applied sciences

MDPI

Article

A Comprehensive Evaluation of Machine Learning Algorithms for Digital Soil Organic Carbon Mapping on a National Scale

Dorijan Radočaj *, Danijel Jug, Irena Jug and Mladen Jurišić

Faculty of Agrobiotechnical Sciences Osijek, University of Josip Juraj Strossmayer in Osijek, Vladimira Preloga 1, 31000 Osijek, Croatia; djug@fazos.hr (D.J.); ijug@fazos.hr (I.J.); mjurisic@fazos.hr (M.J.)
* Correspondence: dradocaj@fazos.hr; Tel.: +385-31-554-965

Abstract: The aim of this study was to narrow the research gap of ambiguity in which machine learning algorithms should be selected for evaluation in digital soil organic carbon (SOC) mapping. This was performed by providing a comprehensive assessment of prediction accuracy for 15 frequently used machine learning algorithms in digital SOC mapping based on studies indexed in the Web of Science Core Collection (WoSCC), providing a basis for algorithm selection in future studies. Two study areas, including mainland France and the Czech Republic, were used in the study based on 2514 and 400 soil samples from the LUCAS 2018 dataset. Random Forest was first ranked for France (mainland) and then ranked for the Czech Republic regarding prediction accuracy; the coefficients of determination were 0.411 and 0.249, respectively, which was in accordance with its dominant appearance in previous studies indexed in the WoSCC. Additionally, the K-Nearest Neighbors and Gradient Boosting Machine regression algorithms indicated, relative to their frequency in studies indexed in the WoSCC, that they are underrated and should be more frequently considered in future digital SOC studies. Future studies should consider study areas not strictly related to human-made administrative borders, as well as more interpretable machine learning and ensemble machine learning approaches.

Keywords: random forest; web of science core collection topic search; LUCAS dataset; environmental covariates; digital soil mapping; remote sensing

Citation: Radočaj, D.; Jug, D.; Jug, I.; Jurišić, M. A Comprehensive Evaluation of Machine Learning Algorithms for Digital Soil Organic Carbon Mapping on a National Scale. *Appl. Sci.* **2024**, *14*, 9990. https://doi.org/10.3390/app14219990

Academic Editors: Vassilis J. Inglezakis and Atsushi Mase

Received: 10 September 2024
Revised: 28 October 2024
Accepted: 30 October 2024
Published: 1 November 2024

1. Introduction

Accurate digital soil organic carbon (SOC) mapping on a national scale is vital for a multitude of ecological and environmental reasons, including soil health, nutrient availability, water retention, and overall soil fertility, which directly impact agricultural productivity and sustainability [1]. Moreover, SOC is a significant component of the global carbon cycle, affecting greenhouse gas emissions and climate change mitigation efforts [2]. National SOC maps particularly enable informed land management and policy decisions, enabling the identification of carbon hotspots and guiding reforestation, conservation, and carbon sequestration projects [3]. Additionally, such mapping supports compliance with international climate agreements by providing essential data for carbon accounting and reporting [4]. Therefore, the development of high-resolution, accurate SOC maps is imperative for advancing research and practices aimed at enhancing soil conservation, improving food security, and fostering climate resilience. The primary difference between digital SOC mapping at a national scale and mapping performed at other scales is the spatial resolution and the level of detail [5]. The most common spatial resolution for national-scale SOC maps is of the order of 250 m–1 km [6]. For achieving higher spatial resolutions, soil sampling is time-consuming, labor-intensive, and costly, which can limit the availability of soil samples for mapping purposes [7]. In addition, the distribution of soil samples may not be uniform or representative of the entire area of interest, which can further limit the accuracy and resolution of the resulting SOC maps [8]. The presence of datasets such as LUCAS (Land Use/Cover Area frame statistical Survey) is highly beneficial for digital soil organic carbon

(SOC) mapping at a national scale due to its comprehensive and systematic approach to soil sampling across diverse landscapes. LUCAS provides standardized soil data, including SOC content, collected at predefined locations that can enhance the representativeness of soil information [9]. Furthermore, the extensive coverage and methodological rigor of LUCAS facilitate the development of predictive models that can extrapolate SOC distribution across wider areas, ultimately contributing to better-informed agricultural practices, carbon stock assessments, and environmental policies at the national level.

Knowing which machine learning algorithms to use for digital soil mapping is critical as different algorithms possess unique strengths and limitations that can significantly influence the accuracy and reliability of soil property predictions, including SOC content [10]. Machine learning techniques, such as Random Forests, support vector machines, and neural networks, differ in their ability to handle various data types, manage high dimensionality, and model the complex non-linear relationships that are inherent in soil data. Selecting the appropriate algorithm can enhance the model's performance by optimizing predictive accuracy and minimizing overfitting, which is particularly important in heterogeneous landscapes [11]. Thus, the informed selection of machine learning algorithms not only improves the robustness of SOC predictions, but also fosters informed decision-making for sustainable land use and effective climate action strategies. The integration of machine learning algorithms in digital soil mapping represents a state-of-the-art approach, surpassing conventional geostatistical methods in several aspects [12]. Unlike traditional geostatistical techniques, which rely on spatial autocorrelation and Gaussian processes, machine learning algorithms can effectively handle high-dimensional and heterogeneous datasets, capturing complex non-linear relationships between soil properties and environmental factors [13]. This enables more accurate predictions of SOC content, particularly in areas with diverse landscapes and soil types. Furthermore, machine learning algorithms can automatically select relevant predictors, reducing the need for manual feature selection and minimizing the risk of overfitting. In contrast, geostatistical approaches often require the explicit modeling of spatial relationships and may struggle with high-dimensional data, leading to a reduced accuracy and interpretability [14]. Environmental covariates are crucial in digital soil mapping as they provide relevant information about the soil–landscape relationships, enabling the creation of more accurate predictive models [15]. These covariates can be topographic, climatic, biophysical, or anthropogenic factors that indirectly influence soil properties through their effects on soil-forming processes [16]. Remote sensing has a significant role in digital soil mapping by providing consistent, large-scale, and high-resolution data on various environmental covariates, facilitating the accurate prediction of soil properties in data-scarce regions and enabling the identification of spatial patterns and trends that may be difficult to detect using conventional ground-based measurements alone [17]. The integration of remote sensing data with machine learning algorithms has the potential to further enhance the predictive accuracy and interpretability of digital soil mapping models, ultimately contributing to better-informed land use and climate action strategies.

While the evaluation of machine learning algorithms for digital SOC mapping has been a focus of several previous studies, there is a predominant conclusion that the optimal machine learning algorithm and its hyperparameters are heavily dependent on the input soil sample properties [10]. To ensure the optimal selection of predictive approaches, it is essential to evaluate several machine learning algorithms in each study [18]. However, a significant research gap exists regarding which specific algorithms should be prioritized for evaluation. Selecting from the vast array of available algorithms can be computationally inefficient, especially when conducting national-scale digital SOC mapping, a process that is often utilized in land management.

The aim of this study is to address this research gap by providing a comprehensive assessment of the prediction accuracy of 15 commonly used machine learning algorithms in digital SOC mapping. This assessment is based on studies indexed in the Web of Science

Core Collection (WoSCC), thereby establishing a foundation for algorithm selection in future research.

2. Materials and Methods

The workflow of the comprehensive evaluation of machine learning algorithms for digital soil organic carbon mapping on a national scale consisted of the following four fundamental steps: (1) the collection and preprocessing of soil samples; (2) the modeling and preprocessing of environmental covariates; (3) the evaluation of 15 machine learning regression algorithms for digital SOC mapping; and (4) an accuracy assessment (Figure 1). These results were compared with the frequency of evaluated machine learning algorithms in previous studies indexed in the WOSCC, leading to relationships between their prediction results from this study and their popularity in previous studies.

