A Comparative Study of Hyperparameter Optimization Methods for Soil Organic Carbon Stock Prediction Using Machine Learning

Alexandre Pardelinha[1], Douglas O. Cardoso[2], Laura Assis[1], Kennedy Fernandes[3], Gabriela R. da Costa[4], Jorge Soares[1], Marcos Bacis Ceddia[4], and Diego Brandão[1(✉)]

[1] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rio de Janeiro, Brazil
{alexandre.pardelinha,laura.assis,jorge.soares,diego.brandao}@cefet-rj.br
[2] CeBER, Faculty of Economics, Univ. Coimbra, Av. Dias da Silva 165, 3004-512 Coimbra, Portugal
docardoso@uc.pt
[3] Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil
kennedy.fernandes@ufjf.br
[4] Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, Brazil

Abstract. To understand how carbon moves in soil, we need to know how it is stored there, the called soil organic carbon (SOC). This knowledge is fundamental given that it facilitates the formulation of climate change mitigation policies. More and more people are using machine learning models for this task because they are able to uncover nonlinear and complex relationships among environmental variables. However, getting improved accuracy models requires carefully choosing their hyperparameters, which is often a hard, time-consuming task. This article analyzes the influence of different optimization strategies on hyperparameter identification for machine learning models used to estimate SOC in Brazilian soils. Data from approximately 3,404 Brazilian soil profiles, originating from different biomes, were used. The optimization methods evaluated were grid search, random search, and successive partitioning, using different tree-based models. The results indicate that hyperparameter optimization significantly affects model performance, with slight predictive superiority for the Gradient Boosting method using Grid Search.

Keywords: Soil Organic Carbon · Machine Learning · Hyperparameter Optimization · Environmental Modeling

1 Introduction

Attempting to determine soil organic carbon stocks (SOCS) is essential for understanding and combating climate change, since soil represents one of the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16762, pp. 231–242, 2027.
https://doi.org/10.1007/978-3-032-30524-4_15

largest carbon reservoirs in terrestrial ecosystems [5]. Soil carbon dynamics determine how much carbon is stored or released to the atmosphere, directly affecting atmospheric CO_2 concentrations and thereby influencing aspects of global warming. In countries such as Brazil, where the agricultural sector accounts for a significant portion of the economy, a more accurate estimate of SOCS is particularly relevant, not only to improve land management practices but also to ensure compliance with international commitments, such as the United Nations Sustainable Development Goals (UN SDGs), especially SDG 13 (Climate Action) [4].

Despite its importance, estimating SOC content is a complex task due to the intrinsic heterogeneity of soils, the large geographic area typically monitored, and the temporal variability of carbon levels driven by human activities and weathering [18]. Another complicating factor in estimating SOCS is the use of traditional measurement methods, known as direct methods, which are typically expensive, labor-intensive, and prone to measurement errors during sampling, processing, and laboratory analysis [1]. As a result, large-scale, high-resolution monitoring of soil carbon is significantly limited.

To address these challenges, indirect measurement methods have been increasingly adopted. These methods use indirect data, such as environmental and soil variables, which are easier to obtain and are combined with statistical or machine-learning techniques to estimate SOCS [8,9]. In this context, classical methods such as pedotransfer functions based on regression models or kriging operators are commonly used. However, the performance of these methods is often limited, as they require linearity assumptions and must be calibrated for each region studied [7,15].

As an alternative to traditional indirect measurement methods, machine learning (ML) models have been widely used in recent years. They have emerged as powerful tools for approximating complex, nonlinear relationships and for handling large, heterogeneous datasets [10]. However, the performance of these models depends not only on data quality but also on the appropriate selection of their hyperparameters [19]. This process involves high computational costs, which can affect the model's accuracy and generalization [6].

Thus, hyperparameter optimization is a step that requires careful attention in the data-driven modeling workflow [20]. Although classical techniques, such as *grid search*, *random search*, and *successive halving*, are commonly employed, there are still few comparative studies evaluating their impact on machine learning models, and even fewer when it comes specifically to SOCS estimation for Brazilian soils.

In this study, we aim to fill this gap by systematically comparing three hyperparameter optimization strategies (*grid search*, *random search*, and *successive halving*) across three machine learning models: Decision Tree, Random Forest, and Gradient Boosting. The main contribution of this study is to conduct an empirical comparative evaluation of the effects of different optimization strategies on the predictive performance and computational efficiency of SOCS estimation models.

