Manuscript sent in revised form on 16 September 2024 and accepted on 19 September 2024

Published manuscript data:

Barca Emanuele, De Benedetto Daniela, Stellacci Anna Maria. Optimization of sampling design for soil total organic carbon assessment in the precision agriculture framework: Impact of different variogram models and potentiality of ground penetrating radar (GPR) covariate information. Computers and Electronics in Agriculture 226 (2024) 109470

Link to the publisher version:

https://www.sciencedirect.com/science/article/pii/S0168169924008615?via%3Dihub

DOI: 10.1016/j.compag.2024.109470

Optimization of sampling design for soil total organic carbon assessment in the precision agriculture framework: Impact of different variogram models and potentiality of ground penetrating radar (GPR) covariate information

Barca Emanuele[1], De Benedetto Daniela[2*], Stellacci Anna Maria[3]

[1]Water Research Institute (IRSA)–National Research Council (CNR), Bari, Italy

[2]Council for Agricultural Research and Economics—Research Centre for Agriculture and Environment, CREA-AA, Via Celso Ulpiani 5, 70125 Bari, Italy

[3]Department of Soil, Plant and Food Sciences, University of Bari "Aldo Moro", via Amendola 165/A, Bari, Italy

*****corresponding author: daniela.debenedetto@crea.gov.it

Abstract

Assessing soil organic carbon (SOC) at the field scale is crucial for efficient environmental and agronomic management, especially within the precision agriculture framework. This enables the implementation of crop and soil management strategies to enhance soil quality, increase carbon sequestration, and improve crop yields. However, the process of sampling and assessing SOC is resource-intensive, demanding both in time and labour. This aspect is particularly relevant also in agronomic research, where the need to assess the spatial correlation of the investigated variables requires the collection of large datasets with georeferenced data and often measurements need to be repeated over time as for long-term field experiment - LTE. Methods for minimizing information loss while reducing the sampling scheme, such as spatial simulated annealing (SSA), are particularly valuable in the scope of SOC assessment, especially when faced with budget and time/labour constraints. Within the structure of the SSA method, two critical components can be identified: i) the inclusion of highly informative covariates for the primary variable (SOC); ii) the selection of the most appropriate variogram model for spatial variability assessment. Covariates strongly correlated with

SOC, such as those obtained from ground penetrating radar (GPR) that can be collected at a higher spatial density compared to SOC data, along with a well-performing model, can significantly enhance the efficiency of the sampling scheme reduction process. We conducted a study using data from an agronomic field experiment, which included 71 georeferenced sampling locations and, through an iterative downsizing process utilizing SSA, we progressively reduced the number of sampling points by removing 10, 15, and 20 observations. The sampling scheme was refined according to two distinct variogram models: the spherical and Gaussian-Matérn models, both for the primary variable (SOC) and the covariate GPR variable. This process allowed us to identify the optimal variogram model, which has a key role in maximizing the reduction of redundant points while preserving those with valuable information, and to assess the role of the covariate variable both in improving the optimization of the sampling scheme for SOC and in replacing the primary variable to optimize the sampling scheme. Finally, to assess the impact of sampling scheme reduction, a validation process was performed by estimating the dropped points by means of the remaining points using ordinary kriging and regression kriging and analysing the accuracy of such estimation. Our analysis demonstrated that it was possible to reduce the original sampling scheme by approximately 20% equivalent to eliminating 15 sampling points out of 71 without compromising its predictive capability.

Keywords: Ground Penetrating Radar, Precision agriculture, reduction of sampling scheme, regression kriging, soil organic carbon, spatial simulated annealing,variogram model.

1. Introduction

1.1 *The sampling issue*

Evaluating soil properties, particularly organic carbon content, is crucial for both environmental preservation and effective agricultural management. Understanding the spatial distribution of soil total organic carbon (SOC) is essential for modelling spatio-temporal dynamics and assessing the impact of agricultural practices on soil fertility and carbon sequestration (Sherpa et al., 2016; Zhao et

al., 2016; De Benedetto et al., 2022). In addition, the assessment of the field-scale spatial variability of soil organic carbon (SOC) is critical in precision agriculture applications. Precision agriculture is an environmental-aware approach that supports farmers in optimizing field activities and agronomic techniques, including soil management, irrigation, and fertilization, by offering a range of innovative technological tools and applying advanced data analysis methods (Abo-Habaga et al., 2024; Colovic et al., 2024; Mzid et al., 2023; Zaman, 2023). The primary goal of precision agriculture is to enhance crop production and resource use efficiency through site-specific practices, demanding a precise assessment of spatial variations of key variables such as soil organic carbon or soil water content at fine resolutions. This information aids farmers in identifying homogeneous zones for tailored agronomic input application (Popolizio et al., 2022a-b; Zaman, 2023).

The assessment of the spatial structure of the investigated variables is also an important objective of agronomic research in order to adopt appropriate statistical models for data analysis. This requires the collection of large datasets with georeferenced data (Lawrence et al., 2020; Popolizio et al., 2022a); in addition, often measurements need to be repeated over time as in the case of long-term field experiments -LTEs. However, obtaining accurate soil and crop property samples can be challenging and expensive. Therefore, optimizing sampling strategies is essential to provide precise information at minimal costs. Selecting optimal sampling approaches and data analysis methods is critical for addressing fine-scale variability and understanding spatial dependencies between observations and residuals, preventing misleading conclusions about soil property estimations (Littell, 2006). The process of predicting soil variables typically begins with designing an appropriate sampling scheme to capture soil heterogeneity effectively (Stein and Ettema, 2003; Warrick et al., 2002; Jalilian et al., 2024). Common sampling methods include grid-based, simple random, and stratified random schemes. Grid-based sampling is prevalent at the plot scale, but it can become cost-prohibitive when applied over large areas (Atkinson and Lloyd, 2007). Optimizing the sampling design can reduce the number of sampling points with minimal accuracy loss. Critical considerations in sampling design optimization involve choosing between deterministic (e.g., inverse distance

weighted) and stochastic (e.g., kriging) approaches and selecting the appropriate objective or cost function. When the stochastic approach is preferred, optimizing spatial sampling includes selecting the optimal model for spatial dependence, typically represented by a variogram model when using ordinary kriging variance as the objective function.

1.2 GPR as a source of valuable covariate information and sample redundancy discarding

Another important aspect is the incorporation of covariate information, which can help explain variability in soil parameters (Tarr et al., 2005). Utilizing covariate variables, such as data from geophysical sensors, can mitigate the limitations imposed by low sampling density and enhance the accuracy of soil organic carbon (SOC) estimation.

Among these geophysical-based methods, in particular, Ground Penetrating Radar (GPR) has gained attention for its applications in agriculture and in soil studies to assess various soil properties and their spatial and temporal variability (Pathirana et al., 2023). Compared to traditional methods, the non-destructive nature of GPR offers several advantages in agricultural landscape applications, such as making repeated measurements and providing more extensive spatial coverage with geo-referenced data. GPR is particularly attractive due to its high resolution and the ability to provide three-dimensional (3D) information (Liu et al., 2019). The main limits of this technique are the difficulty of real-time data processing and interpretation. In fact, GPR does not directly measure soil properties (Pathirana et al., 2023), but its 3D data can be correlated with soil attributes. Incorporating GPR data as covariates into models like regression kriging and multivariate adaptive regression splines has proven effective for spatially predicting SOC at the field scale (De Benedetto et al., 2022).

In order to reduce sample redundancy, a predictive model is needed. When a stochastic approach is preferred, the model is composed by: i) the trend component and ii) the stochastic component. The first component is arranged as a linear model with SOC as response, while the second component is arranged through the assessment of the variogram model representing the spatial variability of SOC observations. Since the finding of the best variogram assessment is crucial to maximize the predictive

capability of the interpolation model, different data-driven approaches are available, such as Weighted Least Squares (WLS) approach and Restricted Maximum Likelihood (REML).

Once defined the model to predict SOC, the next step is to proceed at reducing the complete sample design by discarding an increasing number of sampling points. This task can be carried out by means of Spatial Simulated Annealing (SSA), a largely recognized optimization method. SSA is a heuristic method based on Markov chain properties, designed to increase the probability of converging towards the global optimum (Metropolis et al. 1953). SSA checks the predictive capability of any generated sub-sample by computing the average kriging variance of an estimation grid superimposed over the study area. The sub-set that achieves the minimum average kriging variance is selected as the best sub-sample. The R function "spsann" has been run to apply SSA (Samuel-Rosa, 2017).

In this study, our primary goal was to reduce an initially oversized sampling plan for soil total organic carbon assessment. It is a common practice to initially oversample an area and subsequently optimize the sampling scheme, for instance in the case of multi-temporal sampling experimental designs (Popolizio et al., 2022b). Moreover, the study pursed two additional goals: (i) the comparison of two variogram models in order to select the one with the best capability in reducing SOC sampling scheme; and (ii) the potential use of GPR information both as covariate variable and in the place of the primary variable (SOC) within the optimization process. The study was carried out within an agronomic field experiment where the effect of different irrigation and fertilization management on soil properties was evaluated. This specific case study was chosen considered the importance of the assessment of spatial correlation in agronomic research both for adopting appropriate statistical data analysis approaches and for supporting farmers in delineating management zones and fostering the application of precision agriculture strategies.

2. Materials and Methods

2.1. Study area and data collection

Soil data derived from a field experiment (96 m x 20 m) carried out in Southern Italy (40° 50' N, 17° 25' E 26 m above sea level) where the effect of the irrigation and fertilization management on olive tree were compared. The following treatments were applied: irrigation with fresh water and full fertilization supply (FW); irrigation with treated municipal wastewater (TWW) and full fertilization supply (R1); irrigation with TWW and fertilizer supply reduced by the amount provided by TWW (R2), (De Benedetto et al., 2022). The fertilizer rates corresponding to the full fertilization regime (FW, R1) were computed considering soil nutrient availability, plant uptakes and target yield and were equal to: 100 kg ha^{-1} of N, 60 kg ha^{-1} of P_2O_5 and 80 kg ha^{-1} of K_2O in 2015 and 110 kg ha^{-1} of N, 60 kg ha^{-1} of P_2O_5 and 85 kg ha^{-1} of K_2O in 2016. In both years, the treated wastewaters contributed with about 20%, 2% and 70% of the olive fertilizer requirements for N, P_2O_5 and K_2O, respectively; the fertilizer rates of treatment R2 were therefore reduced accordingly. Treatments were arranged in a randomized complete block design (RCBD) with four replicates (Fig. 1). Soil samples with absolute coordinates were collected on a regular grid (April 2017) on six locations (sub-replicates, SR) per plot at a 0-0.20 m depth for a total of 72 observations (3 Treatments x 4 Replicates x 6 SR). For the data analysis, 71 observations were considered since one was falling outside the sampling grid as shown in Figure 1. Total organic carbon (SOC) was quantified on air-dried and sieved samples through dry combustion. Further details about the experimental trial are reported in De Benedetto et al. (2022).

