Catena 243 (2024) 108180

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena

Comparison of field and imaging spectroscopy to optimize soil organic carbon and nitrogen estimation in field laboratory conditions

Ashfak Mahmud [a,*], Markku Luotamo [b], Kristiina Karhu [c], Petri Pellikka [a,e], Juuso Tuure [d], Janne Heiskanen [a,f]

[a] Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland
[b] Department of Computer Science, University of Helsinki, Helsinki, Finland
[c] Department of Forest Sciences, University of Helsinki, Helsinki, Finland
[d] Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland
[e] University of Nairobi, Kenya
[f] Finnish Meteorological Institute, Helsinki, Finland

ARTICLE INFO

Keywords:
Soil Organic Carbon
Soil nitrogen
Machine learning
Hyperspectral
Imaging spectroscopy
Spectroradiometer

ABSTRACT

Regenerative agriculture (RA) aims to improve soil health, water retention capacity, and resilience through sustainable regeneration and retention of soil organic carbon (SOC) and soil nitrogen (SN). Although laboratory analysis offers a reliable method for measuring SOC and SN, access to such facilities can be limited in remote areas and inconvenient, particularly if a large number of samples need to be analyzed. To address this, we compared two hyperspectral sensors, SVC HR-1024i field spectroradiometer (FS) and Specim IQ imaging spectrometer (IS), to estimate SOC and SN using 157 soil samples collected from agricultural sites in Taita Hills, Kenya. Reference SOC and SN content (%) were analyzed in the laboratory, and spectral measurements and images of the air-dried soil samples were generated using protocols suitable for field laboratories. Finally, we employed partial least squares regression (PLSR), Gaussian process regression (GPR), and least absolute shrinkage and selection operator (LASSO) to estimate SOC and SN from the preprocessed soil spectra over different wavelength ranges. Our best models yielded better accuracy for SOC estimation (R^2 = 0.83 and RMSE = 0.36 %) and an R^2 of 0.70 with an RMSE of 0.07 % for SN using the field spectroradiometer. Both SOC and SN modeling over the full wavelength and shortwave-infrared (SWIR) regions achieved considerably better predictive accuracy than visible to near-infrared regions. These results suggest that FS with the SWIR region is best suited for SOC and SN estimation to support the planning and monitoring needs of RA initiatives.

1. Introduction

Soil properties are important indicators of efficient farming and agricultural productivity. Soil organic carbon (SOC) influences the physical, chemical, and biological functions of the soil, thereby improving its capacity to retain water and nutrients (Haynes, 2005; Ontl & Schulte, 2012). Furthermore, SOC is a key factor of soil fertility which is directly associated with soil microbial activity, erosion reduction, and ultimately increased crop productivity in agricultural settings (Haynes, 2005; Hong et al., 2018; Ontl & Schulte, 2012; Powlson et al., 2011). Soil nitrogen (SN) is an essential mineral macronutrient, which ensures the physiochemical balance for plant growth (B. Tan et al., 2022). During plant development, plants harness usable nitrogen by decomposing organic nitrogen followed by microbial-mediated mineralization (nitrification and ammonification (Ehrenfeld, 2013).

Regenerative agriculture (RA) aims to play an important role in managing and storing SOC and SN through combinations of different crop management practices (e.g., reducing tillage, cover crops, and planting nitrogen-fixing crops) (Abdalla et al., 2019; Dick et al., 1998; Poeplau & Don, 2015). Moreover, RA has the benefits of reducing SN input costs, greenhouse gas emissions, and water (ground and surface) contamination (Loisel et al., 2019; Prescott et al., 2021). The success of carbon market projects in agriculture to tackle climate change relies on the integration of RA practices to limit carbon emissions and increase SOC stock. Monitoring, reporting, and verification are crucial in these projects which need robust quantitative scientific evidence outlining the long-term SOC change resulting from these interventions to encourage farmers to adopt RA practices (Tan & Kuebbing, 2023).

