land

Article

Prediction of Soil Organic Carbon at Field Scale by Regression Kriging and Multivariate Adaptive Regression Splines Using Geophysical Covariates

Daniela De Benedetto [1], Emanuele Barca [2,*], Mirko Castellini [1], Stefano Popolizio [3], Giovanni Lacolla [4] and Anna Maria Stellacci [3]

[1] Council for Agricultural Research and Economics-Agriculture and Environment Research Center (CREA-AA), 70126 Bari, Italy; daniela.debenedetto@crea.gov.it (D.D.B.); mirko.castellini@crea.gov.it (M.C.)
[2] Water Research Institute (IRSA)—National Research Council (CNR), 70185 Bari, Italy
[3] Department of Soil, Plant and Food Sciences, University of Bari "A. Moro", 70126 Bari, Italy; stefano.popolizio@uniba.it (S.P.); annamaria.stellacci@uniba.it (A.M.S.)
[4] Department of Agricultural and Environmental Science, University of Bari "A. Moro", 70126 Bari, Italy; giovanni.lacolla@uniba.it
* Correspondence: emanuele.barca@ba.irsa.cnr.it

Abstract: Knowledge of the spatial distribution of soil organic carbon (SOC) is of crucial importance for improving crop productivity and assessing the effect of agronomic management strategies on crop response and soil quality. Incorporating secondary variables correlated to SOC allows using information often available at finer spatial resolution, such as proximal and remote sensing data, and improving prediction accuracy. In this study, two nonstationary interpolation methods were used to predict SOC, namely, regression kriging (RK) and multivariate adaptive regression splines (MARS), using as secondary variables electromagnetic induction (EMI) and ground-penetrating radar (GPR) data. Two GPR covariates, representing two soil layers at different depths, and X geographical coordinates were selected by both methods with similar variable importance. Unlike the linear model of RK, the MARS model also selected one EMI covariate. This result can be attributed to the intrinsic capability of MARS to intercept the interactions among variables and highlight nonlinear features underlying the data. The results indicated a larger contribution of GPR than of EMI data due to the different resolution of EMI from that of GPR. Thus, MARS coupled with geophysical data is recommended for prediction of SOC, pointing out the need to improve soil management to guarantee agricultural land sustainability.

Keywords: SOC spatial distribution; regression kriging (RK); multivariate adaptive regression splines (MARS); secondary variables; electromagnetic induction technique (EMI); ground-penetrating radar (GPR)

check for updates

Citation: De Benedetto, D.; Barca, E.; Castellini, M.; Popolizio, S.; Lacolla, G.; Stellacci, A.M. Prediction of Soil Organic Carbon at Field Scale by Regression Kriging and Multivariate Adaptive Regression Splines Using Geophysical Covariates. *Land* **2022**, *11*, 381. https://doi.org/10.3390/land11030381

Academic Editor: Manuel López-Vicente

Received: 18 January 2022
Accepted: 2 March 2022
Published: 4 March 2022

1. Introduction

Soil organic carbon (SOC) is one of the most important indicators for assessing soil quality and overall soil health [1]. SOC plays a key role in unveiling soil structure development, nutrient turnover and stability, soil water retention, regulation of greenhouse gases, and susceptibility or resilience to land degradation [2]. SOC stock is thus a main factor in soil health, fertility, quality, and productivity [3] and supports important soil-derived ecosystem services (ESs) including water filtration and erosion control, soil strength and stability, nutrient conservation, and climate change adaptation and mitigation by sequestration of atmospheric CO_2 [4]. By selecting key soil indicators under different land use and management practices, Shukla et al. [5] concluded that SOC was the main soil quality indicator and suggested using SOC to monitor soil quality changes [6].

SOC distribution is influenced by many factors, including climate variables (temperature and rainfall), topographical features, soil texture, parent material, vegetation, land-use types, and human management at different spatial scales [7].

Agronomic management strategies, with particular regard to fertilization, soil tillage, and irrigation, may significantly modify SOC content and its labile fractions, mainly in the shallower soil layers [6,8–10]. Because of the interaction of the factors described, SOC spatial variation is often wide and complex, and the knowledge of its spatial distribution is the key information in agricultural productivity to improve food security, enhance crop production [11], and predict the effects of different agronomic management strategies. Among these strategies, irrigation with treated municipal wastewater can be considered important for saving limited freshwater resources and protecting the environment, but its effects should be monitored to avoid soil fertility decline in the medium to long term [9].

Conventional laboratory methods for quantifying this soil variable are destructive, time consuming, expensive, and hazardous for the environment. In addition, because of the associated costs, soil is sampled at relatively few spatial locations, which are often irregularly distributed over the study area. The small sample size does not allow meeting the criteria for soil quality assessment for precision farming or for using statistical methods taking into account residual autocorrelation [12]. Making a short review, a number of samples ranging from 50 [13] to 100 [14] is considered well suited for an accurate spatial analysis.

A strategy to enhance the quality of the estimation of SOC content and to reduce the spatial sampling intensity consists of incorporating secondary information correlated to the primary variable [15,16]. This multivariate approach allows utilization of secondary information, such as that derived from proximally and remotely sensed data, that is often much more abundant than information deriving from the primary target variable [17,18].

Proximal sensing data could provide strong support for characterizing the spatial variability at the field or even regional scale. These data are very attractive because of their high resolution, their noninvasive nature, the relatively low cost of data acquisition, the possibility for a mobile survey configuration, and their three-dimensional (3D) information, although their outcome is not a direct measurement of soil properties [19].

Among the geophysical methods, electromagnetic induction (EMI) and ground-penetrating radar (GPR) have been widely applied. EMI methods measure apparent electrical conductivity (EC_a), an integrated value of soil physical, chemical, and biological properties [20] that can capture soil spatial variability and characterize soil organic carbon distribution [21,22]. However, since soil properties vary in both the horizontal and vertical domains, soil needs to be described in three dimensions, and EMI sensors may have limitations when highly contrasting horizons are present [23]. Ground-penetrating radar (GPR) technology allows overcoming this limitation by measuring large volumes of soil (about cubic decimetres to cubic meters). Thus, GPR is suggested for field-scale determinations rather than for pointwise measurements, provides higher resolution of subsurface features, and is particularly suited to visualizing soil in two or three dimensions [24]. One of the most useful presentations of GPR data is to display horizontal maps of recorded reflection amplitudes, called "time slice" (or depth slice) maps [25]. There have been several studies involving GPR to determine thickness and characterize depths of organic soil materials [26,27], but few studies have been devoted so far to the potentiality of GPR to study the spatial variation of soil organic carbon.

The use of geophysical proximal sensor data as auxiliary information to effectively support an irregularly sampled target variable is not free from practical difficulties and experimental limits. This is because proximal sensing data are often massive, need to be collected on different spatial and temporal scales, and use different measurement supports. Several statistical methods are able to incorporate secondary information; for example, a multivariate extension of kriging, known as cokriging, is used for improving the prediction of a primary variable by using secondary information [28,29]. This technique assumes intrinsic stationarity, both of the target variables and of more intensively measured secondary variables, supposing a strong correlation between primary and secondary information [30].

These conditions are not always verified. Another way of taking into account the secondary variable is by checking for a spatial trend in the primary variable with respect to the secondary variable(s) and combining the deterministic part and the stochastic component, as in "hybrid methods" [29], or by adopting complex multivariate nonlinear approaches. In recent times, a number of hybrid interpolation techniques, which combine kriging with methods that use auxiliary information (covariates), have been developed and applied. Several authors have compared some of the techniques to incorporate trends and account for nonstationarity [31,32]. Two possible methods of nonstationary interpolation are regression kriging (RK) [28,33] and multivariate adaptive regression splines (MARS) [34]. In many cases, these techniques have been proven superior to common geostatistical methods, yielding more detailed results and higher accuracy of prediction, because they take advantage of being linear hybrid (RK) or nonlinear (MARS) [35]. MARS is a nonparametric predictive method that intrinsically models nonlinearities and interactions between variables, suitably managing local nonstationarity [34]. This method has been successfully applied in various fields, such as estimating the collapse potential for compacted soil, underground gas storage in bedded salt formations, and lateral spreading induced by earthquakes [36,37].

The regression kriging (RK) method is of straightforward use and often performs better than cokriging [38–40].

In this study, we compared the performance of RK and MARS to achieve the following objectives: (i) to prove that there are preferential nonlinear relationships between SOC and geophysical measurements, and (ii) to compare the performance of two nonstationary interpolation methods to effectively model SOC at the field scale. Machine learning techniques may open new perspectives to modelling SOC spatial distribution at the field and regional scales. The study was performed on a dataset deriving from a field experiment in which water of different qualities was used for irrigation.

To the best of authors' knowledge, no comparison between these methods has been presented before; therefore, it can be considered a novelty.