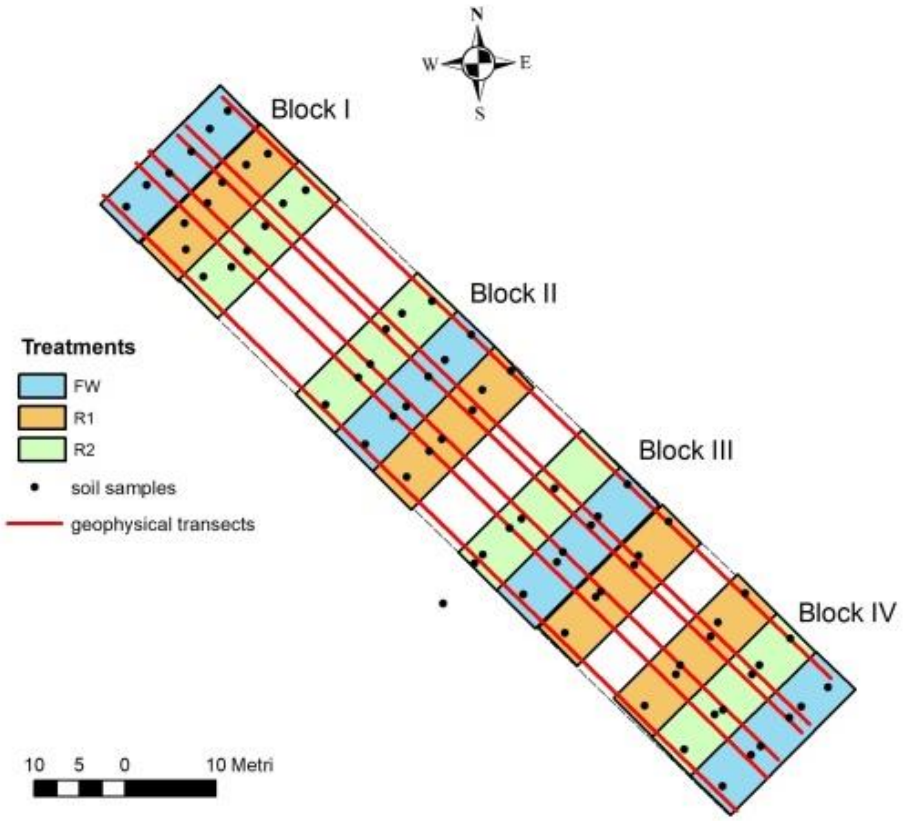


Fig. 1. The field experiment with the soil sampling locations (black dots) and Ground Penetrating Radar (GPR) acquisitions along transects (red lines). The treatments were: irrigation with fresh water and full fertilization supply (FW, blue colour), irrigation with treated municipal wastewater (TWW) and full fertilization supply (R1, orange colour) and irrigation with TWW and fertilizer supply reduced by the amount provided by TWW (R2, green colour).

A geophysical survey was carried out a GPR sensor (connected to a DGPS) along 6 parallel transects by sliding the sensors on the surface (Fig. 1).

GPR data used in this study were collected with the common offset reflection method, using a monostatic system 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 sampling interval of 0.05 ns; successive traces were collected every 0.024 m. The successive results will be focused on frequency of 600MHz because the maps were less noisy compared with the ones of the 1600 MHz antenna, because of their spatial resolution. The GPR produces a short pulse of high

frequency (10–1000 MHz) electromagnetic energy which propagates through the sub-surface materials. The propagation of the radar signal depends on the electrical properties of the soil and variations in the electrical properties of soils are usually associated with changes in volumetric water content, porosity, texture, with regard to clay types and clay mineralogy, CEC and dissolved ions in the soil water present in soil pores (McNeill, 1980). The theoretical aspects of radar components and their working principles can be found in detail in Davis and Annan (1989).

2.2.Pre-processing of GPR data

Processing the raw data consisted in extracting quantifiable variables, such as attenuation, and to display GPR data in horizontal maps at a specified time (or depth), called amplitude maps or time slices. The pre-processing of GPR signal amplitude data included the application of a set of filters (De Benedetto et al., 2015) and extracting quantifiable variables. Detailed descriptions of the above procedures can be found in the historical publications by Reynolds (1997) and Cassidy (2009).

The enveloped amplitude maps (time slices) were built 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 (2 ns). The total time interval was of 20 ns because this time was comparable with the depth of soil (at 0–0.30 m depth). The time slices were then transformed in depth slices using the velocity of the radar waves determined through the analysis of hyperbolae (Daniels, 2004).

2.3.Mapping data using Geostatistics

SOC map was estimated from 71 data points by ordinary kriging (OK) method (Matheron, 1963) on 0.5 × 0.5-m grid. The experimental variograms were calculated and fitted using a linear model of coregionalization (LMC) for GPR data. The fitting parameters, tested by using cross-validation and calculating mean error (ME) and mean squared standardized error (MSSE) within a typical used

tolerance of 1±3Sqrt(2/N) (where N is the number of observations), were used to interpolate the data on the same grid using ordinary cokriging (CK) (Goovaerts, 1997).

All geostatistical analyses were performed with ISATIS software (Geovariances, 2015) and with the following R libraries {automap, spsann}.

2.4.Spatial simulated annealing (SSA)

The Spatial Simulated Annealing (SSA) is an optimization technique designed to efficiently extract a subset of samples from a larger sampling scheme among those contained in a vast solution space. Its objective is to minimize a specific objective function while taking advantage from the spatial auto-correlation of the sampled variable. The theory underlying Spatial Simulated Annealing (SSA) draws inspiration from the organizational dynamics of a metal atomic network during rapid heating and gradual cooling processes, known as *annealing* see Barca et al. (2015). During annealing, atoms within the metal reconfigure to a state of minimal energy maintenance cost. In this analogy, the arrangement of atoms corresponds to the configuration of sampling points, while the Objective Function (OF) mirrors the energy state of the system (Pardo-Igùzquiza 1998; Deutsch and Cockerham 1994). Figure 2 illustrates the flowchart of SSA's operation, which can be understood as comprising a pre-processing stage followed by two nested loops. In the pre-processing phase, essential parameters required to start the optimization method are estimated, including the initial configuration (S0) and the initial temperature (T0).

The outer loop, reminiscent of temperature control in the physical analogy, stops as the temperature variable approaches zero. Conversely, the inner loop, linked to problem size (k), iterates k times for a given temperature value. Within this inner loop, candidate solutions are generated and subjected to a decision rule. Consequently, the maximum number of iterations of the method is k*NTemp, where NTemp represents the total different temperature values assumed during the run. It is worth noting that the total number of iterations corresponds to the total number of processed network configurations. SSA evaluates a neighbouring configuration Si of the current optimal configuration

S* at each step i, probabilistically deciding whether to move to Si or remain in S*. The transition probability PT from the current configuration S* to the candidate new configuration Si is determined by an acceptance probability function PT(S*→Si), which involves the OF values and temperature (T). The decision rule allows SSA to avoid being trapped in local optima by occasionally accepting worse configurations. Typically, the neighbourhood search results in minimal changes to the current optimal configuration, wherein only one randomly chosen monitoring location si* within its neighbourhood is altered. The initial configuration S0 is typically randomized by removing k locations from the original sampling scheme.

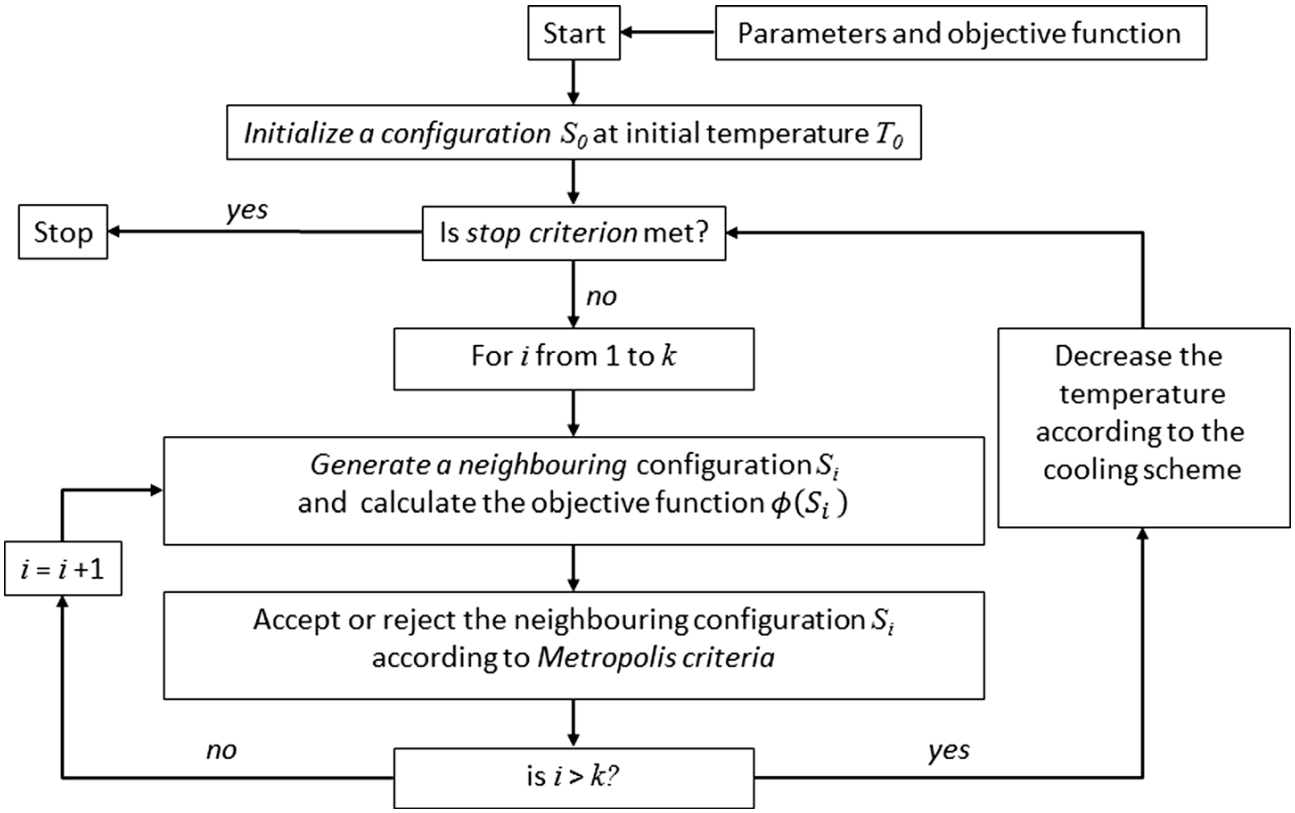


Fig. 2. The SSA working flowchart (adapted from Lobato et al. 2012)

Alternatively, a smart initialization, involving the application of a Greedy heuristic (Drosou and Pitoura 2009) before starting SSA, can be applied. This heuristic provides a sub-optimal configuration quickly, which can be used as S0 by SSA. The computational load increase due to the application of the Greedy heuristic is offset by the faster convergence of SSA. SSA continues to run until a specified

stop criterion is met such as maximum iterations, persistence of the optimal transient solution, and temperature threshold of the cooling scheme.

