International Journal of
Geo-Information

Article

Prediction and Uncertainty Capabilities of Quantile Regression Forests in Estimating Spatial Distribution of Soil Organic Matter

Melpomeni Nikou [1,2] and Panagiotis Tziachris [2,*]

[1] Department of Meteorology-Climatology, School of Geology, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece; nikomelp@geo.auth.gr

[2] Soil and Water Resources Institute, Hellenic Agricultural Organization—Demeter, 570 01 Thessaloniki, Greece

* Correspondence: p.tziachris@swri.gr

Abstract: One of the core tasks in digital soil mapping (DSM) studies is the estimation of the spatial distribution of different soil variables. In addition, however, assessing the uncertainty of these estimations is equally important, something that a lot of current DSM studies lack. Machine learning (ML) methods are increasingly used in this scientific field, the majority of which do not have intrinsic uncertainty estimation capabilities. A solution to this is the use of specific ML methods that provide advanced prediction capabilities, along with innate uncertainty estimation metrics, like Quantile Regression Forests (QRF). In the current paper, the prediction and the uncertainty capabilities of QRF, Random Forests (RF) and geostatistical methods were assessed. It was confirmed that QRF exhibited outstanding results at predicting soil organic matter (OM) in the study area. In particular, R^2 was much higher than the geostatistical methods, signifying that more variation is explained by the specific model. Moreover, its uncertainty capabilities as presented in the uncertainty maps, shows that it can also provide a good estimation of the uncertainty with distinct representation of the local variation in specific parts of the area, something that is considered a significant advantage, especially for decision support purposes.

Keywords: quantile regression forests; random forests; geostatistics; machine learning; soil organic matter; prediction uncertainty

check for updates

Citation: Nikou, M.; Tziachris, P. Prediction and Uncertainty Capabilities of Quantile Regression Forests in Estimating Spatial Distribution of Soil Organic Matter. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 130. https://doi.org/10.3390/ijgi11020130

Academic Editor: Wolfgang Kainz

Received: 7 December 2021
Accepted: 8 February 2022
Published: 11 February 2022

1. Introduction

Digital soil mapping (DSM), also known as predictive soil mapping or pedometric mapping, refers to the creation of digital maps that include spatial soil information, such as soil type or soil properties. These maps are created from the combination of multiple parameters (soil, climate, relief etc.) and usually depict the spatial distribution of soil phenomena along with relative information (e.g., estimation uncertainty).

DSM makes extensive use of geographic information systems (GIS), global positioning systems (GPS), remotely sensed spectral data, topographic data derived from digital elevation models (DEMs), predictive or inference models, and software for data analysis. To cope with the large amount of data used in DSM, semi-automated techniques and technologies are used to acquire, process, and visualize these data. Machine learning (ML) and artificial intelligence (AI) are some innovative state of the art technologies that are increasingly used in soil mapping and their uptake is transforming the way soil scientists produce their maps [1]. ML that emerged in the 1990s as a tool for DSM [2] is defined as the computer-assisted practice of using data-driven (and mostly non-linear) statistical models which resorts to a large amount of input data to learn a pattern and make a prediction [1].

According to Leo Breiman [3] two statistical modeling paradigms were distinguished: a data model and an algorithmic model. A data model is an abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities, whereas an algorithmic model is a model that uses mathematical algorithms based on the elements of data and estimates the parameters. One broadly used algorithm for that kind of model is Random Forests (RF). The RF is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and is extensively used in DSM [4]. For example, it provided the best results in estimating soil OM [5], shortened the training time during the soil OM modeling process and improved the model's accuracy and its predictive ability [6]. Finally, according to John et al. [7], RF was the best performing model among other ML algorithms such as artificial neural network (ANN), support vector machine (SVM), and cubist regression.

Most of the time, DSM products represent estimates of spatially distributed soil properties. These estimations comprise an element of uncertainty that is not evenly distributed over the area covered by DSM. If we quantify the uncertainty spatially explicitly, this information can be used to improve the quality of DSM by optimizing the sampling design [8]. Wadoux et al. [1] stated that while the (spatial) cross-validation results might show strong agreement between predicted and measured soil property or class and therefore validate a ML model with very high predictive abilities, an uncertainty quantification would show unrealistic predictions characterized by a large uncertainty. However, most ML methods RF included do not provide uncertainty estimates by default and only 30% of the recent soil studies in their literature review quantified the uncertainty associated with the prediction.

One ML method that intrinsically addresses the lack of uncertainty estimates is Quantile Regression Forests (QRF). The QRF is an extension of RF developed by Nicolai Meinshausen [9] that provides non-parametric estimates of the median predicted value as well as prediction quantiles. It therefore allows spatially explicit non-parametric estimates of model uncertainty by providing information for the full conditional distribution of the response variable, and not only about the conditional mean [10]. As a result, QRF can potentially combine the high accuracy of RF with the built-in uncertainty estimates. However, QRF is not broadly used in soil studies, despite its advantages.

Vaysse and Lagacherie [11], for example, conducted an experiment in which they employed QRF in a temperate Mediterranean area with a comparable soil organic carbon (SOC) dataset in terms of areal extent, observation density, and distribution homogeneity. They claim that QRF outperforms RK when it comes to interpreting uncertainty patterns and is better suited than other modeling methods when spatial sampling is sparse. In Dharumarajan's study [12] the QRF model was used for the estimation of several important soil qualities of Northern Karnataka according to GlobalSoilMap criteria. The QRF model caught maximum variability for most of the soil parameters, and the predicted soil values were dependable with minimum errors. The QRF was also used for the production of global maps of soil properties explicitly highlighting the importance of quantitative and qualitative evaluation and uncertainty communication [13]. Finally, in Veronesi's study in 2019 [14], RF and QRF generated the most reliable confidence intervals for predicting SOC. Even though this is potentially important for practical uses, the confidence intervals were also very wide, so they suggest that these intervals should be handled carefully.

In the current study, the prediction capability along with the uncertainty assessment capacity of the QRF is examined. The popular geostatistical methods of Ordinary Kriging (OK) and Kriging with External Drift (KED) were compared with the ML methods of RF and QRF, in the case of soil OM. Prediction maps of soil OM along with maps of uncertainties were also produced and presented. The study area that was chosen is in northern Greece, at the regional unit of Kastoria and next to the shore of Lake Orestiada. A total number of 414 samples of soil were collected in randomly sampled unique locations in autumn during a six-year period. GPS receivers were used to identify the sampling positions. A high-resolution Digital Elevation Model (DEM) was used to derive topographic products such as aspect, slope, altitude etc. along with Sentinel-2 imagery, for each year from our study period, to produce Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) that were used as input data. Finally, the effect of the above-mentioned covariates to the prediction of soil OM was assessed based on the importance score of the applied machine learning methods.

2. Materials and Methods

2.1. Area of Study and Soil Sampling

The study area was chosen in northern Greece, near the shore of Lake Orestiada, in the regional unit of Kastoria (Figure 1). Its coordinates in the World Geodetic System of 1984 (WGS84) include the area between 40°28′42.41″ N and 40°32′35.61″ N latitudes and longitudes of 21°19′4.01″ E and 21°23′8.18″ E longitudes.

