Journal of Agricultural Engineering

https://www.agroengineering.org/

Exploiting soil spectroscopy in the VNIR-SWIR range to estimate soil organic carbon stock: what role can sampling density and spacing play?

Caterina Mazzitelli,[1] Nunzio Romano,[1] Eyal Ben-Dor,[2] Lis Wollesen de Jonge,[3] Cecilie Hermansen,[3] Paolo Nasta[1]

[1]Department of Agricultural Sciences, University of Naples Federico II, Portici (NA), Italy
[2]The Remote Sensing Laboratory, Tel Aviv University, Israel
[3]Department of Agroecology, Aarhus University, Denmark

Corresponding author: Paolo Nasta, Department of Agricultural Sciences, University of Naples Federico II, Piazza Carlo di Borbone 1, 80055 Portici (NA), Italy. E-mail: paolo.nasta@unina.it

Please cite this article as doi: 10.4081/jae.2026.1995

Submitted: 9 October 2025
Accepted: 19 June 2026

Exploiting soil spectroscopy in the VNIR-SWIR range to estimate soil organic carbon stock: what role can sampling density and spacing play?

Caterina Mazzitelli,[1] Nunzio Romano,[1] Eyal Ben-Dor,[2] Lis Wollesen de Jonge,[3] Cecilie Hermansen,[3] Paolo Nasta[1]

[1]Department of Agricultural Sciences, University of Naples Federico II, Portici (NA), Italy
[2]The Remote Sensing Laboratory, Tel Aviv University, Israel
[3]Department of Agroecology, Aarhus University, Denmark

Corresponding author: Paolo Nasta, Department of Agricultural Sciences, University of Naples Federico II, Piazza Carlo di Borbone 1, 80055 Portici (NA), Italy. E-mail: paolo.nasta@unina.it

Contributions: all the authors made a substantive intellectual contribution, read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: The authors declare no competing interests, and all authors confirm accuracy.

Availability of data and materials: the data used to support the findings of this study are available from the corresponding author on reasonable request.

Acknowledgments: This work was supported by the European Union - Next-GenerationEU - National Recovery and Resilience Plan (NRRP) – MISSION 4 COMPONENT 2, INVESTIMENT N. 1.1, CALL PRIN 2022 PNRR D.D. 1409 14-09-2022 – (Assessing and mapping novel agroecosystem vulnerability and resilience indicators in southern Italy-ASAP, CUP N. E53D23021880001).

Abstract

Soil organic carbon stock (SOCS) is a key indicator of soil fertility, ecosystem functioning, and climate change mitigation, yet its direct measurement remains labor-intensive, time-consuming, and costly. Visible, near-infrared, and short-wave infrared (VNIR-SWIR) soil reflectance spectroscopy offers a rapid and cost-effective alternative for SOCS estimation, although its predictive performance may strongly depend on sampling design and landscape heterogeneity. This study investigated the extent to which sampling density, spatial scale and spacing affect the accuracy of spectroscopy-based SOCS prediction. An ensemble modeling framework integrating partial least squares regression (PLSR), random forest (RF), and artificial neural networks (ANN), combined with five spectral pre-processing techniques, was applied to two contrasting datasets

collected in southern Italy. The first dataset represented the entire heterogeneous region of Campania (CAM), comprising 2,957 topsoil samples collected on a sparse irregular grid with spacing ranging from 1 to 4 km. The second dataset corresponded to a local experimental field (MFC2, 8 ha), where 135 topsoil samples were collected on a dense regular grid with 25 m spacing. Model performance was evaluated using independent validation datasets through the coefficient of determination (R^2), root mean square error (RMSE), and residual predictive deviation (RPD). Marked differences emerged between spatial scales. A fair predictive performance was achieved at MFC2 (best model: $R^2 = 0.65$; RPD = 1.71), whereas all models performed poorly at the regional CAM scale (best model: $R^2 = 0.24$; RPD = 1.14). Variogram analysis showed greater unresolved spatial variability (nugget effect) in the CAM dataset, indicating that the sparse regional sampling design failed to adequately capture fine-scale SOCS heterogeneity. To further investigate the role of sampling density, a denser subset of 130 CAM soil samples, with approximately 1 km spacing, was analyzed. Although model performance improved moderately (best $R^2 = 0.35$), predictive accuracy remained unsatisfactory (RPD <1.5), confirming that increased sampling density alone can be insufficient in highly heterogeneous regional landscapes. Overall, these findings demonstrate that sampling design and spatial heterogeneity are dominant factors controlling the reliability of spectroscopy-based SOCS estimation, emphasizing the adoption of spatially optimized sampling strategies to support robust regional soil carbon monitoring and assessment.

Key words: Soil reflectance spectra; partial least square regression; random forest; artificial neural network; pre-processing method.

Highlights:

- We compiled a soil spectral library of 3,092 soil samples in Campania
- SOCS was estimated by combining different pre-processing methods with machine learning
- The unexplained variance of SOCS decreases by reducing the spacing
- Reducing the spacing helps improve the estimation of SOCS when employing spectral information

Introduction

The growing threat of climate change demands new approaches to reduce atmospheric greenhouse gas concentrations. Agricultural soils offer significant potential for carbon sequestration, which can help mitigate climate change while enhancing soil health and fertility (Panagos *et al*., 2025; Mason *et al*., 2025; Omer *et al*., 2026). The accurate assessment of soil organic carbon stock (SOCS), calculated from soil organic carbon (SOC) content, oven-dry bulk density (BD), and soil depth, is crucial for monitoring these sequestration efforts (Lang *et al*., 2025; Vasu *et al*., 2026). However, conventional field and laboratory methods for measuring these properties are labor-intensive, time-consuming, and costly (Nasta *et al*., 2020b).

The use of soil diffuse reflectance spectroscopy in the visible, near-infrared, and short-wave infrared range (VNIR-SWIR; 400 nm - 2,500 nm) provides rapid and cost-effective estimation of soil properties using air-dried soil samples sieved to 2.0-mm (Bellon-Maurel and McBratney, 2011). Soil constituents exhibit distinct spectral signatures within the VNIR-SWIR range, allowing the development of empirical relationships between reflectance spectra and soil properties. Machine learning (ML) models facilitate the calibration of these empirical relationships, commonly referred to as spectro-transfer functions (STFs) (Askari *et al*., 2015). Despite its considerable potential, soil spectroscopy still faces several limitations that constrain its operational application for large-scale soil monitoring. First, soil reflectance is influenced not only by soil organic carbon, but also by multiple confounding factors (i.e., soil texture, mineralogy, iron oxides) which can obscure the spectral response associated with the target property. Second, STFs are often site-specific, and their transferability across contrasting pedological and environmental conditions remains limited. Third, the predictive performance of ML models strongly depends on the representativeness of the calibration dataset and the spatial structure of the sampled variable. In heterogeneous landscapes, sparse sampling designs may fail to capture fine-scale variability, thereby introducing unresolved variance that reduces model robustness. Even though extensive research has focused on spectral pre-processing techniques and ML algorithms, comparatively less attention has been paid to the role of sampling design, particularly sampling density and spacing, in controlling model performance for SOCS estimation across multiple spatial scales (Shi *et al*., 2023; Wang *et al*., 2024).