2. Materials and Methods

2.1. Study Area

Soil data were derived from a field experiment carried out in an olive grove located in Fasano (Apulia region, Southern Italy). The climate of the study area is "accentuated thermo-Mediterranean", as classified by UNESCO FAO [41,42], characterized by rather mild and rainy winters and warm and dry summer months. The soil of the experimental site is classified as loam (USDA classification), with an average content of silt, clay, and sand fractions of 35.28%, 21.74%, and 42.98%, respectively.

Olive trees were irrigated with treated municipal wastewater (TWW), and the following treatments were applied: irrigation with fresh water and full fertilization supply (FW); irrigation with TWW and full fertilization supply (R1); and irrigation with TWW and fertilizer supply reduced by the amount provided by TWW (R2) [10]. Treatments were arranged in a randomized complete block design (RCBD) with four replicates (Figure 1). Unit plot size was 108 m^2, with 3 plants per plot and a plant spacing of 6 m × 6 m; field size was 1296 m^2 (whole experimental area was 1728 m^2).

2.2. Soil Sampling and Soil Analysis

Soil samples with absolute coordinates were collected on a regular grid (April 2017) at 6 locations (subreplicates) per plot at a 0–0.20 m depth for a total of 72 observations (Figure 1); only 71 were used in this study. Soil organic carbon (SOC) was quantified on air-dried and sieved samples through dry combustion [43]. Further details about the experimental trial were reported by Barca et al. [44] and Stellacci et al. [10].

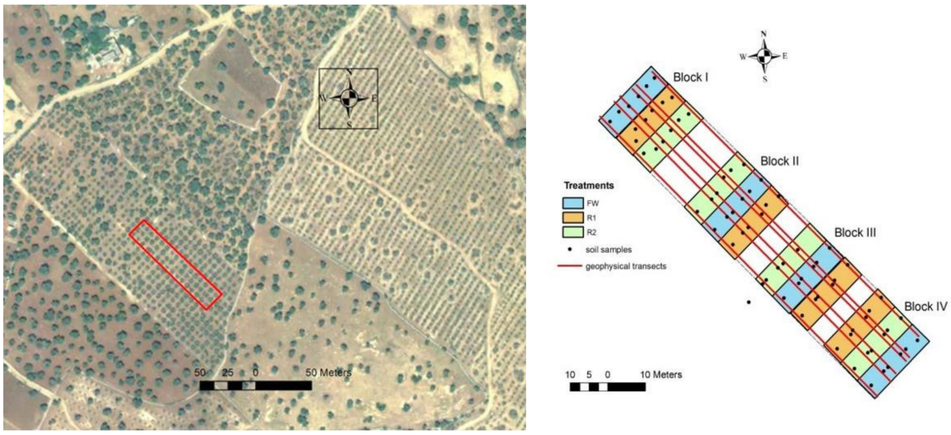

Figure 1. Location of the field experiment (Google Earth Pro, 2021) the soil sampling locations (black dots), and the electromagnetic induction (EMI) and ground-penetrating radar (GPR) acquisitions along transects (red lines).

2.3. Acquisition and Preprocessing of Auxiliary Information

A geophysical survey was carried out using an EMI sensor (EM38DD, Geonics Limited, Mississauga, ON, Canada) and a Georadar (RIS 2k-MF Multifrequency Array Radar-System, manufactured by IDS SpA, Italy) connected to the DGPS along 6 parallel transects by sliding the sensors on the surface (Figure 1) on the same day as soil sampling.

EMI soil survey is based on the principle that a transmitter coil in contact with the soil surface produces a time-varying primary magnetic field in the subsoil. The eddy currents induced in the soil generate a secondary magnetic field, which is recorded by a receiver coil in the EM unit. The apparent conductivity near the receiver is determined by the ratio of the magnitude of the secondary magnetic field to that of the primary magnetic field [22]. The EMI sensor used herein consisted of two perpendicularly superposed EM38 sensors that simultaneously measured apparent electrical conductivity (EC_a, expressed in mSm^{-1}) near the soil surface (0–0.75 m depth) with the horizontal mode (EC_a-H) and up to 1.5 m depth with the vertical mode (EC_a-V) [22]. Before operation, the instrument was set to zero at a height of 1.5 m, according to the manufacturer's instructions, and at the end of the survey, the zeroing was checked to detect possible drift. The survey was performed using a nonmetallic platform with wood cover, and the sensor was towed behind a tractor/The EC_a was recorded every second, with spatial resolution of 0.5 m, on average, along each transect.

Immediately after the EMI survey, the GPR survey was carried out by sliding the sensor along the surface. GPR data were collected with the common offset reflection method, using a monostatic system (the transmitting and receiving antenna placed in the same box) with two central frequencies of 600 and 1600 MHz (IDS Ing-manufactured, RIS 2k-MF Multifrequency Array Radar-System). The GPR worked with a time window of 60 ns and a temporal sampling interval of 0.05 ns; successive traces were collected every 0.024 m. GPR used electromagnetic pulse energy in the frequency range of 10 MHz to 1000 MHz. The transmitter component of the GPR system allowed the passage of generated pulse energy, which propagated through the subsurface materials, and the interactions with the material were sensed by the receiver component. Traditional surveys employ reflections of electromagnetic waves from boundaries between environments of different electromagnetic properties [45]. Theoretical aspects and working principles of radar components can be found in detail in Davis and Annan [46].

Both the data quality check and cleaning procedure characterized the preliminary data analysis. For EMI data, the points at which the instrument was stationary and any negative values were removed.

Processing the raw GPR data consisted of extracting quantifiable variables, such as attenuation, and displaying GPR data in horizontal maps at a specified time (or depth), called amplitude maps or time slices. The preprocessing of GPR signal amplitude data included the application of a set of filters [47] and the extraction of quantifiable variables.

The enveloped amplitude maps (time slices) were built by averaging the amplitude (or the square amplitude) of the radar signal, expressed in digital number (DN), within overlapping time windows of width Δt equal to the order of the dominant period of each antenna (2 and 1 ns for the 600 and 1600 MHz antennas, respectively). The total time interval was of 10 ns for the 600 MHz antenna because this time was comparable with the depth of the soil, and it was 6 ns for 1600 MHz because of the attenuation of radar signal. The time slices were then transformed in depth slices using the velocity of the radar waves determined through the analysis of hyperbolae [48]. Data preprocessing was performed with ReflexW Software [49].

In order to estimate the geophysical covariates at the same locations as the SOC measurements, geostatistical procedures were separately applied to EMI and GPR data by using a multivariate approach and fitting a linear model of coregionalization (LMC) to the experimental variograms. Each group of geophysical data was interpolated with ordinary cokriging (ck) on a 0.5 m × 0.5 m grid. The estimated covariates, migrated at the sample locations, were: the EC_a in horizontal (EC_aH) and vertical (EC_aV) modes; the amplitude for the 600 MHz antenna at ten depths from 0.05 m to 0.50 m with a step of 0.05 m (Amp600MHz_0.05 m-Amp600MHz_0.50 m); and the amplitude for 1600 MHz frequency antenna at eleven depths from 0.025 m to 0.275 m with a step 0.025 m (Amp1600MHz_0.025 m-Amp1600MHz_0.275 m).

Finally, 25 covariates were considered, namely, the 23 geophysical covariates plus the (two) geographical coordinates expressed in the WGS84 coordinate system.

2.4. Regression Kriging (Residual Kriging)

In the present paper, kriging combined with linear regression (RK), a hybrid interpolation technique, was applied [35,39] (see Figure 2). In mathematical terms, RK can be described as the sum of a deterministic (regression) component and kriging as shown in the following equation:

$$\hat{z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) = \sum_{k=1}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{N} \lambda_i \cdot e(s_i) \qquad (1)$$

where s_0 is the spatial location associated with the desired prediction, $\hat{m}(s_0)$ is the trend, $\hat{e}(s_0)$ is the interpolated residual, $\hat{\beta}_k$ are the estimated regressive coefficients, q_k are the covariates, p is the number of coviariates, λ_i are kriging weights, N is the number of observations, and $e(s_i)$ is the residual (i.e., the difference between the regression estimation minus the observation) at the generic observational location s_i.

From a practical standpoint, once the trend component has been estimated, the residual can be interpolated with kriging and then added to the previously estimated component. The prediction of the residual is a very critical step, because in principle, only the autocorrelated components should be estimated, neglecting the purely random component. Unfortunately, it is very difficult to separate the overall residual into the autocorrelated and the noncorrelated components. There are many different opinions about the best way to accomplish this issue [50,51]. In the present paper, the variography directly performed on the residuals provided results that did not depart much from those obtained with more sophisticated statistical methods; in other words, this approach did not significantly bias the final predictions. Therefore, the more straightforward approach, which brutally separates observations from trend values to obtain residuals, was preferred [29,52]. The validation of the RK method is usually carried out by means of the cross-validation procedure, and specifically the leave-one-out method [53]. Cross-validation is structured as a two-stage procedure. In the first stage, a leave-one-out method is applied, which consists of dropping an observation from the dataset and predicting this omitted value using the remaining data. Leave-one-out is iterated for each value in the dataset, and each time, a residual is computed as the difference between the observed and predicted values. The second stage of the cross-validation consists of making inferences about the residuals' distribution [54,55]. The R library [56] used to perform the aforementioned analysis was {Automap version 1.0–14}.