* Corresponding author.

https://doi.org/10.1016/j.catena.2024.108180
Received 9 February 2024; Received in revised form 4 June 2024; Accepted 10 June 2024
Available online 14 June 2024

Soil has a complex composition of organic and inorganic traits that show substantial variability even at an intra-field level (Jandl et al., 2014). The spatial heterogeneity of SOC is influenced by several spatial and temporal factors, including climate, land use, topography, vegetation type, underlying geological composition, and human interventions (Johnson et al., 2000; Keesstra et al., 2016; Mulder et al., 2016; Parton et al., 1987). Therefore, accurate estimation and instantaneous monitoring of SOC and SN require optimized and reliable frameworks. Commonly used methods for SOC retrieval are Walkley-Black wet oxidation, as proposed by Walkley (1935), and the dry combustion method introduced by Nelson and Sommers (1983). Similarly, for SN retrieval, Kjeldahl (Sáez-Plaza et al., 2013) is used conventionally. These laboratory methods are costly and time-consuming for running long test processes (Conant et al., 2011; Sinfield et al., 2010; B. Tan et al., 2022). Thus, scientists have been looking for alternatives and complementary approaches for extracting soil properties (e.g., SOC and SN) to reduce dependency on complicated laboratory methods, especially for remote areas where laboratory access is limited (Angelopoulou et al., 2019; Hong et al., 2018; Nocita et al., 2015; Omran, 2017).

In recent years, a proximal soil sensing technique of visible (VIS, 400–700 nm), near-infrared (NIR, 700–1100 nm), and shortwave infrared (SWIR, 1100–2500 nm) spectroscopy has been widely used to retrieve soil properties as complementary to laboratory retrievals (Ben-Dor et al., 2009; Ben-Dor & Banin, 1995b; Morellos et al., 2016). Condit (1970) and Stoner & Baumgardner (1981) were the first researchers to systematically analyze the relationship between spectral signals from soil and its properties. In VIS, absorption bands resulting from soil color were impacted by chromophores. In the NIR-SWIR (700–2500) spectral region, the interaction between carbon, oxygen, hydrogen, phosphorus, and nitrogen causes vibration, which is translated into reflective spectral patterns (Ben-Dor & Banin, 1995a; Dalal & Henry, 1986; Stuart, 2004; Viscarra Rossel et al., 2006a; Viscarra Rossel & Behrens, 2010). To indirectly quantify soil properties, multivariate statistics are usually employed to extract complex absorption patterns and correlation is established with various soil parameters (Stenberg et al., 2010).

SOC and SN retrieval using imaging spectrometer (IS) and field spectroradiometers (FS) is becoming more accurate with the recent development of technology and machine learning (ML) methods (Ben-Dor et al., 2009; Pellikka et al., 2023). Imaging spectrometers (IS) acquire images of soil samples (i.e. multiple measurements simultaneously) and can capture soil texture, unlike FS, which makes point-wise measurements (O'Rourke and Holden, 2011). In contrast, measurement with FS involves less image processing while image processing with IS can be an intensive task. In addition to the acquisition technique, variations in the spectral regions covered by the instruments can have an impact on the accuracy of soil parameter estimation. Although Pellikka et al. (2023) recently demonstrated the use of Specim IQ IS, the comparability of IS and FS to assess SOC and SN remains untested.

Many scientists have highlighted the potential of regression models, such as the partial least squares regression (PLSR), Gaussian process regression (GPR), and least absolute shrinkage and selection operator (LASSO) in predicting soil properties (Bao et al., 2017; Bricklemyer et al., 2018; Chen et al., 2011; Demattê et al., 2017; dos Santos et al., 2023; Dyar et al., 2012; He et al., 2005; Hong et al., 2018; Jakab et al., 2018; Li et al., 2019; Pellikka et al., 2023; Viscarra Rossel et al., 2006b). Recently, Pellikka et al. (2023) utilized IS for the first time for SOC and SN quantification by employing different ML approaches. However, different ML model performances over different wavelength ranges using FS and IS remained unexplored.

In this study, we compared FS and IS, different wavelength ranges, and several ML methods for predictive modeling of SOC and SN in cultivated soils. The instruments included SVC HR-1024i FS with full wavelength range (FWR, 350–2500 nm) and Specim IQ IS with visible near-infrared (VNIR, 400–1000 nm) range. Hence, while FS provides a broader spectral range, including the SWIR region, IS provides image-format sampling and the instrument is less expensive. The soil samples were collected from different types of land cover (cropland, sisal farm, agroforestry, and shrubland) in Kenya with an aim to develop a model for rapid analysis of a larger number of soil samples in field laboratory conditions. To model SOC and SN, we tested PLSR, GPR, and LASSO on different spectrum ranges and used LASSO to study which spectral bands are most influential for SOC and SN modeling. Finally, we discuss the measurement protocols and modeling pipelines for both sensors in assessing soil properties especially for the needs of RA initiatives in Africa to effectively manage carbon and nitrogen in the cultivated soils.