In this context, a novel aspect of the present study lies in explicitly evaluating how sampling density, spacing, and spatial scale affect the reliability of spectroscopy-based SOCS estimation across contrasting agricultural domains.

The objective of this study was to assess whether limitations in SOCS prediction accuracy using VNIR-SWIR spectroscopy are primarily associated with the spectral method itself or with inadequate characterization of spatial variability due to sampling design. An ensemble of predictive models, including partial least squares regression (PLSR), random forest (RF), and artificial neural network (ANN), was developed and applied to datasets retrieved from two different agroecosystems of Campania. The first dataset corresponds to the entire Campania (CAM), a large and heterogeneous region sampled on a sparse grid with spacing ranging from 1 km to 4 km (Romano *et al*., 2025). The second dataset corresponds to a small experimental field (MFC2, 8 ha) sampled on a dense regular grid with 25 m spacing (Romano *et al*., 2018). Unlike previous studies primarily focused on optimizing spectral pre-processing techniques or model selection, this study explicitly investigates the extent to which the predictive performance of spectroscopy-based SOCS estimation is controlled by sampling design and the ability to adequately capture the spatial variability of the target variable.

Materials and Methods

Study areas

Two study areas at different spatial scales were considered in this study (Figure 1a): i) the entire administrative region of Campania (CAM) in southern Italy, and ii) a small sub-catchment near the village of Monteforte Cilento (MFC2).

The Campania region covers an area of 13,671 km², of which 50.8% is classified as hilly, 34.6% as montane (southern Apennines), and 14.6% as lowland. A particular feature of the region is the presence of several volcanic complexes, including Roccamonfina, Mt. Somma-Vesuvius, the Campi Flegrei, and the volcanic islands of Ischia, Procida, and Vivara. This geological diversity results in a wide range of soil types and characteristics. A field campaign was organized in 2017 across Campania, during which 3,316 topsoil samples were collected from an irregular grid with a spacing between 1 and 4 km (Figure 1b) at a depth of 10 cm (Palladino *et al*. 2022; Allocca *et al*., 2023; Romano *et al*., 2025). Other 88 disturbed soil samples and intact cores were collected in the topsoil at a depth of 10 cm in the Sele river alluvial plain to validate the Airborne Visible InfraRed

Imaging Spectrometer-Next Generation (AVIRIS-NG) sensor (Francos *et al*., 2024). From this collection, a total of 2,957 samples were selected for this study.

The second study area, MFC2, is a small sub-catchment covering an area of 0.08 km² (8 ha). Within this area, a high-density, regular 25-m grid was established for sampling, resulting in a total of 135 topsoil samples at a depth of 10 cm (Figure 1c). More details on the MFC2 site can be found in Nasta *et al*. (2020a).

Soil measurements

The following soil properties were measured in the laboratory: soil organic carbon (SOC) content (%) and oven-dry soil bulk density (BD, g cm^{-3}). SOC was determined using the Walkley-Black dichromate method (Mebius, 1960). BD was measured using undisturbed cores collected in sharpened steel cylinders (7.2 cm inner diameter, 7.0 cm height), which were dried at 105°C for at least 48 hours.

The soil organic carbon stock (SOCS, kg m^{-2}) was computed using the following relation (Popleau *et al*., 2017):

$$SOCS = BD \cdot SOC \cdot D \cdot 10 \quad \text{(Eq. 1)}$$

where D (cm) is the soil upmost layer thickness. For this study, D was assumed to be a constant value of 30 cm to standardize the SOCS calculation across all samples. The benchmark SOCS values used for model training and validation were obtained from the direct laboratory measurements of SOC and BD.

Spectral data acquisition and pre-processing

Spectral measurements were executed in the laboratory on bulk soil samples (~40 g) that were air-dried and sieved to <2.0 mm. An Analytical Spectral Devices (ASD) FieldSpec® 3 spectrometer (model FSP 350-2500P) was used to measure diffuse reflectance (R) across 2,151 bands in the visible to shortwave-infrared (VNIR-SWIR) range (350-2,500 nm). The spectral resolution (full-width-half-maximum, FWHM) is approximately 3 nm in the VNIR region (350-1,000 nm) and 10 nm in the SWIR regions (1,001-2,500 nm). Each recorded spectrum was an average of 30 internal readings. Reflectance was calculated relative to a Halon white reference panel (Spectralon®), and measurements were conducted following the internal soil standard (ISS) protocol (Ben-Dor *et al*., 2015). The noisy spectral regions at the lowest edges of the spectrum (350-399 nm) were removed

prior to analysis, resulting in 2101 spectral bands. To correct for background effects and scattering and to enhance absorption features, five different pre-processing strategies were applied to the spectral data, which were subsequently used as inputs for the machine learning models (Barra *et al*., 2021):

- Raw Reflectance (RR): No transformation applied.
- Continuum Removal Transformation (CRT): A baseline correction method that normalizes spectra by fitting a convex hull over the data, enhancing the comparison of individual absorption features from a common baseline.
- 1st Derivative Savitzky-Golay (FSG): A Savitzky-Golay filter (polynomial order of 2, window size of 11 bands) was applied to calculate the first derivative, which helps to resolve overlapping peaks and remove baseline offsets.
- 2nd Derivative Savitzky-Golay (SSG): The same Savitzky-Golay filter was used to calculate the second derivative, which can further resolve spectral features and remove both constant and linear baseline drift.
- Standard Normal Variate (SNV): This transformation centers and scales each individual spectrum (subtracting its mean and dividing by its standard deviation), effectively reducing particle size and scattering effects.

These pre-processing methods have been widely used to reduce nonlinearities in soil spectra (Dotto *et al*., 2018; Ge *et al*., 2019; Wang *et al*., 2023).

Model development

Partial least squares regression (PLSR) is a linear regression method particularly well-suited for spectral data, where the number of predictor variables (bands) is large and the variables are highly collinear (multicollinearity). It reduces the high-dimensional spectral data into a smaller number of orthogonal latent variables (LVs) that maximize the covariance between the predictors (spectra) and the response (SOCS) (Viscarra Rossel, 2007; Carrascal *et al*., 2009). The optimal number of LVs was determined within the training set via 10-fold cross-validation to prevent overfitting (Wold *et al*., 2001).

Random forest (RF) is a non-linear, ensemble machine learning algorithm that builds a multitude of decision trees during training (Breiman, 2001). For regression, the final prediction is the average of the predictions from all individual trees. RF operates by constructing each tree on a bootstrapped

subset of the training samples ("bagging") and, at each node, considers only a random subset of the predictor variables for the split. This dual randomization process makes RF robust against overfitting and capable of capturing complex, non-linear relationships between the spectra and SOCS without extensive hyperparameter tuning.