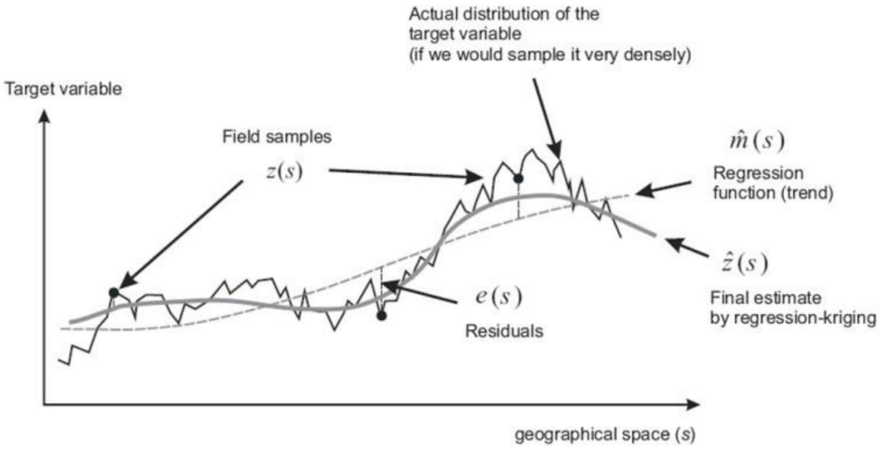

Figure 2. An example of the regression-kriging approach shown by means of a cross-section of the spatial random field (after Hengl, [35]).

2.5. *Multivariate Adaptive Regression Splines (MARS)*

MARS is a nonparametric and nonlinear predictive method that automatically models nonlinearities and interactions between variables managing suitably local nonstationarity [34]. Datasets are split into piecewise curves (splines) of differing slopes. Splines consist of two branches, i.e., left-sided (Equation (2)) and right-sided (Equation (3)) truncated functions, separated by a point called the *knot* [57].

$$\mathrm{b}_\mathrm{q}^-(\mathrm{x}-\mathrm{t}) = [-(\mathrm{x}-\mathrm{t})]_+^\mathrm{q} = \begin{cases} (\mathrm{t}-\mathrm{x})^\mathrm{q} \text{ if } \mathrm{x} < \mathrm{t} \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

$$\mathrm{b}_\mathrm{q}^+(\mathrm{x}-\mathrm{t}) = [+(\mathrm{x}-\mathrm{t})]_+^\mathrm{q} = \begin{cases} (\mathrm{x}-\mathrm{t})^\mathrm{q} \text{ if } \mathrm{x} > \mathrm{t} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

$\mathrm{b}_\mathrm{q}^-(\mathrm{x}-\mathrm{t})$ and $\mathrm{b}_\mathrm{q}^+(\mathrm{x}-\mathrm{t})$ are splines describing the regions on the right and left sides of the knot (t), respectively, and q is the degree of the polynomial. The subscript "+" indicates that the result of the function is 0 outside the local definition domain. For each of the covariate variables, MARS selects the couple of splines and the knot location more in accordance with the response variable. In a next stage, the different splines are added up in a single multivariate model, which describes the response as a function of the covariates. The result is a nonlinear model assuming the form:

$$\hat{\mathrm{y}} = \mathrm{a}_0 \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{a}_\mathrm{m} \mathrm{B}_\mathrm{m}(\mathrm{x}) \tag{4}$$

where $\hat{\mathrm{y}}$ is the prediction of the response variable; a_0 is the known term; M is the number of basic splines; and B_m and a_m are the m-th basic spline and its coefficient, respectively [58].

Overall, a MARS analysis consists of three stages. Specifically, (i) the variable that best describes the response by means of the splines in terms of R^2 is selected. Afterwards, (ii) other covariates are added stepwise, always using splines, to build a multivariate model (i.e., the global MARS model). The aim of this addition is the improvement of model in terms of performance (R^2). The performance is computed on the training set. Since the global MARS model is usually affected by *overfitting*, it needs to be "pruned" in a further stage, for which iterations of the generalized cross-validations (GCV) alternated with 10-fold cross-validation are used [59]. The GCV index is a sum of squared errors (observations minus predictions) adjusted by embodying a penalty for reducing the model complexity. This criterion is used to prevent overfitting derived from an excessively accurate model with respect to the training set:

$$\mathrm{GCV} = \frac{\frac{1}{n}\sum_{m=1}^{n}\left(y_i - \hat{f}_m(x_i)\right)^2}{(1 - C(M)/n)^2} \tag{5}$$

where C(M) is a parameter that penalizes models involving a large number of splines, defined as follows:

$$C(M) = (M+1) + dM \tag{6}$$

where M is the number of nonconstant splines (i.e., all terms of Equation (4) except a_0) in the MARS model and d is a user-defined penalty value for each spline optimization. Increases in the cost d cause the exclusion of splines. Substantially, d is increased during the pruning step in order to obtain smaller models. Besides its use during the pruning phase, GCV index is essential to rank covariates based on their importance in the model. The definition of the final model is reached in a third phase. This phase (iii) is performed by cross-validation or a new independent test set. The R library used to perform the aforementioned analysis herein is {earth} [59].

3. Results

3.1. Exploratory Data Analysis

Descriptive statistics showed that SOC data were normally distributed as confirmed by skewness and kurtosis values (Table 1) and by Shapiro–Wilk test (p = 0.656); for this reason, they were not subjected to a normal transform. The reported bubble plot (Figure 3) shows the spatial distribution of the SOC observations, evidencing some clusters of similar values.

Table 1. Summary statistics for SOC (g 100 g^{-1}).

Variable	N	Mean	Std	Min	Max	Skewness	Kurtosis
SOC	71	1.85	0.28	1.19	2.43	−0.21	−0.29

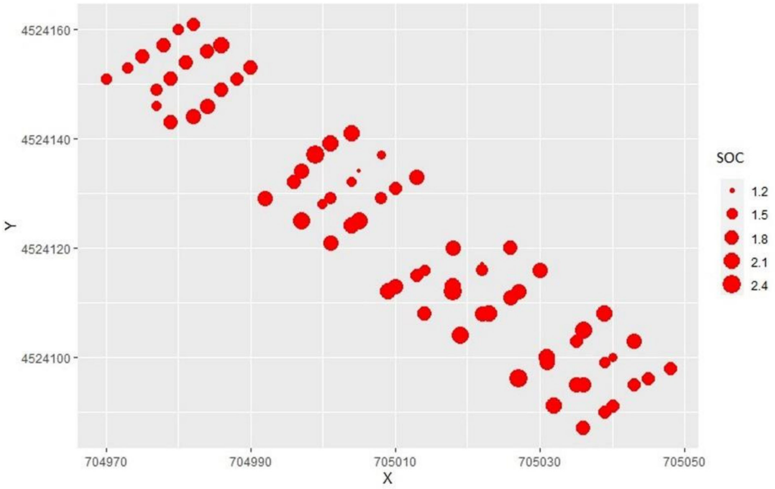

Figure 3. Bubble plot of spatial distribution of SOC values (g 100 g^{-1}).

The global Moran index provided an assessment of the spatial autocorrelation strength over the study area and is reported in Table 2. The result (I = 0.42) indicated a significant spatial autocorrelation (p = 0.00034). In addition to the global Moran index, the peak of the Moran index (local Moran index) was estimated by means of the computation of the mean of nearest neighbours. Afterwards, a lagged scatterplot provided the Moran computation at such distance lag. For the considered case, the mean of nearest neighbours was 2.63 m, and Figure 4 shows the Moran value corresponding to that distance, indicating a greater spatial correlation at short range (r = 0.75).

Table 2. Assessment of the global Moran index.

Spatial Autocorrelation Analysis (Original Data)			
Moran I	**Variance**	**Expectation**	***p*-Value**
0.42	0.017	−0.014	0.00034

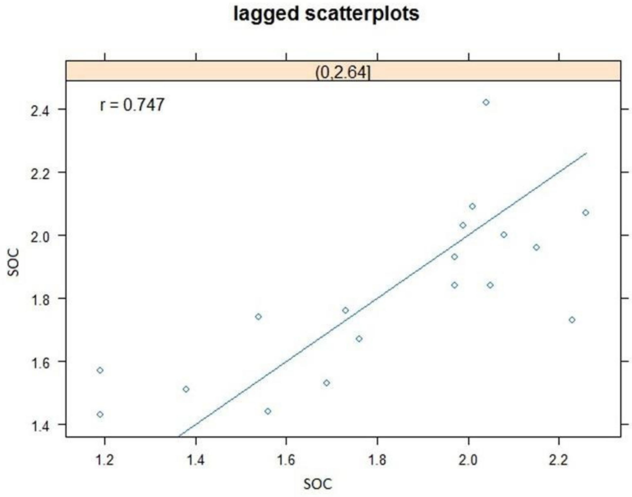

Figure 4. h-scatterplot for assessing local Moran I.