2. Material and methods

2.1. Study area

The study material was collected from Taita Hills in Taita Taveta County in the southeastern part of Kenya (Fig. 1). The region can be broadly separated into dry lowland plains (altitude range 500 to 1000 m a.s.l.) and highlands featuring a variety of land covers that reach heights of up to 2200 m (Pellikka et al., 2013). The rainy season in the Taita Hills is bimodal in nature, characterized by short rains occurring between November and December and long rains from March to June. According to the meteorological observations by Taita research station, the average annual rainfall and temperature is 1125 mm and 18.6 °C in Wundanyi (2013–2022) at 1400 m a.s.l. and 756 mm and 22 °C in Dembwa (2011–2020) at 1114 m a.s.l., while it was 612 mm and 25.2 °C (1990–2018) at 844 m a.s.l. in the Teita Sisal Estate (Wachiye et al., 2021). Taita Hills has nine agroecological zones, where soil conditions exhibit large differences between zones and are strongly influenced by altitude (Jaetzold et al., 2012). A significant shift to agriculture around Taita Hills and foothills has been observed, replacing mostly natural vegetation cover and bushlands (Pellikka et al., 2018). Agriculture, especially smallholder cropping systems and the production of cash crops like sisal (*Agave sisalana* 'Hildana' and *Agave sisalana* 'Hybrid 11648') (Vuorinne et al., 2021), is the most common land use in the study area (Autio et al., 2021; Pellikka et al., 2023).

According to the Farm Management Handbook of Kenya (Jaetzold et al., 2012; Jaetzold & Schmidt, 1983), soils in foothills and lowlands are characterized by the presence of well-drained low fertile red Ferralsols, Luvisols, and Arenosols. In contrast, the soil found in the smaller hills of the lowlands includes differing degrees of fertility levels in the Regosols and Cambisols. In the Taita Hills, the soil is eutric Regosols and calcic Cambisols in an inconsistent fertility situation and orthic Rendzinas in a high fertility situation. SOC and SN in Taita Hills are lowest in the Ferra sols and highest in the Umbrisols (Njeru et al., 2017). Njeru et al. (2017) and Pellikka et al. (2023) reported that SOC and SN increase with altitude in Taita Hills. Furthermore, Pellikka et al. (2018) studied above-ground carbon (AGC) for each land cover type and found that montane and exotic forest had the highest AGC, whereas cropland and shrubland have lower AGC.

2.2. Soil sampling

We utilized soil samples from (Pellikka et al., 2023), who followed the soil sampling strategy from Njeru et al. (2017) for studying SOC and SN stock in Taita Hills. The soil samplings were performed at a lowland to highland transect (30 km), starting from the town of Mwatate (~800 m a.s.l.) and ending in a montane forest of the Taita Hills on Vuria mountain (~2000 m a.s.l.). There were 157 sampling sites, which included soil from cropland (n = 59), sisal plantation (n = 44), agroforestry (n = 39), and shrubland (n = 17). We did not include soil samples from the forest (n = 32), as our aim was to develop SOC and SN models only for agricultural environments; forests generally have higher SOC and SN levels. Each sample consisted of soil from three points, taken from 20-cm depth using a soil auger within an area range of 1 m^2. Each soil sample weighed around 100–200 g, was sieved with a 2-mm sieve net, mixed on site, and stored in a plastic zip bag until analysis. The