Artificial neural networks (ANNs) are computational models inspired by the structure and function of biological neural networks. For this study, a standard feed-forward neural network, also known as a multi-layer perceptron, was implemented. The network consisted of an input layer with neurons corresponding to the number of spectral bands, a single hidden layer of neurons using a rectified linear unit activation function, and a single output neuron for the SOCS prediction. The network "learns" by adjusting the connection weights between neurons through a process called backpropagation, minimizing the error between its predictions and the true SOCS values in the training set (Haykin, 2009).

The modelling workflow adopted in this study is illustrated in Figure 2. To systematically evaluate the influence of spectral treatment and predictive algorithm selection on SOCS estimation, five spectral pre-processing methods were combined with three predictive ML approaches (PLSR, RF, and ANN), generating a total of 15 model configurations. For each dataset, soil samples were first sorted in ascending order according to SOCS values, and every fourth sample was assigned to the validation subset, resulting in a 75% calibration set and a 25% independent validation set. The remaining samples were used for model calibration. This stratified sampling strategy ensured a balanced representation of the SOCS range in both subsets, while a Student's *t*-test confirmed that the calibration and validation datasets were statistically comparable in terms of soil properties (results not shown for brevity). Each model configuration was then independently calibrated and validated, allowing a comparative assessment of predictive robustness and performance across contrasting spatial scales.

Variogram of SOCS

A variogram is a fundamental tool in geostatistics used to quantify the spatial correlation or spatial dependence between observations of a spatially distributed target variable (y). The variogram, $\gamma(h)$, provides a measure of dissimilarity between data points, as a function of the lag distance, h (the separation distance between pairs of observations):

$$\gamma(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [y(x_i) - y(x_i + h)]^2 \quad \text{(Eq. 2)}$$

where $N(h)$ is the number of pairs of observations separated by h, and x_i is the location of the i^{th} observation.

A theoretical mathematical model is used to fit the experimental variogram. We used the simplest models, i.e., spherical, exponential, Gaussian, or power models. The selection of the appropriate model is typically done by visually inspecting the experimental variogram and then using statistical criteria (e.g., the sum of squares of residuals, described below) to find the best fit. The key components of any variogram model are nugget, sill, and range. The nugget is the value of the variogram at a lag distance of zero and represents the unexplained variance when the distance between samples is infinitesimally small. This jump at the origin accounts for two main sources of variability: 1) Errors introduced during data collection, laboratory analysis, or sampling; 2) Spatial variation that occurs at distances smaller than the smallest sampling interval. This variability cannot be resolved by the sampling scheme. We crudely assumed that measurement errors are zero. The range is the lag distance at which the variogram reaches its maximum value (the sill) and then typically flattens out. It represents the distance beyond which samples are no longer spatially correlated. In other words, samples separated by a distance greater than the range are considered statistically independent. The sill represents the total variance of the data, including both the spatially correlated variance and the nugget effect. Once the variogram reaches the sill, increasing the lag distance further does not increase the variance, indicating that the maximum dissimilarity (and thus the absence of spatial correlation) has been reached.

Model performance

All model performance metrics were computed exclusively on the independent validation dataset (25% of the total samples), while the remaining 75% were used for model calibration. This approach ensured an unbiased evaluation of model generalization performance. The performance of the predictive models was assessed using the root mean square error (RMSE), coefficient of determination (R^2), and the ratio of performance to deviation (RPD), defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{(Eq. 3)}$$

$$R^2 = 1 - \left(1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right) \quad \text{(Eq. 4)}$$

$$RPD = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}} \quad \text{(Eq. 5)}$$

where y and $\bar{y}$ denote the observed and mean of observed SOCS, $\hat{y}$ refers to the predicted values of SOCS.

Optimal prediction is characterized by a value of 0 for RMSE and 1 for R^2. The RPD, also known as the residual predictive deviation, is a normalized index to evaluate model performance. According to Chang *et al*. (2001), the following ranges indicate classes of model performance: RPD<1.5 (poor performance), 1.5 ≤ RPD < 2.0 (fair performance), 2.0 ≤ RPD < 3.0 (good performance), 3.0 ≤ RPD < 4.0 (very good performance), RPD≥4.0 (excellent performance) (Ge *et al*., 2019).

Results

Data analysis

Descriptive statistics for BD, SOC, and SOCS datasets for both the regional scale of Campania (CAM) and local scale of Monteforte Cilento (MFC2) are summarized in Table 1. The corresponding frequency distributions for these properties are illustrated in Figure 3.

In the regional CAM dataset (N=2,957), BD showed a mean value of 1.28 g cm^{-3} and a low coefficient of variation (CV) of 16%, indicating relatively low variability across the entire study area. The frequency distribution for BD is approximately symmetric, as confirmed by a low skewness value of −0.30 (Figure 2a). In contrast, both SOC and SOCS exhibited significantly greater variability, with CVs of 67% and 63%, respectively. These variables displayed highly positively skewed distributions, with skewness values of 2.05 for SOC and 1.92 for SOCS. This is visually evident in their histograms (Figure 3 b,c), where the mean values (2.17% for SOC, 811.73 kg m^{-2} for SOCS) are considerably larger than their respective medians (1.81% and 691.70 kg m^{-2}) and modes (1.45% and 111.90 kg m^{-2}). This indicates a prevalence of lower values with a long tail extending towards higher values.

The dataset (n=135) of the MFC2 catchment, representing a small and more homogeneous study area, showed markedly different characteristics for carbon-related properties. Although the mean BD of 1.27 g cm^{-3} was nearly identical to that of the CAM dataset, its variability was lower (CV = 10%). SOC and SOCS distributions were nearly symmetric, with very low skewness values of 0.30

and 0.40, respectively. This is further supported by the close agreement between their mean and median values (Table 1) and the symmetric shape of their histograms (Figure 3 e,f). Consequently, the variability of SOC and SOCS in MFC2 was substantially lower (CVs of 32% and 31%) than in the regional CAM dataset.

A direct comparison highlights that the heterogeneity of the large CAM region leads to much greater variability and positive skewness in SOC and SOCS when compared to the local MFC2 field. However, bulk density remains a remarkably consistent property across both spatial scales.

Prediction performance in CAM and MFC2

The predictive performance of the 15 models, derived from five pre-processing techniques and three ML algorithms (see Figure 2), was evaluated for both the Campania (CAM) and Monteforte Cilento (MFC2) datasets. Model performance was assessed using independent validation datasets, with results summarized in Table 2 for CAM and Table 3 for MFC2.

For the regional CAM dataset, prediction performance was generally poor across all model combinations (Table 2). The optimal results were obtained using RF applied to spectra pre-processed with SSG), yielding a R^2 of 0.24, a RMSE of 441.1 kg m^{-2}, and a RPD of 1.14. According to the classification by Chang *et al*. (2001), RPD values below 1.5 indicate poor predictive capability. This is consistent with the scatter plot in Figure 4a, which shows a wide dispersion of points around the 1:1 identity line and a weak correlation between observed and predicted values. None of the models for the CAM dataset achieved a RPD equal to or greater than 1.5, highlighting the difficulty of developing robust spectro-transfer functions at this large and heterogeneous spatial scale.