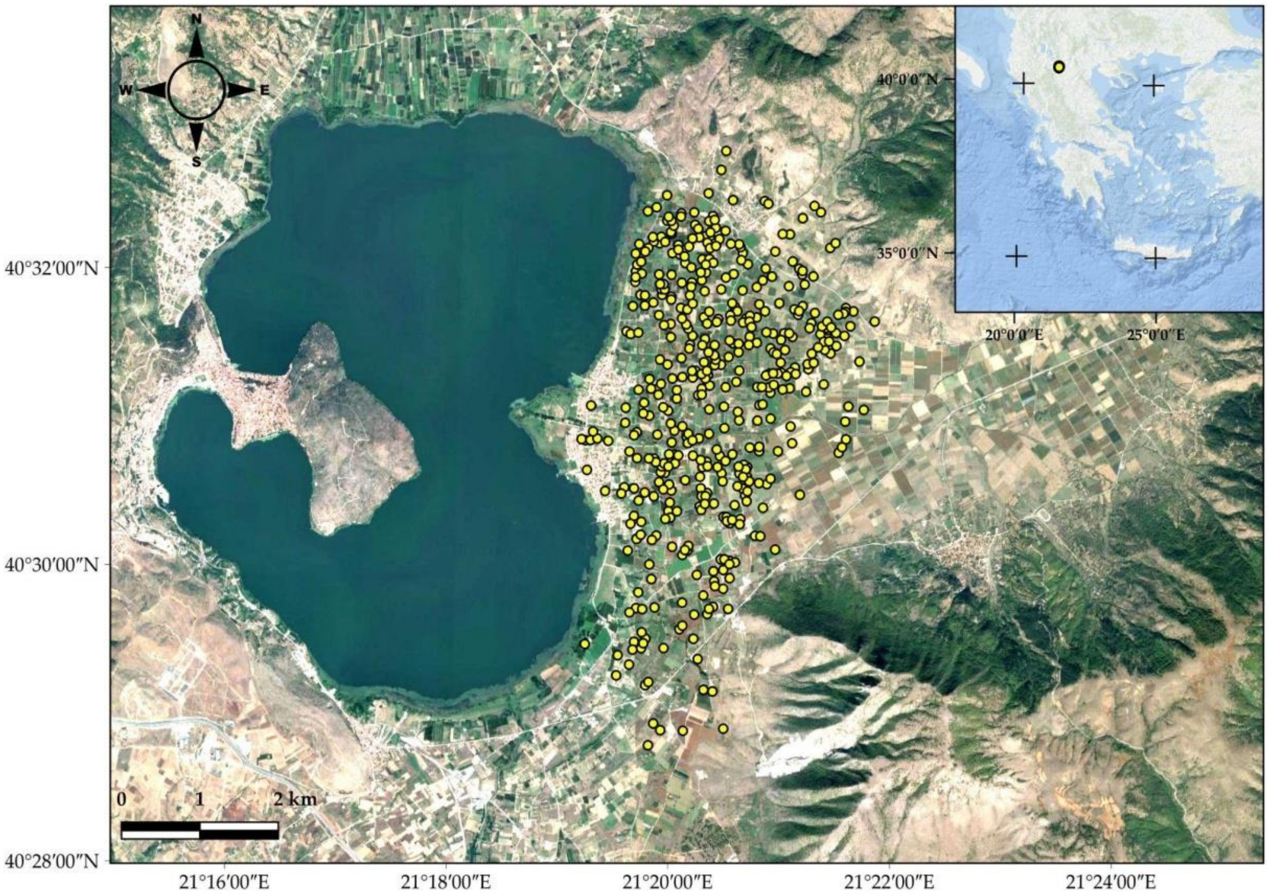

Figure 1. The study area on the shore of Lake Orestiada, in the regional unit of Kastoria, Greece.

While the region of interest is flat, the mean altitude is around 640 m above sea level, ranging from 620 m near the lake to 700 m further north. The climate is temperate and often warm, with harsh winters that frequently keep the temperature below zero throughout the day. The average annual temperature is 11.5 °C, with 636 mm of precipitation. Summers are hot and dry, with a relative humidity of 50% to 55%. Apple trees and beans are the primary agricultural crops.

During a six-year period, a total of 414 soil samples were collected in randomly sampled distinct places around the study area (2012 to 2019). A total of 30 cm of top soil was gathered during late fall season (around end of November). The sampling positions were determined using Global Positioning System (GPS) devices. The minimum distance between two sampling places varies between 60 and 480 m, with an average of 90 m.

2.2. Soil, Environmental and Satellite Covariates

In this study, soil variables, environmental variables and satellite images were selected (Table 1) as potential inputs in the models. Regarding soil covariates, the 414 soil samples that were collected from the area were analyzed for Clay (C) with soil hydrometer

(Bouyoucos) method [15], Magnesium (Mg) with ammonium acetate method and Zinc (Zn) with DTPA method [16]. Moreover, an organic matter (OM) analysis (wet oxidation method) from the same locations was conducted, to calibrate the models and to assess the prediction results. In more detail, during the soil sampling procedure, a composite soil sample consisting of several sub-samples up to a depth of 30 cm was obtained from each field parcel and the soil samples were dried and analyzed in the laboratory of the Soil and Water Resources Institute in Thessaloniki, Greece.

Table 1. Variables used in the study.

	Variables	Category
1	Clay (C)	soil
2	Organic Matter (OM)	soil
3	Magnesium (Mg)	soil
4	Zinc (Zn)	soil
5	Elevation (altitude)	environmental
6	Slope	environmental
7	Aspect	environmental
8	SAGA wetness index (TWI)	environmental
9	Negative Topographic Openness (openn)	environmental
10	Positive Topographic Openness (openp)	environmental
11	Deviation from Mean Value (devmean)	environmental
12	Multiresolution index of Valley Bottom Flatness (vbf)	environmental
13	NDVI 2016	satellite
14	NDWI 2016	satellite
15	NDVI 2017	satellite
16	NDWI 2017	satellite
17	NDVI 2018	satellite
18	NDWI 2018	satellite
19	NDVI 2019	satellite
20	NDWI 2019	satellite

The environmental covariates were derived from the second version of the Advanced Spaceborne Thermal Emission Radiometer-Global Digital Elevation Model version 2 (ASTER GDEM2). The release of the ASTER GDEM2 has enriched the availability of free-of-charge DEM sources, which are especially useful for developing countries, and prompted users to assess its quality and accuracy [17]. The ASTER GDEM2 is consisted of $1^\circ \times 1^\circ$ tiles (30 m resolution) in the World Geodetic System 1984 (WGS84), that was reprojected to Greek Geodetic Reference System 1987 (GGRS87) for this study [18].

Moreover, satellite indices were derived from Sentinel-2 imagery. More specifically Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) from 2016 to 2019 were collected at approximately the same time period that the soil data were collected (end of November). The well known and widely used NDVI is a simple, but effective index for quantifying green vegetation. Red light is actively absorbed by healthy plants, while near infrared is reflected. To determine the state of a plant's health, we must compare the values of red and infrared light absorption and reflection [7,19]. The NDWI is a vegetation index sensitive to the water content of vegetation and is complementary to the NDVI. High NDWI values indicate a high plant water content and a high plant fraction coating. Low vegetation content and cover with low vegetation correspond to low NDWI values. The NDWI rate will drop during times of water stress [20].

2.3. Data Preparation and Assessment

Initially, the topographic data and satellite indices along with the soil analysis data were combined and spatially overlayed on the sampling locations. The overall dataset was assessed for outliers and missing values. From the initial 414 points, 403 points remained at the end that were used as inputs for the models in the study.

From the full set of variables, only a subset was used in the study. The variables were eliminated using Akaike Information Criteria (stepAIC) technique and Principal Component Analysis (PCA) and were also assessed for multicollinearity. The remaining variables were C, OM, ZN, MG, Vdepth, Altitude, NDVI_2016, NDVI_2017, and NDWI_2019 (Table 2).

Table 2. Descriptive statistics of the auxiliary variables from the 403 locations in study area.

	C (%)	OM (%)	ZN (%)	MG (%)	Vdepth (m)	Altitude (m)	NDVI_2016	NDVI_2017	NDWI_2019
mean	17.43	2.13	1.91	305.88	31.64	635.33	0.49	0.52	0.63
sd	7.07	0.76	1.43	148.04	7.06	7.39	0.13	0.14	0.10
median	16.00	2.00	1.51	275.00	31.46	634.00	0.50	0.55	0.66
trimmed	16.89	2.06	1.69	290.45	31.64	634.42	0.50	0.53	0.65
mad	5.93	0.68	1.01	127.50	7.14	5.93	0.11	0.13	0.07
min	2.00	0.64	0.12	44.00	4.67	624.00	0.10	0.12	0.24
max	48.00	5.24	10.75	905.00	46.58	680.00	0.75	0.90	0.79
skew	0.89	0.98	2.12	1.05	−0.17	2.49	−0.80	−0.57	−1.44
kurtosis	1.49	1.11	6.86	1.16	0.15	10.30	0.75	−0.05	1.96

The maps of the spatial distributions of the environmental covariates that were derived from ASTER GDEM2 and were used in the study (Vdepth and Altitude), are presented next (Figure 2).

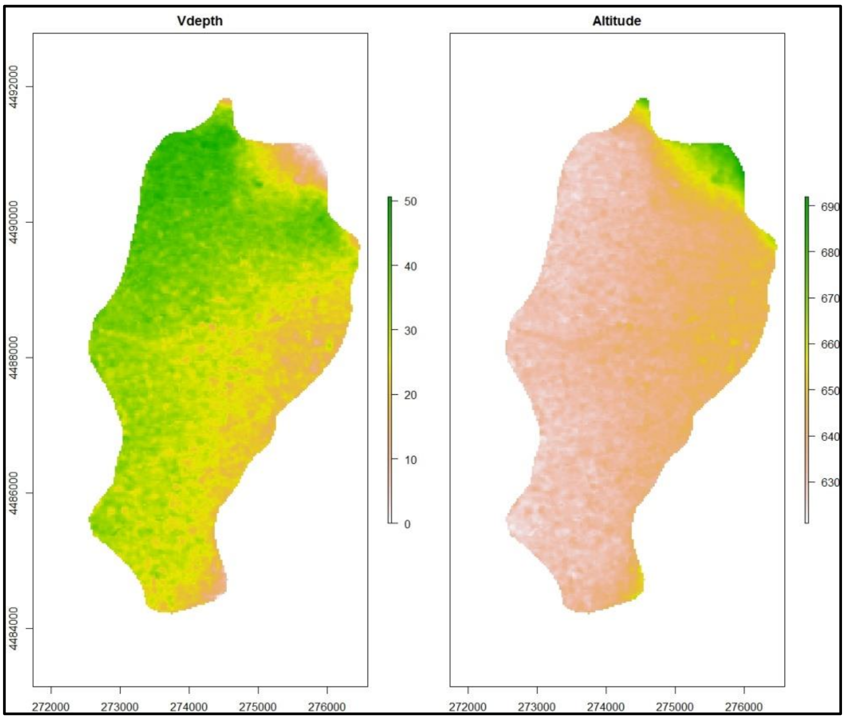

Figure 2. Topographic covariates of the study area.

The maps of the satellite covariates that were used in the study (NDVI_2016, NDVI_2017, and NDWI_2019) are the following (Figure 3).

Finally, the soil covariates (C, MG, ZN) were interpolated from the known point locations of the full dataset with the use of OK and their spatial distribution for the overall study area was estimated (Figure 4).