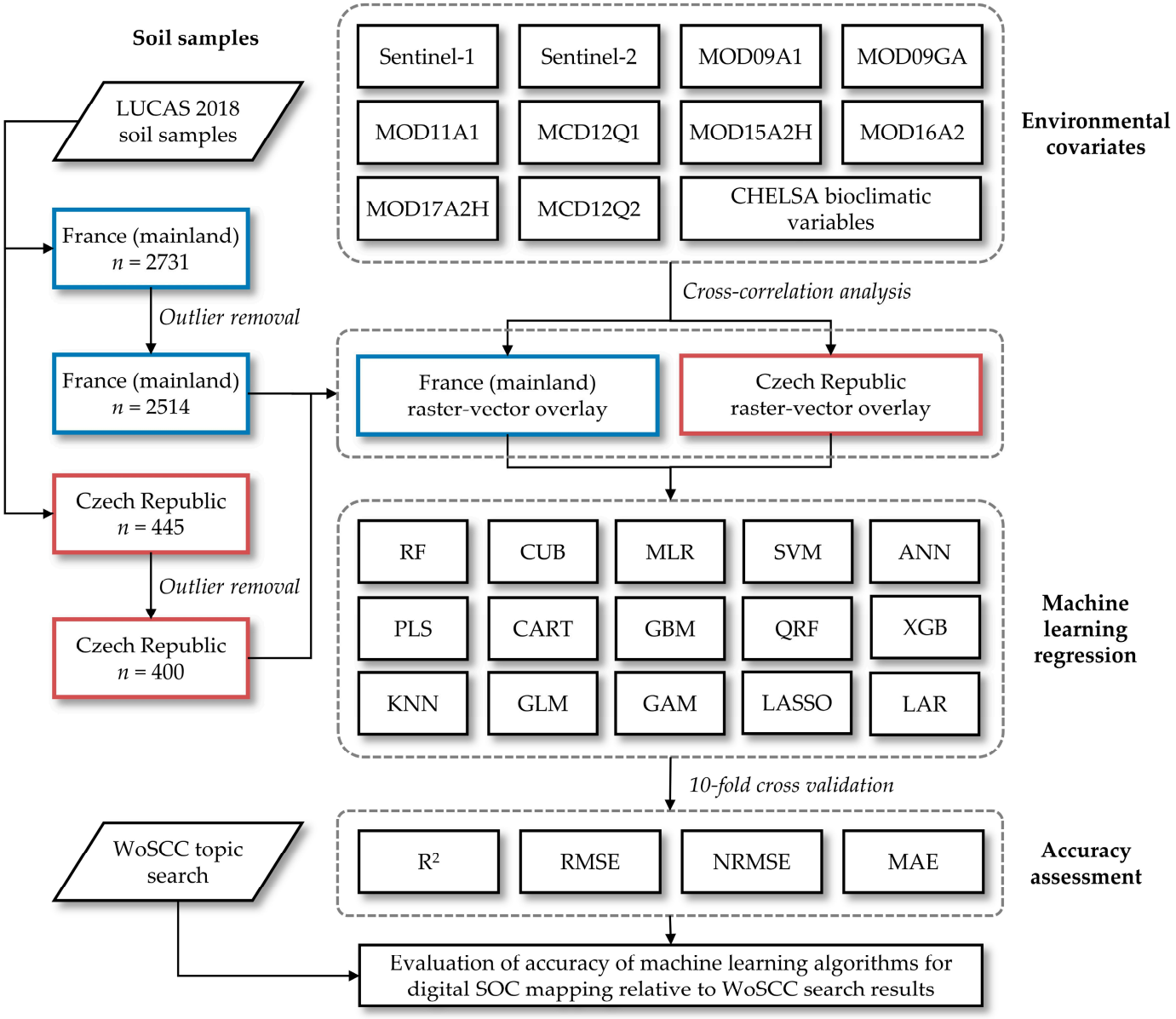

Figure 1. Workflow of the comprehensive evaluation of machine learning algorithms for digital soil organic carbon mapping on a national scale.

2.1. Study Areas and Soil Samples

France (mainland) and the Czech Republic were selected as the two study areas, being representative countries within the European Union for digital SOC mapping at a national scale due to their diverse geography, climate, and topography (Figure 2). France's varied landscapes, ranging from the fertile plains of the Loire Valley to the mountainous regions of the Alps and Pyrenees, provide a wide range of soil types and land uses, allowing for comprehensive organic carbon assessments across different ecological contexts. Conversely, the Czech Republic, characterized by its hilly terrain and continental climate, showcases a distinct range of soil types affected by specific agricultural and forestry practices.

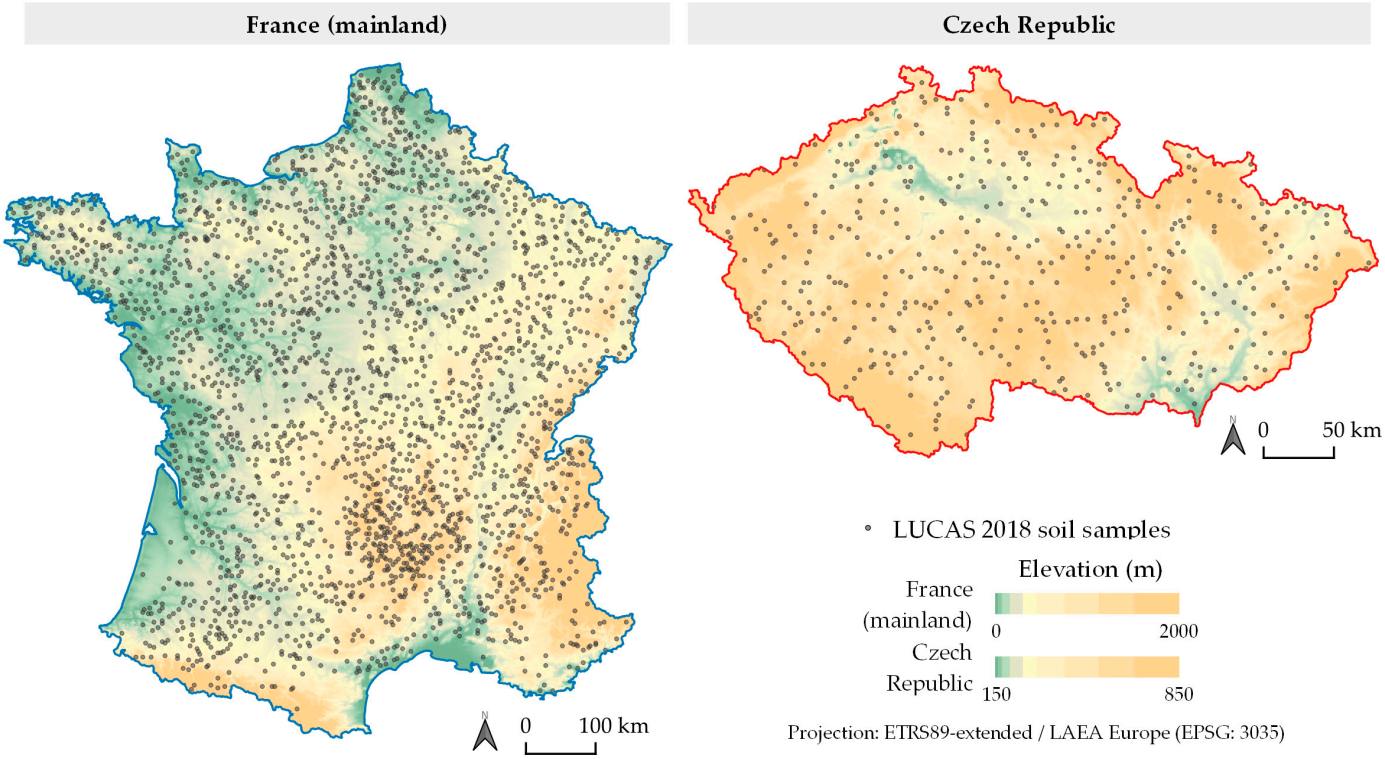

Figure 2. Two study areas, including mainland France and the Czech Republic.

The LUCAS (Land Use/Cover Area Frame Statistical survey) 2018 SOC database is a comprehensive resource developed by the European Commission's Joint Research Centre, aimed at assessing soil quality across Europe [9]. This database provides critical information on the spatial distribution and quantity of SOC, which is fundamental for understanding soil health, as well as carbon sequestration potential and its implications for climate change. The soil samples from the entire LUCAS 2018 dataset at 0–20 cm soil depth measured in g kg^{-1} were filtered according to the geographical coverage of two study areas. Outlier removal from the raw soil sampling datasets was performed using an interquartile range approach, having an interquartile range of 1.5 as the threshold below the first and above the third quartile for outlier removal.