3.2. Linear Model Outcomes

The correlation matrix between SOC and the 25 covariates (23 geophysical variables plus the geographical coordinates) was first computed, and different sets of highly correlated covariates were derived and used to fit SOC data.

The following equation shows the first attempt to model SOC with the most correlated variables:

SOC ~ ckAmp0.05m_600MHz+ ckAmp0.1m_600MHz + ckAmp0.4m_600MHz

The five-point summary statistics and the coefficients of the linear model are reported in Tables 3 and 4. The outcomes seemed to indicate a larger contribution of the GPR data than of the EMI sensor data. The covariates related to the higher frequency antennae (1600MHz frequency) were therefore excluded.

Table 3. Five-point table of the linear model’s residuals.

Min	1Q	Median	3Q	Max
−0.47	−0.16	0.01	0.15	0.63

In particular, the GPR data representations for both frequencies showed a first discontinuity in the radar signal at 0.1 m depth, a high level of spatial continuity along the soil profile at least to 0.30 m, and a second discontinuity after 0.30 m depth. Therefore, the selected covariates were representative of information derived by two different layers.

Table 4. Coefficients of the linear model.

	Estimate	Std_Error	*t*_Value	Pr(>|t|)
(Intercept)	−7.056e−01	1.14e+00	−0.62	0.54
ckAmp0.05m_600MHz	4.63e−06	7.06e−05	0.07	0.95
ckAmp0.1m_600MHz	2.07e−04	8.12e−05	2.55	0.013 *
ckAmp0.4m_600MHz	−6.34e−04	1.51e−03	−0.42	0.68

Signif. codes: 0.01, "*"; 0.05, ".".

The model was significant (F-statistic: 4.80 on 3 and 67 DF, *p*-value: 0.004) and showed a residual standard error of 0.26 with 67 degrees of freedom; multiple R-squared and adjusted R-squared were 0.177 and 0.14, respectively. Analysing Table 4, it was evident that there was a unique significant covariate, ckAmp0.1m_600MHz. The result showed the distribution of SOC to be significantly affected by the shallower layer, probably because it was comparable with the portion of sampled soil.

After many other attempts (not reported), a model was developed with the following optimal arrangement of the covariates:

$$\text{SOC} \sim X + Y + \text{ckAmp0.35m\_600MHz}$$

This model included the geographical coordinates and a unique geophysical covariate, ckAmp0.35m_600MHz (see Tables 5 and 6). This model was better that the aforementioned one, with all the covariates significant, a better value of R-squared (multiple R-squared: 0.26, adjusted R-squared: 0.22), and a more significant F-statistic *p*-value (F = 7.9 on 3 and 67 DF, *p*-value: 0.00018). Residual standard error was 0.24 with 67 degrees of freedom.

Table 5. Five-point table of the second linear model's residuals.

Min	1Q	Median	3Q	Max
−0.47	−0.16	−0.02	0.13	0.66

Table 6. The second linear model's coefficients with related statistics.

Coefficients	Estimate	Std. Error	*t* Value	Pr(>|t|)
(Intercept)	7.3e+04	2.1e+04	3.4	0.00153 **
X	−1.0e−02	3.9e−03	−2.7	0.01270 *
Y	−1.4e−02	4.1e−03	−3.5	0.00118 **
ckAmp0.35m_600MHz	−2.4e−03	6.0e−04	−4.0	0.00018 ***

Signif. codes: 0, "***"; 0.001, "**"; 0.01, "*"; 0.05, ".".

The model's residuals were then analysed. The Shapiro–Wilk Gaussianity test showed a nonsignificant departure from the normal distribution (W = 0.98567, *p*-value = 0.598); as a consequence, the Gaussian hypothesis was accepted. Afterwards, spatial autocorrelation analysis was performed to check at what extent the linear model filtered out the autocorrelation present in the raw data.

From Table 7, it was evident that in the linear model's residuals, there was still a significant quantity of spatial autocorrelation (*p*-value = 0.0012). Therefore, it made sense to apply regression kriging (RK) to exploit the residual autocorrelation with the aim of improving the goodness of fit.

Table 7. Linear model coefficients with related statistics.

Spatial Autocorrelation Analysis (Linear Model's Residuals)			
Moran I	Variance	Expectation	*p*-Value
0.29	0.01	−0.014	0.0012

3.3. Regression Kriging (RK)

Geostatistical analysis was then applied to the linear model's residuals with the aim of finding in them a structure that could represent their spatial variability.

The goodness of fit between the selected variogram model and the empirical variogram was evaluated by means of the SSErr index, which provides a value that helps user to judge the quality of the final model. For the case at hand, the value was SSErr = 0.00050, which appeared to be a satisfactory result. Moreover, by analysing the variogram parameters, reported in Table 8, it was possible to figure out the strength of the model by computing the nugget-to-sill ratio index [60], also called the spatial dependence index (SDI; [61]). For the case at hand, the observed value was 0.075, indicating high descriptive capability for the variogram model.

Table 8. Variogram model and parameters.

Model	Psill *	Range
Nugget	0.0042	0.0
Spherical	0.056	8.64

* Psill = Partial sill.

In Figure 5, the experimental variogram and the fitted nested model (nugget + spherical) are reported.

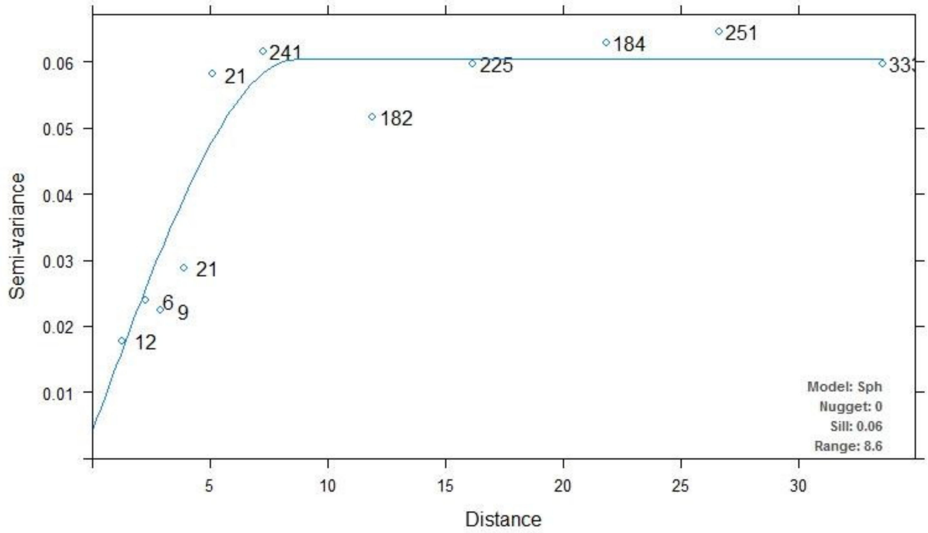

Figure 5. Experimental variogram and fitted variogram model.

Cross-validation statistics showed an MAE to RMSE ratio of 0.76, indicating a very good outcome. Mathematically, RMSE is always larger than MAE, because large errors are magnified by the square contained in the formula; therefore, the ratio between MAE and RMSE is always less than 1. However, the closer to 1 the ratio is, the fewer large errors made are by the model. This positive result was confirmed by a MAPE value far lower than 10% (Table 9). Computing the Lin coefficient (CCC) between observations and predictions, the outcomes were 0.65 for overall CCC, 0.68, for overall precision, and 0.95 for overall accuracy. The scatterplot of predicted versus observed values qualitatively showed the adequacy between the two data series (Figure 6).

Table 9. Accuracy metrics to assess the goodness of fit of the RK model.