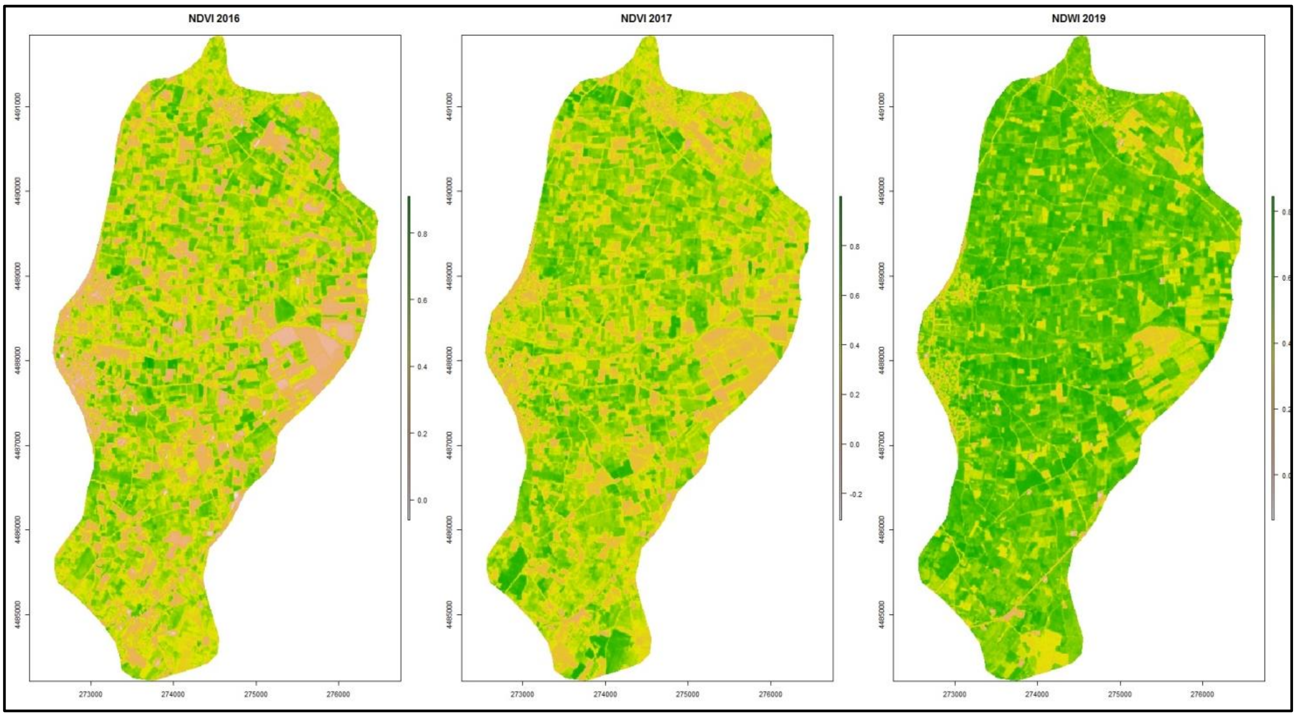

Figure 3. Satellite covariates of the study area.

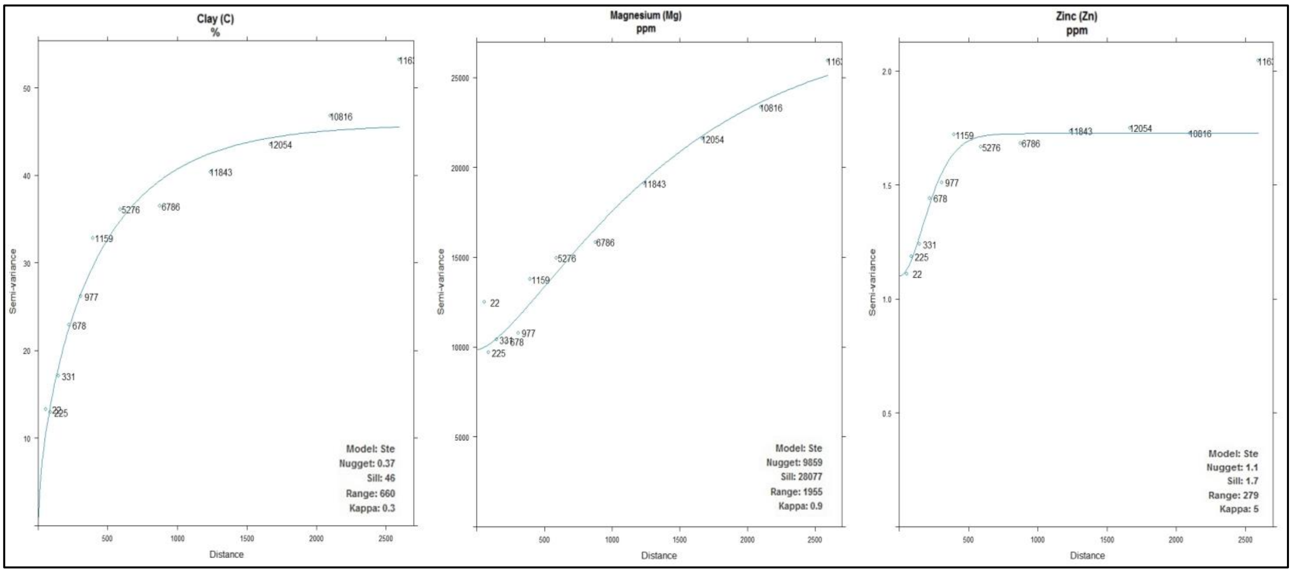

Figure 4. Semivariograms and fitted model of soil covariates of the study area.

For all the soil parameters, the Matern semivariogram model with M. Stein's parameterization (Ste) was applied as the fitted model using gstat's default parameters. Regarding C, its range was 660 m and exhibited a strong spatial dependence with a nugget to sill ratio of 0.8% [21]. The Mg had a range of 1955 m with strong spatial dependence (nugget to sill 3.5%) whereas Zn had a range of 279 m with moderate spatial dependence with a nugget to sill close to 65%.

The produced kriging maps from the soil covariates were used for the estimation of soil OM in the area by the models of the current study (KED, RF, QRF) (Figure 5).

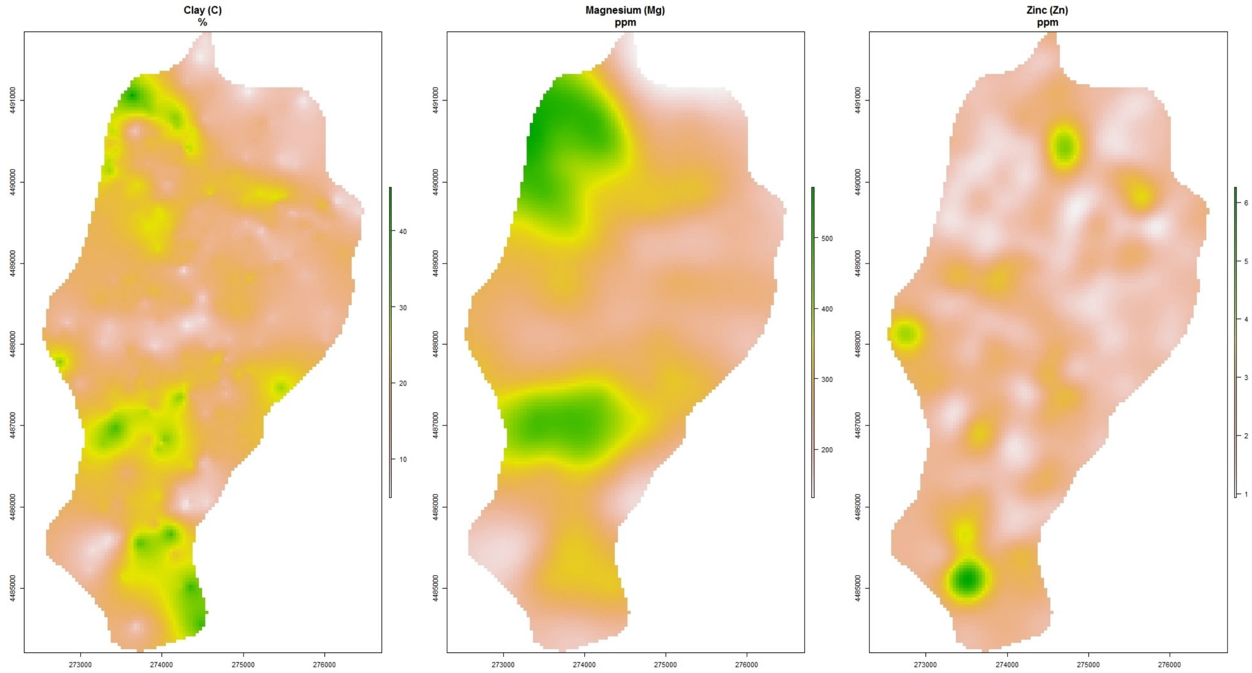

Figure 5. Spatial distribution of soil covariates of the study area.

2.4. Ordinary Kriging (OK) and Kriging with External Drift (KED)

Ordinary Kriging is a type of kriging in which the values' weights sum to one. It is linear because its estimates are a linear combination of the available data. It is also unbiased because it attempts to keep the mean residual to be zero and it tries to minimize the residual variance [22]. The OK implicitly evaluates the mean in a moving neighborhood with local second-order stationarity and its variance is equal to the sum of the simple kriging variance (assuming a known mean) plus the variance due to uncertainty about the true mean value [23].