2.5.Variograms

Under the name of (semi)variogram, two different types of functions related to geographically referenced data are usually denoted. The first type of function, the *experimental variogram*, is discrete and represents the plot of the semivariance computed at different lags. The second type of function, the *variogram model*, is parametric, continuous and conditionally negative definite (Castellini et al., 2019). Model's parameters are the *nugget effect*, the *partial sill*, σ^2 (that is the maximum variability degree of the process net of noise) and the *range* (α) is the distance beyond which the spatial correlation disappears. The Matèrn model has a further parameter with respect the other models, k, called the *shape* while $\mathcal{K}_k$ is the modified Bessel function of the second kind. All the reported parameters are positive for definition. The models used in the present study are reported in the following table in their analytical form.

Table 1. Theoretical variograms reported in the present study.

Model	Variogram	Parameter Restrictions
Spherical	$\gamma(h) = \sigma^2\left[(1 - h/\alpha)^2 + (1 + h/2\alpha)\right]$	
Gaussian	$\gamma(h) = \sigma^2(1 - exp\,(-({}^{h}/_{\alpha})^2))$	
Matérn	$\gamma(h) = \sigma^2[1 - (\alpha h)^k \mathcal{K}_k(\alpha h)]$	$k > 0$

2.6.Automated variogram modelling

In the sampling downsizing optimization, a key role is played by the variogram model. In this study, the model selection was based on an optimization tool which lists the theoretical models that best fit

the experimental variogram according to an objective function which expresses the difference between the experimental variogram and the candidate model.

Automated, omnidirectional variogram model fitting takes place according to the following steps:

1. Let D be 0.35 times the length of the diagonal of the box that spans the data, i.e.

$$D = 0.35\sqrt{dx^2 + dy^2} \quad \text{(eq. 1)}$$

with dx the range of x coordinates and dy the range of y coordinates of the data points;

2. for p distance intervals with boundaries 0, 2, 4, 6, 9, 12, 15, 25, 35, 50, 65, 80, and 100% of D, the classical omnidirectional sample semi-variogram values $\gamma_{exp}(h_i)$ at each h_i are computed as the average distance of all point pairs falling within the distance interval i-th;

3. if the number of point pairs for lag i, N_i, in the ith lag interval is smaller than 5, two distance intervals are merged, the sample variogram recomputed and p lowered by one;

4. for the variogram model parameters as initial values the following are considered for the fit:

- sill: mean of the maximum and median value of the $\gamma_{exp}(h_i)$;
- nugget: minimum value of the $\gamma_{exp}(h_i)$;
- range: D/3.5, which is 10% of the length of the diagonal of the rectangle enveloping the data;
- a list of candidate values for the smoothness parameter k of the Matérn variogram model is provided. For each variogram model type contained in a predefined list, the candidate variogram model is tuned using a Gauss–Newton algorithm by minimising the weighted error sum of squares

$$SSErr = \sum_{i=1}^{N} \frac{N_i}{h^2{}_i} \left(\gamma_{exp}(h_i) - \gamma_{theor}(h_i)\right)^2 \quad \text{(eq. 2)}$$

with $\gamma_{theor}(h_i)$ the variogram model value for distance h_i;

5. Finally, the model (model type, nugget, sill, range, k in case of the Matérn model) with the smallest *SSerr* is selected.

These heuristic steps have been presented earlier in Hiemstra et al. (2009) and some of them in Pebesma (2004, 2005) and Pebesma and Wesseling (1998). In the present work, the model selection is based on a facility available in the R environment (library automap, function autofitvariogram).

2.7.Sampling scheme optimization process

A function of the Average Kriging Variance (AKV) was considered as objective function of the sampling scheme optimization process. Kriging estimation variance is a value provided by the kriging system in association to the predicted value and represents the uncertainty attached to the prediction. The adopted objective function follows the lines provided by Delhomme (1978) who described a suitable function for assessing an optimal reduced configuration on the basis of the areal average uncertainty. The relative change of the average kriging variance estimation is defined as:

$$\hat{\phi}_{AKV} = \frac{AKV_{Init} - AKV_i}{AKV_{Init}} \qquad \text{(eq. 3)}$$

where AKV_i is the average kriging variance assessed on a previously defined estimation grid and associated to the i-th reduced configuration and AKV_{Init} is the corresponding variance associated to the original complete sampling scheme. The relative change of average kriging variance can be interpreted as the quantity of information lost due to the scheme reduction.

A useful indicator to quantify the degradation degree of the information downstream the reduction process, i.e. the deviation from the full sampling design, is the ***efficiency index***,

$$Efficiency\ Index = \frac{AKV_{Init}}{AKV_i} \qquad \text{(eq. 4)}$$

which is a value lesser than 1 because we expect that after a reduction the AKV will grow.

In this study, we evaluated a progressive reduction in the sampling points starting from a number of 10 to 20 observations. Moreover, in order to assess the effect of information loss due to the sampling scheme reduction, theoretical variograms derived from the reduced sets of observations were computed and the goodness of fitting evaluated and analysed.

294 *2.8.Validation process*

295 After the reduction of the sampling scheme, a check is due to control the actual redundancy of the
296 excluded points by predicting them from the reduced scheme (Barca et al., 2016). For this aim, a set
297 of error metrics are computed, which provides the means to analyse the impact of the reduction.
298 Generally, it is difficult to assign a qualitative degree (excellent, good or poor) to the goodness of fit
299 since error indices generally lack a superior limit, this problem can be overcome by means of Equation
300 5:

$$\mathrm{MAPE\%} = \frac{1}{n}\sum_{i=1}^{n}\frac{|z^*(x_i) - z(x_i)|}{z(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{|r_i|}{z(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i^a}{z(x_i)} * 100 \qquad \text{(eq. 5)}$$

301 The Mean absolute percentage error (MAPE) index is dimensionless since any residual is rescaled in
302 terms of the corresponding observed value. Since it is an average relative value, it is possible to define
303 thresholds of goodness of fit; for example, if MAPE is less than 0.1 (or 10% if it is expressed as a
304 percentage), this can be considered a good result according to the Lewis scale (Lawrence and
305 Klimberg, 2016; Barca et al., 2017). The R library used for using the error metrics mentioned above
306 was {Metrics}.
307 The results of the validation allow also judging the kind of variogram model better suited to drive the
308 reduction process. Moreover, the regression-kriging was applied (Barca et al., 2019) as it enables
309 exploiting the ancillary information, and thus further improves the prediction of the excluded points.
310 In this way, the effectiveness of the secondary information provided by proximal sensing could be
311 exploited.
312

313 *2.9.Lin's agreement (concordance) coefficient*

314 To assess the strength of the concordance between two series of data, the Lin's agreement coefficient
315 has been applied. Lin's agreement coefficient provides a measure of overall accuracy which takes
316 into account both bias correction and precision. Bias correction is calculated from two bias measures:
317 constant bias (location-shift) and systematic bias (scale-shift) (Barnhart et al., 2002). The precision

is provided by the Pearson's (r) correlation coefficient. The formula is the following (Stevenson et al., 2018):

$$\rho_c = \frac{2 \cdot r \cdot \sigma_X \cdot \sigma_Y}{\sigma_Y^2 + \sigma_X^2 + (\mu_Y - \mu_X)^2} \quad \text{(eq. 6)}$$

3. Results

3.1.Soil Organic Carbon map

Descriptive statistics for the primary variable, SOC, considering the complete dataset (71 points) are reported in Table 2. Basic statistics showed that the assumption of a normal distribution for the SOC data was reasonable as confirmed by skewness and kurtosis values and by Shapiro-Wilk test (p = 0.678). For this reason, data were not subjected to a normal transform.

Table 2. Summary statistics for SOC (g 100 g^{-1})

Variable	N	Mean	Sd	Min	Max	Skewness	Kurtosis
SOC	71	1.85	0.28	1.19	2.43	-0.21	-0.29

A spherical model with a range of 7 m was fitted to the experimental variogram of SOC. The results of the cross-validation test were satisfactory because Mean Error was quite close to 0 (ME = -0.00626) and the Mean Square Standardized Error was 0.85. The MSSE was out of the tolerance interval (0.91 - 1.08) probably due to the smoothing of the modelling (Carroll and Cressie, 1996; Chilès and Delfiner, 1999).

The spatial distribution of SOC estimated by ordinary kriging (OK) showed larger average values in the southern part of the field (Fig. 3). The coarse sampling scale likely determined the high degree of smoothness.

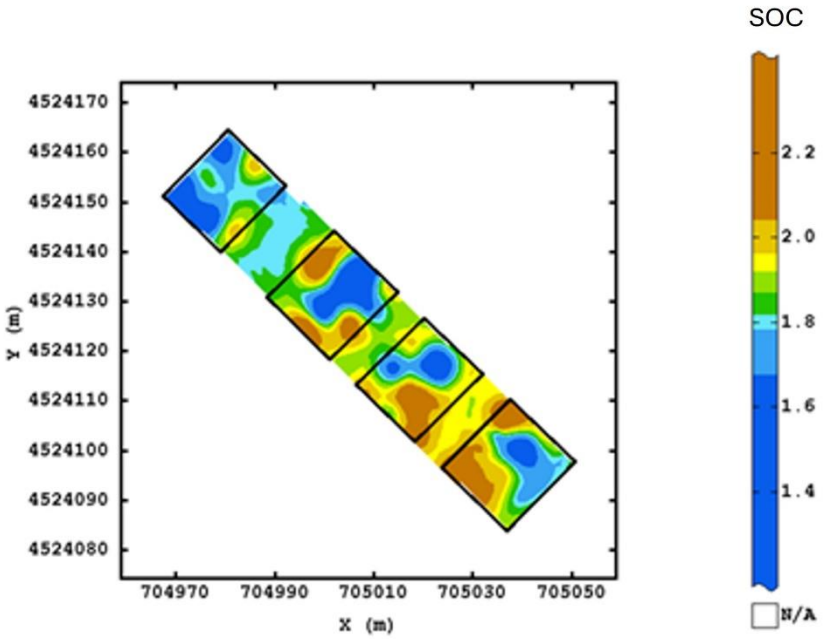


Fig. 3. Spatial estimates of SOC (g 100 g^{-1}). The black polygons indicate the four blocks in the RCB experimental design.

3.2.Comparison of the impact of different theoretical variogram models to assess spatial dependence of SOC values in the reduction of a previously defined sampling design

In order to compare the effect of the variogram model on the sampling scheme reduction, different theoretical models were compared. The variogram model optimization criterion (SS_{ERR}; eq. 2) found, for the SOC variable, Matérn and spherical as the best theoretical models for the complete dataset. Both models showed good adaptation to the experimental variogram, with a slightly lower SS_{ERR} for the Matérn model (0.00304 vs 0.00307). In accordance, slightly higher values of Lin's correlation concordance coefficients (eq. 6) and overall accuracy were observed for the Matérn model (0.59 and 0.91 vs 0.59 and 0.90, for Matérn and spherical models respectively).