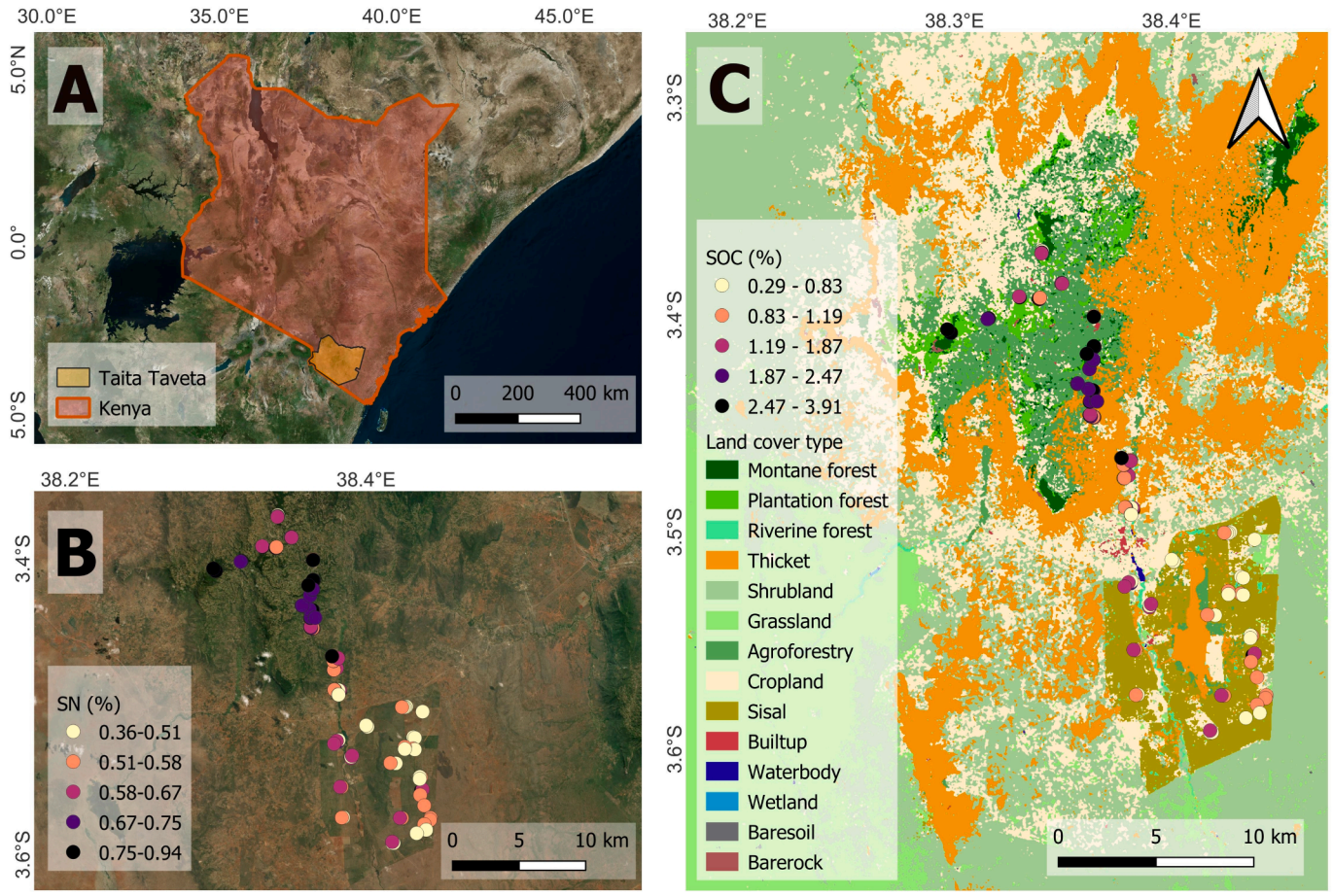

Fig. 1. (A) Boundary of Kenya and Taita Taveta county, (B) sampling locations illustrated with low to high SN (%) in graduated color, and (C) sampling locations illustrated with low to high SOC (%) in graduated color overlaid on a land cover map from Abera et al. (2022).

locations of the center of each square meter plot were recorded using a Garmin handheld GPS receiver (GPSMAP 64x). The sampling strategy aimed to bring variation in geographic position, elevation, and land cover, indirectly causing variations in soil color, texture, and SOC (Pellikka et al., 2023).

2.3. Laboratory analysis

The soil samples were transported to the University of Helsinki and then oven-dried at 40 °C for 4–6 days in open plastic bags. Samples were ground in a mortar before analysis in a Leco CN828 dry oxidization analyzer for SOC and SN content. To avoid contamination from the mortar and pestle, thorough cleaning with ethanol was performed before progressing to the next soil sample. Following the instrument's (Leco CN828) guide, a weight of 300 mg from each soil sample was analyzed for SOC and SN assessment for 2-minute duration per sample (Pellikka et al., 2023).