Metric	MBE	MAE	RMSE	MAE/RMSE	MAPE	MIN	MAX
value	0.0013	0.15	0.20	0.76	8.47%	−0.49	0.42

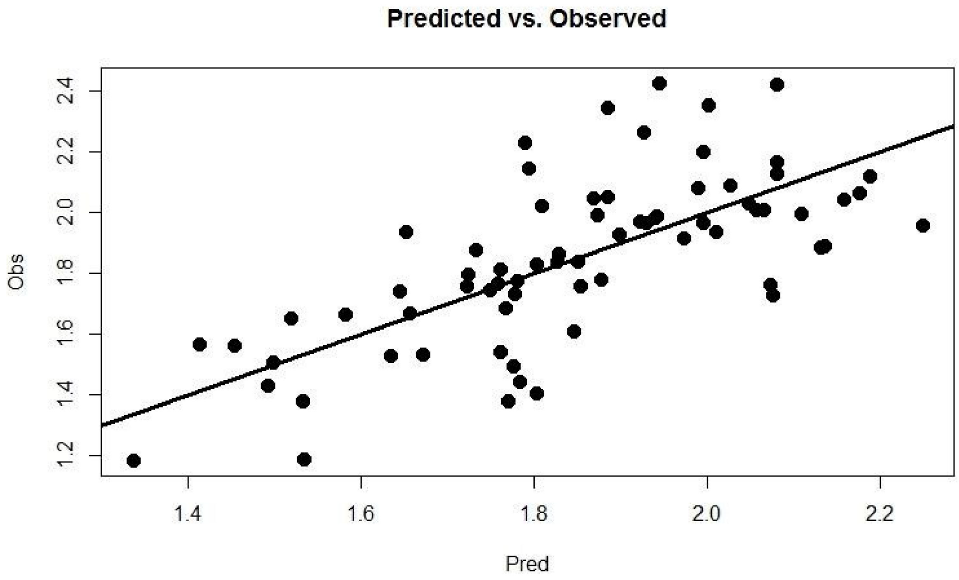

Figure 6. Scatterplot of predicted (RK model) vs. observed values.

The spatial distribution of SOC obtained through RK is reported in Figure 7.

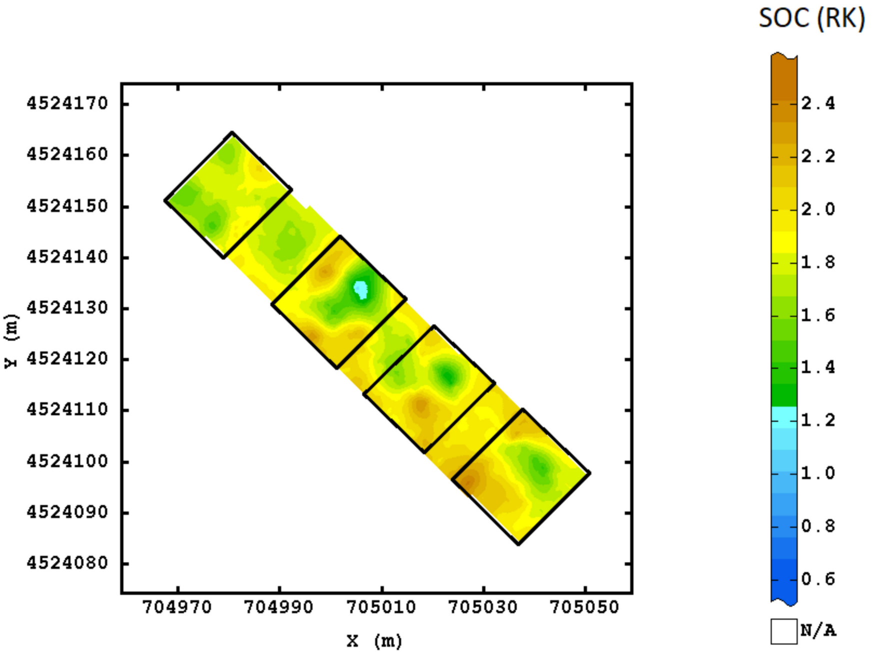

Figure 7. Map of SOC obtained with regression kriging. The black polygons indicate the four blocks in the RCB experimental design.

3.4. MARS Model Assessment

The original dataset was split into two complementary subsets, namely, training and test, corresponding to 80% and 20% of the original data, respectively.

Since the model is calibrated by means of the training dataset with the aim to predict the test data, the two subsets should be (statistically) similar at some extent. For this reason, after the splitting, subsets were subjected to the t-test for mean homogeneity and the Levene test for variance homogeneity. In addition, a univariate cluster analysis, carried out to assess the presence of clusters among data, showed that observations could be split into four groups. This represents another constraint about the splitting that has to be taken into account, i.e., the training and test subsets should be formed by a balanced quantity of elements extracted from all the clusters. Subsets were both checked for Gaussianity by means of Shapiro–Wilk test; results showed for both subsets a nonsignificant departure from normal distribution (W = 0.99, *p*-value = 0.90, for the training set; W = 0.97, *p*-value = 0.81, for the test set).

A Welch two-sample t-test showed that the means of the two subsets were not statistically different (t = −0.25, df = 20.36, *p*-value = 0.81). In addition, a boxplot confirmed the equality of the two means of the SOC variable subsets (Figure 8).

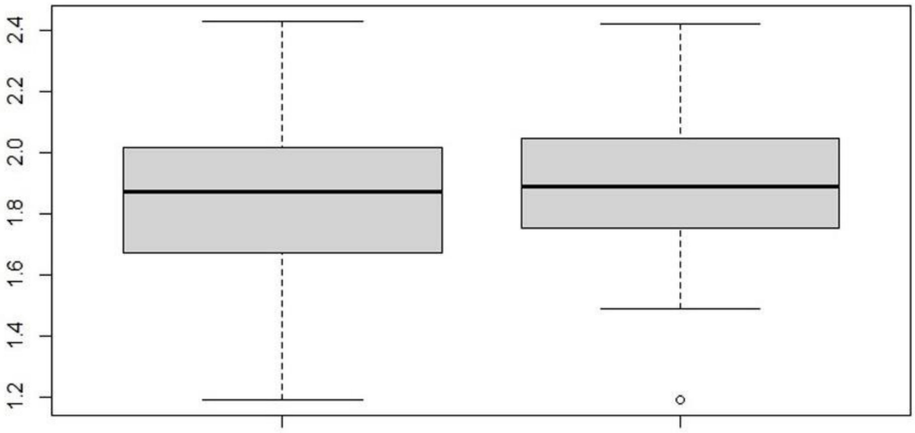

Figure 8. Boxplot for SOC comparison between training and test sets.

A Levene test, based on the absolute deviations from the median with a modified structural zero removal method and correction factor, showed the homogeneity of the group variances (test statistic = 0.059, *p*-value = 0.81). In Figure 9, the placement of the observations for the training (red points) and the test (green points) sets is reported.

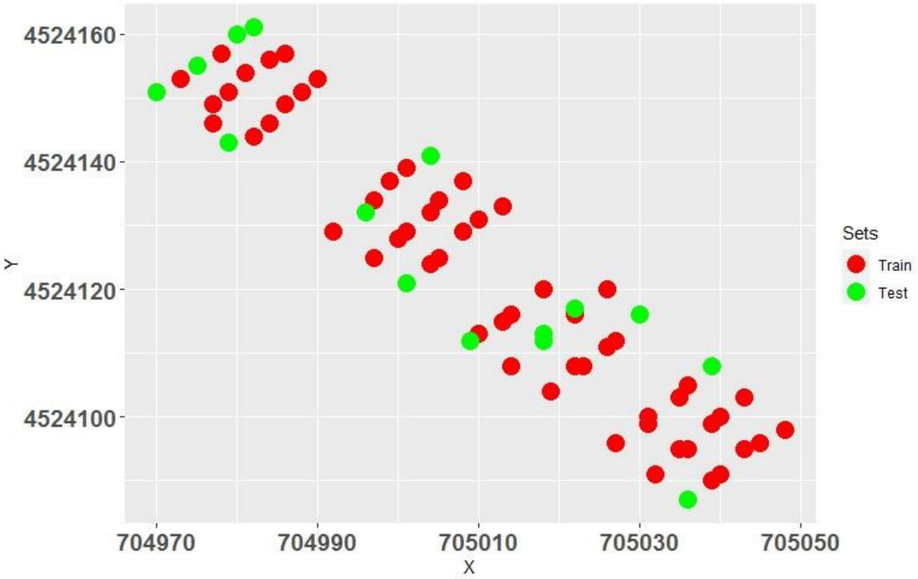

Figure 9. Spatial distribution of training and test sets points.

In summary, the two subsets could be considered similar according to the distribution, mean value, and variance comparisons. Therefore, the training set seemed to be appropriate to calibrate the model and the test set to check for overfitting.

The MARS model selected only 4 out of 25 predictors, namely, ckAmp0.35m_600MHz, X, ckEC$_a$Ver, and ckAmp0.1m_600MHz.

The model included the main GPR covariates selected previously. Regarding EMI data, only the apparent electrical conductivity measured in vertical polarization was selected because the two electrical conductivity variables were strongly correlated and therefore redundant. In addition, the sensor in vertical polarization had a maximum sensitivity approximately at a depth of 0.40 m, which was comparable with the time slices of GPR repeatedly selected (0.35 m).

From Table 10, it can be drawn that the MARS model was formed by four terms; apart from the intercept, the first was linear, and the remaining two were interactions between couples of covariates. After importance analysis was applied, by using the GCV and raw residual sum of squares (Rss) indices, the selected predictors were ranked accordingly (Table 11).