Nugget variance for the nested spherical model fitted was zero. The two theoretical models found close sill variance (approximately 0.08) as shown in Table 3.

352 *Table 3. Parameters of the two theoretical nested models fitted to the experimental variogram of*
353 *SOC complete dataset.*

	Nested model	Partial sill	Range	Shape (k)
1	Nugget	0.01	0.0	
	Matérn	0.07	4.0	10
2	Spherical	0.08	7.7	

354

355 To initialize the iterative downsizing optimization process the parameters of the variogram model are
356 needed as an input.

357 The process was set up with a starting number of points to be dropped from the complete dataset (71
358 points) equal to 10, 15 and 20. The convergence graphs plot the values of the objective function
359 (average kriging variance, eq. 3) against the (log) temperature value assumed at each run. Inspecting
360 the convergence graphs within-models and between-models, it was possible to ascertain an average
361 behaviour of the convergence scheme for the two different models. Matérn model tended to get in
362 stuck to a suboptimal solution; in fact, better solutions were found at the initial steps of the process,
363 apart for the -20 downsizing stage. On the contrary, the spherical model appeared to aid the
364 convergence, showing to find always the global optimum. In Figure 4 the convergence graphs related
365 to the -15 downsizing stage are reported.

366

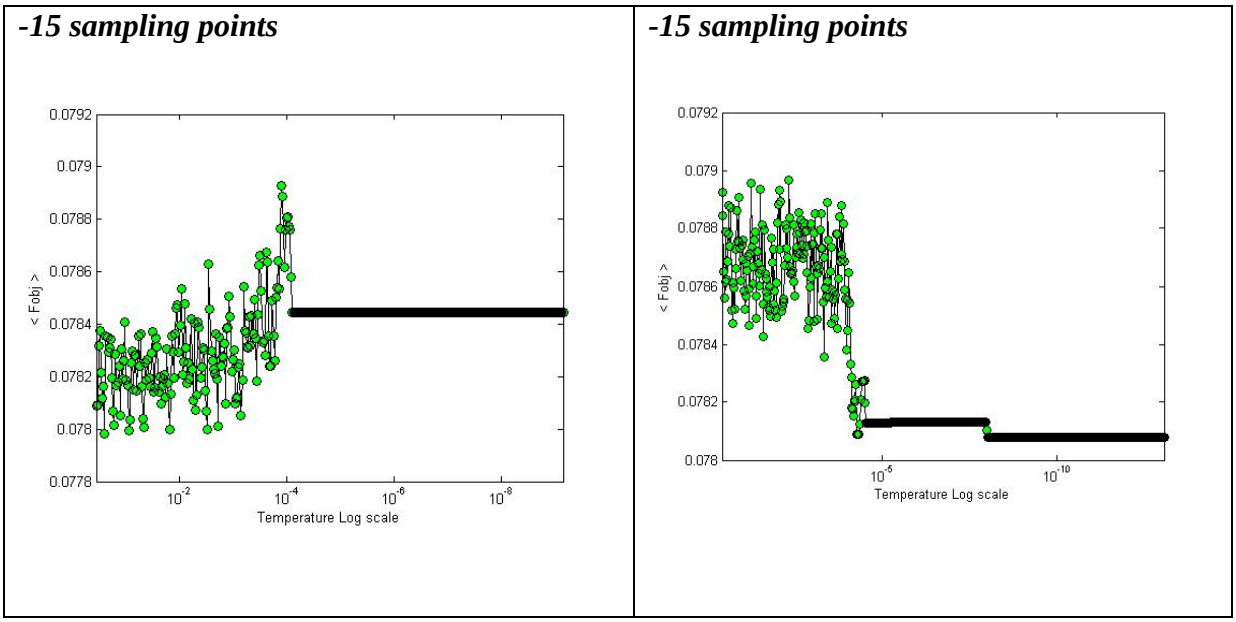


Fig. 4. Convergence graphs related to the two different variogram models, Matérn (left side) and spherical (right side), in the -15 sampling points downsizing process performed considering the primary variable (SOC).

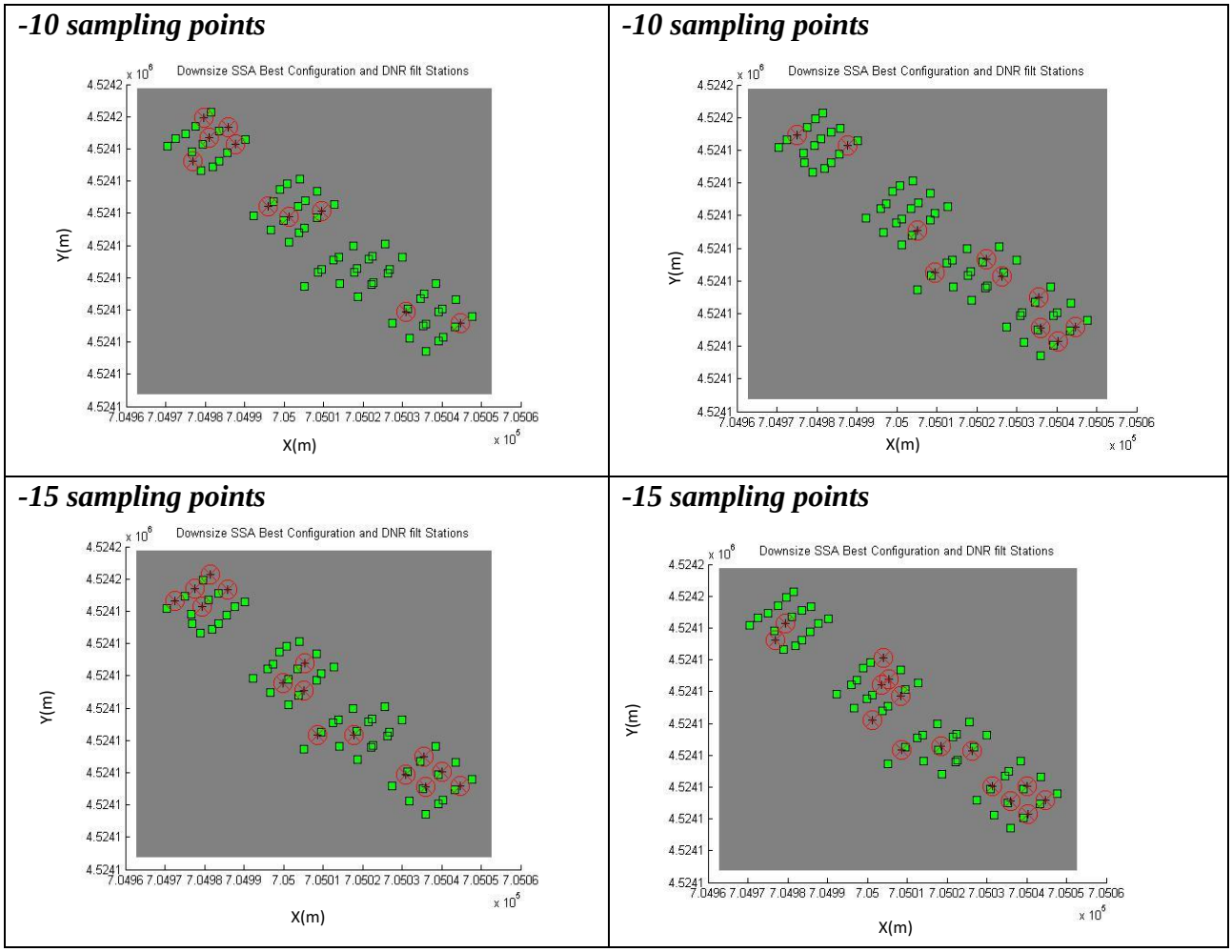
These results were confirmed by the values assumed by the efficiency indices (EI, eq. 4), which were slightly better for the spherical models and with an increasing deviation between the models at each run (Table 4). Therefore, the reduced schemes appeared to have a similar and narrow information loss.

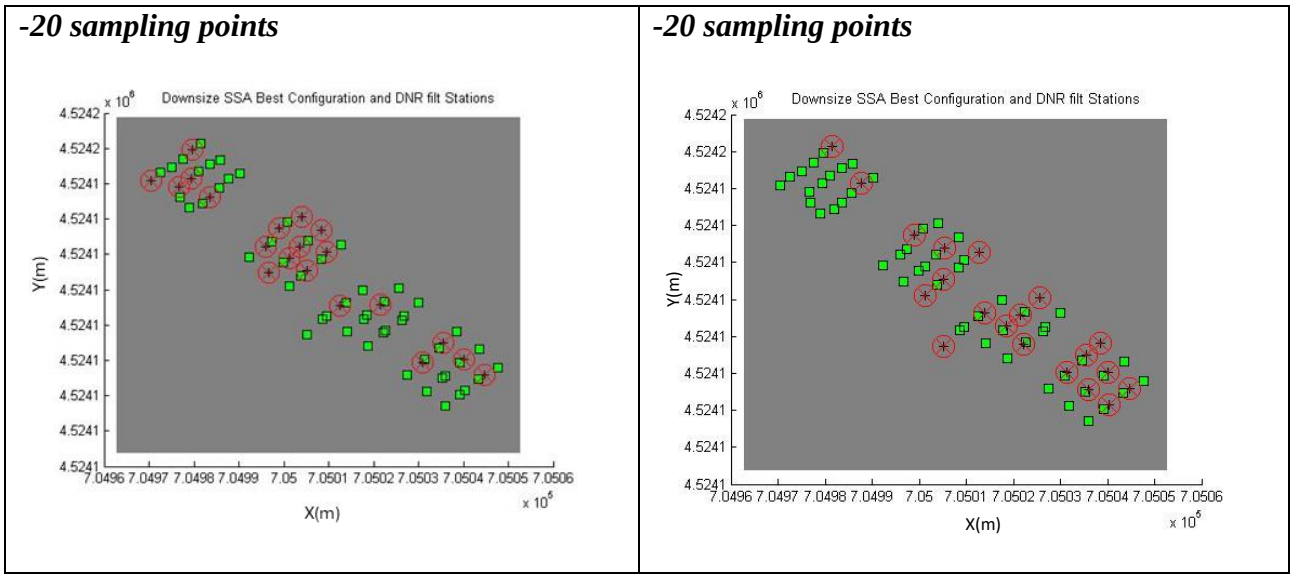
Table 4. Values of the efficiency index (EI) at convergence.

Nested models	**Downsizing processes (EI)**		
	-10	**-15**	**-20**
1 Nugget+Matérn	0.992	0.988	0.981
2 Spherical	0.996	0.991	0.988

In Figure 5 the optimized (reduced) sampling schemes, for the two theoretical models and to the increase of dropped sampling points, are reported; by comparing the two dropped sets of points there can be found low similarity between the schemes.

