the data and noise in the accuracy of the predictive models [17]. Data variations amplify imprecision, particularly in more complex predictive models [18]. However, it is

essential to validate data analyses with accurate predictive models. In this context, the RT predictive model provided better accuracy for S and Ca than did the actual analyzed data.

Traditional methods for estimating soil nutrients, such as laboratory analysis, are recognized for their accuracy, but they have significant limitations in terms of time and cost [10]. These methods require specialized labor, and the use of chemical reagents, in addition to being expensive, can pose environmental risks due to the generation of potentially contaminating waste [10]. This process becomes unfeasible for large-scale and timely application, which contrasts with the growing demand for fast, economical, and sustainable methodologies in the agricultural sector. In this context, the use of multispectral sensors combined with machine learning techniques has emerged as a promising alternative for estimating soil nutrients in a more efficient and environmentally friendly way. The combination of these tools allows data to be obtained noninvasively and in real time, facilitating continuous and large-scale monitoring of soil attributes. In this study, three specific multispectral bands were used to predict soil nutrient levels, employing different machine learning algorithms [6]. This approach offers a potentially faster and more accessible methodology that can contribute to more sustainable and precise agricultural practices, enabling more effective soil management in response to the needs of agricultural production [7].

Among the investigated nutrients, S had the highest predictive value, close to 0.80 for the correlation coefficient I, and notably low values of the mean absolute error (MAE), confirming its ability to offer highly accurate predictions (Figure 2). High prediction values were achieved by the M5P and RF algorithms, highlighting the robustness and reliability of the algorithms in the task of predicting S content. Both algorithms performed well because of their high r and low MAE values, and their use in other agricultural tasks, such as predicting soil organic carbon, stands out [19]. With satisfactory results in predicting soil nutrients, RF can be used to infer soil fertility [20]. The soil nutrients significantly influence the distribution of soil organic carbon [21].

The other predictions yielded median values, with a certain significance, highlighting the complexity of predicting potassium content through the evaluated algorithms, which is particularly relevant in agronomic and soil fertility contexts, where the precise estimation of Mg, K, and Ca contents is crucial for effective and sustainable agricultural practices. In magnesium prediction, the RF algorithm stands out as a superior choice, guaranteeing greater accuracy. These results highlight the relevance of careful selection of the appropriate algorithm for a given task. Dharumarajan et al. [22] reported that RF was the best model for most soil properties, from macro- to micronutrients, indicating that the RF model is better for solving multivariate adjustment problems since RF combines many trees to form an accurate prediction mechanism. In addition, the RF algorithm, when necessary, has fewer parameters to adjust.

In the potassium prediction results, there was homogeneity in the results regarding r, suggesting that the analyzed algorithms exhibited a moderate correlation with the real observations of potassium content, although they did not reach higher levels of correlation, indicating that the prediction of this element can be challenging. In a similar study, Forkuor et al. [23] reported that no machine learning algorithm works best for all global situations, and models must be tested to calibrate them to identify an accurate model for predicting soil properties, optimizing data processing.

The analysis of the calcium prediction (Ca) content revealed that the M5P algorithm performed better than the other algorithms did, as evidenced by correlation coefficient values that approached 0.3. This result suggests that M5P established a moderate and consistent r with actual observations of Ca content, indicating its ability to provide meaningful estimates.

The superior performance of M5P in predicting S and Ca, for which the algorithm stands out, highlights its applicability and usefulness in agronomic contexts, where accurate estimation of this nutrient is crucial for adequate soil management and the development of effective agricultural practices. This contributes to increased productivity and sustainability

of agricultural activities. According to previous studies, this model performs well in several tasks in different areas, such as predicting cadmium in agricultural soils [24]. M5P also presents good results in the physical and chemical prediction of soil and water due to its greater accuracy and speed than the regression model [25]. This diversity of accuracies in different situations demonstrates the generalizability and robustness of the algorithm.

The complexity of S forms in soils contributes to variability in spectral measurements. Sulfur occurs in organic and inorganic forms in soil and is transformed by microorganisms in the soil [26]. In this sense, organosulfur compounds are stable over time, whereas others may decompose or convert to other forms, leading to variability in spectral measurements [27].