2.2. Environmental Covariates Used for Digital SOC Mapping

The remote sensing data from the Sentinel-1, Sentinel-2, MOD09A1, MOD09GA, MOD11A1, MCD12Q1, MOD15A2H, MOD16A2, MOD17A2H, MCD12Q2, and CHELSA bioclimatic variables were utilized for the modeling of the environmental covariates for prediction in 1000 m spatial resolution. CHELSA v2.1 data were manually downloaded from their official website, while Google Earth Engine was used for downloading the remaining remote sensing data. All multitemporal imagery represented a 2018 annual median, which coincided with the soil sampling time frame. Sentinel-1 radar data were used

to estimate soil moisture and texture [19], while multispectral Sentinel-2 data were used to estimate vegetation health and its phenological properties [20]. MOD09A1 and MOD09GA surface reflectance data were additionally used to estimate a broad scope of vegetation and soil properties based on multispectral imaging. MOD11A1 land surface temperature data were used to estimate soil moisture and thermal properties [21]. MCD12Q1 and MCD12Q2 land cover and land cover change data estimated the influence of land use and possible land use change on SOC concentrations [22], while MOD15A2H leaf area index data were used to estimate vegetation productivity and biomass [23]. Additionally, MOD16A2 gross primary productivity data were used to estimate carbon sequestration potential, and MOD17A2H net primary productivity data were used to estimate carbon cycling dynamics. CHELSA bioclimatic variables provided climate data, which were used to estimate the influence of climate on SOC concentrations and were proven to have high importance in digital soil carbon mapping [24]. Any covariates with cross-correlations quantified by a correlation coefficient higher than 0.9 were removed during preprocessing to avoid multicollinearity issues that could negatively impact model performance.

2.3. Machine Learning Regression Algorithms Evaluated for Digital SOC Mapping

The selection of machine learning regression algorithms to be evaluated in this study was performed according to total indexing in the WoSCC based on the following topic search: "TS = (digital soil mapping AND soil organic carbon AND machine learning algorithm)" (Table 1). The machine learning prediction was performed in R v4.3.2 [25] and the caret library [26]. Prior to modeling, the data were pre-processed using centering and scaling to standardize the variables and reduce the effects of differing scales. All models were tuned using a grid search approach, where 10 different combinations of tuning parameters were evaluated to optimize model performance according to the hyperparameters listed in Table A1. The final model was selected based on its performance on the training data, and its predictive accuracy was evaluated using the specified resampling method.

In traditional linear models such as MLR, GLM and GAM, predictions are made by fitting a linear equation to the data, relying on assumptions about the linear relationship between input variables and the response [39]. As the simplest prediction algorithm of those evaluated in this study, MLR estimated the linear relationship between the sampled SOC and environmental covariates by estimating coefficients to compute the predicted SOC value for a new observation. GLM is a generalization of linear regression that allowed for non-normal response variables and non-linear relationships between the SOC values and environmental covariates [40]. GAM estimated the additive relationship between SOC and environmental covariates using smooth functions [41]. Regularization techniques such as LASSO and LAR enhanced these models by introducing penalties, enabling feature selection, and managing overfitting. LASSO is a linear regression model with an L1 penalty term that encouraged sparsity in the estimated coefficients to compute SOC based on environmental covariate data [42]. Similarly, LAR is a linear regression model that estimates the coefficients using a forward stagewise procedure that adds one variable at a time to the model [43]. The last of the traditional prediction methods, PLS concatenated centroids of predictors and responses to effectively explain variance in settings with high-dimensional data, making it suitable in regression contexts where the predictors exceed the number of observations [44].

On the other hand, tree-based methods like RF, CART, and GBM are non-parametric and model interactions between predictors through hierarchical tree structures, as well as by averaging, boosting, or bagging model predictions to enhance accuracy [45]. CART is a decision tree algorithm that recursively partitioned the feature space into subspaces and fitted a simple model to each subspace, with the output as the value of the constant at the leaf node [46]. GBM is an ensemble learning method that built multiple decision trees and combined their predictions using a gradient descent algorithm, with the output as the weighted sum of the predictions from all the trees [47]. Similarly, Cubist is a rule-based regression model that built a series of if–then rules to predict the response variable [48].

It partitioned the feature space into rectangular regions and fitted a linear model to each region. During prediction, a new observation was passed through the rules, and the output was the weighted sum of the linear models that matched the observation. As the most frequently used machine learning method for digital SOC mapping based on WoSCC search results, RF is an ensemble learning method that built multiple decision trees and averaged their predictions to produce a final output, utilizing bootstrap aggregation (bagging) in the process, having a new observation passed through each decision tree, as well as the average of the predictions from all the trees as the output [49]. QRF is a variation of RF that estimated the conditional quantiles of the response variable, operating similarly to RF; however, instead of averaging the predictions, it outputs the specified quantile estimate from each tree [50]. SVM for regression applied a margin-based approach, aiming to find the best hyperplane that minimizes prediction error while allowing for some deviations, effectively capturing non-linear relationships through kernel functions [51]. Similarly, KNN relied on the local structure of the data, predicting outputs based on the average outcome of the k-nearest data points. As another advanced ensemble method, XGB optimized tree learning with regularization and control overfitting [52], while ANN leveraged neuron layers to capture complex, non-linear patterns through numerous interconnected layers [53].

Table 1. The machine learning regression algorithms used for the evaluation of digital SOC mapping in this study.

Machine Learning Algorithm	Abbreviation	Total Papers Indexed in WoSCC (–2023)	Library	Reference
Random Forest	RF	347	randomForest	[27]
Cubist	CUB	92	Cubist	[28]
Multiple Linear Regression	MLR	85	/	[25]
Support Vector Machines	SVMs	82	kernlab	[29]
Artificial Neural Networks	ANNs	72	brnn	[30]
Partial Least Squares	PLS	61	pls	[31]
Classification and Regression Trees	CARTs	40	rpart	[32]
Gradient Boosting Machine	GBM	35	gbm	[33]
Quantile Random Forest	QRF	31	quantregForest	[34]
Extreme Gradient Boosting	XGB	23	xgboost	[35]
K-Nearest Neighbors	KNN	15	/	[25]
Generalized Linear Model	GLM	12	/	[25]
Generalized Additive Model	GAM	11	gam	[36]
Least Absolute Shrinkage and Selection Operator	LASSO	9	elasticnet	[37]
Least Angle Regression	LAR	2	lars	[38]

Algorithms with a non-indicated library were available in base R.

2.4. Accuracy Assessment

A 10-fold cross-validation was employed to evaluate the performance of each evaluated predictive machine learning regression model, where the dataset was randomly partitioned into 10 subsets, with 9 subsets being used for training and 1 subset for testing [54]. This process was repeated 10 times, with each subset serving as the test set once. The model's performance was assessed using four metrics—coefficient of determination (R^2), root mean squared error (RMSE), normalized root mean squared error (NRMSE), and mean absolute error (MAE). R^2 measured the proportion of variance in the dependent variable explained by the model, while RMSE and NRMSE provided an indication of the

model's accuracy, with lower values indicating a better fit. MAE, on the other hand, evaluated the model's ability to predict the actual values, with lower values indicating smaller absolute errors [55]. By using 10-fold cross-validation, the reliability and generalizability of the model's performance were ensured, as the results were averaged across multiple iterations, providing a robust estimate of the model's predictive capabilities [56].

3. Results and Discussion

The median SOC values for the raw soil sampling datasets range from 19.9 to 23.0, indicating a high level of variability (Table 2). However, the coefficient of variation (CV) reveals a higher degree of dispersion in the entire LUCAS 2018 dataset (1.72) compared to the two study areas (0.90 and 1.07). The skewness and kurtosis values suggest that the distributions are positively skewed and leptokurtic, with the France (mainland) dataset exhibiting the most extreme values (skewness = 4.88, kurtosis = 37.72). Overall, the descriptive statistics indicate that the entire LUCAS 2018 dataset has a more dispersed and variable distribution compared to the two study areas used in this study. However, it should be considered that input samples for the two study areas were preprocessed, in contrast to the raw observations in the LUCAS 2018 dataset, achieving a notably lower CV, as well as a more normal SOC value distribution, as indicated by the skewness and kurtosis values.

Table 2. Descriptive statistics of SOC values from the two study areas and the entire LUCAS 2018 dataset.