Therefore, the downsize following the two different variogram models resulted in two quite different optimized schemes, highlighting the role of the theoretical model applied in ruling the sampling reduction. Comparing the figures emerges major tendency of spherical variogram to almost equally distribute the dropped points over the different blocks with respect to the Matèrn variogram. The reduced sampling schemes showed in Figure 5 are gained without using the information contained in the GPR covariates but only SOC observations.





386 *Fig. 5. Optimal sampling configuration schemes obtained with the Matérn (left side) and spherical*

387 *(right side) models using only SOC observations.*

388 Table 5, where the number of sample points eliminated for each block is reported, clearly confirms

389 this result. An index of agreement to compare the different reduced configurations has been defined

390 as the cumulated absolute difference between the numbers of points dropped in each block (called

391 *delta (abs)*). These findings highlight that the selection of the appropriate variogram model is a critical

392 passage upstream the optimization process when a model-based optimization is pursued.

393

394 *Table 5. Comparison of the number of sampling points eliminated in each block in the reduced*

395 *sampling schemes obtained with the two theoretical models.*

	Block I	**Block II**	**Block III**	**Block IV**	Delta (abs)
Matérn (-10)	5	3	0	2	
Spherical (-10)	2	1	3	4	
deviations	3	2	3	2	10
Matérn (-15)	5	3	2	5	
Spherical (-15)	2	5	3	5	
deviations	3	2	1	0	6
Matérn (-20)	5	9	2	4	
Spherical (-20)	2	4	6	7	
deviations	3	5	4	3	15

396

3.3.Assessment of the possibility of using covariate information, substituting the primary variable, for the optimization process

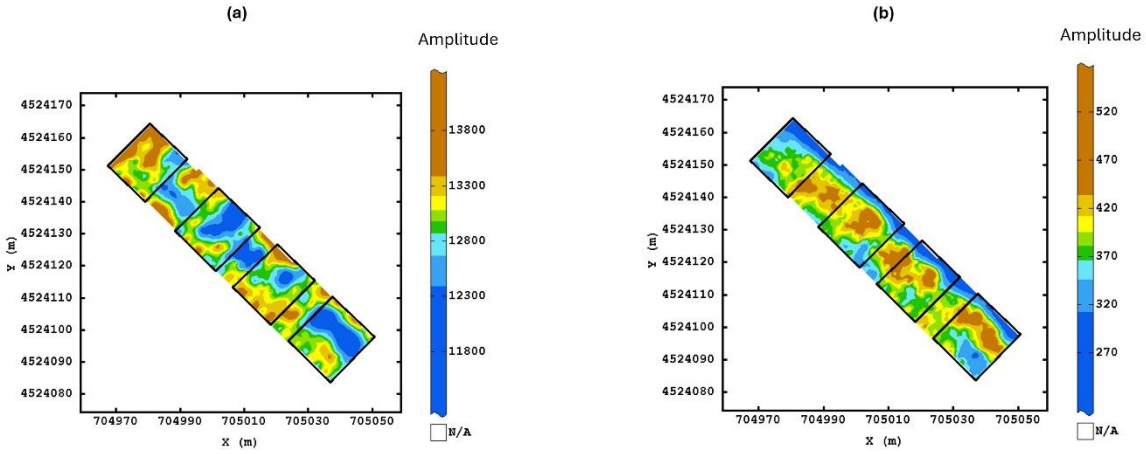
With the aim of assessing the possibility of using covariate information to support the optimization process for the primary variable, GPR data were considered.

The time slices for 600 MHz antenna were built within consecutive time windows of 2 to 10 ns, since the antenna showed a sensible attenuation of signal at about 10 ns (equal to about 0.50 m depth). GPR data at any time interval were correlated among them and this justified the multivariate approach. The data were previously transformed into standard Gaussian variables (since the hypothesis of normality was rejected) and a Linear Model of Co-regionalization was fitted to all experimental direct and cross-variograms with the following basic structures: Nugget effect and a Spherical model with a range of 9m. Mean error was quite close to 0 and the mean square standardized error was almost 1.

In Figure 6, only the results related to time slice at 2 ns (0.1 m depth) and 7 ns (0.35 m depth) were reported, because a discontinuity in the radar signal after 0.1m depth was observed and these time/depth maps were comparable with the depth of soil samples.

The amplitude map estimated at 0.10 m depth displayed lower values in the central part of the area. Starting from the 4ns time slice (0.2m-depth) until the 10ns time slice (0.50m-depth), the maps looked quite consistent with a wide central structure characterised by higher values. This may be indicative of some discontinuity along the vertical profile at 0.2m-depth.

The colour scale is always an iso-frequency scale to display the spatial structure, but the values of amplitude signal (in digital number) show the signal attenuation with the depth.



419 *Fig. 6. Time slices at 2ns and 7ns, corresponding to 0.1 m and 0.35 m depth, respectively (a and b).*
420 *Colour scale uses iso-frequency classes.*

421 From the visual inspection of the maps, apart the block in the northern area, a similar behaviour was
422 observed between the amplitude map estimated at 0.1 m depth and SOC spatial distribution (Fig. 3
423 and 6).
424 In order to identify the covariate variables most related to the primary variable (SOC), bivariate and
425 multivariate analysis approaches were used. To this aim, firstly the geostatistical estimates of the
426 selected geophysical variables were collocated into the point file by assessing, to each location and
427 for each variable, the nearest estimate. The relationships between SOC and GPR variables were first
428 explored through correlation analysis; afterwards a stepwise multiple linear regression analysis was
429 carried out to select the covariates most related to the primary soil property.
430 The analysis showed a highly significant relationship of SOC with amplitudes at 0.05m-depth, 0.1m-
431 depth and 0.5m-depth and a significant correlation with amplitudes at 0.15m-depth, 0.35m-depth and
432 0.4m-depth (Table 6).
433 Stepwise regression analysis results selected only amplitude at 0.1m-depth as significant variable for
434 explaining SOC variation ($F=14.64$, $P=0.0003$). For this reason, only this amplitude variable was
435 selected as proxy covariate for designing optimal scheme for SOC sampling.
436

437 *Table 6. Correlation matrix between SOC complete dataset and GPR variables.*

Pearson Correlation Coefficients, N=71, Prob > |r| under H0: Rho=0

	SOC	ckAmp 0.05m	ckAmp 0.1m	ckAmp 0.15m	ckAmp 0.2m	ckAmp 0.25m	ckAmp 0.3m	ckAmp 0.35m	ckAmp 0.4m	ckAmp 0.45m	ckAmp 0.5m
SOC	1										
ckAmp 0.05m	0.314 0.007	1									
ckAmp 0.1m	0.418 0.0003	0.668 <.0001	1								
ckAmp 0.15m	0.279 0.018	-0.043 0.716	0.705 <.0001	1							
ckAmp 0.2m	0.189 0.114	-0.308 0.009	0.483 <.0001	0.95 <.0001	1						
ckAmp 0.25m	0.186 0.120	-0.354 0.002	0.424 0.0002	0.928 <.0001	0.990 <.0001	1					
ckAmp 0.3m	0.079 0.507	-0.571 <.0001	0.144 0.2285	0.746 <.0001	0.876 <.0001	0.923 <.0001	1				
ckAmp 0.35m	-0.300 0.010	-0.891 <.0001	-0.544 <.0001	0.098 0.4126	0.334 0.0044	0.379 0.0011	0.607 <.0001	1			
ckAmp 0.4m	-0.236 0.046	-0.687 <.0001	-0.426 0.0002	0.084 0.482	0.260 0.028	0.251 0.034	0.318 0.006	0.791 <.0001	1		
ckAmp 0.45m	-0.125 0.2956	-0.829 <.0001	-0.289 0.0143	0.412 0.0004	0.614 <.0001	0.626 <.0001	0.707 <.0001	0.827 <.0001	0.830 <.0001	1	
ckAmp 0.5m	-0.308 0.008	-0.936 <.0001	-0.550 <.0001	0.152 0.203	0.395 0.0007	0.423 0.0002	0.587 <.0001	0.898 <.0001	0.780 <.0001	0.918 <.0001	1

438 In grey shadow, the correlations of the GPR variables with SOC are highlighted.

439

The variogram model optimization criterion (SS_{ERR}) found, for the GPR variable, the Gaussian and the spherical as the best theoretical models.

Both models showed good adaptation to the experimental variogram, with a slightly better SS_{ERR} for the Gaussian model (3.38 e+10 vs 4.92 e+10). In accordance, slightly higher values of Lin's correlation concordance coefficients and overall accuracy were observed for the Gaussian model (0.87 and 0.99 vs 0.83 and 0.96, for Gaussian and spherical models respectively).

The parameters of the two theoretical nested models are reported in Table 7 and were used to optimize the sampling design. Nugget variance was null for both nested models fitted.

Table 7. Parameters of the two theoretical nested models fitted to the experimental variogram of GPR data.

	Nested model	Partial sill	Range
1	Gaussian	257413.4	4.73
2	Spherical	249483.2	10.28

As for SOC sampling points' optimization, an iterative downsizing optimization process was set up with a starting number of points to be dropped equal to 10.

The downsize following the two different variogram models resulted in two quite different optimized schemes. Gaussian model, as observed also for Matérn model with the primary variable, seems to reach suboptimal solutions in comparison to the spherical model. The different behaviour is underlined by the baseline of the ordinate axis, which was always lower for the spherical model. The convergence graphs related to the -15 downsizing stage are reported as an example in the Figure 7.

-15 sampling points	***-15 sampling points***

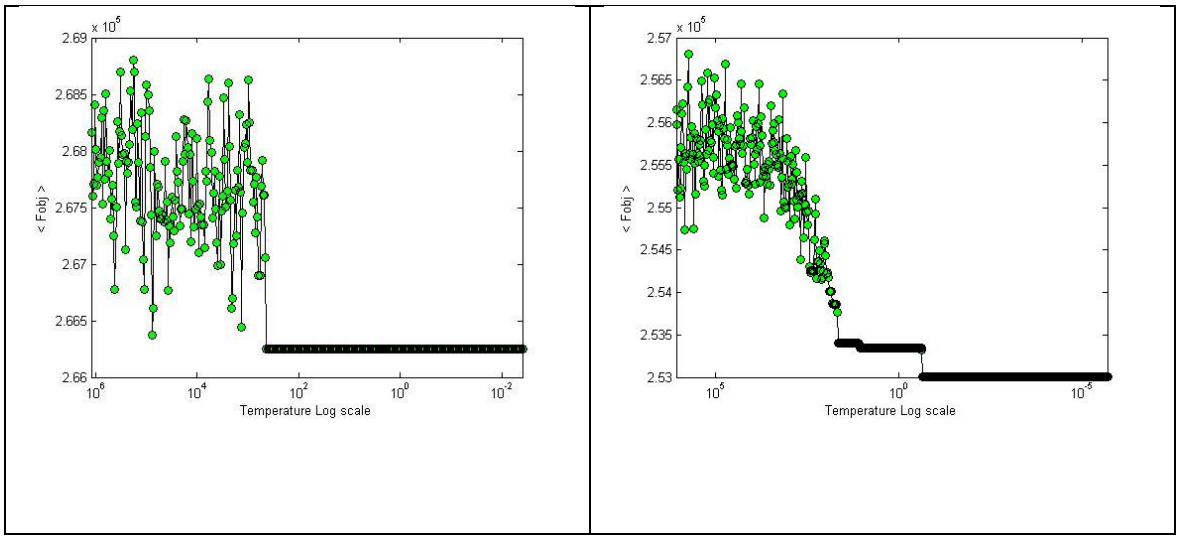


Fig. 7. Convergence graphs related to the two different variogram models, Gaussian (left side) and spherical (right side), in the -15 sampling points downsizing process, performed considering the covariate GPR variable.