Summary statistics of laboratory-analyzed SOC and SN for soil samples are presented in Table 1. Across all samples, the mean SOC content was 1.65 % (range 0.29 % to 3.9 %, standard deviation 0.89 %). The mean SN content was 0.6 % (range 0.35 % to 0.93 %, standard deviation 0.13 %). Fig. 2 indicates that SOC and SN values found in sisal are concentrated within a relatively narrow range compared with other land cover types. The altitudinal gradient has a clear influence on the distribution of SOC values, with a notable increase in SOC concentrations towards higher altitudes, and a corresponding decline in SOC values towards lower altitudes (Pellikka et al., 2023). Furthermore, the spatial representation of SOC in Fig. 1B indicates that the agroforestry regions (Fig. 1C) exhibit the highest SOC values, whereas the lowest SOC values are predominantly observed in sisal sites.

2.4. Imaging and spectral measurement protocol

2.4.1. Field spectroradiometer

The portable SVC HR-1024i spectroradiometer has 1024 spectral channels distributed across three detectors, capturing wavelengths from

Table 1
Summary statistics for soil organic carbon (SOC) and soil nitrogen (SN) samples.

Land Cover	SOC (%)					SN (%)				
	Mean	SD	Min	Max	Skew	Mean	SD	Min	Max	Skew
All (n = 157)	1.65	0.89	0.29	3.90	0.52	0.60	0.13	0.35	0.93	0.30
Agroforestry (n = 39)	2.07	0.88	0.40	3.91	0.02	0.66	0.12	0.36	0.89	−0.08
Field (n = 57)	1.95	0.88	0.29	3.86	0.16	0.67	0.13	0.42	0.90	−0.08
Shrubland (n = 17)	1.57	0.76	0.42	2.83	−0.16	0.74	0.12	0.54	0.94	−0.17
Sisal (n = 44)	0.92	0.37	0.42	2.34	1.9	0.53	0.07	0.39	0.73	0.91

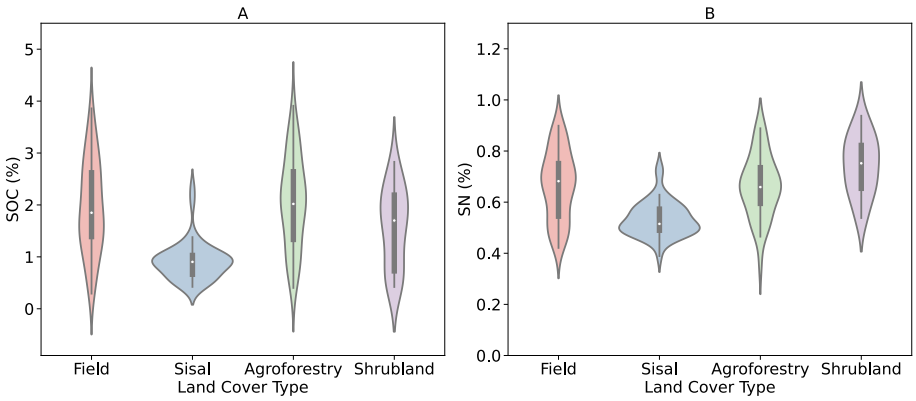

Fig. 2. Distribution of SOC (A) and SN (B) analyzed in the laboratory across different land cover types in Taita Hills, Kenya. The shape of the violin plot shows distribution of SOC and SN range.

350 nm up to 2500 nm (SVC, 2021). Spectral resolution (FWHM) is 3.3 nm, 9.5 nm, and 6.5 nm at 700 nm, 1500 nm, and 2100 nm, respectively. The instrument was equipped with a Leaf-Clip Reflectance-Probe (LC RP PRO), which allows contact measurements using an internal tungsten halogen lamp (SVC, 2020, 2021). We developed our own protocol for soil spectral measurements with SVC FS and LC RP PRO, which is adopted from instrument manuals (SVC, 2020, 2021) and the literature (Ben Dor et al., 2015; Gholizadeh et al., 2021). The measurements were carried out at the University of Helsinki in April 2023, where the dried soil samples from Pellikka et al. (2023) were stored in air-tight zip bags at room temperature.