Dataset	Preprocessed	*n*	Median	Min	Max	CV	Skewness	Kurtosis
France (mainland)	No	2731	23.0	3.2	473.0	1.07	4.88	37.72
	Yes	2514	25.6	3.2	74.1	0.59	1.08	0.56
Czech Republic	No	445	19.9	3.2	208.6	0.90	3.46	15.60
	Yes	400	21.2	3.2	51.2	0.44	1.19	1.03
Entire LUCAS 2018	No	18,984	21.8	2.1	723.9	1.72	3.97	16.62

The accuracy assessment results suggest that RF and QRF perform with superior accuracy compared to the other evaluated methods on both datasets, with R^2 values of 0.411 and 0.409, respectively, for France, and 0.249 and 0.223, respectively, for the Czech Republic (Tables 3 and A1). There are several possible causes for their outperformance, one of them being their ability to effectively reduce overfitting by averaging the predictions from many trees. Since training on each tree was performed on different subsets of the data and features, this decreased the correlation between individual trees and enhanced its ability to generalize to unseen data [57]. Managing high-dimensional datasets with numerous features, typical for digital soil mapping studies [58], was automatically performed through feature selection by considering only a random subset of features for each split in the decision trees. Additionally, the ensemble nature of RF makes it resilient to noise and outliers in the data [59].

In contrast, GAM and CART exhibited a poor performance, with R^2 values of 0.373 and 0.307, respectively, for France, and 0.135 and 0.152, respectively, for the Czech Republic. Overall, the results suggest that ensemble methods (RF, QRF, and GBM) tend to outperform individual models (ANN, SVM, and MLR) on these datasets. This observation agrees with previous studies [10] but also suggests that methods like ANN and SVM tend to heavily depend on the quantity and properties of input soil sampling datasets [60]. The accuracy metrics were generally in agreement, with RF and QRF consistently ranking among the top three algorithms in both countries. However, there are some discrepancies between the metrics, with RMSE and NRMSE suggesting that KNN and GBM were also highly accurate, while R^2 and MAE indicate that SVM and PLS are among the top performers. The results suggest that ensemble methods, such as RF and GBM, tend to outperform traditional algorithms, such as MLR and GLM, in both study areas. Accuracy assessment metrics were also in slight disagreement regarding which algorithms are the most accurate; while

QRF has the highest R^2 value in mainland France, it has a lower R^2 value than RF in the Czech Republic. This observation agrees with a previous digital soil mapping study of total carbon in the majority of Europe, in which QRF was the second-most-accurate algorithm behind deep learning [60]. Predicted SOC maps according to the most accurate machine learning methods are displayed in Figure 3. The scatterplots of the final predictions are displayed in Figure 4 for France (mainland) and Figure 5 for the Czech Republic.

Table 3. Accuracy assessment results of digital SOC mapping based on the 15 evaluated machine learning regression methods.

Algorithm	France (Mainland)				Czech Republic			
	R^2	RMSE	NRMSE	MAE	R^2	RMSE	NRMSE	MAE
RF	0.411	11.58	0.453	8.61	0.249	8.19	0.386	6.22
CUB	0.378	11.97	0.468	8.76	0.191	8.55	0.403	6.48
MLR	0.375	11.95	0.468	8.93	0.216	8.45	0.398	6.43
SVM	0.388	11.99	0.469	8.44	0.227	8.39	0.395	6.14
ANN	0.378	11.91	0.466	8.86	0.230	8.27	0.390	6.22
PLS	0.375	11.93	0.467	8.92	0.256	8.12	0.383	6.19
CART	0.307	12.63	0.494	9.50	0.152	8.89	0.419	6.69
GBM	0.390	11.79	0.462	8.78	0.239	8.26	0.389	6.30
QRF	0.409	11.80	0.462	8.29	0.223	8.41	0.396	6.09
XGB	0.341	12.46	0.488	9.20	0.212	8.66	0.408	6.51
KNN	0.393	11.79	0.461	8.60	0.217	8.40	0.396	6.29
GLM	0.375	11.95	0.468	8.93	0.206	8.50	0.401	6.48
GAM	0.373	11.97	0.468	8.89	0.135	9.32	0.439	7.05
LASSO	0.376	11.92	0.466	8.92	0.229	8.22	0.387	6.23
LARS	0.375	11.91	0.466	8.91	0.241	8.23	0.388	6.20

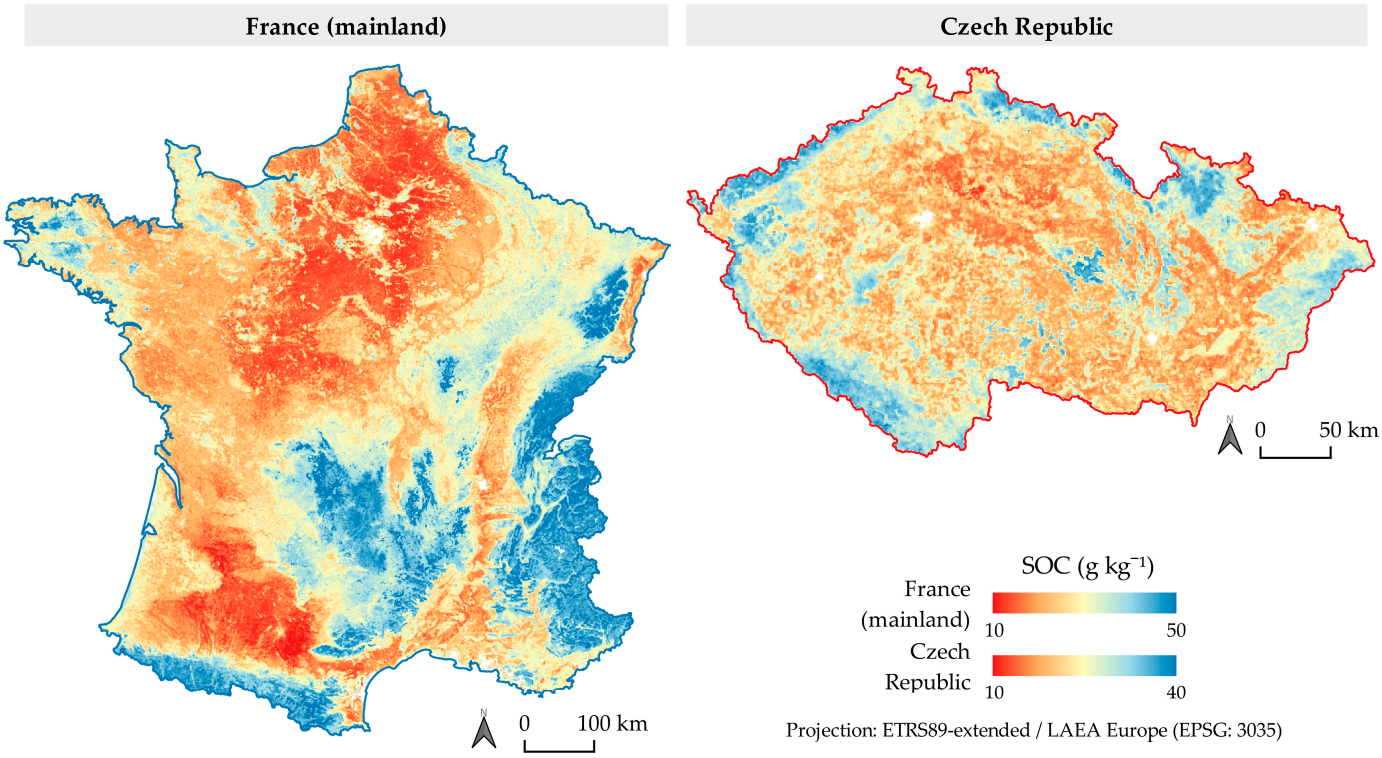

Figure 3. The prediction results of SOC in two study areas based on the most accurate machine learning methods; RF for France (mainland) and PLS for the Czech Republic.