In contrast, models developed for the local MFC2 dataset demonstrated considerably better performance (Table 3). The most accurate predictions were achieved using PLSR combined with FSG pre-processing. This model produced a R^2 of 0.65, a RMSE of 116.1 kg m^{-2}, and a RPD of 1.71. This RPD value is indicative of a fair prediction performance, capable of discriminating between high and low values. The corresponding scatter plot (see Figure 4b) confirms this result, with points tightly clustered around the 1:1 identity line, indicating a strong agreement between observed and estimated SOCS values. Other model combinations performed satisfactorily for the MFC2 dataset, including PLSR applied to SNV processed spectra, which yielded an RPD of 1.68.

Overall, the results clearly demonstrate that the predictive accuracy of the models strongly depends on the spatial scale and sampling density of the dataset. The best-performing model for the local, densely sampled MFC2 dataset (FSG + PLSR, $R^2 = 0.65$) significantly outperformed the best model for the regional, sparsely sampled CAM dataset (SSG + RF, $R^2 = 0.24$).

Variograms for measured SOCS were compared for both CAM and MFC2 datasets (Figure 5). The variogram parameters help explain the poor predictive performance observed for the CAM dataset. The nugget value for SOCS in the CAM dataset is 1.68×10^5, within a range of approximately 5 km. This relatively large nugget, accounting for about one-third of the sill (2.58×10^5), indicates the presence of substantial unexplained variance at short distance. This suggests that a sampling spacing ranging from 1.0 km to 4.0 km was rather inadequate to adequately capture the spatial variability of SOCS across the regional scale. In contrast, the MFC2 dataset exhibited a nugget-to-sill ratio of 73%, suggesting that a larger proportion of the spatial variability of SOCS was explained at the local scale compared with CAM. Therefore, the sampling scheme with a 25-meter spacing within a range of approximately 150 meters was more effective in capturing the features of the SOCS spatial structure. Anyway, according to Cambardella *et al*. (1994), the nugget-to-sill ratios fall between the 25% and 75% classes, meaning that measured SOCS is moderately spatially dependent in both cases.

The impact of spacing on the prediction performance in a subset of CAM

To further investigate the impact of sampling spacing on the prediction performance, we extracted a subset of 130 soil samples with the highest sampling density in the CAM dataset in the Sele alluvial plain (Francos *et al*., 2024; Figure 6). This subset represents the area with the greatest sampling density in the regional dataset, with most sample pairs separated by approximately 1.0 km. The primary goal of analyzing this subset was to assess whether increased sampling density can better capture the spatial variability of SOCS and, consequently, mitigate the problem observed with the sparsely sampled CAM dataset. Figure 6 presents the experimental variogram of measured SOCS based on this densely sampled subset of the CAM dataset, which shows a typical pure nugget behavior. Compared with the full CAM dataset (Figure 5a; n=2,957), the nugget value decreased substantially from 1.68×10^5 to 7.60×10^4 and this reduction indicates that the 1 km sampling spacing of this subset at least reduced the unresolved short-range variability that was not captured in the original regional sampling scheme, but SOCS still appears spatially uncorrelated at this reduced

sampling scale. Any spatial structure that might exist should occur at a spatial scale finer than the sampling interval of 1 km. Apart from the inevitable measurement errors and their propagation in the determination of SOCS, this also suggests the presence of dominant short-range variability below the sampling resolution or substantial random noise due to coarse sampling interval relative to actual spatial variability or strong local disturbances (i.e., tillage, fertilizer application, traffic, irrigation). Therefore, a pure nugget does not substantiate the random nature of the variable; rather, it means that the selected sampling scheme cannot resolve any spatial continuity. Next, the same 15 combinations of pre-processing and ML models were applied to this subset. The results, summarized in Table 4, show an improvement in prediction performance compared to the full CAM dataset. The best-performing models were an ANN and a PLSR, both applied to spectra pre-processed with the FSG and RR, respectively. These models yielded an identical, improved R^2 of 0.35 and a RMSE of 333.8 kg m^{-2}.

Despite this improvement, the best RPD value achieved was only 1.26, which is still classified as poor performance (RPD <1.5). The corresponding scatter plot (Figure 7) confirms this modest enhancement, showing points that are still considerably scattered and not well-aligned with the 1:1 identity line. This result suggests that while increasing sampling density helps, a spacing of ~1 km is still insufficient to adequately model the heterogeneity present at a regional scale.

Discussion

The results of this study highlight the profound impact that spatial scale, sampling density, and landscape heterogeneity exert on the ability of soil spectroscopy to accurately predict SOCS (Viscarra Rossel and Webster, 2012). A stark contrast in model performance between the regional CAM dataset and the local MFC2 dataset clearly illustrates these challenges.

The consistently poor performance of all 15 model combinations for the CAM dataset (Table 2) can be mainly attributed to the pronounced variability inherent in such a large study region. Campania encompasses a wide range of soil-forming factors, including diverse parent materials ranging from volcanic to carbonate complexes, varied topography from mountainous areas to lowlands, and multiple land-use systems. This heterogeneity is quantitatively captured by the variogram of measured SOCS (Figure 5a), which revealed a very high nugget effect. This large nugget indicates that a substantial component of the SOCS variability occurs at distances smaller

than the 1-4 km sampling interval, rendering it as unresolvable spatial noise that fundamentally limits the predictive capacity of any STF.

Conversely, the fair prediction accuracy achieved at the MFC2 site (Table 3), where the best-performing model (FSG + PLSR) yielded an RPD of 1.71, demonstrates the effectiveness of soil spectroscopy under more homogeneous environmental conditions and denser sampling. The lower SOCS variability in the MFC2 variogram indicates that the 25-meter sampling grid adequately captured the dominant spatial structure of SOCS, facilitating the development of a more robust predictive model. These findings suggest that predictive performance was governed not only by the choice of ML algorithm or spectral information, but also by the extent to which the sampling strategy captured the spatial heterogeneity of SOCS. The poorer performance observed across the larger and more heterogeneous CAM domain compared with the more homogeneous MFC2 site indicates that sampling design represents a critical but often overlooked determinant of spectroscopy-based prediction accuracy. This finding extends previous studies that primarily emphasized model optimization and spectral pre-processing, highlighting instead the fundamental role of spatial representativeness in operational soil monitoring applications.