Universal kriging (UK), Kriging with External Drift and Regression-Kriging (RK) belong to the group of the so-called 'hybrid' [24], i.e., non-stationary geo-statistical methods [23]. In classical geostatistics, spatial prediction for non-stationary processes is accomplished by taking into account a spatial trend (also known as "drift") that is either modeled solely as a function of the coordinates (in UK) or defined "externally" through some auxiliary variables (in KED) [25].

KED solves kriging weights by extending the covariance matrix with auxiliary variables so that the universality conditions are integrated into the kriging system; here, the difficulty is obtaining satisfactory residual variogram in the presence of drift [26].

The implementations of both OK and KED in the current study was done with the gstat package in R.

2.5. Random Forests (RF) and Quantile Regression Forests (QRF)

The idea of developing the RF method is based on the combination of the Bagging and Random Subspace Method, utilizing their advantages and compensating their disadvantages, with impressive results [3,27].

According to Breiman (2001) [3], in the case of classification, "A random forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θk), k = 1, . . . } where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x". In case of regression Breiman states that "...random forests for regression are formed by growing trees depending on a random vector Θ such that the tree predictor h(x, Θ) takes on numerical values as opposed to class labels. The output values are numerical, and we assume that the training set is independently drawn from the distribution of the random vector Y, X".

The RF for regression is extensively used in DSM (e.g., [28–33]) with very positive results in the prediction of different soil parameters. More importantly though, it works equally well with skewed and normally distributed variables, without the need of statistical assumptions or restrictions that other methods demand. Therefore, it is easier and more straightforward to use. It just requires special attention in optimizing the hyperparameters to get the best results. One major drawback of some of the well-known ML methods (RF, ANN etc.) is their lack of intrinsic uncertainty estimation capabilities. So, apart from prediction maps, no prediction error variance can be estimated, contrary to classical geostatistics methods. The main reason for this is that most ML methods, RF included, provide only mean value predictions.

A possible solution to this deficiency came from Nicolai Meinshausen [9], who generalized the standard RF to provide information for the full conditional distribution of the response variable, and not only about the conditional mean. This ML algorithm is called Quantile Regression Forests (QRF) and gives a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables. The key difference between QRF and RF is as follows: for each node in each tree, RF keeps only the mean of the observations that fall into this node and neglect all other information. In contrast, QRF keeps the values of all observations in this node, not just their mean, and assesses the conditional distribution based on this information.

In the current study, the ranger package in R language was used to implement the ML models. Ranger is a rapid implementation of RF or recursive partitioning that is especially well suited for high-dimensional data.

The assessment of the optimal hyperparameters of a ML model is a crucial step for the estimation of the best ML models for each specific use case. The ideal hyperparameter settings have a direct impact on the model's performance. Although there are various automatic optimization methods, their strengths and drawbacks change when applied to different types of situations [34]. In the current study, the random search method was performed (a 10 k-fold with 3 repeats), in which random combinations of parameters were employed from a range of values and used as hyperparameters. The ML model with the set of parameters that gave the highest accuracy was considered to be the best and used for prediction. The overall dataset (403 samples) was split in two distinct datasets: the training dataset (70% of the data) that was used for estimating the models hyperparameters and the testing dataset (30% of the data) that was used for assessing the different models. The specific hyperparameters for RF that were optimized are presented in Table 3.

Table 3. RF and QRF hyperparameters.

Hyperparameters	Packages	Description
mtry	ranger	The number of random features used in each tree.
num.trees	ranger	The number of grown trees.
min.node.size	ranger	Minimal node size.
splitrule	ranger	A switch for linear output units.

2.6. Uncertainty

The DSM products represent estimates of spatially distributed soil properties. These estimations comprise an element of uncertainty that is not evenly distributed over the area covered by DSM [8]. These flaws are being addressed by combining soil data at sites with spatially exhaustive environmental factors using quantitative models (e.g., [35–37]) DSM products are repeatable and enable continuous data display due to their quantitative nature. Models can also be updated for multiple reasons, and uncertainty can be measured [38,39].

Measurements, digitisation, data input, interpretation, categorization, generalisation, and interpolation are all common sources of mistake [40]. Modeling bias, parameterization, or even measurement mistakes connected with the input data can all cause uncertainty in digital soil maps [41]. Nelson et al. [42] recommends doing an error budget to assess the contribution of each error using a combination of geostatistical and Monte-Carlo simulations to gain a better understanding of uncertainty. The distinction between model error and spatially explicit uncertainty must also be considered [43]. The average squared difference between the estimated value and the actual value is known as model error, which is frequently assessed as the mean square error (MSE) [44,45]. However, spatially explicit uncertainty, often known as "local error", refers to the quantification of model output prediction intervals (e.g., [11,44,46]).

The prediction is linked to an explicit measure of the uncertainty. In many circumstances, such as in a decision-making process, it is just as important to quantify prediction uncertainty as it is to make the prediction itself, thus uncertainty maps are necessary (e.g., [47,48]). In DSM, uncertainty analysis is crucial in deciding whether the predicted soil map is dependable enough to be applied in agricultural production systems or decision-making. Uncertainty analysis also involves acknowledging the model's limitations, which is a step toward model interpretability [1]. As Heuvelink [49] states, we are very interested in prediction intervals in soil mapping, i.e., the range that is likely to contain the value yet to be measured. However, a very small amount of DSM studies estimate uncertainty. According to Wadoux [1], only around 30% of the studies presented in their paper quantified the prediction's uncertainty.

In the current study the uncertainty was estimated for only three out of four methods: OK, KED, QRF. The RF does not provide uncertainty estimation capabilities per se. The geostatistical methods of OK and KED provide by default the variance that it was used for the uncertainty assessment. Mainly, the standard deviation was calculated and its range was depicted in the maps. For the QRF the range was defined as one standard deviation above and below the median value. This range was used to create the uncertainty map of the study area

2.7. Error Assessment

Different metrics (Table 4) were employed to estimate model performance based on the difference between the observations and the predictions at the testing data set.

Table 4. Measurements to assess model performance.

Metrics	Equation	
Mean absolute error (*MAE*)	$MAE = \frac{\sum_{i=1}^{n} \lvert y_i - x_i \rvert}{n}$	(1)
Root mean square error (*RMSE*)	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} [y_i - x_i]^2}{n}}$	(2)
Coefficient of determination (R^2)	$R^2 = 1 - \frac{SSE}{SSTO}$	(3)
Mean bias error (*MBE*)	$MBE = \frac{\sum_{i=1}^{n} (y_i - x_i)}{n}$	(4)

The root mean square error (RMSE) and the mean absolute error (MAE) were estimated, based on the measured value y_i and its prediction x_i in y_i locations of the samples (Equations (1) and (2)). The MAE is the average of the absolute values of the differences between the forecast and the corresponding observation over the verification sample. Since the MAE is a linear score, all individual differences are weighted equally in the average. The RMSE is a quadratic scoring rule that calculates the average magnitude of the error. Because errors are squared before being averaged, the RMSE gives large errors a relatively high weight. As a result, the RMSE is most useful when large errors are especially undesirable. The MAE and RMSE both have a range of 0 to ∞. They're negative scores, thus the lower the number, the better. The coefficient of determination (R^2) (Equation (3)) represents a model's ability to predict or explain an outcome. The R^2 indicates the percentage of variance in the predicted variable and the measured variable where SSE is the sum of squares of errors and SSTO the total sum of squares. The coefficient of determination ranges from 0 to 1, where in 0 (zero) no variation is explained by the model and in 1 (one) all variation is explained by the model. A high R^2 value, in general, implies that the model is a good fit for the data, though fit interpretations vary depending on the context of analysis. Finally, the mean bias error (MBE) was used as a measurement of the bias estimation of the models (Equation (4)).

2.8. Software

For the statistical analysis of the current study, the R (version 4.0.3) statistical software and the caret package were used [50]. Also, the ranger package [51] was utilized for RF and QRF. The geostatistics were implemented with the gstat package [52]. Finally, the Saga-GIS software (https://saga-gis.sourceforge.io/en/index.html (accessed on 18 November 2021)) was used for the environmental indices.

3. Results

3.1. Semivariograms and Fitting Parameters of OK and KED

Initially, OK and KED were implemented using the training dataset for the prediction of the soil OM in the study area. In the case of OK, based on the empirical semivariogram, the Matern semivariogram model with M. Stein's parameterization (Ste) was fitted (Figure 6) using the weighted least square fit of gstat package. The range was 207 m with a nugget at 0.35 and sill at 0.49. There was a moderate spatial dependence based on the nugget to sill ratio (71%).