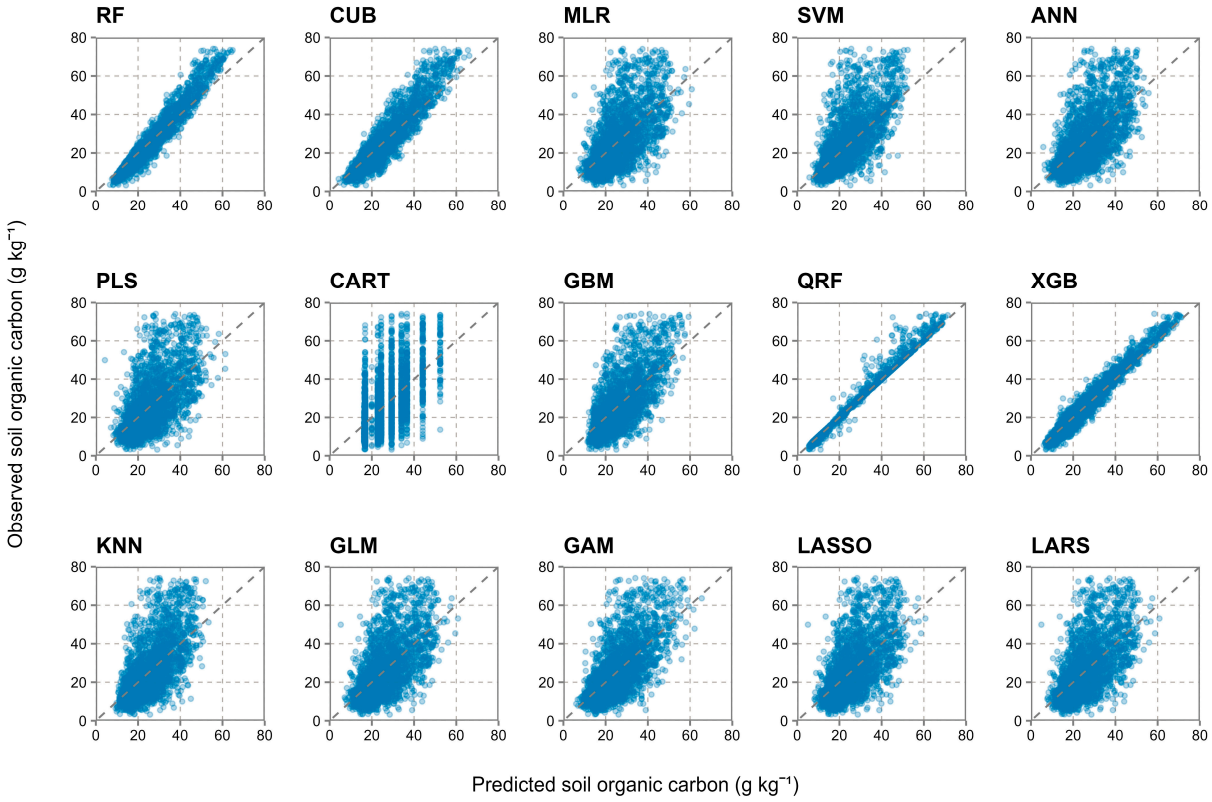

Figure 4. The scatterplots of predicted and observed SOC values based on the 15 evaluated machine learning regression algorithms in France.

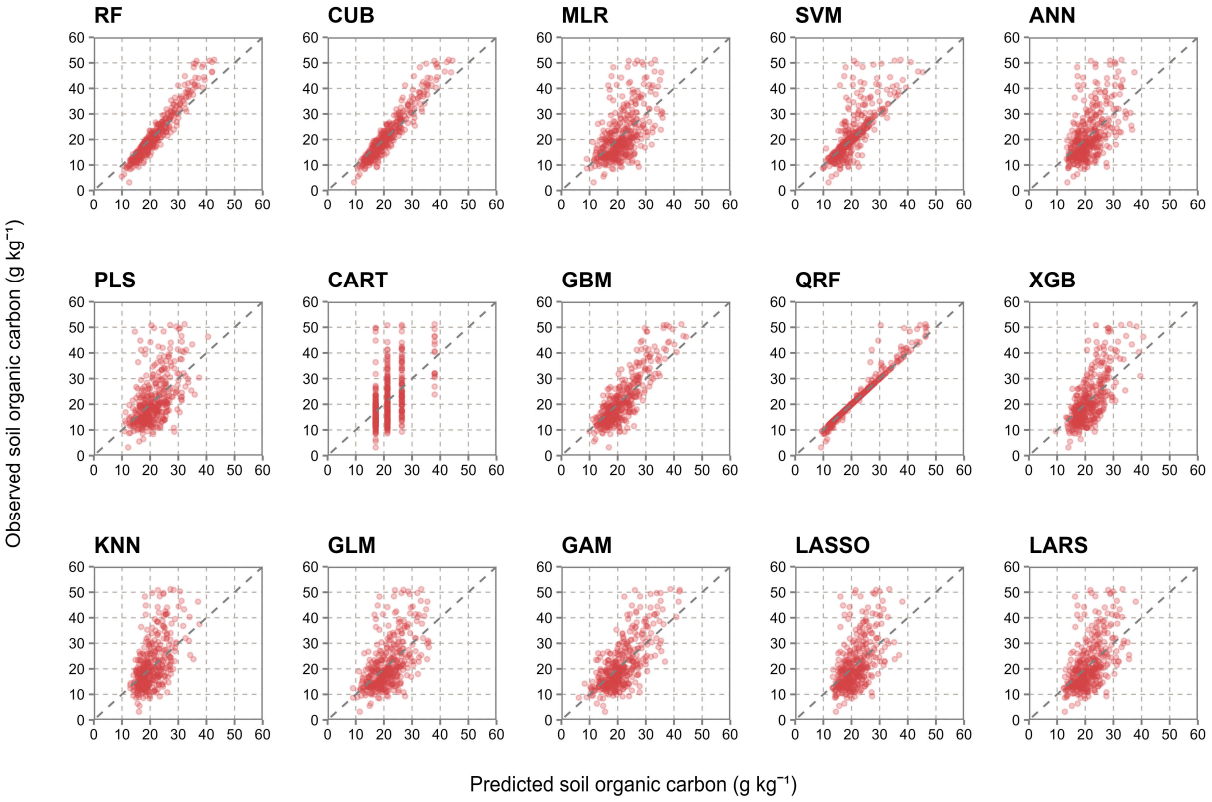

Figure 5. The scatterplots of predicted and observed SOC values based on the 15 evaluated machine learning regression algorithms in the Czech Republic.

NRMSE, as an absolute accuracy metric, offered significant interpretability advantages in evaluating model performance, particularly in comparative studies across different datasets or scales [61]. By normalizing the RMSE against the range or mean of the observed data, NRMSE provided a dimensionless measure that enhances the contextual understanding of prediction accuracy [62]. This is particularly useful when dealing with heterogeneous data that can vary widely in magnitude, as it allows for a more meaningful comparison of model efficacy. Despite these benefits, NRMSE remains underrepresented in the literature compared to more commonly used metrics like R^2 or traditional RMSE. This oversight hinders the development of a more nuanced understanding of model performance across disciplines, as NRMSE's ability to convey relative error in a standardized format can facilitate better decision-making in model selection and refinement. Moreover, it does not have specific computational requirements, as the vast majority of previous studies calculated RMSE during the accuracy assessment, also leading to the simple calculation of NRMSE. Overall, the relative difference between the most and least accurate evaluated machine learning methods was 9.1% for France (mainland) and 14.8% for the Czech Republic.

The covariates which produced the highest feature importance for the most accurate machine learning model for France (mainland), ranked by importance, were the digital elevation model, land cover classes, the MODIS surface reflectance at 483–493 nm (blue) band, and two CHELSA bioclimatic variables, including the mean daily air temperatures of the coldest quarter and the mean monthly precipitation amount of the wettest quarter. For the Czech Republic, the five most important covariates according to most accurate model were the bare soil index and green coverage index from Sentinel-2 data, the Sentinel-2 (near-infrared) band at 842 nm central wavelength, the Sentinel-1 annual median raster with VH polarization, and the MODIS surface reflectance at 545–565 nm (green) band.