Ca, Mg, and K interact with each other and with the soil matrix, which changes their chemical bonding structures with other elements present and their spectral expressions [28]. Calcium and magnesium have the capacity to generate carbonates and additional compounds that affect the reflectance of soil samples [29]. Potassium influences the spectral properties of soil through interactions with clay [28].

Our findings demonstrated the effectiveness of the M5P and RF algorithms in predicting soil nutrients, particularly the S content, where the accuracy reached notable levels. The ability of these models to provide estimates close to real values has significant implications for agriculture and soil management, promoting decision-making on the basis of reliable data and contributing to more efficient and sustainable agricultural practices. The greatest contribution of these technologies is to reduce the work involved in analyzing samples and reducing the use of reagents in laboratory analyses, making this part of the procedure faster, requiring and dispensing with the use of expensive reagents from laboratories that report adequate disposal, which are not always served. With future research applying the algorithms found here, it will be possible to adapt and use such technologies with remote sensors or prototypes to be used in situ. Soil samples from other locations and a larger number of samples can be used to increase the accuracy of the algorithms. The use of hyperspectral sensors can also improve predictive value in addition to their application in agriculture, which enables real-time monitoring in agricultural scenarios.

5. Conclusions

In the present study, the use of the CROP CIRCLE ACS-470 multispectral sensor associated with machine learning was demonstrated to be a promising approach for predicting soil macronutrients, especially sulfur, with correlations between actual and estimated values above 0.6. However, regarding the macronutrients P, K, Ca, and Mg, the prediction accuracy reached values of approximately 0.3, indicating a moderate and coherent correlation with actual observations; however, the development of more refined models is needed to improve the results. These findings highlight the reliability and accuracy of predictions, thus strengthening the usefulness and effectiveness of the sensor in the context of soil analysis and agricultural decision-making.

The use of multispectral sensors and data prediction analysis via the M5P and RF algorithms derived from our results are directly applicable to areas with characteristics similar to those of Rhodic Ferralsol and Arenosol soils. For regions with soils of different compositions, we suggest conducting complementary studies to adapt the proposed practices to local conditions.

Author Contributions: Conceptualization, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B. and P.E.T.; methodology, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B., P.E.T., D.C.S. and C.E.d.S.S.; software, P.R.N.V., L.P.R.T., F.H.R.B., P.E.T. and D.C.S.; validation, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B., P.E.T., D.C.S., C.E.d.S.S., A.M.Z. and J.G.A.; formal analysis, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B., P.E.T., D.C.S. and C.E.d.S.S.; investigation, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B., P.E.T., D.C.S. and C.E.d.S.S.; resources, R.F.R., A.M.Z. and J.G.A.; data curation, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B., D.C.S., C.E.d.S.S., A.M.Z. and J.G.A.; writing—original draft preparation, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B., P.E.T., D.C.S., C.E.d.S.S., A.M.Z. and J.G.A.; writing—review and editing, R.F.R., P.R.N.V., L.P.R.T., F.H.R.B., P.E.T., D.C.S., C.E.d.S.S., A.M.Z. and J.G.A.; visualization, R.F.R., P.R.N.V., L.P.R.T. and F.H.R.B. All authors have read and agreed to the published version of the manuscript.