The remainder of this paper is organized as follows. Section 3 presents related work. Section 2 reviews the theoretical background on machine learning and hyperparameter optimization. Section 4 describes the dataset and methodology. Section 5 presents and discusses the results. Finally, Sect. 6 concludes the paper and outlines directions for future research.

2 Theoretical Background

This section describes aspects of soil's role in the global carbon cycle, the use of machine learning models to describe complex environmental processes, and the impact of hyperparameter optimization on predictive models' performance.

2.1 Global Carbon Cycle

The global carbon cycle describes the continuous exchange of carbon between Earth's major reservoirs, including the atmosphere, oceans, biosphere, and lithosphere [2,3]. Soil is one of the largest terrestrial repositories of carbon. It stores organic carbon in many forms, such as humified organic matter, plant residues, and microbial biomass. The interactions between biological, chemical, and physical processes, like mineralization, decomposition, and stabilization, are what control how soil carbon moves. These processes happen at different times and places, which makes soil carbon different from other types of carbon. Human activities like intensive farming, cutting down trees, and changing how land is used can throw this balance off, which releases more CO2 and makes climate change worse. To understand how carbon moves around and make soil sustainability strategies easier to use, it's important to get an accurate estimate of SOC.

2.2 Machine Learning Techniques

Machine learning (ML) offers a wide range of tools to assist in identifying or determining subtle relationships between predictor variables and target variables, particularly in domains characterized by high variability, noise, and nonlinear effects, such as soil systems [13]. The goal of supervised learning is to derive a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from labeled data, where $\mathcal{X}$ represents the input feature space (variables characterizing the soil, such as sediment, texture, etc.) and $\mathcal{Y}$ denotes the target variable (SOCS). To obtain the learned mapping, we minimize the loss function $\mathcal{L}(y, \hat{y})$ over the training data. This means it performs well on samples we have not seen before.

Here, we focus on tree-based ensemble models due to their trustworthiness and strong performance on tabular environmental datasets, where feature associations together with nonlinearity are prevalent.

Decision Trees (DT) partition the feature space recursively by selecting splits that minimize a given impurity measure, typically the reduction of variance in

regression problems. At each node, the algorithm selects the feature and threshold that maximize the decrease in prediction error. Formally, the split is chosen to minimize:

$$\sum_{i \in R_1} (y_i - \bar{y}_{R_1})^2 + \sum_{i \in R_2} (y_i - \bar{y}_{R_2})^2,$$

where R_1 and R_2 are the resulting regions after the split. Although Decision Trees are highly interpretable and can detect nonlinear relationships, they usually exhibit high variance and are sensitive to small changes in the data, resulting in overfitting.

Random Forest (RF) tackles this shortcoming by constructing an ensemble of decision trees using bootstrap aggregation (bagging). Each tree is trained on a random subset of the data, and at each split, a random subset of features is considered. The final prediction is obtained by averaging the outputs of all trees:

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x),$$

where T is the number of trees. This process reduces variance whilst holding low bias, thereby increasing generalization performance. Key hyperparameters, such as the number of trees, the maximum tree depth, and the number of features considered at each split, critically affect model results.

Gradient Boosting (GBoost) follows a different paradigm by building models sequentially rather than independently. Each new learner is trained to minimize the residual errors of the current ensemble, typically using gradient descent in function space. The model is updated as:

$$f_m(x) = f_{m-1}(x) + \gamma_m h_m(x),$$

where $h_m(x)$ is a weak learner fitted to the negative gradient of the loss function and γ_m is the learning rate. This progressive refinement enables the model to detect complex patterns and interactions with high accuracy. However, Gradient Boosting is more sensitive to hyperparameter choices–such as learning rate, number of estimators, and tree depth–and may overfit without proper regularization.

Overall, tree-based models are particularly suitable for SOCS estimation because they naturally handle nonlinear associations, high-order interactions, and heterogeneous distributions of characteristics, which are commonly observed in soil datasets. Additionally, they are relatively robust to scaling issues and can accommodate mixed types of predictors, making them fitting for environmental modeling tasks.

2.3 Hyperparameter Optimization

Hyperparameters are predefined configuration parameters that control a model's learning process, such as tree depth, the number of estimators, and the learning rate. Unlike model parameters, hyperparameters are not directly learned from data and must be optimized to balance bias and variability.

Formally, hyperparameter optimization can be defined as:

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \mathcal{L}(f_\lambda),$$

where λ represents a set of hyperparameters, Λ is the search space, and $\mathcal{L}$ is a validation loss function.