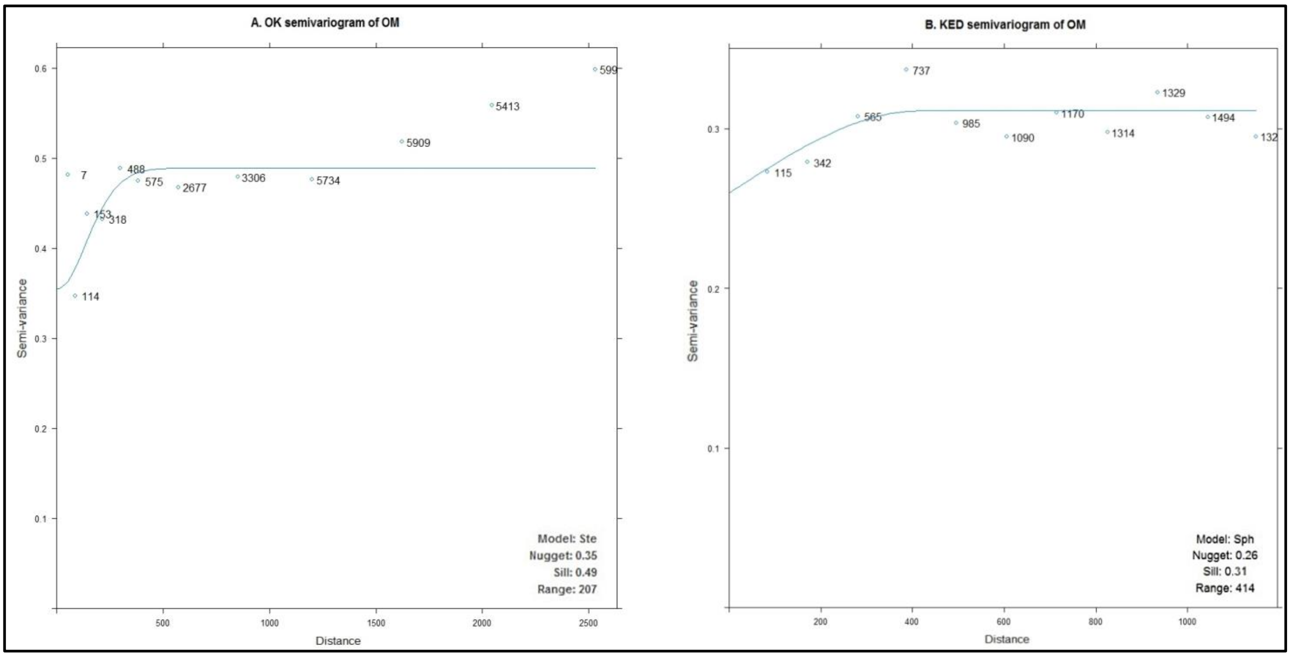

Figure 6. Empirical semivariogram and fitted model of OK and KED.

Regarding KED, the spherical model was used for fitting with the default method of gstat based on weighted least squares fit. The range was at 414 m, double than the OK range. In this case there was a weak spatial dependence with a nugget to total sill ratio of 83%.

3.2. RF and QRF Hyperparameters' Optimization Results

As it is already stated (Section 2.5), ML models need the assessment of their optimal hyperparameters in order to provide the best prediction results. In the case of RF and QRF as defined by the ranger library, four hyperparameters need to be estimated (Table 3). An iterative process (trial and error) was used with the random search optimization method, where different random values of these parameters were introduced from a range of values. The R^2 of the ML models was assessed using a 10-fold cross-validation method that was repeated 3 times in the training data set (Figure 7). The hyperparameters that returned the highest R^2 were finally chosen (Table 5).

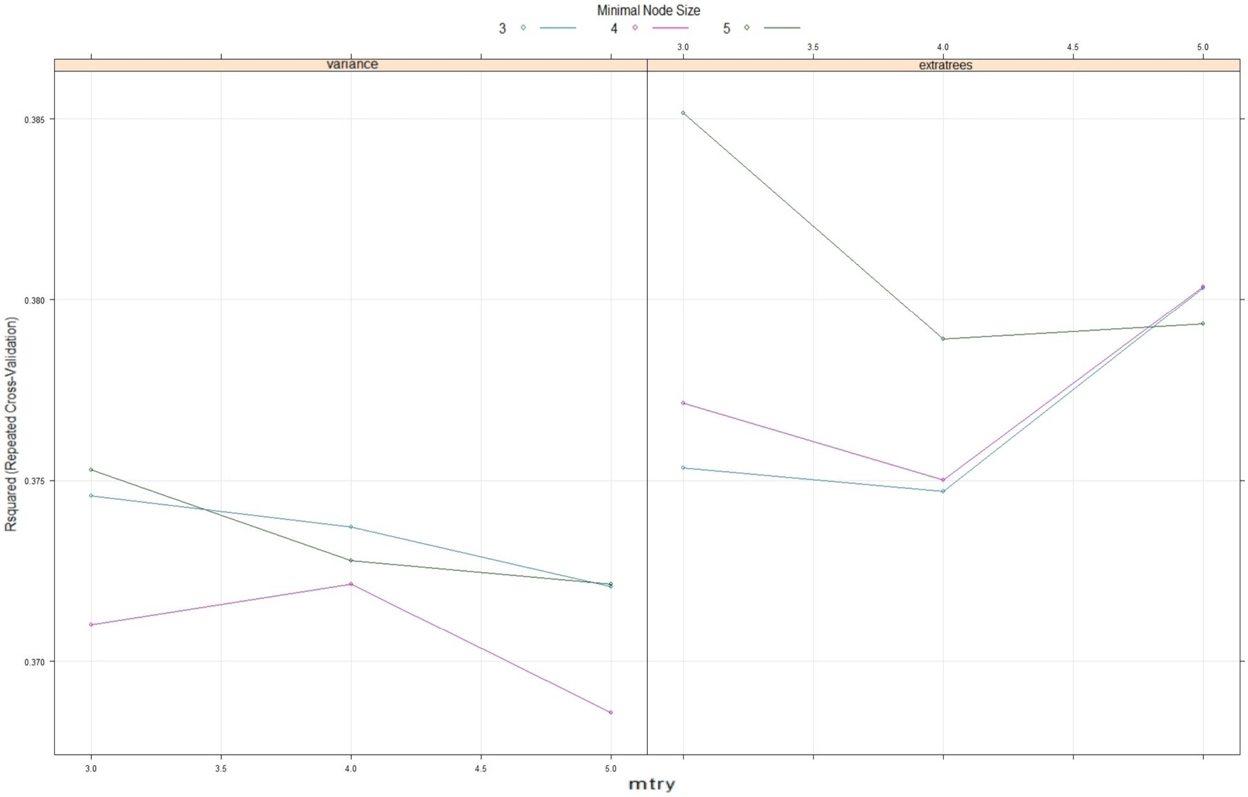

Figure 7. Hyperparameters assessment results based on R^2 in the training dataset.

For the splitrule hyperparameter, only "variance"and "extratrees" methods were used due to unrecoverable errors from "maxstat" and "beta" values.

Table 5. Hyperparameters' range and values used in the current study for RF and QRF.

Parameter	Range	Value Used
mtry	3–5	3
num.trees	500, 1000	500
min.node.size	3–5	5
splitrule	"variance", "extratrees", "maxstat" *, "beta" *	extratrees

* "Maxstat" and "beta" resulted unrecoverable errors and they were not used.

The specific optimal hyperparameters were introduced in the RF and QRF models and used to estimate their prediction capabilities on the testing dataset.

3.3. Feature Importance of the ML Models

The feature importance of the RF and QRF was estimated (Figure 8) with the permutation technique [3], that is defined as the decrease in the model score when a single feature value is randomly shuffled. A feature is "important" if shuffling its values increases the model error (strong effect on the prediction) and "unimportant" if shuffling its values leaves the model error unchanged (low or no effect on the prediction).

Regarding the importance scores, both RF and QRF concede that the soil covariates exhibited the highest importance, something that was expected due to comparable findings of a previous study [28] in a nearby area. In the current study specifically, Zn had the highest score with C second and Mg third. The Altitude from the topographic indices was next, along with the NDVI of 2016 and Vdepth. The last positions were occupied by NDVI of 2017 and NDWI of 2019.

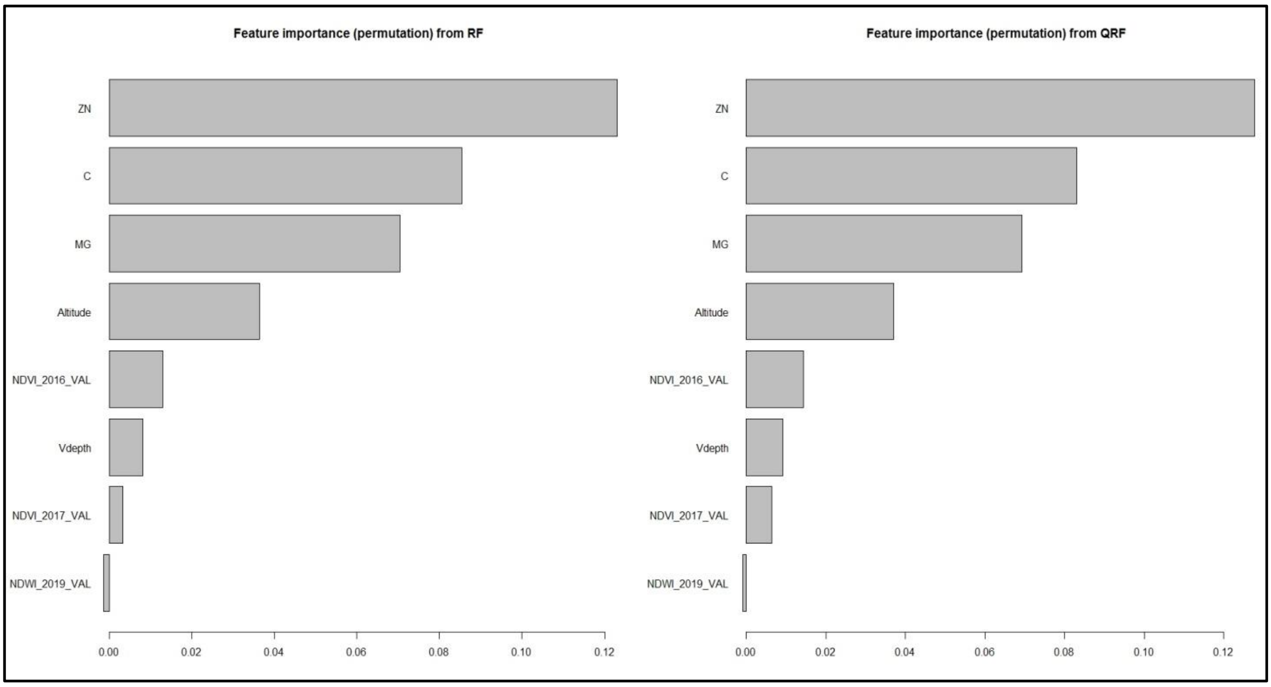

Figure 8. Hyperparameters assessment results based on R^2 in the training dataset.