Testing the core hypothesis that increasing sampling density would improve regional-scale predictions confirmed this relationship, albeit with an important limitation. By isolating a denser subset within CAM, the nugget effect was significantly reduced (Figure 6), and the best-model R^2 value improved from 0.24 to 0.35. This directly confirms that better characterization of spatial variability translates into more accurate spectral predictions. However, the performance for this subset remained poor (RPD <1.5), indicating that even a ~1 km spacing is insufficient to overcome the high heterogeneity of a regional landscape. This finding suggests that achieving fair-to-good predictive performance with soil spectroscopy at regional scales should require sampling densities approaching those commonly adopted at the hectare scale. The challenges identified in this study further underscore the need for standardized laboratory and spectral acquisition protocols to support the development of large, high-quality soil spectral libraries (SSLs). The sharing of well-characterized SSLs across the research community is essential for developing robust and transferable models that can be integrated with continental-scale resources, such as LUCAS, and effectively applied at regional scales (Orgiazzi *et al*., 2018).

Conclusions

This study evaluated the performance of VNIR-SWIR soil spectroscopy combined with different predictive modeling approaches for estimating soil organic carbon stock (SOCS) across contrasting spatial scales. The results demonstrated that predictive accuracy is governed primarily by spatial heterogeneity and sampling design, rather than by the choice of machine learning algorithm alone. While fair prediction performance was achieved at the local scale in the densely sampled and relatively homogeneous experimental field (MFC2), all model configurations performed poorly at the regional Campania scale, where the sparse sampling design failed to adequately capture fine-scale SOCS variability. Increasing sampling density within a regional subset improved model performance, confirming the importance of spatial representativeness, although prediction accuracy remained limited under heterogeneous conditions. These findings indicate that the successful operational application of soil spectroscopy for regional SOCS monitoring requires not only robust spectral modeling, but also sampling strategies specifically designed to resolve the spatial structure of the target variable. Future research efforts should prioritize the integration of optimized sampling schemes, standardized large-scale soil spectral libraries, and multi-scale modeling frameworks to enhance the reliability of spectroscopy-based soil carbon assessment.

References

Allocca C, Castrignanò A, Nasta P, Romano N, 2023. Regional-scale assessment of soil functions and resilience indicators: accounting for change of support to estimate primary soil properties and their uncertainty. Geoderma 431:116339.

Askari MS, O'Rourke SM, Holden NM, 2015. Evaluation of soil quality for agricultural production using visible–near-infrared spectroscopy. Geoderma 243-244:80-91.

Barra I, Haefele SM, Sakrabani R, Kebede F, 2021. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: recent advances — a review. Trends Anal Chem 135:116166.

Bellon-Maurel V, McBratney A, 2011. Near infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils: critical review and research perspectives. Soil Biol Biochem 43:1398-410.

Ben-Dor E, Ong C, Lau IC, 2015. Reflectance measurements of soils in the laboratory: standards and protocols. Geoderma 245-246:112-124.

Breiman L, 2001. Random forests. Mach Learn 45:5-32.

Cambardella CA, Moorman TB, Novak JM, Parkin TB, Karlen DL, Turco RF, Konopka AE, 1994. Field-scale variability of soil properties in central Iowa soils. Soil Sci Soc Am J 58:1501-11.

Carrascal LM, Galván I, Gordo O, 2009. Partial least squares regression as an alternative to current regression methods used in ecology. Oikos 2118:681-90.

Dotto AC, Dalmolin RSD, ten Caten A, Grunwald S, 2018. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma 314:262-74.

Francos N, Nasta P, Allocca C, Sica B, Mazzitelli C, Lazzaro U, et al., 2024. Mapping soil organic carbon stock using hyperspectral remote sensing: a case study in the Sele River plain in southern Italy. Remote Sens 16:897.

Ge Y, Morgan CLS, Wijewardane NK, 2019. Visible and near-infrared reflectance spectroscopy analysis of soils. Soil Sci Soc Am J 84:1495-1502.

Haykin SS, 2009. Neural networks and learning machines. 3rd ed. London, Pearson Education.

Hermansen C, Knadel M, Moldrup P, Greve MG, Karup D, de Jonge LW, 2017. Complete soil texture is accurately predicted by visible near-infrared spectroscopy. Soil Sci Soc Am J 81:758-69.

Knadel M, Thomsen A, Schelde K, Greve MH, 2015. Soil organic carbon and particle sizes mapping using vis–NIR, EC and temperature mobile sensor platform. Comput Electron Agric 114:134-144.

Lang AK, Pastore MA, Walters BF, Domke GM, 2025. Bulk density calculation methods systematically alter estimates of soil organic carbon stocks in United States forests. Biogeochemistry 168:44.

Ludwig B, Murugan R, Parama VRR, Vohland M, 2019. Accuracy of estimating soil properties with mid-infrared spectroscopy: implications of different chemometric approaches and software packages related to calibration sample size. Soil Sci Soc Am J 83:1542-1552.

Mason E, Cornu S, Arrouays D, Fantappiè M, Jones A, Götzinger S, et al., 2025. Monitoring systems of agricultural soils across Europe regarding the upcoming European soil monitoring law. Eur J Soil Sci 76:e70163.

Mebius LJ, 1960. A rapid method for the determination of organic carbon in soil. Anal Chim Acta 22:120-4.

Nasta P, Bogena HR, Weuthen A, Sica B, Vereecken H, Romano N, 2020a. Integrating invasive and non-invasive monitoring sensors to detect field-scale soil hydrological behavior. Front Water 2:26.

Nasta P, Palladino M, Sica B, Pizzolante A, Trifuoggi M, Toscanesi M, et al., 2020b. Evaluating pedotransfer functions for predicting soil bulk density using hierarchical mapping information in Campania, Italy. Geoderma Reg 21:e00267.

Ogen Y, Zaluda J, Francos N, Goldshleger N, Bon-Dor E, 2019. Cluster-based spectral models for a robust assessment of soil properties. Geoderma 340:175-184.

Omer E, Szlatenyi D, Csenki S, Tünde G, Chhetri G, Veres Z, Láng V, 2026. Soil health for sustainable agriculture: a bibliometric review of EU current scientific findings and research trends. Soil Adv 5:100097.

Orgiazzi A, Ballabio C, Panagos P, Jones A, Fernández-Ugalde O, 2018. LUCAS soil, the largest expandable soil dataset for Europe: a review. Eur J Soil Sci 69:140-153.

Palladino M, Romano N, Pasolli E, Nasta P, 2022. Developing pedotransfer functions for predicting soil bulk density in Campania. Geoderma 412:115726.

Panagos P, Jones A, Lugato E, Ballabio C, 2025. A soil monitoring law for Europe. Glob Chall 9:2400336.

Poeplau C, Vos C, Don A, 2017. Soil organic carbon stocks are systematically overestimated by the misuse of the parameters bulk density and rock fragment content. Soil 3:61-66.

Romano N, Mazzitelli C, Nasta P, 2025. Root-zone water-storage capacity and uncertainty: an intrinsic factor affecting agroecosystem resilience to drought. Water Resour Res 61:e2024WR037719.

Romano N, Nasta P, Bogena HR, De Vita P, Stellato L, Vereecken H, 2018. Monitoring hydrological processes for land and water resources management in a Mediterranean ecosystem: the Alento River Catchment observatory. Vadose Zone J 2018;17:180042.