In this study, three optimization strategies are considered:

Grid Search (GS) performs an exhaustive evaluation over a predefined grid of hyperparameter combinations. Although simple and deterministic, it leads to an exponential computational cost with the number of parameters.

Random Search (RS) samples hyperparameter configurations from predefined distributions. It is more efficient in high-dimensional search spaces, as it explores a wider set of configurations with fewer evaluations.

Successive Halving (SH) is a resource-efficient approach based on iterative elimination. It starts by evaluating many configurations with limited computing resources and progressively allocates more resources to the most promising candidates. This strategy diminishes computational cost while preserving competitive performance.

Selecting an appropriate optimization strategy is particularly important in SOCS estimation, where model effectiveness is highly sensitive to hyperparameter settings.

2.4 Evaluation Metrics

Model performance is assessed using two regression metrics [17]:

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |e_i| \tag{1}$$

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} e_i^2} \tag{2}$$

where $e_i = \hat{y}_i - y_i$ represents the prediction error for sample i. These two metrics are interesting because, while the MAE is less sensitive to outliers, the RMSE penalizes larger errors more severely, making it useful for assessing whether the model remains stable under observations with high variance.

The k-fold cross-validation ($k = 5$) was used both during model training and hyperparameter optimization to prevent overfitting. The idea behind this procedure is to partition the dataset into k approximately equal, mutually exclusive subsets (folds). At each iteration, one fold is used as the validation set, while the remaining $k - 1$ folds are used for training. This process is repeated k times, so that each fold is used exactly once for validation.

The overall performance is then obtained by averaging the evaluation metric across all folds:

$$\hat{\mathcal{L}} = \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_j,$$

where $\mathcal{L}_j$ represents the loss computed on the j-th validation fold.

The use of cross-validation provides a more robust estimate of model performance than traditional holdout strategies, in which a single split is made between training and testing, as it reduces the variance caused by the data split and allows all samples to be used for both training and validation, making it even more valuable in scenarios with limited data availability. In addition to these advantages, it is important to consider that when dealing with heterogeneous datasets, such as soil datasets, spatial variability and non-uniform distributions could otherwise distort the evaluation.

In hyperparameter optimization, cross-validation is used to guide the selection of model configurations by evaluating each candidate set of hyperparameters across multiple folds. This prevents overfitting to a specific subset of the data and leads to more reliable and generalizable models.

It is important to avoid data leakage during model evaluation. Data leakage occurs when information from the validation set is inadvertently used during training, leading to overly optimistic performance estimates. In this study, cross-validation is applied strictly during training, ensuring that each validation fold remains unseen during model fitting.

Furthermore, because the dataset may exhibit inherent structure and potential ordering effects, cross-validation helps mitigate biases arising from a single data split. However, it is acknowledged that in datasets with possible spatial or regional dependencies–such as soil data–additional validation strategies (e.g., spatial cross-validation) could further improve the robustness of performance estimates. This aspect is considered a direction for future work.

3 Related Work

This section reviews related studies on estimating soil carbon stock using machine learning techniques. A search conducted on Google Scholar using the query (“Carbon stocks” AND “machine learning” AND “hyperparameter” AND (“Brazil” OR “Brazilian”)) limited to publications from 2019 onward returned 14 studies. Notably, prior to 2018, only two studies were identified, none of which focused on Brazilian soils.

Among the 14 studies, four used data from China, one focused on landfills, and another addressed image classification using convolutional neural networks. Two studies used machine learning to analyze the impact of straw on soil carbon and phosphorus stocks. Two others used remote sensing and machine learning for Italian soils.

Safanelli (2020) [12] developed a cloud platform for estimating soil carbon stocks in Brazilian and European soils using machine learning and remote sensing data, employing *grid search* for hyperparameter tuning.

Barbosa *et al.* (2023) [11] applied machine learning and remote sensing to estimate both soil carbon and nitrogen stocks, also using *grid search* for hyperparameter optimization.

Santos *et al.* (2023) [14] used satellite vegetation data and machine learning to estimate soil carbon stocks in Brazil, again using *grid search*.

Thus, no prior studies were found comparing different hyperparameter optimization techniques for machine learning models applied to soil carbon estimation in Brazilian soils.

4 Methodology

This section describes the proposed methodology, including the dataset used and the workflow adopted for training machine learning models with hyperparameter optimization.