The lower efficiency observed with the Gaussian variogram model is not completely confirmed by the values assumed by the efficiency indices, reported in Table 8, which show always slightly better values for the Gaussian in comparison to the spherical model.

Table 8. Values of the efficiency index at convergence.

Nested models	**Downsizing processes (EI)**		
	-10	**-15**	**-20**
1Gaussian	0.997	0.992	0.989
2Spherical	0.993	0.987	0.980

In Figure 8 the reduced sampling schemes, for the Gaussian and spherical models to the increase of dropped sampling points, are reported; by comparing the two dropped sets of points there can be found a sufficient agreement between the schemes particularly with a lower number of points eliminated. Table 9, where the number of sample points eliminated in each block is reported, confirms this result, with only 4 and 6 deviations when 10 and 15 points were dropped, respectively, as shown by the delta (abs) index of agreement.

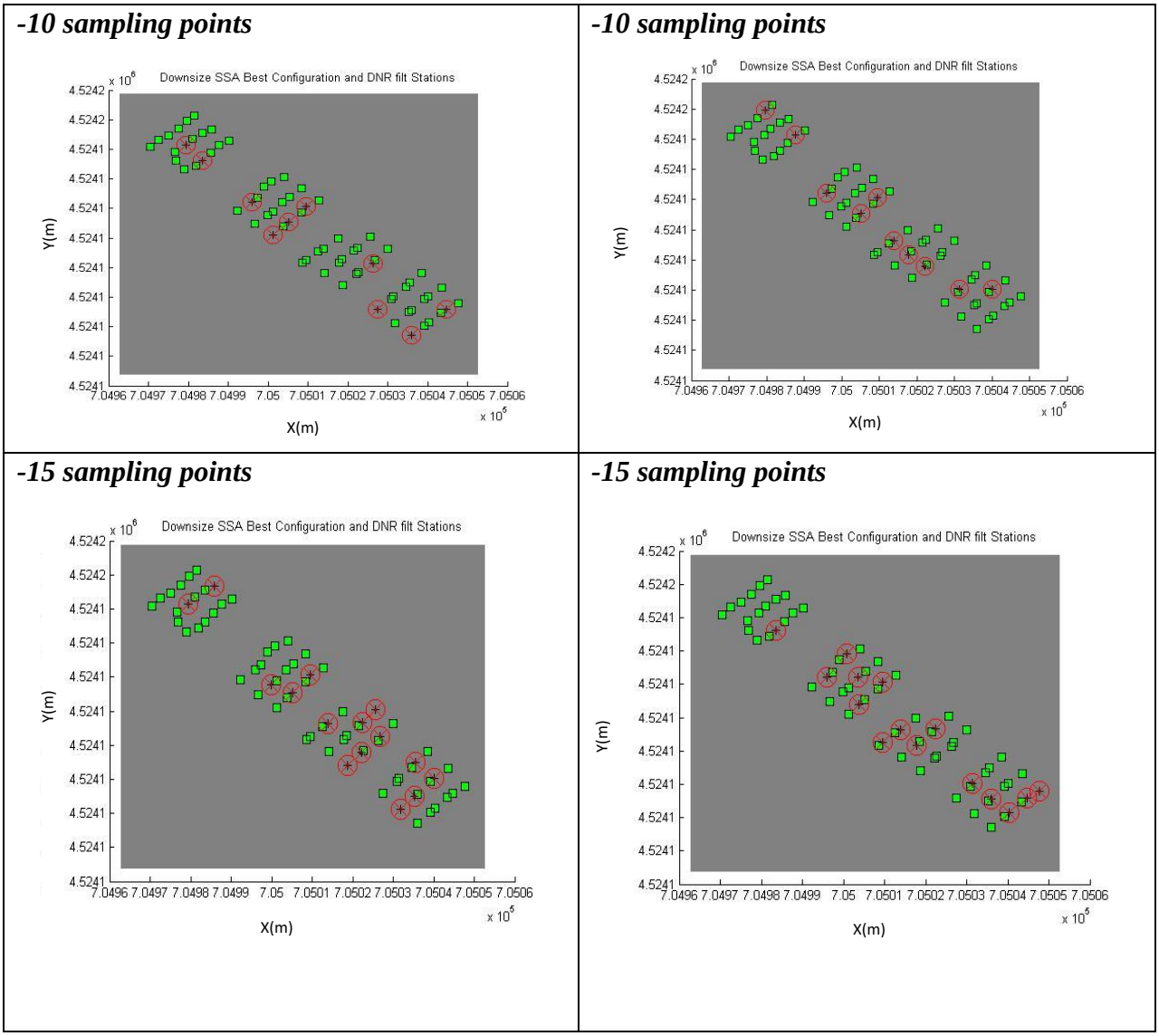


Fig.8. Optimal sampling configuration schemes obtained with the Gaussian (left side) and spherical (right side) models using the covariate GPR variable.

Table 9. Comparison of the number of sampling points eliminated in each block in the reduced sampling schemes obtained with the two theoretical models using the covariate GPR variable.

	Block I	**Block II**	**Block III**	**Block IV**	Delta (abs)
Gaussian (-10)	2	4	1	3	

Spherical (-10)	2	3	3	2	
deviations	0	1	2	1	4
Gaussian (-15)	2	3	6	4	
Spherical (-15)	1	5	4	5	
deviations	1	2	2	1	6
Gaussian (-20)	9	4	2	5	
Spherical (-20)	5	5	6	4	
deviations	4	1	4	1	10

Also, for the covariate, these findings highlight the critical role of the selection of the appropriate variogram model upstream the optimization process, when a model-based optimization is pursued. Comparing the reduced configurations garnered with the same variogram models but starting from the information provided by the primary (SOC) and the covariate (GPR) variables, it emerged a more similar behaviour with the use of the spherical model (Table 10) rather than that observed with the Matérn-Gaussian models (Table 11).

In particular, the first two configurations (-10 and -15 points), obtained with the spherical model, resulted very close each other indicating that the covariate variable can play an important role in substituting the primary variable and opening interesting perspectives for further development in this research field. The Gaussian-Matérn models, notwithstanding the good results obtained at the optimization stage, underlined by the efficiency indices values, showed different optimized configurations, confirming once more the impact of the variogram model adopted in the sampling scheme optimization.

Table 10. Number of sampling points eliminated in each block in the reduced sampling schemes. Comparison between SOC e GPR reduction schemes obtained with the Spherical model..

	Block I	**Block II**	**Block III**	**Block IV**	Delta (abs)

Spherical SOC (-10)	2	1	3	4	
Spherical GPR (-10)	2	3	3	2	
deviations	0	2	0	2	4
Spherical SOC (-15)	2	5	3	5	
Spherical GPR (-15)	1	5	4	5	
deviations	1	0	1	0	2
Spherical SOC (-20)	2	4	6	7	
Spherical GPR (-20)	5	5	6	4	
deviations	3	1	0	3	7

Table 11. Number of sampling points eliminated in each block in the reduced sampling schemes obtained with the Gaussian/Matérn models with the primary, SOC, and the covariate, GPR, variable.

	Block I	**Block II**	**Block III**	**Block IV**	Delta (abs)
Matérn SOC (-10)	5	3	0	2	
Gaussian GPR (-10)	2	4	1	3	
deviations	3	1	1	1	6
Matérn SOC (-15)	5	3	2	5	
Gaussian GPR (-15)	2	3	6	4	
deviations	3	0	4	1	8
Matérn SOC (-20)	5	9	2	4	
Gaussian GPR (-20)	9	4	2	5	
deviations	4	5	0	1	10

Finally, to assess the impact of sampling scheme reduction, theoretical variograms were fitted on the reduced sets of points using the same models selected in the previous data analysis step.

By comparing the theoretical nested models fitted to the experimental variogram of SOC data with a progressively reduced sampling points number (full dataset, -10, -15, -20 sampling points), as a

general behaviour, a constant decrease of the concordance indices was observed for both the considered models (Table 12). In details, an average decrease from 0.59 to 0.14 was recorded for the Lin's concordance coefficient for the two theoretical models. In addition, although starting from very close initial values (full dataset, 0.59 and 0.5847 for the Matérn and spherical model, respectively), to the reduction of sampling point number, the Matérn model showed a better performance with values always larger than those observed for the spherical model (Table 12). Matérn model showed thus to be more resilient to modifications of the operative conditions. The results show that, under our experimental conditions, the downsizing can be considered acceptable up to a discard of 15 sampling points (about 20% less), which caused a decrement of the Lin's CCC of about 24%. A further downsizing of other 5 sampling points caused a severe fall of the goodness of fitting, with a dramatic reduction of Lin's CCC of about 78%, in comparison to the full dataset.

Table 12. Parameters of the two theoretical nested models (Spherical, on the left side, and Matérn models, on the right side) fitted to the experimental variogram of SOC data with a progressively reduced sampling points number (full dataset, -10, -15, -20 sampling points).