Shi L, O'Rourke S, de Santana FB, Daly K, 2023. Prediction of soil bulk density in agricultural soils using mid-infrared spectroscopy. Geoderma 434:116487.

Vasu D, Laxmanarayanan M, Tiwary P, 2026. Long-term soil organic carbon stock changes in croplands of India. Catena 268:110041.

Viscarra Rossel RA, 2007. Robust modelling of soil diffuse reflectance spectra by bagging-partial least squares regression. J Near Infrared Spectrosc 15:39-47.

Viscarra Rossel RA, Webster R, 2012. Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. Eur J Soil Sci 63:848-860.

Wang X, Sun H, Wang C, Liu J, Guo Z, Gao L, et al., 2024. Predicting the soil bulk density using a new spectral PTF based on intact samples. Geoderma 449:117005.

Wang Y, Yang S, Yan X, Yang S, Feng M, Xiao J, et al. 2022. Evaluation of data pre-processing and regression models for precise estimation of soil organic carbon using Vis-NIR spectroscopy. J Soils Sediments 2022;23:634-45.

Wold S, Sjöström M, Eriksson L, 2001. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109-130.

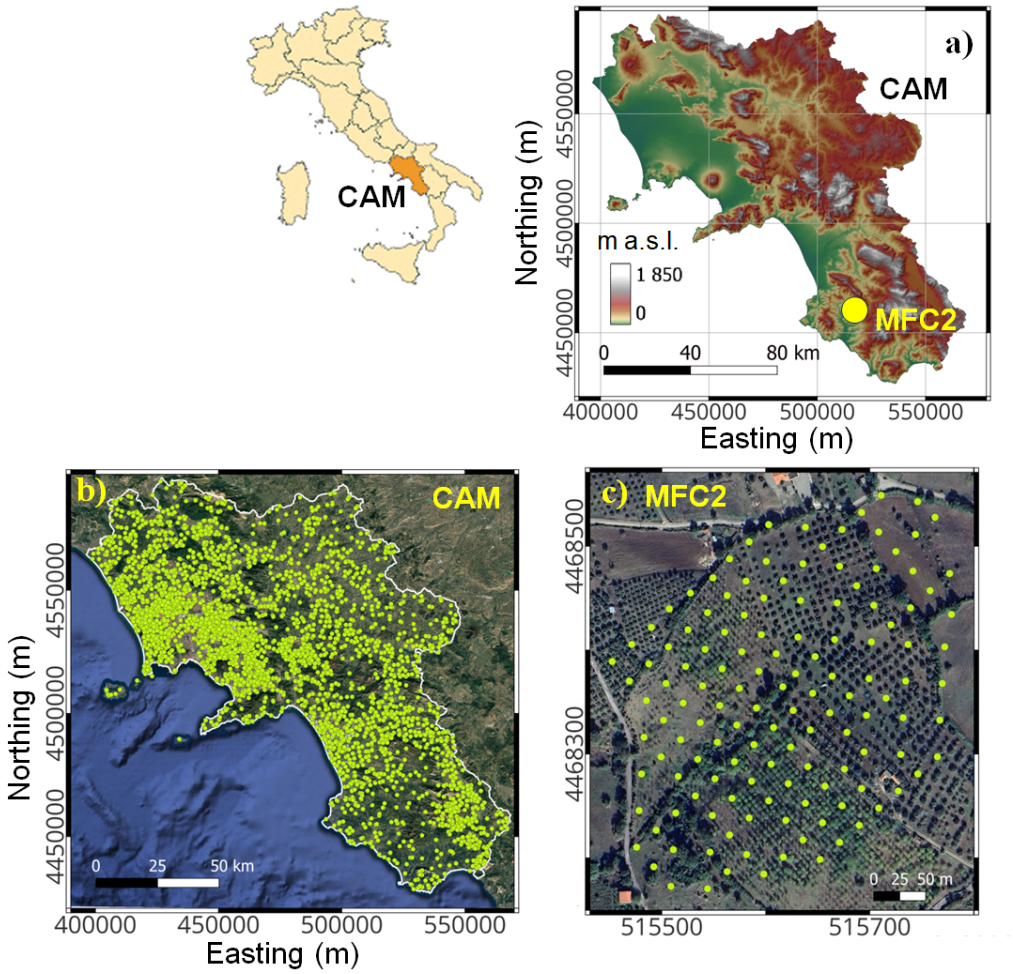

Figure 1. a) Geographic position of the two study areas: namely Campania region (CAM), and the experimental sub-catchment near the town of Monteforte Cilento (MFC2). **b)** Soil sampling positions in CAM (2,957 yellow circles). **c)** Soil sampling positions in MFC2 (135 yellow circles).

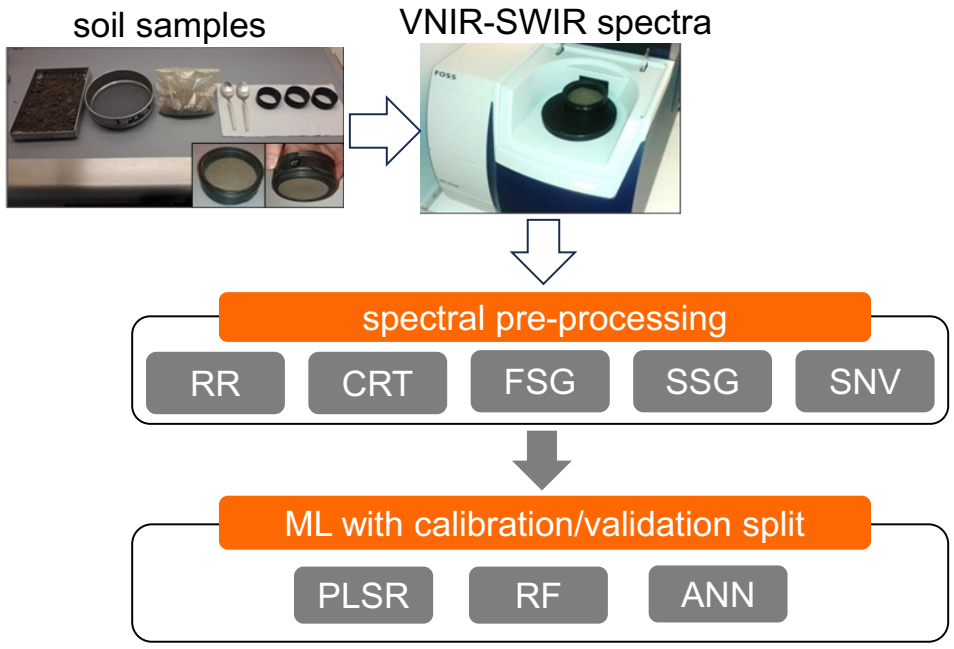

Figure 2. Workflow of the ensemble modelling framework used to estimate soil organic carbon stock (SOCS) from VNIR-SWIR reflectance spectra. Five spectral pre-processing methods were combined with three machine learning (ML) models, yielding 15 model configurations evaluated using independent validation metrics. RR, CRT, FSG, SSG, SNV indicate raw reflectance, continuum removal transformation, 1st Derivative Savitzky-Golay, 2nd Derivative Savitzky-Golay, and standard normal variate, respectively. PLSR, RF, and ANN indicate partial least squares regression, and artificial neural network, respectively.