3.4. Prediction Results

The dataset was partitioned into two random but spatially balanced sets of 70% for training the models and 30% for testing. The difference between the observations of the soil OM and their predictions in the testing data set was used to assess the prediction accuracy of the different models and they are presented in Table 6 and Figure 9.

Table 6. Prediction results for soil OM.

Model	RMSE	R^2	MAE	MBE
Ordinary Kriging (OK)	0.783	0.127	0.586	−0.002
Kriging with external drift (KED)	0.618	0.452	0.455	−0.022
Random Forests (RF)	0.615	0.538	0.453	−0.020
Quantile Regression Forests (QRF)	0.635	0.532	0.459	−0.046

As it is presented in the results (Table 6), OK was the least accurate model with very low R^2 (0.127) and high RMSE and MAE, something that was expected due to its lack of capacity to incorporate auxiliary information. The OK prediction capability is based only on the variable's (OM) spatial autocorrelation, hence the poor current results. It presented the smaller bias though based on the MBE (−0.002).

The KED combines the predictive capabilities of the trend that is based on the auxiliary variables, along with the kriging interpolation. Therefore, the results are decent, with low RMSE (0.618) and MAE (0.455) that are very close to RF and even slightly better than QRF. However, the coefficient of determination (0.452) is much worse than the ML methods. The bias was also small close to zero (−0.022), however higher than the OK.

The ML models exhibited higher prediction capabilities than geostatistical models. More specifically, the best results were achieved by the RF. Especially its R^2 was the highest (0.538) among the models with an improvement of around 20% from KED. Regarding RMSE and MAE, RF's results were best with the lowest values overall. The model bias was low (−0.020) close to KED's value.

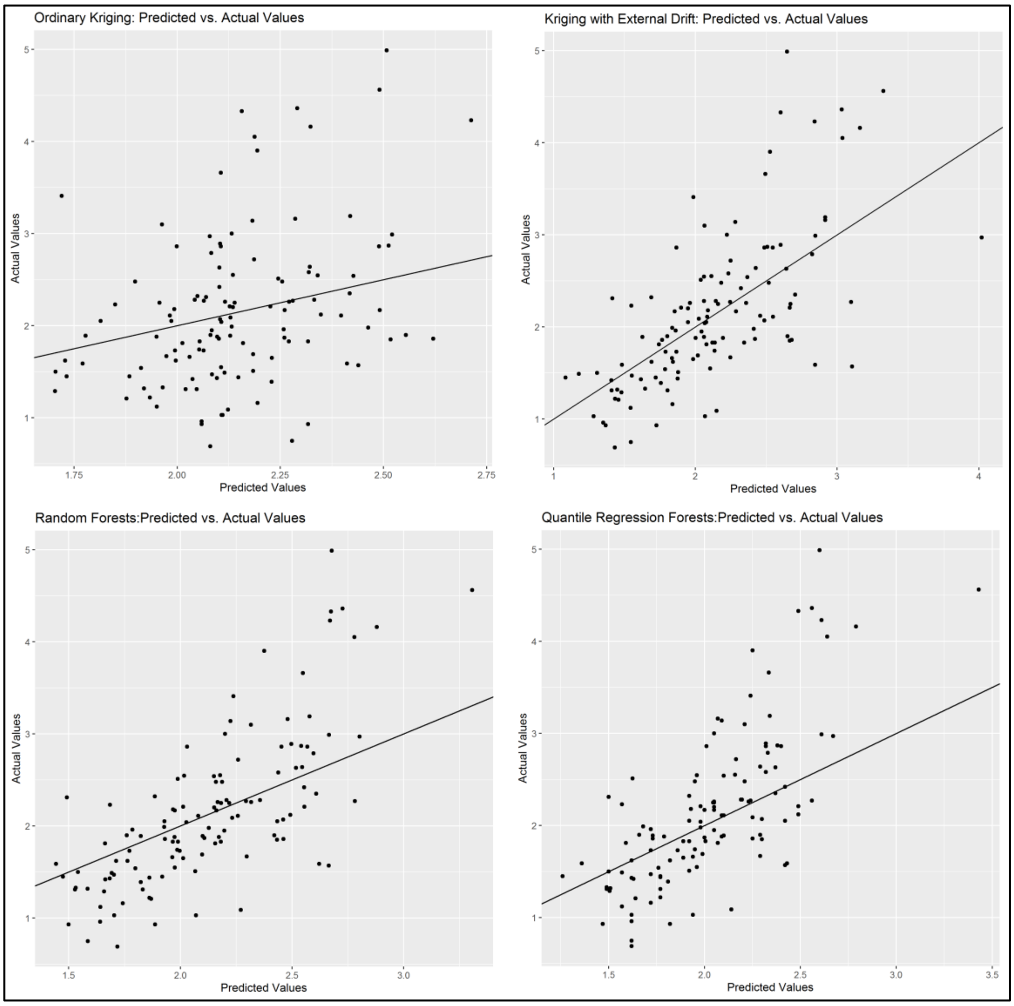

Figure 9. Predicted vs observed scatter plots.

The QRF model also exhibited very good prediction capabilities with high R^2 (0.532) very close to RF and quite low RMSE and MAE, close to RF and KED. The MBE was higher than the other models (−0.046), however still low and close to zero. Thus, RF and QRF can both be used interchangeably for predicting soil OM with similar results in the current study.

3.5. Maps of Prediction and Uncertainty

Next, two sets of maps were produced from the different models of the current study. The first set consists of four prediction maps that present the spatial distribution of soil OM in the area, one for each method: OK, KED, RF and QRF (Figure 10). The second set of maps consist of three maps with OK, KED and QRF methods that depict the spatial distribution of prediction's uncertainty in the area (Figure 11). In this case the RF was not used due to its lack of uncertainty capabilities.

The prediction map of OK exhibited interpolation results with prediction patterns that are relatively uniform all over the study area. The main reason for this is that OK is based solely on the spatial autocorrelation of OM using global model parameters that smooths the results. The KED model produces a map that changes more abruptly due to its covariates' effect, leading to multiple areas with higher and lower local values than the OK. Regarding ML methods (RF, QRF) their maps had even more contrast than the OK and KED, due to their capability of producing patterns that match data as much as possible by better fitting to the dataset. Among them, the QRF seem to present slightly more abrupt patterns than RF with areas with slightly lower and higher values.