4.1 Dataset

The dataset used in this study is based on the work of [16] and is extended with additional information on geographical regions according to the classification of the Brazilian Institute of Geography and Statistics (IBGE). To ensure consistency between training and testing subsets, the dataset was reorganized prior to the experiments.

The complete dataset consists of 3,404 records, covering all five geographical regions of Brazil (North, Northeast, Central-West, Southeast, and South). The database was compiled from 69 different sources of publications on soil data in Brazil. This range of sources makes certain that our database is more representative and reliable, allowing a more robust modeling process.

The variables used in this study are described in Table 1.

4.2 Modeling Pipeline

The proposed training pipeline is summarized in Fig. 1, which presents an overview of the main stages: *(i)* Dataset, *(ii)* Preprocessing, *(iii)* Model Training, *(iv)* Hyperparameter Optimization, and *(v)* Performance Evaluation.

Three machine learning models were evaluated: Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GBoost). For each model, three hyperparameter optimization techniques were applied, as described in Sect. 2.3: *Grid Search*, *Random Search*, and *Successive Halving*.

The input data correspond to the training subset described in Sect. 4.1. A preprocessing stage was applied to ensure that the data were suitable for machine learning.

The first step is to choose features based on Pearson's correlation. We only kept variables that had an absolute correlation with the target variable ($SOCS$) that was greater than 0.1. This process makes the problem less complex by getting rid of features that don't add much predictive value.

Table 1. Definition of variables used in the model.

Variable	Description
$SOCS$	Soil organic carbon stock
ds	Soil bulk density
$sand$	Proportion of sand particles in the soil
$silt$	Proportion of silt particles in the soil
$clay$	Proportion of clay particles in the soil
pH	Soil pH in water
sb	Base saturation
al	Aluminum content
h	Hydrogen content
SOC	Soil organic matter content
$Espessura$	Soil thickness

The next step is to use a normalization method on all of the numbers. This process makes sure that all variables are in the same numerical range. This stops differences in magnitude between variables from affecting the learning process, which makes the training algorithm converge faster. We decided to use the rescaling method, which is also called *min-max* normalization, in this study.

Finally, the three tree-based machine learning models were used with each of the three hyperparameter optimization strategies, yielding evaluations under the same experimental conditions, thereby allowing for a fair comparison of the impact of each strategy on predictive performance.

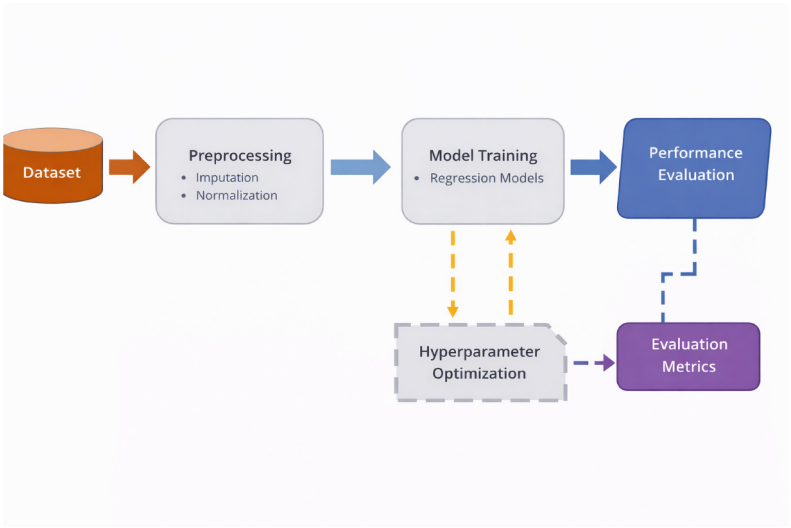

Fig. 1. Overview of the proposed machine learning pipeline.

5 Results

The experiments were conducted using computational routines implemented in Python version 3, on a Dell Inspiron machine equipped with an Intel Core i7-1255U CPU at 3.20 GHz, 10 cores, and 16GB of RAM. The libraries used were *pandas*[1] version 1.5.3, *numpy*[2] version 1.24.2, and *scikit-learn*[3] version 1.2.2. The dataset was split into 80-20% for training and testing, preserving the original order of the data. The hyperparameter optimization methods evaluated were *Random Search* (RS), *Successive Halving* (SH), and *Grid Search* (GS), all combined with cross-validation (k-*Fold* = 5).