Table 10. MARS model structure.

MARS Terms	Coefficients
(Intercept)	2.0
h(ckAmp0.35m_600MHz-408)	−1.12e−02
h(13011-ckAmp0.1m_600MHz)*h(408-ckAmp0.35m_600MHz)	−5.87e−06
h(704990-X)*ckECaVer	−2.68e−03

Table 11. Covariates of the MARS model listed according to their importance rank with respect to GCV (generalized cross-validation) and Rss (raw residual sum of squares).

	GCV	Rss
ckAmp0.35m_600MHz	100.0	100.0
X	63.4	66.5
ckECaVer	63.4	66.5
ckAmp0.1m_600MHz	48.2	47.9

As first step, the Gaussianity of the residuals after the training was tested using the Shapiro–Wilk test; the residuals distribution could be considered Gaussian with a distribution ~ $N(0.0,\ 0.036)$ (W = 0.98, p-value = 0.50).

By applying a blind cross-validation with k-fold = 10, the resulting R^2 was 0.51, but it should be borne in mind that this was a pessimistic result, as the extractions of blocks of 10 elements (k-fold with k = 10) from the original dataset was performed 200 times in a purely random fashion, neglecting similar subsets. Moreover, the original dataset was relatively small and represents a not-very-homogeneous reality. Finally, the results in terms of goodness of fit were averaged.

The first step consisted of checking the correlation between predicted and observed values for the training set; the results showed a certain agreement (r = 0.72, p-value $\approx$ 0.0). In addition, correlation between residuals and predicted values of training subset was checked and was close to zero, as expected.

Afterwards, the MARS model calibrated on the training set was applied to predict SOC data from the test set, which was independent from the model calibration (training) set.

As a first step, the correlation between observations and (test set) predictions was analysed. This resulted in a highly significant correlation (r = 0.87, p-value $\approx$ 0.0). The value gained after the validation step surprisingly outperformed that of the training set, which is a rare event. The correlation between residuals and (test-set) predicted was not significant.

The residuals, according to the Shapiro–Wilk test, were Gaussian, with a distribution ~ $N(0.027,\ 0.025)$ (W = 0.93, p-value = 0.23).

Computing the Lin coefficient (CCC) between observations and predictions, the outcomes showed very good agreement (overall CCC, 0.81; overall precision, 0.88; overall accuracy, 0.93).

Since the observations were available, it was possible to compute the error metrics, which are reported in Table 12.

Table 12. Accuracy metrics to assess the goodness of fit of the MARS model.

Metric	MBE	MAE	RMSE	MAE/RMSE	MAPE	MIN	MAX
value	−0.03	0.13	0.16	0.80	6.7%	−0.42	0.19

The error indices were good overall; in particular, MAPE was below 10%, which value has been indicated in literature as a critical threshold. Another very interesting result concerned the ratio between MAE and RMSE, which was larger than that obtained with regression kriging (0.8 vs. 0.76). In conclusion, the MARS model could be considered effective whenever the coefficients of the covariates were not constant over the study domain and the covariates were intertwined in more complex ways than additively.

By comparing the error indices and Lin's coefficients of both methods, it became evident that MARS performed better than RK. The two methods were linear (RK) and nonlinear (MARS), respectively. The main difference concerns the interaction terms, since the MARS model has one linear term and two multiplicative terms (interactions) that represent the added value that allowed improving the predictive capability of MARS with respect to that of RK.

In Figure 10, the map of SOC predictions obtained with the MARS model is reported. Comparing the RK and MARS maps, they showed overall agreement, with a cluster of lower values in the northern part of the study area, a central part with the lowest values, and finally, a southern part with two clusters of larger values and a cluster of lower values.

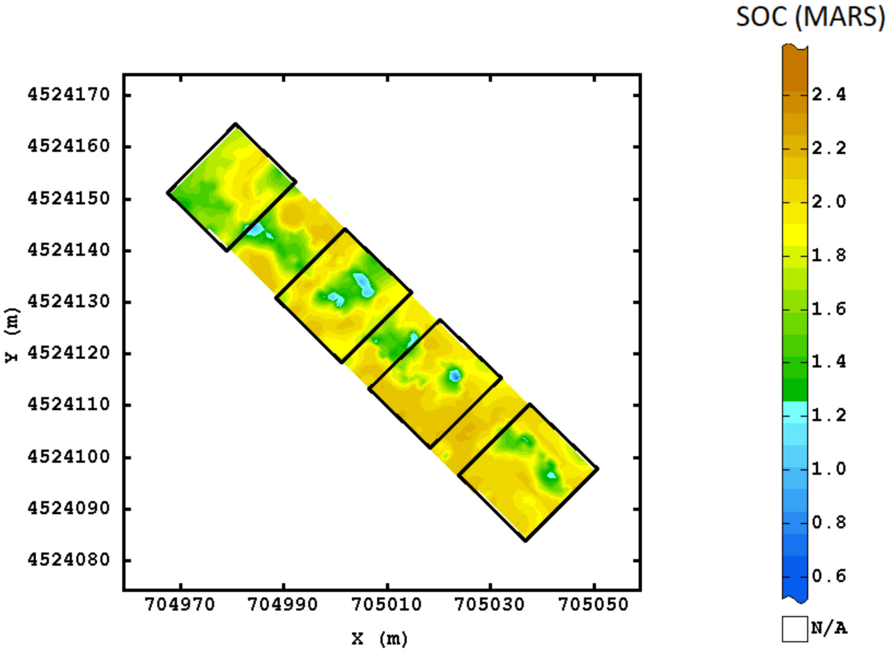

Figure 10. Map of SOC obtained with MARS model. The black polygons indicate the four blocks in the RCB experimental design.

Finally, to quantitatively compare the maps obtained by the two methods, a cross-correlogram was computed. The result was a value of 0.67 at the distance 0. Therefore, the map gained from RK can be considered a first approximation of that from MARS. This result underlines the reliability of the SOC spatial distribution predicted by the MARS model.

4. Discussion

Spatial prediction of SOC is critical for assessing the effect of agronomic management strategies on soil quality and crop productivity. In this scope, the sample size is a value that plays a key role in SOC prediction. Thus, it needs to be balanced between economic and predictive constraints. In fact, increasing the sample size may allow the application of statistical methods that take residual autocorrelation into account and thereby reduce the probability of inflation of the type I error rate [12], but at large cost. Regression kriging and MARS, incorporating covariate information often available at a finer resolution than primary data, such as proximal and remote sensor data, may improve the quality of SOC estimation without increasing the sampling size of the primary variable [62,63].

Outcomes obtained from linear models seem to highlight a larger informative contribution of GPR than of EMI data. From a physical standpoint, this result can be explained by the different nature of sensors' outcomes. In fact, GPR information results are more sensitive to near-surface effects than EMI data, which are integrated values over all soil layers [15]. However, unexpectedly [64], the covariates related to the higher-frequency antenna (1600 MHz frequency) were excluded, probably because they did not add further information or were redundant in this study case.

Two GPR covariates, namely, ckAmp0.35m_600MHz and ckAmp0.1m_600MHz, were selected by the MARS model. The same variables were also chosen by the final RK model (ckAmp0.35m_600MHz) and the preliminary RK model (ckAmp0.1m_600MHz). Similar importance was also assigned to the selected variables by both statistical methods, as shown by the ranking defined by GCV and Rss in MARS model, suggesting that their significance was physically based. In fact, the selected covariates were representative of information derived by two soil layers with different physical properties influencing radar signal and soil organic carbon distribution. The two methods also had the X geographical coordinate in common, indicating a larger continuity along this direction.

The main difference between the two approaches concerned the selection of the EMI covariate in vertical polarization performed only by the MARS model, indicating the different explanatory power of information brought by the two sensors. This result was tied to the intrinsic capability of the MARS model to intercept the interactions among variables and highlight nonlinear features underlying the data [34]. In addition, the coefficients of the MARS model were not constant but piecewise linear (splines), and therefore, their gradient varied over the studied domain [57]. This explains the larger descriptive capability of the MARS model and its ability to select hidden features with respect to regression kriging. Although MARS is not explicitly a spatial method, its capability of modelling covariate coefficients by means of flexible functions allows, when the geographic variables are included in the analysis, filtering out the spatial autocorrelation contained in the data, which makes it substantially a spatial method [65]. A confirmation of this was the statistical nonsignificance of the Moran I index obtained from the MARS residuals.

Studies on the spatial variability of SOC in agricultural soils remain a central theme in assessing the environmental sustainability of agricultural systems [66], because agronomic inputs could be rationalized in order to not impoverish the soil's fertility. Therefore, our results represent a knowledge contribution for future studies aimed at detecting the spatial distribution of soil organic carbon at the field scale. Geophysical methods show new applicative potentialities for environmental sciences (see, among others, [67]) and can represent support for research in this field. However, because of the complex interactions with soil properties, the use of geophysical measurements as covariates needs to be investigated in more detail to draw more precise conclusions. A limit of the present work could be its potential site-specificity, which could not be quantified in advance. Therefore, further experiments in different study areas and agroenvironments should be performed to test the performance of the methods under different conditions.