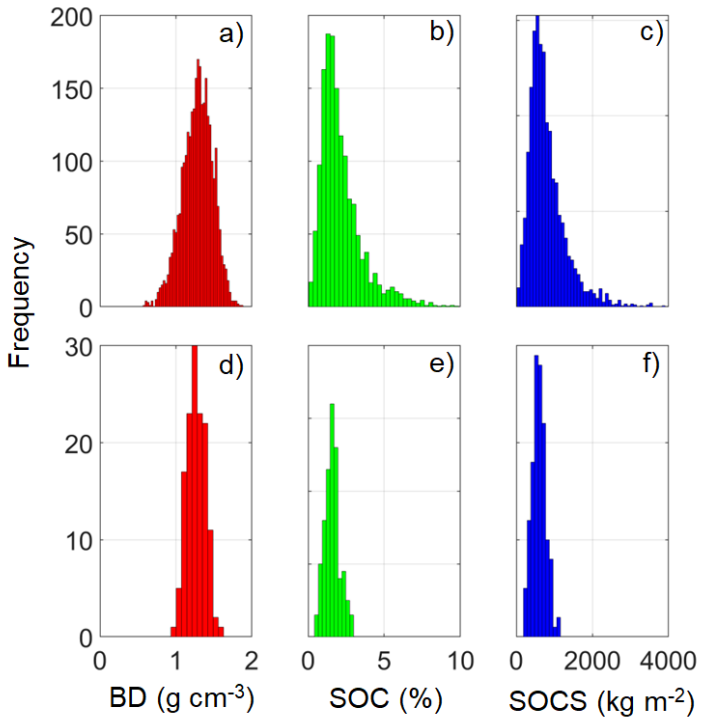

Figure 3. Frequency distributions in the region of Campania (CAM) of: **a)** soil bulk density (BD in g cm^{-3}); **b)** soil organic carbon content (SOC in %); **c)** soil organic carbon stock (SOCS in kg m^{-2}). Frequency distributions in the region of in the experimental sub-catchment near the town of Monteforte Cilento (MFC2) of: **d)** soil bulk density (BD in g cm^{-3}); **e)** soil organic carbon content (SOC in %); **f)** soil organic carbon stock (SOCS in kg m^{-2}).

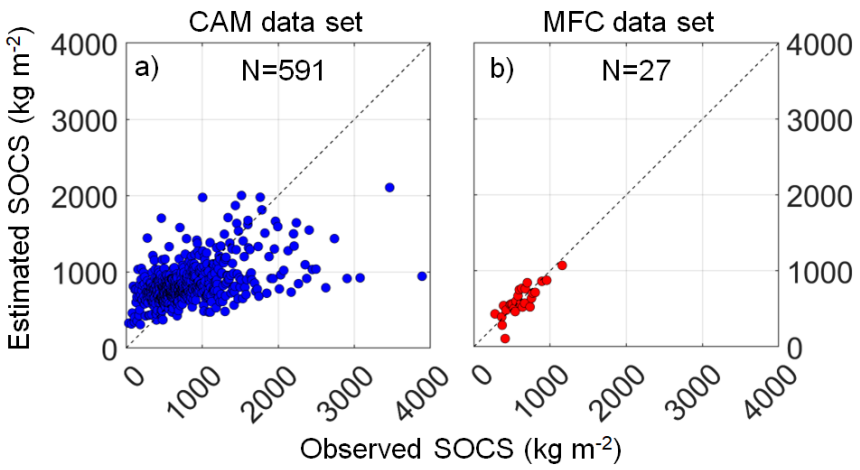

Figure 4. Comparison between observed and estimated (using the best combination between machine learning model and preprocessing method) soil organic carbon stock (SOCS) using **a)** validation data set (574 soil samples) in the region of Campania (CAM), and **b)** validation data set (27 soil samples) in the experimental sub-catchment near the town of Monteforte Cilento (MFC2). The diagonal dashed line depicts the identity line (1:1 line).

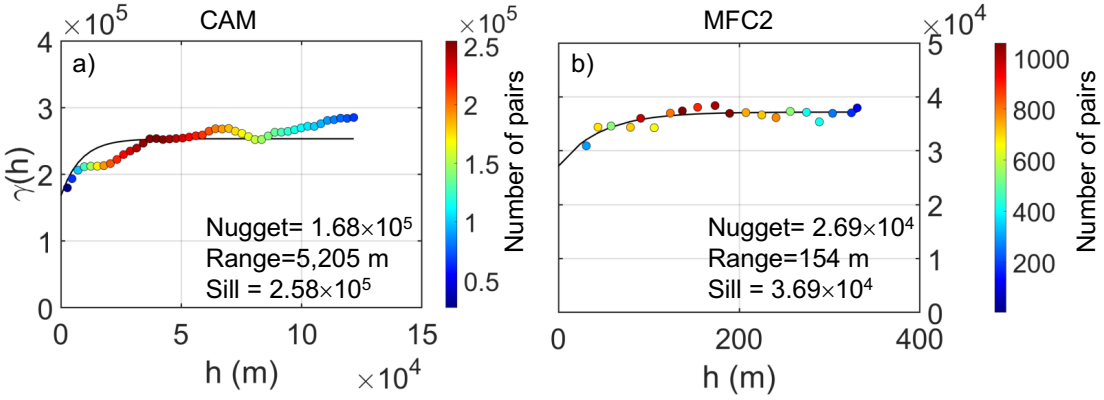

Figure 5. Exponential model fitted on the experimental variograms based on the soil organic carbon stock (SOCS) measured in **a)** the region of Campania (CAM) and **b)** the experimental sub-catchment near the town of Monteforte Cilento (MFC2). The data pairs are color-coded by the number of pairs. The nugget, range, and sill are reported in each subplot.

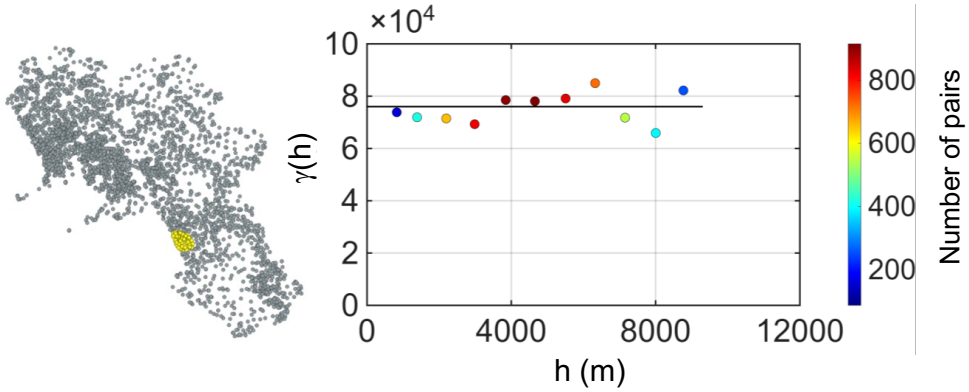

Figure 6. Exponential model fitted on the experimental variogram based on the soil organic carbon stock (SOCS) measured in a subset of data (130 soil samples) sampled in Campania region. The data pairs are color-coded by the number of pairs.