	Full dataset				**Full dataset**			
	Nested model	partial sill	range		Nested model	partial sill	range	shape
1	Nugget	0.0	0.0	1	Nugget	0.0108	0.0	
2	Spherical	0.080	7.673	2	Matérn	0.071	3.986	10
	Overall CCC	0.584			Overall CCC	0.590		
	Overall precision	0.653			Overall precision	0.650		
	Overall accuracy	0.894			Overall accuracy	0.907		
	- 10 sampling points				**- 10 sampling points**			
	Nested model	partial sill	range		Nested model	partial sill	range	shape
1	Nugget	0.010	0.0	1	Nugget	0.018	0.0	
2	Spherical	0.075	6.951	2	Matérn	0.071	4.046	5
	Overall CCC	0.457			Overall CCC	0.504		
	Overall precision	0.588			Overall precision	0.603		
	Overall accuracy	0.777			Overall accuracy	0.835		
	- 15 sampling points				**- 15 sampling points**			
	Nested model	partial sill	range		Nested model	partial sill	range	shape
1	Nugget	0.0	0.0	1	Nugget	0.007	0.0	
2	Spherical	0.053	8.506	2	Matérn	0.062	4.332	10
	Overall CCC	0.418			Overall CCC	0.447		
	Overall precision	0.521			Overall precision	0.545		
	Overall accuracy	0.802			Overall accuracy	0.821		
	- 20 sampling points				**- 20 sampling points**			
	Nested model	partial sill	range		Nested model	partial sill	range	shape
1	Nugget	0	0	1	Nugget	0	0	0
2	Spherical	0.068	8.129	2	Matérn	0.069	3.577	10
	Overall CCC	0.127			Overall CCC	0.148		
	Overall precision	0.163			Overall precision	0.180		
	Overall accuracy	0.780			Overall accuracy	0.820		

By comparing the theoretical nested models fitted to the experimental variogram of GPR data, a similar general behaviour was observed (Table 13). In details, a decrease from an average value of 0.84 to 0.60 was recorded for the Lin's concordance coefficient. Even in this case, although starting from very close initial values (full dataset, 0.87 and 0.83 for the Gaussian and spherical model, respectively), to the reduction of sampling point number, the Gaussian model showed a better performance with values of the coefficients always larger than those observed for the spherical model (Table 13). Gaussian model, similarly to the Matérn model, showed thus to be more resilient to modifications of the operative conditions with respect to the spherical model. The downsizing showed optimal conditions up to a discard of 15 sampling points, with an average decrement of the Lin's CCC of about 14%. A further downsizing caused a decrease of the goodness of fitting, with a reduction of Lin's CCC of about 28%, in comparison to the full dataset. The last reduction impacted more severely on the spherical model, where a decrement of about 36% was recorded.

Table 13. Parameters of the two theoretical nested models (Spherical, on the left side, and Matérn models, on the right side) fitted to the experimental variogram of GPR data with a progressively reduced sampling points number (full dataset, -10, -15, -20 sampling points).

	Full dataset				***Full dataset***		
	Nested model	*partial sill*	*range*		*Nested model*	*partial sill*	*range*
1	*Nugget*	*0.0*	*0.0*	*1*	*Nugget*	*0.0*	*0.0*
2	*Spherical*	*249483.2*	*10.276*	*2*	*Gaussian*	*257413.4*	*4.729*
	Overall CCC	*0.829*			*Overall CCC*	*0.868*	
	Overall precision	*0.865*			*Overall precision*	*0.875*	
	Overall accuracy	*0.958*			*Overall accuracy*	*0.991*	
	- 10 sampling points				***- 10 sampling points***		
	Nested model	*partial sill*	*range*		*Nested model*	*partial sill*	*range*
1	*Nugget*	*0.0*	*0.0*	*1*	*Nugget*	*0.0*	*0.0*
2	*Spherical*	*272137.4*	*10.681*	*2*	*Gaussian*	*294255.2*	*4.823*
	Overall CCC	*0.780*			*Overall CCC*	*0.858*	
	Overall precision	*0.820*			*Overall precision*	*0.870*	
	Overall accuracy	*0.951*			*Overall accuracy*	*0.986*	
	- 15 sampling points				***- 15 sampling points***		
	Nested model	*partial sill*	*range*		*Nested model*	*partial sill*	*range*
1	*Nugget*	*0.0*	*0.0*	*1*	*Nugget*	*0.0*	*0.0*
2	*Spherical*	*265827.9*	*11.574*	*2*	*Gaussian*	*251744.8*	*5.294*
	Overall CCC	*0.703*			*Overall CCC*	*0.745*	
	Overall precision	*0.746*			*Overall precision*	*0.762*	
	Overall accuracy	*0.941*			*Overall accuracy*	*0.978*	
	- 20 sampling points				***- 20 sampling points***		
	Nested model	*partial sill*	*range*		*Nested model*	*partial sill*	*range*
1	*Nugget*	*0*	*0*	*1*	*Nugget*	*0*	*0*
2	*Spherical*	*231990.3*	*8.440*	*2*	*Gaussian*	*288129*	*4.794*
	Overall CCC	*0.533*			*Overall CCC*	*0.668*	
	Overall precision	*0.656*			*Overall precision*	*0.715*	
	Overall accuracy	*0.812*			*Overall accuracy*	*0.934*	

3.4.Final validation

As a conclusive test, the dropped 15 points, representing the optimal downsizing condition among those examined, were predicted starting from the reduced sampling dataset by means of the two different variogram models and the kriging predictor (ordinary kriging). The models used are those showed in Table 12 and 13 associated to the reduced schemes. In Table 14, a set of error metrics, related to the prediction of the dropped points for SOC values, is reported. Results can be considered nearly optimal and substantially equivalent for both models.

Table 14. Error indices of SOC dropped points estimation under different models.

Model	MBE	RMSE	MAE	MAPE%	MAX	MIN
Matèrn	-0.02	0.24	0.20	11	0.35	-0.42
Spherical	0.09	0.24	0.18	11.8	0.54	-0.25

Table 15 reports error metrics related to the reduction of GPR point scheme. The garnered values could be misleading as the variable has a large order of magnitude; therefore, to express a correct judgement the focus should be put on MAPE% (eq. 5). Such index highlights the substantial equivalence between the two models. MAPE% showed relative errors around 1% that can be considered a negligible quantity.

Table 15. Error indices of GPR dropped points estimation under different models.

Model	MBE	RMSE	MAE	MAPE%	MAX	MIN
Gaussian	-76.84	234.32	156.84	1.2	232.69	-747.48
Spherical	11.1	180.38	140.86	1.1	478.49	-202.61

Finally, regression-kriging was applied to exploit the secondary information in predicting the points dropped from SOC scheme. According to the metrics values reported in Table 16, the spherical model seems more capable to take advantage from exploiting secondary information than Matèrn model. A

possible explanation can be that the points being dropped are better selected during downsize process by the spherical model rather than by Matèrn one. Another more interesting explanation highlights larger similarities between the reduced sets of primary and secondary variables, respectively, when spherical model is considered as a driver of the reduction process (Table 10). On the contrary, the reduced sets are quite different between primary and secondary variables if the driver is Matèrn model (Table 11).

Table 16. Error indices of SOC dropped points estimation under different models with regression-kriging.

Model	MBE	RMSE	MAE	MAPE%	MAX	MIN
Matèrn	-0.05	0.25	0.20	11	0.36	-0.53
Spherical	0.06	0.20	0.16	10.3	0.40	-0.32

In conclusion, it appears that the spherical model approximates slightly worse the experimental variogram of the reduced sampling scheme but selects better the redundant points to be reduced, whereas Matèrn seems to approximate better the experimental variogram after the sampling scheme reduction but is less effective to select actually redundant points from the original sampling scheme. These results open interesting perspectives to future studies.

4. Discussion

Appropriate sampling strategy and data analysis play a crucial role in agronomic and environmental field research to assess i) the variability that occurs at a scale smaller than the block size in traditional experimental designs; ii) the spatial dependence of observations and residuals. This information is invaluable to analyse the treatment significance.

This study was aimed at defining advances in sampling design optimization for a key soil variable, SOC, investigating the effect of the choice of the variogram model and of the use of covariate

583 information when a downsizing of a sampling scheme was pursued. In addition, a strategy to assess
584 the optimal downsizing, based on the variogram fitting, was defined.
585 Gaussian and Matérn models showed to be more resilient to modifications of the operative conditions
586 with respect to the spherical model. The better performance of the Matérn and Gaussian models can
587 be attributed to their greater flexibility with respect to the spherical model. These findings highlight
588 the critical role of the selection of the appropriate variogram model upstream the optimization
589 process, when a model-based optimization is pursued.
590 The results showed that, under our experimental conditions, the downsizing can be considered
591 acceptable up to a discard of 15 sampling points (about 20% less), which caused a decrement of the
592 Lin's CCC of about 24% for SOC. A further downsizing of other 5 sampling points caused a severe
593 fall of the goodness of fitting, with a dramatic reduction of Lin's CCC of about 78%, in comparison
594 to the full dataset.
595 Comparing the reduced configurations garnered with the same variogram models but starting from
596 the information provided by the primary (SOC) and the covariate (GPR) variables, it emerged a more
597 similar behaviour with the use of the spherical model rather than that observed with the Matérn-
598 Gaussian models. The configurations (less 10 and 15 points) obtained with the spherical model
599 resulted very close each other indicating that the covariate variable can play an important role in
600 substituting the primary variable and opening interesting perspectives for further development in this
601 research field. The Gaussian-Matérn models, notwithstanding the good results obtained at the
602 optimization stage, underlined by the efficiency indices values, showed different optimized
603 configurations, confirming once more the impact of the variogram model adopted in the optimization
604 of the sampling scheme.
605 The methodology proposed was assessed within an agronomic field experiment. The issue of the
606 optimization of the sampling design, by minimizing at the same time the information loss, is in fact
607 particularly relevant in agronomic research, where the need to assess the spatial correlation of the
608 investigated variables, both for improving data analysis methods and for spatial interpolation, requires

609 the collection of large datasets with georeferenced data. The size of the study area does not affect the
610 methodology proposed whose applicability depends on the presence of a significant spatial
611 correlation structure for the primary variable and on the availability of appropriate covariate variables.
612

613 ## 5. Conclusions

614 The primary aim of this study was to optimally reduce an initially oversized sampling plan for
615 assessing soil total organic carbon (SOC). Our analysis demonstrated that it was possible to reduce
616 the original sampling scheme by approximately 20% — equivalent to eliminating 15 sampling points
617 out of 71 — without compromising the predictive capability of the sampling scheme. As concerns
618 the reduced configuration, the spherical variogram model proved to be the most effective in
619 identifying and removing redundant sampling points. This reduction process preserved the integrity
620 and precision of the SOC assessment. By maintaining robust predictive performance despite the
621 reduction in sample size, we have shown that the optimized sampling plan effectively balances
622 efficiency with informational accuracy.
623 The results of our investigation showed that the GPR covariate had an important role both in
624 improving the optimization of the sampling scheme for SOC and in replacing the primary variable to
625 optimize the sampling scheme. Therefore, the use of relatively easy, accurate, and inexpensive
626 geophysical methods can represent a viable solution to obtain the information needed to advance the
627 precision agriculture.
628

629 **Author Contributions**

630 Conceptualization, A.M.S., E.B. and D.D.B.; methodology, E.B.; software, E.B.; validation, E.B.,
631 D.D.B., A.M.S.; formal analysis, E.B., A.M.S. and D.D.B.; investigation, D.D.B..; data curation,
632 D.D.B.; writing—original draft preparation, E.B.; writing—review and editing, D.D.B., A.M.S.,
633 E.B.; visualization, D.D.B.; supervision, A.M.S. and E.B.; project administration and funding
634 acquisition, A.M.S. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

This study was supported by the EU and MIUR in the frame of the collaborative international consortium DESERT “Low-cost water desalination and sensor technology compact module” under the ERANET 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).