5. Conclusions

The results of our investigation showed that MARS outperformed RK in predicting SOC spatial distribution. The nonlinearity of MARS evidenced the contribution of EMI variables neglected by linear approaches. That result would have to be deepened in future works in consideration of the fact that EMI measures are more easily achievable than GPR ones.

The accuracy reached in mapping SOC with the support of MARS was remarkable and opens interesting perspectives in applying other, more powerful machine learning methods (e.g., deep learning) to even better exploit proximally sensed data. In the future, it is hoped that these machine learning methods will be successfully associated with mapping procedures and then applied at the regional and national level.

The use of relatively easy, accurate, and inexpensive geophysical methods for SOC estimation, together with application of advanced statistical techniques for SOC spatialization, can represent a viable solution to investigate agroecosystem sustainability.

Author Contributions: Conceptualization, A.M.S., E.B. and D.D.B.; methodology, E.B.; software, E.B.; validation, E.B., D.D.B., A.M.S., M.C. and G.L.; formal analysis, E.B., A.M.S. and D.D.B.; investigation, D.D.B.; resources, M.C. and S.P.; data curation, D.D.B.; writing—original draft preparation, E.B.; writing—review and editing, D.D.B., A.M.S., E.B., M.C. and G.L.; visualization, D.D.B.; supervision, A.M.S. and E.B.; project administration and funding acquisition, A.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank the EU and MIUR for funding the present methodological contribution under the framework of the collaborative international consortium DESERT (ID-217) financed under the ERA-NET Cofund WaterWorks2014 Call. This ERA-NET is an integral part of the 2015 Joint Activities developed by the Water Challenges for a Changing World Joint Programme Initiative (Water JPI).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Gregorich, E.G.; Carter, M.R.; Angers, D.A.; Monreal, C.M.; Ellert, B.H. Towards a minimum data set to assess soil organic matter quality in agricultural soils. *Can. J. Soil Sci.* **1994**, *74*, 367–385. [CrossRef]
2. Johnston, C.A.; Groffman, P.; Breshears, D.D.; Cardon, Z.G.; Currie, W.; Emanuel, W.; Gaudinski, J.; Jackson, R.B.; Lajtha, K.; Nadelhoffer, K.; et al. Carbon cycling in soil. *Front. Ecol. Environ.* **2004**, *2*, 522–528. [CrossRef]
3. Lorenz, K.; Lal, R.; Ehlers, K. Soil organic carbon stock as an indicator for monitoring land and soil degradation in relation to United Nations' Sustainable Development Goals. *Land Degrad. Dev.* **2019**, *30*, 824–838. [CrossRef]
4. Adhikari, K.; Hartemink, A.E. Linking soils to ecosystem services—A global review. *Geoderma* **2016**, *262*, 101–111. [CrossRef]
5. Shukla, M.K.; Lal, R.; Ebinger, M. Determining soil quality indicators by factor analysis. *Soil Till. Res.* **2006**, *87*, 194–204. [CrossRef]
6. Stellacci, A.; Castellini, M.; Diacono, M.; Rossi, R.; Gattullo, C. Assessment of Soil Quality under Different Soil Management Strategies: Combined Use of Statistical Approaches to Select the Most Informative Soil Physico-Chemical Indicators. *Appl. Sci.* **2021**, *11*, 5099. [CrossRef]
7. Fang, X.; Xue, Z.; Li, B.; An, S. Soil organic carbon distribution in relation to land use and its storage in a small watershed of the Loess Plateau, China. *CATENA* **2012**, *88*, 6–13. [CrossRef]
8. Ferrara, R.M.; Mazza, G.; Muschitiello, C.; Castellini, M.; Stellacci, A.M.; Navarro, A.; Lagomarsino, A.; Vitti, C.; Rossi, R.; Rana, G. Short-term effects of conversion to no-tillage on respiration and chemical-physical properties of the soil: A case study in a wheat cropping system in semi-dry environment. *Ital. J. Agrometeorol.* **2017**, *1*, 47–58.
9. Leogrande, R.; Stellacci, A.M.; Vitti, C.; Lacolla, G.; Moscelli, S.; Mastrangelo, M.; Vivaldi, G.A. Soil properties as affected by irrigation with treated municipal wastewater. In Proceedings of the XLVII Conference of Italian Society for Agronomy, Marsala, Italy, 12–14 September 2018.
10. Stellacci, A.M.; De Benedetto, D.; Leogrande, R.; Vitti, C.; Castellini, M.; Barca, E. Use of Mixed Effects Models accounting for residual spatial correlation to analyze soil properties variation in a field irrigated with treated municipal wastewater. In Proceedings of the XLVII Conference of Italian Society for Agronomy, Marsala, Italy, 12–14 September 2018.
11. Stevenson, F.J.; Cole, M.A. *Cycles of Soil*, 2nd ed.; Wiley: New York, NY, USA, 1999.
12. Littell, R.C.; Milliken, G.A.; Stroup, W.W.; Wolfinger, R.D.; Schabenberger, O. *SAS for Mixed Models*, 2nd ed.; SAS Institute Inc.: Cary, NC, USA, 2006.
13. Journel, A.G.; Huijbregts, C.J. *Mining Geostatistics*; Academic Press: Waltham, MA, USA, 1978.
14. Webster, R.; Oliver, M.A. How large a sample is needed to estimate the regional variogram adequately? In *Geostatistics Tróia '92*; Springer: Berlin, Germany, 1993; pp. 155–166.