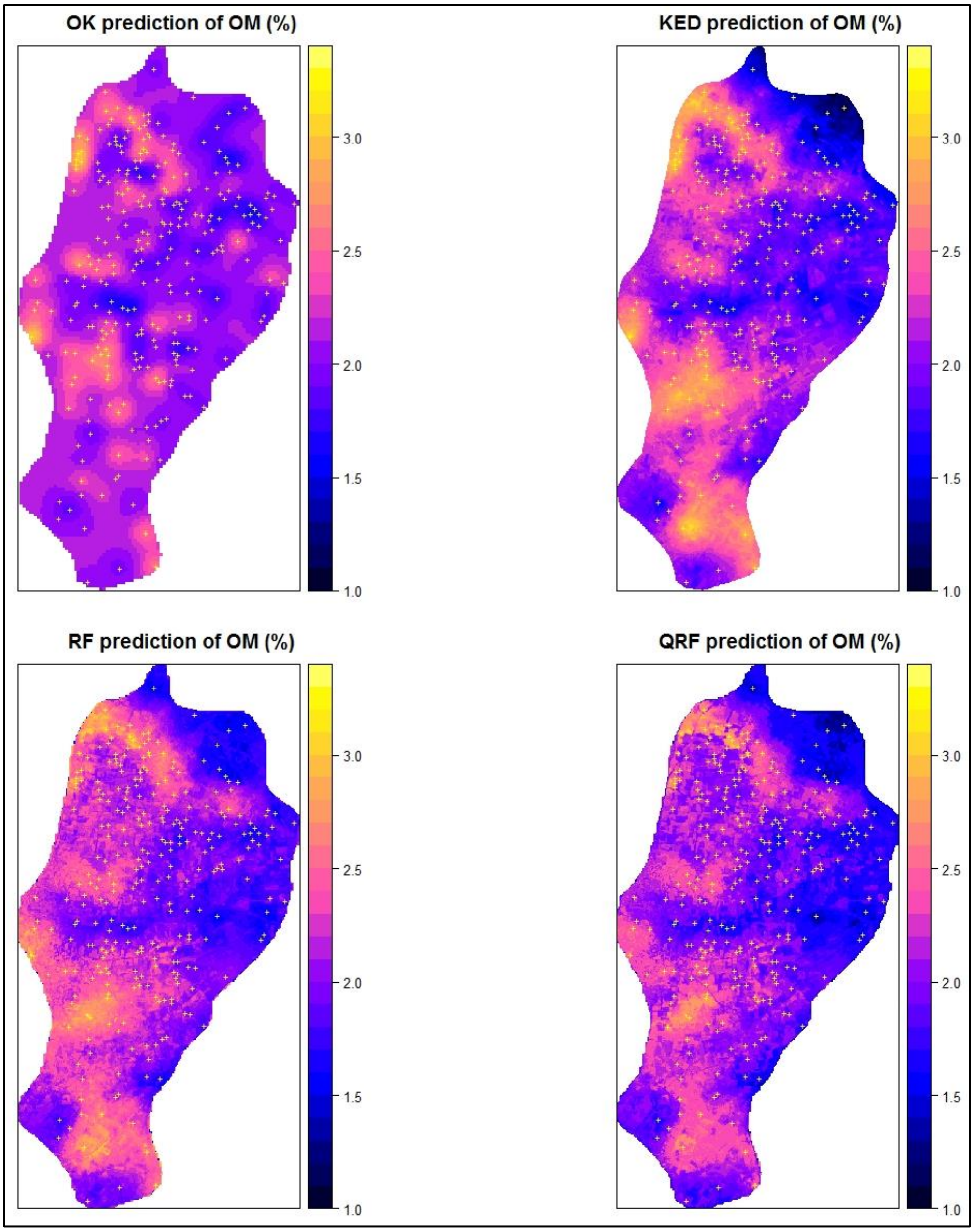

Figure 10. Prediction maps of soil OM.

As it was already mentioned, apart from the prediction results, prediction's uncertainty is a crucial parameter that needs to be estimated in the different locations of the study area. In the current study the uncertainty maps were calculated for only the 3 out of 4 methods. The RF does not support uncertainty estimation. The OK and KED intrinsically provide the error variance by which the standard deviation was calculated and its range was presented in the maps. For QRF the range was defined as one standard deviation above and below the median and it was used to create the uncertainty map of the study area (Figure 11).

Based on the uncertainty maps, it is obvious that OK has a smooth and equally distributed uncertainty range in the area with a mean value approximately at 0.8%. So, in each location the real OM value is approximately ±0.4% above or below the predicted value.

The KED uncertainty map has an overall lower uncertainty range than the OK (around 0.6%), that is almost equally distributed in the overall study area, similarly to OK. There are some slightly increased range values in the northern-west area along with some small patches in between (lighter blue areas) due to the covariates minor effect on the uncertainty results.

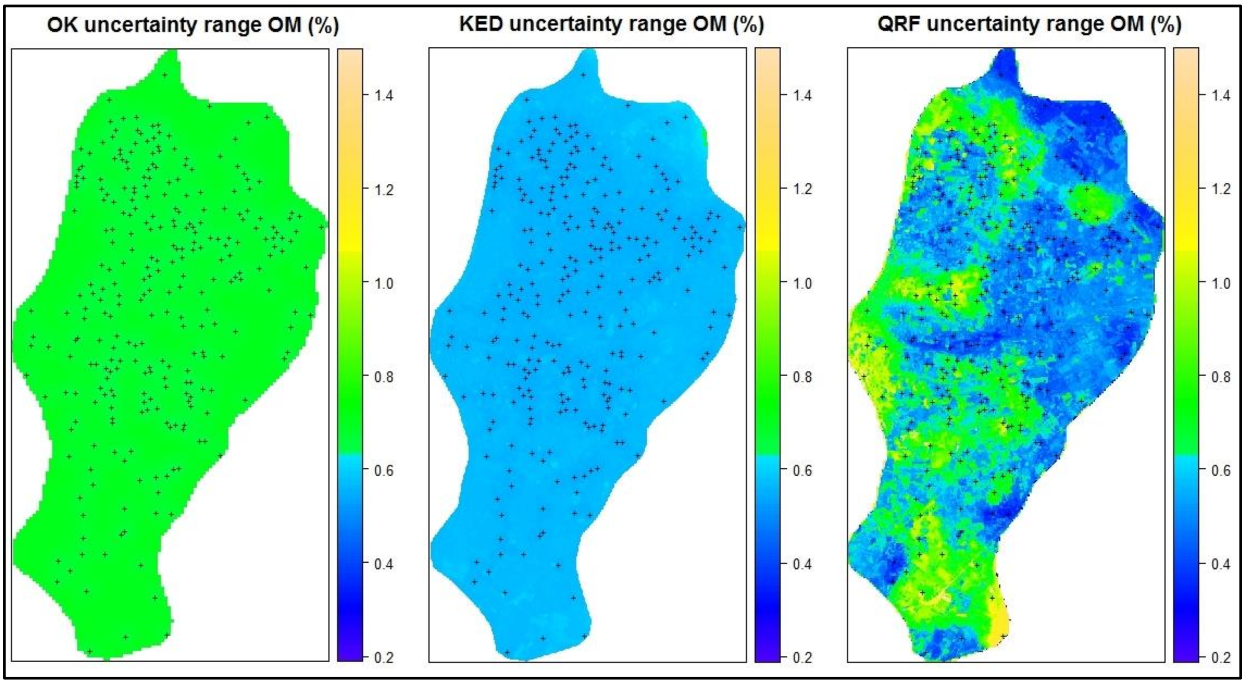
OK uncertainty range OM (%)
KED uncertainty range OM (%)
QRF uncertainty range OM (%)
1.4
1.2
1.0
0.8
0.6
0.4
0.2

Figure 11. Uncertainty maps of soil OM.

In the case of QRF, the uncertainty map is more diversified than the previous ones. There are distinct regions of very low uncertainty like the one in the north or in the center of the area (with dark blue color) and regions of higher uncertainty like the ones in the south or close to the lake (with yellow color). This clear depiction of the uncertainty in a local scale and the straightforward definition of possible uncertainty zones, is a major advantage over the geostatistical methods especially for decision support purposes.

4. Discussion

One of the core tasks in the DSM studies is the estimation and presentation of the spatial distribution of different soil variables in the study area using different interpolation methods. Apart from that though, estimating and presenting the uncertainty of these interpolation methods are equally important in order to assess the overall work, something that is lacking in some of the recent DSM studies, especially the ones that are based on ML.

The ML methods are increasingly used in DSM, based on their outstanding prediction capabilities that outperforms the classic geostatistical methods, without the drawbacks of statistical assumptions and the restrictions of other methods. However, most of them do not have intrinsic uncertainty estimation capabilities. The RF is a very promising ML method used in multiple DSM studies that nevertheless lacks built-in uncertainty estimation capacity. An interesting alternative is QRF that seems to provide advanced prediction capabilities similar to RF along with innate uncertainty estimation metrics.

In the current paper, it was confirmed that QRF exhibited outstanding results at predicting soil OM in the study area, very close to RF method. Especially R^2 was much higher than the geostatistical methods, something that signifies that more variation is explained by the specific model. Moreover, its uncertainty capabilities as presented in the uncertainty maps, shows that it can also provide a very efficient estimation of the uncertainty in the study area. Uncertainty map with QRF exhibit stronger contrast compared to uncertainty maps of OK, and KED, with distinct representation of the local variation of the uncertainty such as small regions with higher or lower uncertainty. Based on this map it is very easy for a user to define clusters of uncertainty zones and categorize its effect locally. This is a real significant advantage, especially for decision support purposes in which users are interested not only in the prediction accuracy but also in the variation of the error range in different parts of the area.

Author Contributions: Conceptualization, Panagiotis Tziachris; Methodology, Panagiotis Tziachris; Software, Panagiotis Tziachris and Melpomeni Nikou; Data Curation, Melpomeni Nikou; Writing—Original Draft Preparation, Panagiotis Tziachris and Melpomeni Nikou; Writing—Review and Editing, Panagiotis Tziachris and Melpomeni Nikou. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to the staff of the Soil and Water Resources Institute of Hellenic Agricultural Organization—Demeter in Thessaloniki, Greece for their support in this study (I. Metaxa, A. Bountla, K. Oikonomou, G. Strikos, C. Tsotras and I. Palavatsios).