Figure 2 presents the correlation among the variables. It can be observed that hydrogen (h), soil thickness ($Espessura$), and soil organic matter are highly correlated with soil organic carbon stock ($SOCS$), while variables such as pH (pH), silt proportion ($silt$), and clay proportion ($clay$) show low correlation. Additionally, it is important to note that the variables related to sand ($sand$) and clay ($clay$) proportions are significantly correlated, suggesting that only one of them may be sufficient for model construction.

Table 2 presents the results considering the evaluation metrics (RMSE and MAE). In terms of MAE, the best result (0.36) was achieved by the Gradient Boosting model (GBoost) using the *Random Search* (RS) technique. This result indicates that the model performs well in minimizing the average absolute prediction error.

Regarding RMSE, the best value (0.93) was also obtained with the GBoost model, using the *Grid Search* (GS) technique for hyperparameter tuning. Considering both metrics, the combination of GBoost with GS achieved the best overall performance, with RMSE and MAE of 0.93 and 0.37, respectively.

Overall, the results indicate that hyperparameter optimization is essential for improving the performance of machine learning models for soil carbon stock prediction.

6 Final Considerations

Accurate estimation of soil organic carbon (SOC) stocks is a crucial step toward understanding and mitigating climate change. Soil carbon plays a key role in the global carbon cycle, as it directly influences atmospheric carbon dioxide concentrations. In Brazil, where the agricultural sector accounts for a significant portion of the national economy, understanding soil carbon dynamics is particularly relevant, as it can support international commitments, such as the United Nations Sustainable Development Goals (SDGs), especially SDG 13, while contributing to sustainable economic development.

As shown, machine learning models have been useful for soil studies due to their ability to handle large datasets and to identify complex relationships among

[1] https://pandas.pydata.org/.
[2] https://numpy.org/.
[3] https://scikit-learn.org/stable/.

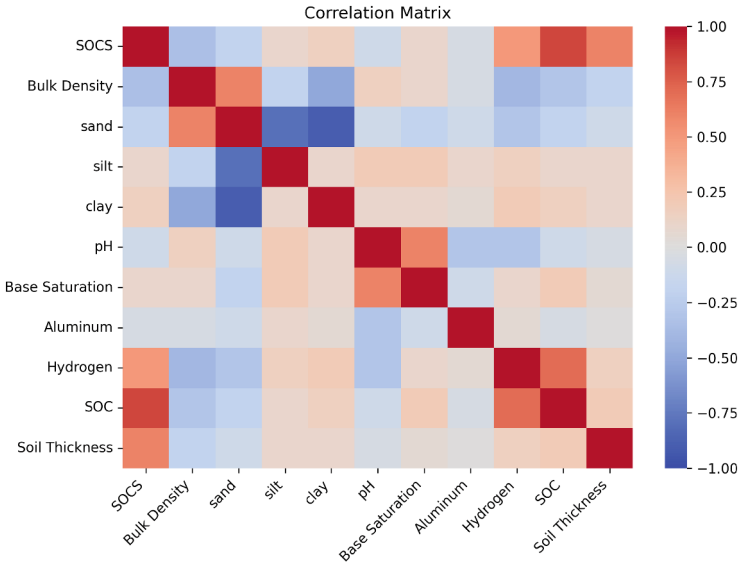

Fig. 2. Correlation matrix of the variables used in the study, highlighting the relationships between soil properties and soil organic carbon stock (SOCS).

Table 2. Performance of machine learning models under different hyperparameter optimization strategies.

Target	Model	Method	RMSE	MAE	Time	Key Hyperparameters
SOCS	RF	RS	0.98	0.50	10 s	$n_iter = 500$, $cv = 5$
	RF	SH	1.14	0.50	10 s	$cv = 5$
	RF	GS	0.94	0.45	10 s	$cv = 5$
	GBoost	RS	0.95	0.36	3 s	$n_iter = 500$, $cv = 5$
	GBoost	SH	0.95	0.41	3 s	$cv = 5$
	GBoost	GS	**0.93**	**0.37**	3 s	$cv = 5$
	DT	RS	1.52	0.50	1 s	$n_iter = 500$, $cv = 5$
	DT	SH	1.54	0.59	1 s	$cv = 5$
	DT	GS	1.58	0.50	1 s	$cv = 5$

variables. These characteristics help create more accurate prediction models, aiding our understanding of the country's soils. However, as noted, achieving good predictive performance from the model often involves a nearly manual process of parameter tuning. Thus, optimization methods for parameter tuning, such as those presented here, are highly relevant for improving model performance and making more accurate predictions.