References

1. IBGE, Portal do IBGE. *Structural Provinces, Relief Compartments, Soil Types and Phytoecological Regions*; IBGE: Rio de Janeiro, Brazil, 2019.
2. Ratke, R.F.; Campos, A.R.; Inda, A.V.; Barbosa, R.S.; Jacques, Y.; Bezerra, A.; César, J.; Nóbrega, A.; Batista, J. Agricultural Potential and Soil Use Based on the Pedogenetic Properties of Soils from the Cerrado-Caatinga Transition. *Semin. Agrar.* **2020**, *41*, 1119–1134. [CrossRef]
3. Brannstrom, C.; Jepson, W.; Filippi, A.M.; Redo, D.; Xu, Z.; Ganesh, S. Land Change in the Brazilian Savanna (Cerrado), 1986-2002: Comparative Analysis and Implications for Land-Use Policy. *Land Use Policy* **2008**, *25*, 579–595. [CrossRef]
4. Dai, G.; Tian, Z.; Fan, J.; Sunil, C.K.; Dewi, C. DFN-PSAN: Multi-Level Deep Information Feature Fusion Extraction Network for Interpretable Plant Disease Classification. *Comput. Electron. Agric.* **2024**, *216*, 108481. [CrossRef]
5. Dai, G.; Fan, J.; Dewi, C. ITF-WPI: Image and Text Based Cross-Modal Feature Fusion Model for Wolfberry Pest Recognition. *Comput. Electron. Agric.* **2023**, *212*, 108129. [CrossRef]
6. Demattê, J.A.M.; Ramirez-Lopez, L.; Marques, K.P.P.; Rodella, A.A. Chemometric Soil Analysis on the Determination of Specific Bands for the Detection of Magnesium and Potassium by Spectroscopy. *Geoderma* **2017**, *288*, 8–22. [CrossRef]
7. Terra, F.S.; Demattê, J.A.M.; Viscarra Rossel, R.A. Spectral Libraries for Quantitative Analyses of Tropical Brazilian Soils: Comparing Vis-NIR and Mid-IR Reflectance Data. *Geoderma* **2015**, *255–256*, 81–93. [CrossRef]
8. Selkowitz, D.J. A Comparison of Multi-Spectral, Multi-Angular, and Multi-Temporal Remote Sensing Datasets for Fractional Shrub Canopy Mapping in Arctic Alaska. *Remote Sens. Environ.* **2010**, *114*, 1338–1352. [CrossRef]
9. Demattê, J.A.M. Characterization and Discrimination of Soils by Their Reflected Electromagnetic Energy. *Pesqui. Agropecu. Bras.* **2002**, *37*, 1445–1458. [CrossRef]
10. Breure, T.S.; Prout, J.M.; Haefele, S.M.; Milne, A.E.; Hannam, J.A.; Moreno-Rojas, S.; Corstanje, R. Comparing the Effect of Different Sample Conditions and Spectral Libraries on the Prediction Accuracy of Soil Properties from Near- and Mid-Infrared Spectra at the Field-Scale. *Soil Tillage Res.* **2022**, *215*, 105196. [CrossRef]
11. Guo, P.; Li, T.; Gao, H.; Chen, X.; Cui, Y.; Huang, Y. Evaluating Calibration and Spectral Variable Selection Methods for Predicting Three Soil Nutrients Using Vis-Nir Spectroscopy. *Remote Sens.* **2021**, *13*, 4000. [CrossRef]
12. Shepherd, K.D.; Ferguson, R.; Hoover, D.; van Egmond, F.; Sanderman, J.; Ge, Y. A Global Soil Spectral Calibration Library and Estimation Service. *Soil Secur.* **2022**, *7*, 100061. [CrossRef]
13. Mohammedzein, M.A.; Csorba, A.; Rotich, B.; Justin, P.N.; Melenya, C.; Andrei, Y.; Micheli, E. Development of Hungarian Spectral Library: Prediction of Soil Properties and Applications. *Eurasian J. Soil Sci.* **2023**, *12*, 244–256. [CrossRef]
14. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.-M.; McBratney, A. Critical Review of Chemometric Indicators Commonly Used for Assessing the Quality of the Prediction of Soil Attributes by NIR Spectroscopy. *TrAC Trends Anal. Chem.* **2010**, *29*, 1073–1081. [CrossRef]
15. Peng, Y.; Zhao, L.; Hu, Y.; Wang, G.; Wang, L.; Liu, Z. Prediction of Soil Nutrient Contents Using Visible and Near-Infrared Reflectance Spectroscopy. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 437. [CrossRef]
16. International Union of Soil Science (IUSS). *World Reference Base for Soil Resources (WRB)*; World Soil; FAO: Rome, Italy, 2015.
17. Awais, M.; Naqvi, S.M.Z.A.; Zhang, H.; Li, L.; Zhang, W.; Awwad, F.A.; Ismail, E.A.A.; Khan, M.I.; Raghavan, V.; Hu, J. AI and Machine Learning for Soil Analysis: An Assessment of Sustainable Agricultural Practices. *Bioresour. Bioprocess.* **2023**, *10*, 90. [CrossRef]
18. Padarian, J.; Minasny, B.; McBratney, A.B. Machine Learning and Soil Sciences: A Review Aided by Machine Learning Tools. *SOIL* **2020**, *6*, 35–52. [CrossRef]
19. Chen, J.; Zhang, H.; Fan, M.; Chen, F.; Gao, C. Machine-Learning-Based Prediction and Key Factor Identification of the Organic Carbon in Riverine Floodplain Soils with Intensive Agricultural Practices. *J. Soils Sediments* **2021**, *21*, 2896–2907. [CrossRef]
20. Jeong, G.; Oeverdieck, H.; Park, S.J.; Huwe, B.; Ließ, M. Spatial Soil Nutrients Prediction Using Three Supervised Learning Methods for Assessment of Land Potentials in Complex Terrain. *CATENA* **2017**, *154*, 73–84. [CrossRef]
21. John, K.; Abraham Isong, I.; Michael Kebonye, N.; Okon Ayito, E.; Chapman Agyeman, P.; Marcus Afu, S. Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil. *Land* **2020**, *9*, 487. [CrossRef]