15. Barca, E.; De Benedetto, D.; Stellacci, A.M. Contribution of EMI and GPR proximal sensing data in soil water content assessment by using linear mixed effects models and geostatistical approaches. *Geoderma* **2019**, *343*, 280–293. [CrossRef]
16. Piccini, C.; Marchetti, A.; Francaviglia, R. Estimation of soil organic matter by geostatistical methods: Use of auxiliary infor-mation in agricultural and environmental assessment. *Ecol. Ind.* **2014**, *36*, 301–314. [CrossRef]
17. Stevens, A.; Udelhoven, T.; Denis, A.; Tychon, B.; Lioy, R.; Hoffmann, L.; van Wesemael, B. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* **2010**, *158*, 32–45. [CrossRef]
18. Nawar, S.; Mouazen, A.M. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *CATENA* **2017**, *151*, 118–129. [CrossRef]
19. Rossel, R.V.; Adamchuk, V.I.; Sudduth, K.A.; McKenzie, N.J.; Lobsey, C. Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. *Adv. Agron.* **2011**, *113*, 243–291.
20. Heil, K.; Schmidhalter, U. The Application of EM38: Determination of Soil Parameters, Selection of Soil Sampling Points and Use in Agriculture and Archaeology. *Sensors* **2017**, *17*, 2540. [CrossRef] [PubMed]
21. Martinez, G.; Vanderlinden, K.; Ordóñez, R.; Muriel, J.L. Can Apparent Electrical Conductivity Improve the Spatial Characterization of Soil Organic Carbon? *Vadose Zone J.* **2009**, *8*, 586–593. [CrossRef]
22. McNeill, J.D. *Electromagnetic Terrain Conductivity Measurement at Low Induction Numbers*; Technical Note TN 6; Geonics Ltd.: Mississauga, ON, Canada, 1980.
23. Sudduth, K.A.; Kitchen, N.R.; Bollero, G.A.; Bullock, D.G.; Wiebold, W.J. Comparison of Electromagnetic Induction and Direct Sensing of Soil Electrical Conductivity. *Agron. J.* **2003**, *95*, 472–482. [CrossRef]
24. Grote, K.; Anger, C.; Kelly, B.; Hubbard, S.; Rubin, Y. Characterization of Soil Water Content Variability and Soil Texture using GPR Groundwave Techniques. *J. Environ. Eng. Geophys.* **2010**, *15*, 93–110. [CrossRef]
25. Conyers, L.B.; Goodman, D. *Ground Penetrating Radar: An Introduction for Archaeologists*; Altamira Press: London, UK, 1997.
26. Collins, M.; Schellentrager, G.; Doolittle, J.; Shih, S. Using ground-penetrating radar to study changes in soil map unit compo-sition in selected Histosols. *Soil Sci. Soc. Am. J.* **1986**, *50*, 408–412. [CrossRef]
27. Winkelbauer, J.; Völkel, J.; Leopold, M.; Bernt, N. Methods of surveying the thickness of humous horizons using ground pen-etrating radar (GPR): An example from the Garmisch-Partenkirchen area of the Northern Alps. *Eur. J. For. Res.* **2011**, *130*, 799–812. [CrossRef]
28. Goovaerts, P. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *J. Hydrol.* **2000**, *228*, 113–129. [CrossRef]
29. Hengl, T.; Heuvelink, G.B.; Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **2004**, *120*, 75–93. [CrossRef]
30. Webster, R.; Oliver, M.A. *Geostatistics for Environmental Scientists*; John Wiley & Sons Ltd.: Chichester, UK, 2001.
31. Bourennane, H.; King, D. Using multiple external drifts to estimate a soil variable. *Geoderma* **2003**, *114*, 1–18. [CrossRef]
32. Castrignanò, A.; Costantini, E.A.; Barbetti, R.; Sollitto, D. Accounting for extensive topographic and pedologic secondary information to improve soil mapping. *CATENA* **2009**, *77*, 28–38. [CrossRef]
33. Odeh, I.; McBratney, A.; Chittleborough, D. Further results on prediction of soil properties from terrain attributes: Heterotopic cokriging and regression-kriging. *Geoderma* **1995**, *67*, 215–226. [CrossRef]
34. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 67. [CrossRef]
35. Hengl, T.A. *Practical Guide to Geostatistical Mapping*; Office for Official Publications of the European Communities: Luxembourg, 2009.
36. Zhang, W.; Goh, A. Multivariate adaptive regression splines for analysis of geotechnical engineering systems. *Comput. Geotech.* **2013**, *48*, 82–95. [CrossRef]
37. Garg, A.; Garg, A.; Tai, K. A multi-gene genetic programming model for estimating stress-dependent soil water retention curves. *Comput. Geosci.* **2014**, *18*, 45–56. [CrossRef]
38. Zhang, S.; Huang, Y.; Shen, C.; Ye, H.; Du, Y. Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information. *Geoderma* **2012**, *171–172*, 35–43. [CrossRef]
39. Minasny, B.; McBratney, A. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. *Geoderma* **2007**, *140*, 324–336. [CrossRef]
40. Eldeiry, A.A.; Garcia, L.A. Comparison of Ordinary Kriging, Regression Kriging, and Cokriging Techniques to Estimate Soil Salinity Using LANDSAT Images. *J. Irrig. Drain. Eng.* **2010**, *136*, 355–364. [CrossRef]
41. United Nations Educational, Scientific and Cultural Organization-Food and Agriculture Organization of the United Nations (UNESCO-FAO). *Bioclimatic Map of the Mediterranean Zone*; UNESCO: Paris, France; FAO: Rome, Italy, 1963; 60p.
42. Legrande, R.; Vitti, C.; Castellini, M.; Mastrangelo, M.; Pedrero, F.; Vivaldi, G.; Stellacci, A. Comparison of Two Methods for Total Inorganic Carbon Estimation in Three Soil Types in Mediterranean Area. *Land* **2021**, *10*, 409. [CrossRef]
43. Vitti, C.; Stellacci, A.M.; Leogrande, R.; Mastrangelo, M.; Cazzato, E.; Ventrella, D. Assessment of organic carbon in soils: A comparison between the Springer-Klee wet digestion and the dry combustion methods in Mediterranean soils (Southern Italy). *Catena* **2016**, *137*, 113–119. [CrossRef]
44. Barca, E.; Stellacci, A.M.; De Benedetto, D. Optimization of Sampling Design for Total Organic Carbon Assessment using Spatial Simulated Annealing: Comparison of Different Variogram Models Performances. In Proceedings of the XLVIII Conference of Italian Society for Agronomy, Perugia, Italy, 18–20 September 2019; pp. 223–224.

45. Annan, A.P. Electromagnetic principles of ground penetrating radar. In *Ground Penetrating Radar: Theory and Applications*; Jol, H.M., Ed.; Elsevier: Amsterdam, The Netherlands, 2009; pp. 3–40. ISBN 978-0-444-53348-7. [CrossRef]
46. Davis, J.L.; Annan, A.P. Ground-penetrating radar for high-resolution mapping of soil and rock stratigraphy. *Geophys. Prospect.* **1989**, *37*, 531–551. [CrossRef]
47. De Benedetto, D.; Quarto, R.; Castrignanò, A.; Palumbo, D.A. Impact of Data Processing and Antenna Frequency on Spatial Structure Modelling of GPR Data. *Sensors* **2015**, *15*, 16430–16447. [CrossRef] [PubMed]
48. Daniels, D.J. *Ground Penetrating Radar*, 2nd ed.; The Institution of Engineering and Technology: London, UK, 2004.
49. *User's Manual Online Version, Sandmeier Scientific Software*, Reflexw, v.6.1.1. Program for Processing and Interpretation of Reflection and Transmission Data. Reflexw: Karlruhe, Germany, 2012.
50. Kitanidis, P.K. Generalized covariance functions in estimation. *Math. Geol.* **1993**, *25*, 525–540. [CrossRef]
51. Lark, R.M.; Cullis, B.R.; Welham, S.J. On spatial prediction of soil properties in the presence of a spatial trend: The empirical best linear unbiased predictor (E-BLUP) with REML. *Eur. J. Soil Sci.* **2005**, *57*, 787–799. [CrossRef]
52. Hengl, T.; Heuvelink, G.B.; Rossiter, D.G. About regression-kriging: From equations to case studies. *Comput. Geosci.* **2007**, *33*, 1301–1315. [CrossRef]
53. Barca, E.; Porcu, E.; Bruno, D.; Passarella, G. An automated decision support system for aided assessment of variogram models. *Environ. Model. Softw.* **2017**, *87*, 72–83. [CrossRef]
54. Myers, J.C. *Geostatistical Error Management: Quantifying Uncertainty for Environmental Sampling and Mapping*; John Wiley and Sons: Hoboken, NJ, USA, 1997.
55. Chiles, J.P.; Delfiner, P. *Geostatistics, Modeling Spatial Uncertainty*; Wiley: Hoboken, NJ, USA, 1999.
56. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.R-project.org/ (accessed on 17 January 2022).
57. Ghasemi, J.B.; Zolfonoun, E. Application of principal component analysis–multivariate adaptive regression splines for the simultaneous spectrofluorimetric determination of dialkyltins in micellar media. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2013**, *115*, 357–363. [CrossRef]
58. Hiemstra, P.H.; Pebesma, E.; Twenhöfel, C.J.; Heuvelink, G.B. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Comput. Geosci.* **2008**, *35*, 1711–1721. [CrossRef]
59. Milborrow, S. Derived from Mda: MARS by T. Hastie and R. Tibshirani. Earth: Multivariate Adaptive Regression Splines. R Package. 2011. Available online: http://www.milbo.users.sonic.net/earth/citing-earth.html (accessed on 17 January 2022).
60. Cambardella, C.A.; Elliott, E.T. Carbon and Nitrogen Dynamics of Soil Organic Matter Fractions from Cultivated Grassland Soils. *Soil Sci. Soc. Am. J.* **1994**, *58*, 123–130. [CrossRef]
61. Pasini, M.P.B.; Dal'Col Lúcio, A.; Cargnelutti, A.F. Semivariogram models for estimating fig fly population density throughout the year. *Pesqui. Agropecu. Bras.* **2014**, *49*, 493–505. [CrossRef]
62. Xie, X.-L.; Li, A.-B.; Mouazen, A. Improving spatial estimation of soil organic matter in a subtropical hilly area using covariate derived from vis-NIR spectroscopy. *Biosyst. Eng.* **2016**, *152*, 126–137. [CrossRef]
63. Mirzaee, S.; Ghorbani-Dashtaki, S.; Mohammadi, J.; Asadi, H.; Asadzadeh, F. Spatial variability of soil organic matter using remote sensing data. *CATENA* **2016**, *145*, 118–127. [CrossRef]
64. De Benedetto, D.; Castrignanò, A.; Sollitto, D.; Modugno, F.; Buttafuoco, G.; Lo Papa, G. Integrating geophysical and geosta-tistical techniques to map the spatial variation of clay. *Geoderma* **2012**, *171–172*, 53–63. [CrossRef]
65. Wang, X.; Yang, C.; Zhou, M. Partial Least Squares Improved Multivariate Adaptive Regression Splines for Visible and Near-Infrared-Based Soil Organic Matter Estimation Considering Spatial Heterogeneity. *Appl. Sci.* **2021**, *11*, 566. [CrossRef]
66. Castellini, M.; Stellacci, A.M.; Tomaiuolo, M.; Barca, E. Spatial variability of soil physical and hydraulic properties in a durum wheat field: An assessment by the BEST-Procedure. *Water* **2019**, *11*, 1434. [CrossRef]
67. Di Prima, S.; Winiarski, T.; Angulo-Jaramillo, R.; Stewart, R.D.; Castellini, M.; Najm, M.R.A.; Ventrella, D.; Pirastru, M.; Giadrossich, F.; Capello, G.; et al. Detecting infiltrated water and preferential flow pathways through time-lapse ground-penetrating radar surveys. *Sci. Total Environ.* **2020**, *726*, 138511. [CrossRef]