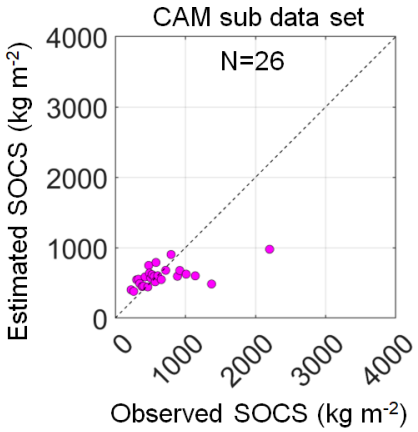

Figure 7. Comparison between observed and estimated (using the best combination between machine learning model and preprocessing method) soil organic carbon stock (SOCS) using a sub dataset (26 soil samples for testing purposes) of Campania region (CAM). The diagonal dashed line depicts the identity line (1:1 line).

Table 1. Mean, median, mode, standard deviation (Std), coefficient of variation (CV), skewness (skew), interquartile range (iqr) of soil organic carbon (SOC), oven-dry bulk density (BD) and soil organic carbon stock (SOCS) in the Campania region (CAM), and in the experimental sub-catchment near the town of Monteforte Cilento (MFC2).

		BD g cm^{-3}	SOC %	SOCS kg m^{-2}
CAM (n=2957)	Mean	1.28	2.17	811.73
	Median	1.29	1.81	691.70
	Mode	1.28	1.45	111.90
	Std	0.20	1.45	508.39
	CV	16%	67%	63%
	Skew	-0.30	2.05	1.92
	iqr	0.28	1.46	528.08
MFC2 (n=135)	Mean	1.27	1.59	595.94
	Median	1.27	1.55	589.22
	Mode	1.27	1.40	186.43
	Std	0.12	0.51	184.66
	CV	10%	32%	31%
	Skew	0.05	0.30	0.40
	iqr	0.18	0.58	245.69

Table 2. Root mean square error (RMSE), coefficient of determination (R^2), and the ratio of performance to deviation (RPD) corresponding to 15 combinations of preprocessing and machine learning methods used to estimate soil organic carbon stock from reflectance data in the validation dataset (n=591) of Campania.

Pre-processing	Machine learning model	R^2	RMSE	RPD
Raw reflectance	PLSR	0.14	468.0	1.08
Raw reflectance	RF	0.05	492.8	1.02
Raw reflectance	ANN	0.17	459.7	1.10
1st Derivative Savitzky-Golay	PLSR	0.15	465.9	1.08
1st Derivative Savitzky-Golay	RF	0.21	449.4	1.12
1st Derivative Savitzky-Golay	ANN	0.18	457.6	1.10
2nd Derivative Savitzky-Golay	PLSR	0.10	477.6	1.06
2nd Derivative Savitzky-Golay	RF	0.24	441.1	1.14
2nd Derivative Savitzky-Golay	ANN	0.10	560.0	0.90
Standard normal variate	PLSR	0.15	465.1	1.09
Standard normal variate	RF	0.15	464.5	1.09
Standard normal variate	ANN	0.04	579.3	0.87
Continuum removal transformation	PLSR	0.13	471.6	1.07
Continuum removal transformation	RF	0.22	446.9	1.13
Continuum removal transformation	ANN	0.14	466.6	1.08

PLSR, partial least squares regression; RF, partial least squares regression; ANN, artificial neural network.

Table 3. Root mean square error (RMSE), coefficient of determination (R^2), and the ratio of performance to deviation (RPD) corresponding to 15 combinations of preprocessing and machine learning methods used to estimate soil organic carbon stock from reflectance data in the validation dataset (n=27) in the experimental sub-catchment near the town of Monteforte Cilento.

Pre-processing	Machine learning model	R^2	RMSE	RPD
Raw reflectance	PLSR	0.41	149.44	1.33
Raw reflectance	RF	0.11	183.73	1.08
Raw reflectance	ANN	0.04	198.67	1.00
1st Derivative Savitzky-Golay	PLSR	0.65	116.12	1.71
1st Derivative Savitzky-Golay	RF	0.53	133.41	1.49
1st Derivative Savitzky-Golay	ANN	0.31	162.09	1.23
2nd Derivative Savitzky-Golay	PLSR	0.32	160.47	1.24
2nd Derivative Savitzky-Golay	RF	0.26	167.94	1.18
2nd Derivative Savitzky-Golay	ANN	0.07	375.41	0.53
Standard normal variate	PLSR	0.63	118.57	1.68
Standard normal variate	RF	0.40	151.64	1.31
Standard normal variate	ANN	0.11	183.79	1.08
Continuum removal transformation	PLSR	0.44	145.88	1.36
Continuum removal transformation	RF	0.26	167.37	1.19
Continuum removal transformation	ANN	0.18	176.84	1.12

PLSR, partial least squares regression; RF, partial least squares regression; ANN, artificial neural network.

Table 4. Root mean square error (RMSE), coefficient of determination (R^2), and the ratio of performance to deviation (RPD) corresponding to 15 combinations of preprocessing and machine learning (ML) methods used to estimate soil organic carbon stock from reflectance data in the validation sub dataset (n=26) of Campania.

Pre-processing	Machine learning model	R^2	RMSE	RPD
Raw reflectance	PLSR	0.35	333.8	1.26
Raw reflectance	RF	0.24	358.5	1.17
Raw reflectance	ANN	0.01	410.1	1.03
1st Derivative Savitzky-Golay	PLSR	0.07	397.6	1.06
1st Derivative Savitzky-Golay	RF	0.15	380.1	1.11
1st Derivative Savitzky-Golay	ANN	0.35	333.8	1.26
2nd Derivative Savitzky-Golay	PLSR	0.10	556.6	0.76
2nd Derivative Savitzky-Golay	RF	0.27	351.7	1.20
2nd Derivative Savitzky-Golay	ANN	0.26	354.1	1.19
Standard normal variate	PLSR	0.05	567.7	0.74
Standard normal variate	RF	0.27	352.8	1.19
Standard normal variate	ANN	0.20	369.4	1.14
Continuum removal transformation	PLSR	0.03	411.9	1.02
Continuum removal transformation	RF	0.17	374.8	1.12
Continuum removal transformation	ANN	0.29	347.0	1.21

PLSR, partial least squares regression; RF, partial least squares regression; ANN, artificial neural network.