The results obtained here highlight the importance of hyperparameter optimization in estimating soil carbon. Experiments comparing different optimiza-

tion strategies show that the chosen method can have a significant impact on both the model's accuracy and the speed at which it runs.

This suggests exploring other methods for optimizing hyperparameters in future work, such as Bayesian hyperparameter tuning and population-based metaheuristics, including Genetic Algorithms and the Biased Random Key Genetic Algorithm (BRKGA).

In addition, the integration of remote sensing data will be investigated to improve the models' predictive capacity and enable more accurate large-scale assessments of soil organic carbon. These extensions are expected to improve both the model's effectiveness and scalability in environmental monitoring applications.

Acknowledgments. This work has been funded by national funds through FCT Fundação para a Ciência e a Tecnologia, I.P., through project UIDB/05037/2025 with DOI 10.54499/UIDB/05037/2025. Additionally, the authors would like to thank the following Brazilian Agencies: FAPERJ, CAPES, and CNPq. DB thanks Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the support through the grants E-26/210.798/2024 and E-26/204.262/2025.

Declaration of Generative AI in Scientific Writing. During the preparation of this work, the authors used Grammarly for English-language refinement. Following the use of these tools, the authors reviewed and edited the manuscript as necessary and assume full responsibility for its content.

References

1. Schierholz, R., et al.: Evaluating different soil compaction measurement techniques. In: Wageningen Soil Conference (2019)
2. Grace, J.: Understanding and managing the global carbon cycle. J. Ecol. **92** (2004)
3. Ussiri, D., Lal, R.: Introduction to Global Carbon Cycling. Springer (2017)
4. Arora, N., Mishra, I.: UN sustainable development goals and environmental sustainability. Environ. Sustain. (2019)
5. Sommer, R., Bossio, D.: Dynamics of soil carbon sequestration. J. Environ. Manage. (2014)
6. Japa, L., et al.: A population-based hybrid approach for hyperparameter optimization. IEEE Access (2023)
7. Szatmari, G., et al.: Countrywide mapping of soil organic carbon using ML. Catena (2023)
8. Wadoux, A., et al.: Machine learning for digital soil mapping. Earth-Sci. Rev. (2020)
9. Wadoux, A., et al.: Shapley values for soil carbon prediction. Soil (2023)
10. Ferreira, A., et al.: Predicting soil carbon stock in the Amazon. Geoderma Regional (2023)
11. Barbosa, K., et al.: Machine learning for soil carbon and nitrogen prediction. SSRN (2023)
12. Safanelli, J.: Leveraging earth observation data for soil mapping. Ph.D. thesis, USP (2020)

13. Aderele, M.O., Srivastava, A.K., Butterbach-Bahl, K., Rahimi, J.: Integrating machine learning with agroecosystem modelling: current state and future challenges. Eur. J. Agron. **168**, 127610 (2025)
14. Santos, E., et al.: Sentinel-1 imagery for soil carbon estimation. Remote Sens. (2023)
15. Ceddia, M., et al.: Pedotransfer functions and soil carbon estimation. Sci. Agricola (2016)
16. Haddad, D., et al.: Soil bulk density prediction using neural regressors. IEEE (2018)
17. Yao, X., Zhang, H., Wang, X., Jiang, Y., Zhang, Y., Na, X.: Which model is more efficient in carbon emission prediction research? A comparative study of deep learning models, machine learning models, and econometric models. Environ. Sci. Pollut. Res. **31**(13), 19500–19515 (2024)
18. Hoffmann, M., et al.: Detecting small-scale spatial heterogeneity and temporal dynamics of soil organic carbon (SOC) stocks: a comparison between automatic chamber-derived C budgets and repeated soil inventories. Biogeosciences **14**(4), 1003–1019 (2017)
19. Pinna, D., et al.: Fault classification of wind turbine: a comparison of hyperparameter optimization methods. In: Pereira, A.I., Mendes, A., Fernandes, F.P., Pacheco, M.F., Coelho, J.P., Lima, J. (eds.) OL2A 2023. CCIS, vol. 1982, pp. 229–243. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-53036-4_16
20. Radzi, S.F.M., Karim, M.K.A., Saripan, M.I., Rahman, M.A.A., Isa, I.N.C., Ibahim, M.J.: Hyperparameter tuning and pipeline optimization via grid search method and tree-based AutoML in breast cancer prediction. J. Pers. Med. **11**(10), 978 (2021)