References

Abo-habaga, M. M., Ismail, Z. E., Moustafa, N. E., & Okasha, M. H. (2024). Development of an Automated Cartesian Arm for Planting Seeds in Pots. Journal of Soil Sciences and Agricultural Engineering, 15(7), 217-224. https://dx.doi.org/10.21608/jssae.2024.305809.1238

Atkinson, P.M., Lloyd, C.D., 2007. Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data. Computers & Geosciences 33(10), 1285-1300. https://doi.org/10.1016/j.cageo.2007.05.011

Barca, E., De Benedetto, D., Stellacci, A.M., 2019. Contribution of EMI and GPR proximal sensing data in soil water content assessment by using linear mixed effects models and geostatistical approaches. Geoderma 343: 280-293. https://doi.org/10.1016/j.geoderma.2019.01.030

Barca, E., Bruno, D. E., Passarella, G., 2016. Optimal redesign of environmental monitoring networks by using software MSANOS. Environmental Earth Sciences 75(14), 1082. https://doi.org/10.1007/s12665-016-5885-x

Barca E., Porcu, E., Bruno, D., Passarella, G., 2017. An automated decision support system for aided assessment of variogram models. Environmental modelling & software, 87, 72-83.

Barca, E., Passarella, G., Vurro, M., Morea, A., 2015. MSANOS: Data-Driven, Multi-Approach Software for Optimal Redesign of Environmental Monitoring Networks. Water Resour. Manage. 29(2), 619–644. https://doi.org/10.1007/s11269-014-0859-9

661 Barnhart, H.X., Haber, M., Song, J., 2002. Overall concordance correlation coefficient for evaluating
662 agreement among multiple observers. Biometrics 58(4), 1020-1027.
663 https://doi.org/10.1111/j.0006-341X.2002.01020.x
664 Carroll, S.S., Cressie, N.A., 1996. Comparison of geostatistical methodologies used to estimate snow
665 water equivalent. Water Resour. Bull. 32 (2), 267–278. https://doi.org/10.1111/j.1752-
666 1688.1996.tb03450.x
667 Chilès, J.P., Delfiner, P., 1999. Geostatististics: Modelling Spatial Uncertainty. Wiley, New York,
668 NY.
669 Cassidy, N.J., 2009. Ground penetrating radar data: processing, modelling and analysis. In Ground
670 Penetrating Radar theory and applications. Edited by Harry M. Jol, Elsevier. ISBN: 978-0-444-
671 53348-7.
672 Castellini, M., Stellacci, A. M., Tomaiuolo, M., Barca, E., 2019. Spatial Variability of Soil Physical
673 and Hydraulic Properties in a Durum Wheat Field: An Assessment by the BEST-Procedure.
674 Water 11(7), 1434. https://doi.org/10.3390/w11102185.
675 Colovic, M., Stellacci, A.M., Mzid, N., Di Venosa, M., Todorovic, M., Cantore, V., Albrizio, R.,
676 2024. Comparative Performance of Aerial RGB vs. Ground Hyperspectral Indices for
677 Evaluating Water and Nitrogen Status in Sweet Maize. AGRONOMY, 14.
678 10.3390/agronomy14030562Daniels, D.J., 2004. Ground Penetrating Radar, 2nd ed.; The
679 Institution of Engineering and Technology: London, UK.
680 Davis, J.L., Annan, A.P., 1989. Ground-penetrating radar for high-resolution mapping of soil and
681 rock stratigraphy. Geophys. Prospect. 37, 531–551. https://doi.org/10.1111/j.1365-
682 2478.1989.tb02221.x.
683 De Benedetto, D., Quarto, R., Castrignano, A., Palumbo, D.A., 2015. Impact of data processing and
684 antenna frequency on spatial structure modelling of GPR data. Sensors 15, 16430-16447.
685 10.3390/s150716430.

De Benedetto, D., Barca, E., Castellini, M., Popolizio, S., Lacolla, G., Stellacci, A.M., 2022. Prediction of Soil Organic Carbon at Field Scale by Regression Kriging and Multivariate Adaptive Regression Splines Using Geophysical Covariates. Land 11(3), 381. https://doi.org/10.3390/land11030381.

Delhomme, J.P., 1978. Kriging in the hydrosciences. Advances in water resources 1, 251-266. https://doi.org/10.1016/0309-1708(78)90039-8.

Deutsch, C.V., Cockerham, P.W., 1994. Practical considerations in the application of simulated annealing to sSOChastic simulation. Mathematical Geology 26(1), 67-82. https://doi.org/10.1007/BF02065876.

Drosou, M., Pitoura, E., 2009. Comparing diversity heuristics. Computer Science Department, University of Ioannina.

Geovariances, 2015. Isatis Technical Ref., ver. 2015.1 Geovariances & Ecole Des Mines De Paris. Avon Cedex, France.

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Applied Geostatistics Series. Oxford Univ. Press, Oxford. 483 pp.

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J., Heuvelink, G.B., 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Computers & Geosciences 35(8), 1711-1721. https://doi.org/10.1016/j.cageo.2008.10.011.

Jalilian, A., Mateu, J., & Sedda, L. (2024). A brief review and guidance on the spatiotemporal sampling designs for disease vector surveillance. Current Research in Parasitology & Vector-Borne Diseases, 100208.

Lawrence, K.D., Klimberg, R.K., 2016. Vol. 11 of Advances in business and management forecasting. Bingley, UK: Emerald Group of Publishing.

Lawrence, P.G., Roper W., Morris T.F., Guillard K., 2020. Guiding soil sampling strategies using classical and spatial statistics: A review. Agronomy Journal. 2020; 112: 493–510. https://doi.org/10.1002/agj2.20048

712 Littell, R.C. et al. 2006. SAS system for mixed models. SAS Inst., Cary, NC.
713 Liu, X., Chen, J., Cui, X., Liu, Q., Cao, X., Chen, X., 2019. Measurement of soil water content using
714 ground-penetrating radar: A review of current methods. Int. J. Digit. Earth 12, 95–118.
715 https://doi.org/10.1080/17538947.2017.1412520
716 Lobato, F.S., Assis, E.G., Steffen, Jr V., da Silva Neto, A.J., 2012. Design and identification problems
717 of rotor bearing systems using the simulated annealing algorithm. In: Sales Guerra Tsuzuki M
718 (ed) Simulated Annealing— single and multiple objective problems. ISBN: 978-953-51-0767-
719 5, InTech. doi: 10.5772/47833
720 Matheron, G., 1963. Principles of Geostatistics. Economic Geology 58, 1246-1266.
721 McNeill, J.D., 1980. Electromagnetic terrain conductivity measurement at low induction numbers.
722 Geonics Limited, Technical Note TN 6, Geonics Ltd. Mississauga, Ontario, Canada.
723 Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state
724 calculations by fast computing machines. The journal of chemical physics 21(6), 1087-1092.
725 https://doi.org/10.1063/1.1699114
726 Mzid, N., Boussadia, O., Albrizio, R., Stellacci, A.M., Braham, M., Todorovic, M., 2023. Salinity
727 Properties Retrieval from Sentinel-2 Satellite Data and Machine Learning Algorithms.
728 AGRONOMY, 13:716. 10.3390/agronomy13030716Pardo-Igúzquiza, E., 1998. Optimal
729 selection of number and location of rainfall gauges for areal rainfall estimation using
730 geostatistics and simulated annealing. Journal of Hydrology 210 (1-4), 206-220.
731 https://doi.org/10.1016/S0022-1694(98)00188-7
732 Pathirana, S., Lambot, S., Krishnapillai, M., Cheema, M., Smeaton, C., Galagedara, L., 2023.
733 Ground-Penetrating Radar and Electromagnetic Induction: Challenges and Opportunities in
734 Agriculture. Remote Sens. 15, 2932. https://doi.org/10.3390/ rs15112932
735 Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences
736 30 (7), 683-691. https://doi.org/10.1016/j.cageo.2004.03.012

737 Pebesma, E.J., 2005. Mapping radioactivity from monitoring data: Automating the classical
738 geostatistical approach. Applied GIS 1(2), 1-10. https://doi.org/10.2104/AG050011
739 Pebesma, E.J., Wesseling, C.G., 1998. Gstat: a program for geostatistical modelling, prediction and
740 simulation. Computers & Geosciences 24(1), 17-31. https://doi.org/10.1016/S0098-
741 3004(97)00082-4
742 Popolizio, S., Barca, E., Castellini, M., Montesano, F.F., Stellacci, A.M., 2022a. Investigating the
743 Spatial Structure of Soil Hydraulic Properties in a Long-Term Field Experiment Using the
744 BEST Methodology. Agronomy 12, 2873. https://doi.org/10.3390/agronomy12112873
745 Popolizio, S., Stellacci, A.M., Giglio, L., Barca, E., Spagnuolo, M., Castellini, M, 2022b. Seasonal
746 and Soil Use Dependent Variability of Physical and Hydraulic Properties: An Assessment under
747 Minimum Tillage and No-Tillage in a Long-Term Experiment in Southern Italy. Agronomy 12,
748 3142. https://doi.org/10.3390/agronomy12123142
749 Reynolds, J.M., 1997. An Introduction to Applied and Environmental Geophysics. Wiley, Chichester.
750 Samuel-Rosa, A., 2017. Spsann: Optimization of Sample Configurations using Spatial Simulated
751 Annealing. Retrieved from https://CRAN.R-project.org/package=spsann R package version
752 2.1–0.
753 Sherpa, S.R., Wolfe, D.W., van Es, H.M., 2016. Sampling and Data Analysis Optimization for
754 Estimating Soil Organic Carbon stocks in Agroecosystems. Soil Sci. Soc. Am. J. 80:1377–
755 1392. doi:10.2136/sssaj2016.04.0113.
756 Stevenson, M., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reiczigel, J., Robison-
757 Cox, J., Sebastiani, P., Solymos, P., Yoshida, K., Jones, G., Pirikahu, S., Firestone, S., Kyle,
758 R., Popp, J., Jay, M., 2018. epiR: Tools for the Analysis of Epidemiological Data. R package
759 version 0.9 - 99. https://CRAN.R-project.org/package=epiR.
760 Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of
761 spatial studies for ecosystem comparisons. Agriculture, Ecosystems & Environment 94(1), 31-
762 47. https://doi.org/10.1016/S0167-8809(02)00013-0

763 Tarr, A.B., Moore, K.J., Burras, C.L., Bullock, D. G., Dixon, P.M., 2005. Improving map accuracy
764 of soil variables using soil electrical conductivity as a covariate. Precision Agriculture 6(3),
765 255-270. https://doi.org/10.1007/s11119-005-1385-9
766 Warrick, A.W., Van Es, H.M., Dane, J.H., 2002. Soil sampling and statistical procedures. Methods
767 of soil analysis: Physical methods. 3rd ed. Madison: Soil Science Society of America, 1-200.
768 Zaman, Q. U. (2023). Precision agriculture technology: A pathway toward sustainable agriculture. In
769 Precision Agriculture (pp. 1-17). Academic Press.
770 Zhao, Y., Xianghua, X., Kang, T., Biao, H., Nan, H., 2016. Comparison of sampling schemes for the
771 spatial prediction of soil organic matter in a typical black soil region in China. Environ Earth
772 Sci (2016) 75:4. https://doi.org/10.1007/s12665-015-4895-4