Table 4 provides a ranking of machine learning algorithms based on their performance in terms of accuracy and their ranking in the WoSCC. The algorithms are ranked using RMSE as the metric for accuracy assessment results. The RF algorithm ranks the highest in terms of accuracy assessment, with a rank of 1 for both France and the Czech Republic, and a rank difference of 0, indicating a consistent performance and a justified place as the most frequently used machine learning algorithm for digital SOC mapping in the literature. CUB ranked second in accuracy assessment but had a significantly lower WoSCC ranking in both countries, resulting in a rank difference of -9 and -10, indicating the most overrated algorithm in terms of prediction accuracy based on the results from this study. The MLR and SVM algorithms had similar accuracy assessment ranks but lower WoSCC rankings, resulting in negative rank differences. However, the accuracy assessment results from this study indicated that GBM and KNN were underrated in terms of prediction accuracy, suggesting that these should be considered more often in future studies. Also, PLS had a high accuracy assessment rank and a moderate WoSCC ranking, resulting in a positive rank difference of 5 in the Czech Republic.

The findings of this study on digital SOC mapping must be interpreted within the context of the specific limitations associated with the research design. The analysis was conducted in two geographical regions—France (mainland) and the Czech Republic—employing soil samples from the LUCAS 2018 dataset as input data. Notably, the sample size in the two administrative units used as study areas, France (mainland) and the Czech Republic, may not sufficiently capture the broader variability of soil types across these regions, potentially impacting the generalizability of the results. Furthermore, the selected study areas may not adequately represent the full spectrum of soil characteristics and environmental conditions present in either country, thereby limiting the applicability of the findings to other regions. The restricted sample sets utilized in this study are a critical factor that may skew the results, as these were related to national borders and not natural phenomena [63]. A limited dataset can lead to an inadequate generalization of the findings, possibly underrepresenting the variability found in larger populations [64]. Given the heterogeneous nature of soil properties, geographic diversity, and climate influences, the results may differ substantially with a larger and more diverse sample. This raises the

concern that the performance of the employed algorithms may not be fully reflective of their efficacy in varied contexts. Moreover, data collection from two distinct countries introduces an additional layer of complexity. Different geographical areas can exhibit distinct soil formation processes, types, and land use practices, which can inherently influence the SOC levels. Consequently, the algorithms developed and validated in this study may exhibit biases when applied to other settings, particularly if trained solely within a single national context.

Table 4. Relative accuracy assessment of the 15 evaluated machine learning regression methods for digital SOC mapping according to their frequency in previous studies indexed in the WoSCC.

Algorithm	Rank per WoSCC Indexing	Accuracy Assessment Rank		Rank Difference	
		France (Mainland)	Czech Republic	France (Mainland)	Czech Republic
RF	1	1	2	0	−1
CUB	2	11	12	−9	−10
MLR	3	10	10	−7	−7
SVM	4	13	7	−9	−3
ANN	5	5	6	0	−1
PLS	6	8	1	−2	5
CART	7	15	14	−8	−7
GBM	8	3	5	5	3
QRF	9	4	9	5	0
XGB	10	14	13	−4	−3
KNN	11	2	8	9	3
GLM	12	9	11	3	1
GAM	13	12	15	1	−2
LASSO	14	7	3	7	11
LARS	15	6	4	9	11

RMSE was used as a metric for ranking machine learning algorithms per accuracy assessment results.

Despite these limitations, the study's insights into the various machine learning algorithms are valuable. The findings suggest that no universally superior algorithm exists for SOC mapping; instead, the effectiveness of each algorithm is contingent upon specific data characteristics and contextual factors. This multiplicity of performance emphasizes the importance of employing an ensemble approach or a suite of algorithms tailored to distinct scenarios in future research. Further studies should prioritize expanding the dataset both in size and geographical representation. Incorporating data from a broader array of countries would enhance the robustness of the results and allow for the identification of global trends in SOC mapping accuracy. Additionally, employing rigorous methodologies that encompass diverse soil types, climates, and land use practices will yield more reliable outcomes. Future research should also focus on the intrinsic factors influencing machine learning algorithm performance, as well as their interpretability [65]. Future studies will also explore the effects of the selection of individual machine learning models in ensemble machine learning, which, in previous studies, had the potential of producing a superior prediction accuracy to individual methods [66,67].

4. Conclusions

While the evaluation of machine learning algorithms for digital SOC mapping has been a focus of several previous studies, it is essential to recognize that these studies often arrived at a predominant conclusion that the optimal machine learning algorithm and its hyperparameters are heavily dependent on the specific properties of the input soil samples. This dependency highlights the necessity for the improved evaluation of various algorithms in each study to ensure the most effective predictive approaches are employed. The multitude of available machine learning algorithms can lead to computational inefficiencies, particularly when conducting national-scale digital SOC mapping, a process that

is crucial for informed land management decisions. This study aimed to address this gap by systematically assessing the prediction accuracy of 15 frequently used machine learning algorithms. To provide an extensive evaluation based on a robust dataset derived from diverse geographical regions, two independent study areas were used, including mainland France and the Czech Republic. The key conclusions from the study are as follows:

- RF ranked first in France (mainland) and second in the Czech Republic for prediction accuracy, confirming its prominence in previous studies. RF should be prioritized for future evaluations in national-scale digital SOC mapping.
- KNN and PLS achieved high prediction accuracy in France and the Czech Republic, respectively, but performed near average in other study areas. Their effectiveness depends on the quantity and distribution of input soil sampling data, warranting situational evaluation.
- The ranking of GBM and KNN suggests they are underrated and should be more frequently considered in future digital SOC studies.
- In contrast, CUB, MLR, and SVM were highly ranked in previous studies but did not justify their popularity based on this study's findings.
- While France and the Czech Republic serve as representative European countries, the study's observations are limited. Future research should explore areas beyond human-made borders and consider more interpretable machine learning approaches.

Author Contributions: Conceptualization, D.R.; methodology, D.R.; software, D.R.; validation, D.R.; formal analysis, M.J.; investigation, D.R.; resources, D.R.; data curation, D.R.; writing—original draft preparation, D.R.; writing—review and editing, D.R., D.J., I.J. and M.J.; visualization, D.R.; supervision, D.J., I.J. and M.J.; project administration, D.R.; funding acquisition, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: This research was supported by the "Soybean cropland suitability prediction based on machine learning regression" project from the "Technical and technological systems in agriculture, GIT, precision agriculture and environment protection" research team of the Faculty of Agrobiotechnical Sciences Osijek.

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

Appendix A

Table A1. Optimal hyperparameters of the 15 evaluated machine learning regression methods for digital SOC mapping.

Algorithm	Hyperparameter	Optimal Hyperparameter Value	
		France (Mainland)	Czech Republic
RF	mtry	18	9
CUB	committees	20	20
	neighbors	9	9
MLR	intercept	TRUE	TRUE
SVM	sigma	0.021	0.019
	C	0.5	2
ANN	neurons	2	1
PLS	ncomp	10	4

Table A1. *Cont.*

Algorithm	Hyperparameter	Optimal Hyperparameter Value	
		France (Mainland)	**Czech Republic**
CART	cp	0.007	0.036
GBM	n.trees	50	50
	interaction.depth	6	5
	shrinkage	0.1	0.1
	n.minobsinnode	10	10
QRF	mtry	34	5
XGB	nrounds	50	50
	lambda	0.0002	0.0075
	alpha	0.1	0.042
	eta	0.3	0.3
KNN	k	19	11
GLM	/	/	/
GAM	select	TRUE	TRUE
	method	1	1
LASSO	fraction	0.722	0.278
LARS	fraction	0.789	0.367