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

References

1. Wadoux, A.M.J.C.; Minasny, B.; McBratney, A.B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Sci. Rev.* **2020**, *210*, 103359. [CrossRef]
2. Lagacherie, P. *Digital Soil Mapping: A State of the Art*; Digital Soil Mapping with Limited Data; Springer: Berlin, Germany, 2008; pp. 3–14. [CrossRef]
3. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
4. Kam, T.H. Random Decision Forests Tin Kam Ho Perceptron training. *Proc. 3rd Int. Conf. Doc. Anal. Recognit.* **1995**, *1*, 278–282.
5. Wiesmeier, M.; Barthold, F.; Blank, B.; Kögel-Knabner, I. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant Soil* **2011**, *340*, 7–24. [CrossRef]
6. Liu, J.; Dong, Z.; Xia, J.; Wang, H.; Meng, T.; Zhang, R.; Han, J.; Wang, N.; Xie, J. Estimation of soil organic matter content based on CARS algorithm coupled with random forest. *Spectrochim. Acta-Part A Mol. Biomol. Spectrosc.* **2021**, *258*, 119823. [CrossRef]
7. John, K.; Isong, I.A.; Kebonye, N.M.; Ayito, E.O.; Agyeman, P.C.; Afu, S.M. 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]
8. Stumpf, F.; Schmidt, K.; Goebes, P.; Behrens, T.; Schönbrodt-Stitt, S.; Wadoux, A.; Xiang, W.; Scholten, T. Uncertainty-guided sampling to improve digital soil maps. *Catena* **2017**, *153*, 30–38. [CrossRef]
9. Meinshausen, N.; Ridgeway, G. Quantile regression forests. *J. Mach. Learn. Res.* **2006**, *7*, 983–999.
10. Freeman, E.A.; Moisen, G.G. An application of quantile random forests for predictive mapping of forest attributes. *For. Invent. Anal. Symp.* **2015**, *931*, 362.
11. Vaysse, K.; Lagacherie, P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* **2017**, *291*, 55–64. [CrossRef]
12. Dharumarajan, S.; Vasundhara, R.; Suputhra, A.; Lalitha, M.; Hegde, R. Prediction of Soil Depth in Karnataka Using Digital Soil Mapping Approach. *J. Indian Soc. Remote Sens.* **2020**, *48*, 1593–1600. [CrossRef]
13. Poggio, L.; De Sousa, L.M.; Batjes, N.H.; Heuvelink, G.B.M.; 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]
14. Veronesi, F.; Schillaci, C. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. *Ecol. Indic.* **2019**, *101*, 1032–1044. [CrossRef]
15. Bouyoucos, G.J. Directions for Making Mechanical Analyses of Soils by the Hydrometer Method. *Soil Sci.* **1936**, *42*, 225–230. [CrossRef]
16. Lindsay, W.L.; Norvell, W.A. Development of a DTPA Soil Test for Zinc, Iron, Manganese, and Copper. *Soil Sci. Soc. Am. J.* **1978**, *42*, 421–428. [CrossRef]
17. Suwandana, E.; Kawamura, K.; Sakuno, Y.; Kustiyanto, E.; Raharjo, B. Evaluation of aster GDEM2 in comparison with GDEM1, SRTM DEM and topographic-map-derived DEM using inundation area analysis and RTK-DGPS data. *Remote Sens.* **2012**, *4*, 2419–2431. [CrossRef]
18. del Rosario Gonzalez-Moradas, M.; Viveen, W. Evaluation of ASTER GDEM2, SRTMv3.0, ALOS AW3D30 and TanDEM-X DEMs for the Peruvian Andes against highly accurate GNSS ground control points and geomorphological-hydrological metrics. *Remote Sens. Environ.* **2020**, *237*, 111509. [CrossRef]
19. Roy, D.P.; Li, Z.; Zhang, H.K. Adjustment of sentinel-2 multi-spectral instrument (MSI) red-edge band reflectance to nadir BRDF adjusted reflectance (NBAR) and quantification of red-edge band BRDF effects. *Remote Sens.* **2017**, *9*, 1325. [CrossRef]
20. Hill, M.J. Vegetation index suites as indicators of vegetation state in grassland and savanna: An analysis with simulated SENTINEL 2 data for a North American transect. *Remote Sens. Environ.* **2013**, *137*, 94–111. [CrossRef]

21. Cambardella, C.A.; Moorman, T.B.; Novak, J.M.; Parkin, T.B.; Karlen, D.L.; Turco, R.F.; Konopka, A.E. Field-Scale Variability of Soil Properties in Central Iowa Soils. *Soil Sci. Soc. Am. J.* **1994**, *58*, 1501–1511. [CrossRef]
22. Peng, X.; Wang, K.; Li, Q. A new power mapping method based on ordinary kriging and determination of optimal detector location strategy. *Ann. Nucl. Energy* **2014**, *68*, 118–123. [CrossRef]
23. Wackernagel, H. *Multivariate Geostatistics*; Springer: Berlin, Germany, 1998; ISBN 9783662035528.
24. McBratney, A.B.; Odeh, I.O.A.; Bishop, T.F.A.; Dunbar, M.S.; Shatar, T.M. An Overview of Pedometric Techniques for Use in Soil Survey. *Geoderma* **2000**, *97*, 293–327. [CrossRef]
25. Ignaccolo, R.; Mateu, J.; Giraldo, R. Kriging with external drift for functional data for air quality monitoring. *Stoch. Environ. Res. Risk Assess.* **2014**, *28*, 1171–1186. [CrossRef]
26. Webster, R.; Oliver, M.A. *Geostatistics for Environmental Scientists*; Wiley: Chichester, UK, 2001; ISBN 0471965537.
27. Dietterichl, T.G. Ensemble Learning. In *The Handbook of Brain Theory and Neural Networks*; Arbib, M., Ed.; MIT Press: Cambridge, MA, USA, 2002; pp. 405–408.
28. Tziachris, P.; Aschonitis, V.; Chatzistathis, T.; Papadopoulou, M. Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. *Catena* **2019**, *174*, 206–216. [CrossRef]
29. Dharumarajan, S.; Hegde, R.; Singh, S.K. Spatial prediction of major soil properties using Random Forest techniques-A case study in semi-arid tropics of South India. *Geoderma Reg.* **2017**, *10*, 154–162. [CrossRef]
30. Wang, D.; Zhu, A.X. Soil mapping based on the integration of the similarity-based approach and random forests. *Land* **2020**, *9*, 174. [CrossRef]
31. Stum, A.K.; Boettinger, J.L.; White, M.A.; Ramsey, R.D. Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah. *Digit. Soil Mapp.* **2010**, 179–190. [CrossRef]
32. Fernandes, D.; Machado, T.; Silva, H.G.; Curi, N.; Duarte De Menezes, M. Soil type spatial prediction from Random Forest: Different training datasets. *Sci. Agric.* **2019**, *76*, 243–254.
33. Shukla, G.; Garg, R.D.; Srivastava, H.S.; Garg, P.K. An effective implementation and assessment of a random forest classifier as a soil spatial predictive model. *Int. J. Remote Sens.* **2018**, *39*, 2637–2669. [CrossRef]
34. Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **2020**, *415*, 295–316. [CrossRef]
35. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
36. MacMillan, R.A. Experiences with Applied DSM: Protocol, Availability, Quality and Capacity Building BT-Digital Soil Mapping with Limited Data. In *Digital Soil Mapping with Limited Data*; Hartemink, A.E., McBratney, A., Mendonça-Santos, M., Eds.; Springer: Dordrecht, The Netherlands, 2008; pp. 113–135; ISBN 978-1-4020-8592-5.
37. Scull, P.; Franklin, J.; Chadwick, O.A.; McArthur, D. Predictive soil mapping: A review. *Prog. Phys. Geogr.* **2003**, *27*, 171–197. [CrossRef]
38. Heuvelink, G.B.M. Identification of field attribute error under different models of spatial variation. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 921–935. [CrossRef]
39. Kempen, B.; Brus, D.J.; Stoorvogel, J.J.; Heuvelink, G.B.M.; de Vries, F. Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. *Soil Sci. Soc. Am. J.* **2012**, *76*, 2097–2115. [CrossRef]
40. Arrouays, D.; McKenzie, N.; Hempel, J.; de Forges, A.R.; McBratney, A.B. *GlobalSoilMap: Basis of the Global Spatial Soil Information System*; CRC Press: Boca Raton, FL, USA, 2014.
41. Minasny, B.; McBratney, A.B. Uncertainty analysis for pedotransfer functions. *Eur. J. Soil Sci.* **2002**, *53*, 417–429. [CrossRef]
42. Nelson, M.A.; Bishop, T.F.A.; Triantafilis, J.; Odeh, I.O.A. An error budget for different sources of error in digital soil mapping. *Eur. J. Soil Sci.* **2011**, *62*, 417–430. [CrossRef]
43. Kasraei, B.; Heung, B.; Saurette, D.D.; Schmidt, M.G.; Bulmer, C.E.; Bethel, W. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. *Environ. Model. Softw.* **2021**, *144*, 105139. [CrossRef]
44. Malone, B.P.; McBratney, A.B.; Minasny, B. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* **2011**, *160*, 614–626. [CrossRef]
45. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]
46. Feizizadeh, B.; Jankowski, P.; Blaschke, T. A GIS based spatially-explicit sensitivity and uncertainty analysis approach for multi-criteria decision analysis. *Comput. Geosci.* **2014**, *64*, 81–95. [CrossRef]
47. Hengl, T.; Toomanian, N. Maps are not what they seem: Representing uncertainty in soilproperty maps. In Proceedings of the Accuracy 2006: 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisbon, Portugal, 5–7 July 2006; pp. 805–813.
48. Goovaerts, P. Geostatistical modelling of uncertainty in soil science. *Geoderma* **2001**, *103*, 3–26. [CrossRef]
49. Heuvelink, G.B.M. Uncertainty quantification of globalsoilmap products. Basis of the global spatial soil information, system. In Proceedings of the First GlobalSoilMap Conference, Orleans, France, 7–9 January 2014; pp. 335–340. [CrossRef]
50. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
51. Wright, M.N.; Ziegler, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
52. Pebesma, E.J. Multivariable geostatistics in S: The gstat package. *Comput. Geosci.* **2004**, *30*, 683–691. [CrossRef]