22. Dharumarajan, S.; Lalitha, M.; Niranjana, K.; Hegde, R. Evaluation of Digital Soil Mapping Approach for Predicting Soil Fertility Parameters—a Case Study from Karnataka Plateau, India. *Arab. J. Geosci.* **2022**, *15*, 386. [CrossRef]

23. Forkuor, G.; Hounkpatin, O.K.L.; Welp, G.; Thiel, M. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLoS ONE* **2017**, *12*, e0170478. [CrossRef] [PubMed]

24. Agyeman, P.C.; Khosravi, V.; Michael Kebonye, N.; John, K.; Borůvka, L.; Vašát, R. Using Spectral Indices and Terrain Attribute Datasets and Their Combination in the Prediction of Cadmium Content in Agricultural Soil. *Comput. Electron. Agric.* **2022**, *198*, 107077. [CrossRef]

25. Morteza, R.; Ghasemnezhad, A.; Ghorbani, K.; Khodayar, H.; Abhari, A. Prediction of Saffron Flower and Stigma Yield Based on the Physical and Chemical Properties of Water and Soil Using Linear Multivariate Regression Models and M5 Decision Tree. *J. Saffron Res.* **2022**, *9*, 352–367.

26. Kar, G.; Schoenau, J.J.; Gillespie, A.W.; Dhillon, G.S.; Peak, D. Sulfur Species Formed in the Seed Row of Sulfur-Fertilized Soils as Revealed by K-Edge X-ray Absorption Near-Edge Structure Spectroscopy. *Soil Sci. Soc. Am. J.* **2019**, *83*, 1324–1332. [CrossRef]

27. Anunciado, M.B.; De Boskey, M.; Haines, L.; Lindskog, K.; Dombek, T.; Takahama, S.; Dillner, A.M. Stability Assessment of Organic Sulfur and Organosulfate Compounds in Filter Samples for Quantification by Fourier- Transform Infrared Spectroscopy. *Atmos. Meas. Tech.* **2023**, *16*, 3515–3529. [CrossRef]

28. Israr, M.A.; Abbas, Q.; Haq, S.-U.; Nadeem, A. Compositional Analysis of Soil Using Calibration-Free Laser-Induced Breakdown Spectroscopy. *Spectrosc. Lett.* **2022**, *55*, 350–361. [CrossRef]

29. Santos, H.S.; Nguyen, H.; Venâncio, F.; Ramteke, D.; Zevenhoven, R.; Kinnunen, P. Mechanisms of Mg Carbonates Precipitation and Implications for CO_2 Capture and Utilization/Storage. *Inorg. Chem. Front.* **2023**, *10*, 2507–2546. [CrossRef]