References

1. Lehmann, J.; Bossio, D.A.; Kögel-Knabner, I.; Rillig, M.C. The Concept and Future Prospects of Soil Health. *Nat. Rev. Earth Environ.* **2020**, *1*, 544–553. [CrossRef] [PubMed]
2. Ramesh, T.; Bolan, N.S.; Kirkham, M.B.; Wijesekara, H.; Kanchikerimath, M.; Srinivasa Rao, C.; Sandeep, S.; Rinklebe, J.; Ok, Y.S.; Choudhury, B.U.; et al. Chapter One—Soil Organic Carbon Dynamics: Impact of Land Use Changes and Management Practices: A Review. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: Cambridge, MA, USA, 2019; Volume 156, pp. 1–107.
3. Paustian, K.; Collier, S.; Baldock, J.; Burgess, R.; Creque, J.; DeLonge, M.; Dungait, J.; Ellert, B.; Frank, S.; Goddard, T.; et al. Quantifying Carbon for Agricultural Soil Management: From the Current Status toward a Global Soil Information System. *Carbon Manag.* **2019**, *10*, 567–587. [CrossRef]
4. Gulluscio, C.; Puntillo, P.; Luciani, V.; Huisingh, D. Climate Change Accounting and Reporting: A Systematic Literature Review. *Sustainability* **2020**, *12*, 5455. [CrossRef]
5. O'Rourke, S.M.; Angers, D.A.; Holden, N.M.; McBratney, A.B. Soil Organic Carbon across Scales. *Glob. Change Biol.* **2015**, *21*, 3561–3574. [CrossRef]
6. Lemercier, B.; Lagacherie, P.; Amelin, J.; Sauter, J.; Pichelin, P.; Richer-de-Forges, A.C.; Arrouays, D. Multiscale Evaluations of Global, National and Regional Digital Soil Mapping Products in France. *Geoderma* **2022**, *425*, 116052. [CrossRef]
7. Chatterjee, S.; Hartemink, A.E.; Triantafilis, J.; Desai, A.R.; Soldat, D.; Zhu, J.; Townsend, P.A.; Zhang, Y.; Huang, J. Characterization of Field-Scale Soil Variation Using a Stepwise Multi-Sensor Fusion Approach and a Cost-Benefit Analysis. *CATENA* **2021**, *201*, 105190. [CrossRef]
8. Radočaj, D.; Jug, I.; Vukadinović, V.; Jurišić, M.; Gašparović, M. The Effect of Soil Sampling Density and Spatial Autocorrelation on Interpolation Accuracy of Chemical Soil Properties in Arable Cropland. *Agronomy* **2021**, *11*, 2430. [CrossRef]
9. Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O. LUCAS Soil, the Largest Expandable Soil Dataset for Europe: A Review. *Eur. J. Soil Sci.* **2018**, *69*, 140–153. [CrossRef]
10. Khaledian, Y.; Miller, B.A. Selecting Appropriate Machine Learning Methods for Digital Soil Mapping. *Appl. Math. Model.* **2020**, *81*, 401–418. [CrossRef]
11. Nussbaum, M.; Spiess, K.; Baltensweiler, A.; Grob, U.; Keller, A.; Greiner, L.; Schaepman, M.E.; Papritz, A. Evaluation of Digital Soil Mapping Approaches with Large Sets of Environmental Covariates. *SOIL* **2018**, *4*, 1–22. [CrossRef]
12. Radočaj, D.; Jurišić, M.; Antonić, O.; Šiljeg, A.; Cukrov, N.; Rapčan, I.; Plaščak, I.; Gašparović, M. A Multiscale Cost–Benefit Analysis of Digital Soil Mapping Methods for Sustainable Land Management. *Sustainability* **2022**, *14*, 12170. [CrossRef]
13. Minasny, B.; McBratney, A.B. Digital Soil Mapping: A Brief History and Some Lessons. *Geoderma* **2016**, *264*, 301–311. [CrossRef]
14. Hengl, T.; Heuvelink, G.B.M.; Stein, A. A Generic Framework for Spatial Prediction of Soil Variables Based on Regression-Kriging. *Geoderma* **2004**, *120*, 75–93. [CrossRef]
15. Broeg, T.; Blaschek, M.; Seitz, S.; Taghizadeh-Mehrjardi, R.; Zepp, S.; Scholten, T. Transferability of Covariates to Predict Soil Organic Carbon in Cropland Soils. *Remote Sens.* **2023**, *15*, 876. [CrossRef]

16. Hengl, T.; de Jesus, J.M.; Heuvelink, G.B.M.; Gonzalez, M.R.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]
17. Radočaj, D.; Gašparović, M.; Jurišić, M. Open Remote Sensing Data in Digital Soil Organic Carbon Mapping: A Review. *Agriculture* **2024**, *14*, 1005. [CrossRef]
18. Pouladi, N.; Gholizadeh, A.; Khosravi, V.; Borůvka, L. Digital Mapping of Soil Organic Carbon Using Remote Sensing Data: A Systematic Review. *CATENA* **2023**, *232*, 107409. [CrossRef]
19. Balenzano, A.; Mattia, F.; Satalino, G.; Lovergine, F.P.; Palmisano, D.; Peng, J.; Marzahn, P.; Wegmuller, U.; Cartus, O.; Dabrowska-Zielinska, K.; et al. Sentinel-1 Soil Moisture at 1 Km Resolution: A Validation Study. *Remote Sens. Environ.* **2021**, *263*, 112554. [CrossRef]
20. Misra, G.; Cawkwell, F.; Wingler, A. Status of Phenological Research Using Sentinel-2 Data: A Review. *Remote Sens.* **2020**, *12*, 2760. [CrossRef]
21. Crosson, W.L.; Al-Hamdan, M.Z.; Hemmings, S.N.J.; Wade, G.M. A Daily Merged MODIS Aqua–Terra Land Surface Temperature Data Set for the Conterminous United States. *Remote Sens. Environ.* **2012**, *119*, 315–324. [CrossRef]
22. Dai, L.; Ge, J.; Wang, L.; Zhang, Q.; Liang, T.; Bolan, N.; Lischeid, G.; Rinklebe, J. Influence of Soil Properties, Topography, and Land Cover on Soil Organic Carbon and Total Nitrogen Concentration: A Case Study in Qinghai-Tibet Plateau Based on Random Forest Regression and Structural Equation Modeling. *Sci. Total Environ.* **2022**, *821*, 153440. [CrossRef] [PubMed]
23. Zhen, Z.; Chen, S.; Yin, T.; Chavanon, E.; Lauret, N.; Guilleux, J.; Henke, M.; Qin, W.; Cao, L.; Li, J.; et al. Using the Negative Soil Adjustment Factor of Soil Adjusted Vegetation Index (SAVI) to Resist Saturation Effects and Estimate Leaf Area Index (LAI) in Dense Vegetation Areas. *Sensors* **2021**, *21*, 2115. [CrossRef] [PubMed]
24. Radočaj, D.; Jurišić, M.; Tadić, V. The Effect of Bioclimatic Covariates on Ensemble Machine Learning Prediction of Total Soil Carbon in the Pannonian Biogeoregion. *Agronomy* **2023**, *13*, 2516. [CrossRef]
25. R: Contributors. Available online: https://www.r-project.org/contributors.html (accessed on 17 August 2024).
26. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Benesty, M.; et al. Caret: Classification and Regression Training. Available online: https://CRAN.R-project.org/package=caret (accessed on 30 May 2022).
27. Cutler, F. Original by L.B. and A.; Wiener, R. port by A.L. and M. RandomForest: Breiman and Cutler's Random Forests for Classification and Regression. Available online: https://CRAN.R-project.org/package=randomForest (accessed on 23 October 2022).
28. Kuhn, M.; Weston, S.; Keefer, C.; Coulter, N.; Quinlan, R. Cubist: Rule- and Instance-Based Regression Modeling. Available online: https://cran.r-project.org/web/packages/Cubist/index.html (accessed on 3 May 2024).
29. Karatzoglou, A.; Smola, A.; Hornik, K.; Maniscalco, M.A.; Teo, C.H. Kernlab: Kernel-Based Machine Learning Lab. Available online: https://CRAN.R-project.org/package=kernlab (accessed on 25 October 2022).
30. Rodriguez, P.P.; Gianola, D. Brnn: Bayesian Regularization for Feed-Forward Neural Networks. Available online: https://cran.r-project.org/web/packages/brnn/index.html (accessed on 14 October 2024).
31. Liland, K.H.; Mevik, B.H.; Wehrens, R.; Hiemstra, P. Pls: Partial Least Squares and Principal Component Regression. Available online: https://CRAN.R-project.org/package=pls (accessed on 21 October 2022).
32. Therneau, T.; Atkinson, B.; Ripley, B. Rpart: Recursive Partitioning and Regression Trees. Available online: https://cran.r-project.org/web/packages/rpart/index.html (accessed on 17 August 2024).
33. Ridgeway, G.; Edwards, D.; Kriegler, B.; Schroedl, S.; Southworth, H.; Greenwell, B.; Boehmke, B.; Cunningham, J. Developers. G.B.M. Gbm: Generalized Boosted Regression Models 2024. Available online: https://github.com/gbm-developers (accessed on 14 October 2024).
34. Meinshausen, N. quantregForest: Quantile Regression Forests. Available online: https://cran.r-project.org/web/packages/quantregForest/index.html (accessed on 17 August 2024).
35. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: Extreme Gradient Boosting. Available online: https://cran.r-project.org/web/packages/xgboost/index.html (accessed on 14 October 2024).
36. Hastie, T. Gam: Generalized Additive Models. Available online: https://cran.r-project.org/web/packages/gam/index.html (accessed on 17 August 2024).
37. Hastie, T. Elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. Available online: https://cran.r-project.org/web/packages/elasticnet/index.html (accessed on 17 August 2024).
38. Hastie, T.; Efron, B. Lars: Least Angle Regression, Lasso and Forward Stagewise. Available online: https://cran.r-project.org/web/packages/lars/index.html (accessed on 17 August 2024).
39. Mondal, R.; Bhat, A. Comparison of Regression-Based and Machine Learning Techniques to Explain Alpha Diversity of Fish Communities in Streams of Central and Eastern India. *Ecol. Indic.* **2021**, *129*, 107922. [CrossRef]
40. Gbur, E.E.; Stroup, W.W.; McCarter, K.S.; Durham, S.; Young, L.J.; Christman, M.; West, M.; Kramer, M. *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2020; ISBN 978-0-89118-182-8.
41. Ravindra, K.; Rattan, P.; Mor, S.; Aggarwal, A.N. Generalized Additive Models: Building Evidence of Air Pollution, Climate Change and Human Health. *Environ. Int.* **2019**, *132*, 104987. [CrossRef]

42. Emmert-Streib, F.; Dehmer, M. High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 359–383. [CrossRef]
43. Krishnan, N.M.A.; Kodamana, H.; Bhattoo, R. (Eds.) Parametric Methods for Regression. In *Machine Learning for Materials Discovery: Numerical Recipes and Practical Applications*; Springer International Publishing: Cham, Switzerland, 2024; pp. 61–83. ISBN 978-3-031-44622-1.
44. Mikulasek, B.; Fonseca Diaz, V.; Gabauer, D.; Herwig, C.; Nikzad-Langerodi, R. Partial Least Squares Regression with Multiple Domains. *J. Chemom.* **2023**, *37*, e3477. [CrossRef]
45. Diaz-Gonzalez, F.A.; Vuelvas, J.; Correa, C.A.; Vallejo, V.E.; Patino, D. Machine Learning and Remote Sensing Techniques Applied to Estimate Soil Indicators—Review. *Ecol. Indic.* **2022**, *135*, 108517. [CrossRef]
46. Hamze-Ziabari, S.M.; Bakhshpoori, T. Improving the Prediction of Ground Motion Parameters Based on an Efficient Bagging Ensemble Model of M5′ and CART Algorithms. *Appl. Soft Comput.* **2018**, *68*, 147–161. [CrossRef]
47. Sahin, E.K. Assessing the Predictive Capability of Ensemble Tree Methods for Landslide Susceptibility Mapping Using XGBoost, Gradient Boosting Machine, and Random Forest. *SN Appl. Sci.* **2020**, *2*, 1308. [CrossRef]
48. Zhou, J.; Li, E.; Wei, H.; Li, C.; Qiao, Q.; Armaghani, D.J. Random Forests and Cubist Algorithms for Predicting Shear Strengths of Rockfill Materials. *Appl. Sci.* **2019**, *9*, 1621. [CrossRef]
49. Hengl, T.; Nussbaum, M.; Wright, M.N.; Heuvelink, G.B.M.; Graeler, B. Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-Temporal Variables. *PeerJ* **2018**, *6*, e5518. [CrossRef] [PubMed]
50. Lagacherie, P.; Arrouays, D.; Bourennane, H.; Gomez, C.; Nkuba-Kasanda, L. Analysing the Impact of Soil Spatial Sampling on the Performances of Digital Soil Mapping Models and Their Evaluation: A Numerical Experiment on Quantile Random Forest Using Clay Contents Obtained from Vis-NIR-SWIR Hyperspectral Imagery. *Geoderma* **2020**, *375*, 114503. [CrossRef]
51. Demir, S.; Şahin, E.K. Liquefaction Prediction with Robust Machine Learning Algorithms (SVM, RF, and XGBoost) Supported by Genetic Algorithm-Based Feature Selection and Parameter Optimization from the Perspective of Data Processing. *Environ. Earth Sci.* **2022**, *81*, 459. [CrossRef]
52. Huber, F.; Yushchenko, A.; Stratmann, B.; Steinhage, V. Extreme Gradient Boosting for Yield Estimation Compared with Deep Learning Approaches. *Comput. Electron. Agric.* **2022**, *202*, 107346. [CrossRef]
53. Baltensweiler, A.; Walthert, L.; Hanewinkel, M.; Zimmermann, S.; Nussbaum, M. Machine Learning Based Soil Maps for a Wide Range of Soil Properties for the Forested Area of Switzerland. *Geoderma Reg.* **2021**, *27*, e00437. [CrossRef]
54. Dutschmann, T.-M.; Kinzel, L.; ter Laak, A.; Baumann, K. Large-Scale Evaluation of k-Fold Cross-Validation Ensembles for Uncertainty Estimation. *J. Cheminformatics* **2023**, *15*, 49. [CrossRef]
55. Chai, T.; Draxler, R.R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?—Arguments against Avoiding RMSE in the Literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
56. Kovačić, Đ.; Radočaj, D.; Jurišić, M. Ensemble Machine Learning Prediction of Anaerobic Co-Digestion of Manure and Thermally Pretreated Harvest Residues. *Bioresour. Technol.* **2024**, *402*, 130793. [CrossRef]
57. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]
58. Poggio, L.; De Sousa, L.M.; Batjes, N.H.; Heuvelink, G.B.; Kempen, B.; Ribeiro, E.; Rossiter, D. SoilGrids 2.0: Producing Soil Information for the Globe with Quantified Spatial Uncertainty. *SOIL* **2021**, *7*, 217–240. [CrossRef]
59. Kalantar, B.; Ueda, N.; Saeidi, V.; Ahmadi, K.; Halin, A.A.; Shabani, F. Landslide Susceptibility Mapping: Machine and Ensemble Learning Based on Remote Sensing Big Data. *Remote Sens.* **2020**, *12*, 1737. [CrossRef]
60. Radočaj, D.; Gašparović, M.; Radočaj, P.; Jurišić, M. Geospatial Prediction of Total Soil Carbon in European Agricultural Land Based on Deep Learning. *Sci. Total Environ.* **2024**, *912*, 169647. [CrossRef]
61. Zhu, Y.; Zhao, C.; Yang, H.; Yang, G.; Han, L.; Li, Z.; Feng, H.; Xu, B.; Wu, J.; Lei, L. Estimation of Maize Above-Ground Biomass Based on Stem-Leaf Separation Strategy Integrated with LiDAR and Optical Remote Sensing Data. *PeerJ* **2019**, *7*, e7593. [CrossRef]
62. Notton, G.; Voyant, C.; Fouilloy, A.; Duchaud, J.L.; Nivet, M.L. Some Applications of ANN to Solar Radiation Estimation and Forecasting for Energy Applications. *Appl. Sci.* **2019**, *9*, 209. [CrossRef]
63. Najwer, A.; Jankowski, P.; Niesterowicz, J.; Zwoliński, Z. Geodiversity Assessment with Global and Local Spatial Multicriteria Analysis. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102665. [CrossRef]
64. Wadoux, A.M.J.-C.; Brus, D.J. How to Compare Sampling Designs for Mapping? *Eur. J. Soil Sci.* **2021**, *72*, 35–46. [CrossRef]
65. Broeg, T.; Don, A.; Gocht, A.; Scholten, T.; Taghizadeh-Mehrjardi, R.; Erasmi, S. Using Local Ensemble Models and Landsat Bare Soil Composites for Large-Scale Soil Organic Carbon Maps in Cropland. *Geoderma* **2024**, *444*, 116850. [CrossRef]
66. Radočaj, D.; Tuno, N.; Mulahusić, A.; Jurišić, M. Evaluation of Ensemble Machine Learning for Geospatial Prediction of Soil Iron in Croatia. *Poljoprivreda* **2023**, *29*, 53–61. [CrossRef]
67. Adeniyi, O.D.; Brenning, A.; Bernini, A.; Brenna, S.; Maerker, M. Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy, Italy. *Land* **2023**, *12*, 494. [CrossRef]