We performed the spectral measurements in a closed, windowless, and dark room (Fig. 3). The surroundings, including the table surface, were covered with a low reflective black sheet to avoid unwanted reflections. Before taking the measurement, the FS was warmed up for 15 min (SVC, 2021) and the LC-RP PRO for at least 90 s (SVC, 2020). The internal tungsten halogen lamp was set to the highest intensity. The soil samples were placed in plastic petri dishes (10-cm diameter, 2-cm depth). Then, the top surface was pressed and labeled with a plastic ruler to make a smooth surface, ensuring better reflection and signal-to-noise ratio (SNR) (Ben Dor et al., 2015; Mouazen et al., 2005). The LC-RP PRO was attached to a tripod (Manfrotto) with a clamp at a fixed position over the soil surface to ensure that the light was not escaping. The scan time was set to 5 s and fore-optic to Lens 4 to use the default Lens 4 calibration configuration during the reflectance measurement, as suggested for LC-RP PRO (SVC, 2020). Each soil sample was measured three times from different parts of the sample surface by moving the petri dish slightly after each scan. The white reference (spectralon) measurements were taken between each sample measurement (three scans) and the dark current was automatically taken by the instrument (SVC, 2021). The petri dishes were cleaned properly before placing another soil sample.

2.4.2. *Imaging spectrometer*

The Specim IQ hyperspectral camera can capture 204 spectral bands ranging from 400–1000 nm at a spectral resolution of 7 nm. The soil samples were imaged by Pellikka et al. (2023). The imaging procedure took place in a windowless, dark room without light contamination from other sources. The camera was fixed at a height of 30 cm above the imaging area with a stand (KAISER RS 1) to ensure the viewing illumination and SNR were uniform for all imaged samples (Fig. 4). Two external light sources (150 W halogen lamps) were fixed at a 45° angle and a distance of 40–50 cm. Before the imaging process, the camera was calibrated in the stable light condition using a white and black reference, which is a rapid and applicable method for a fixed illumination condition (Behmann et al., 2018). The soil samples were prepared in petri dishes (10-cm diameter) and cleaned properly with ethanol between measurements (see Pellikka et al., 2023 for further details).

2.5. *Spectra extraction and treatment*

The spectra from SVC FS come as a signature (SIG) file, a text file containing spectral data, and metadata (SVC, 2021). All soil reflectance and reference spectra were quality-checked before further treatment and analysis. The “Spectrolab” package (Meireles et al., 2023) in the R software environment (R Core Team, 2022) was employed to read and

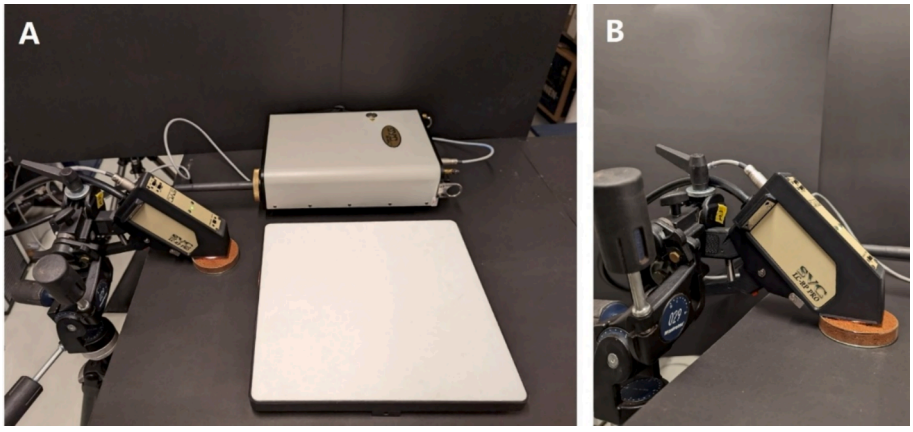

Fig. 3. (A) Soil spectral measurement setup with SVC HR1024i spectroradiometer, Leaf-Clip Reflectance-Probe (LC-RP PRO), and white reference (spectralon). (B) LC-RP PRO mounted on a tripod with a clamp on a soil sample at very close proximity. [Photo by Ashfak Mahmud, 2023].

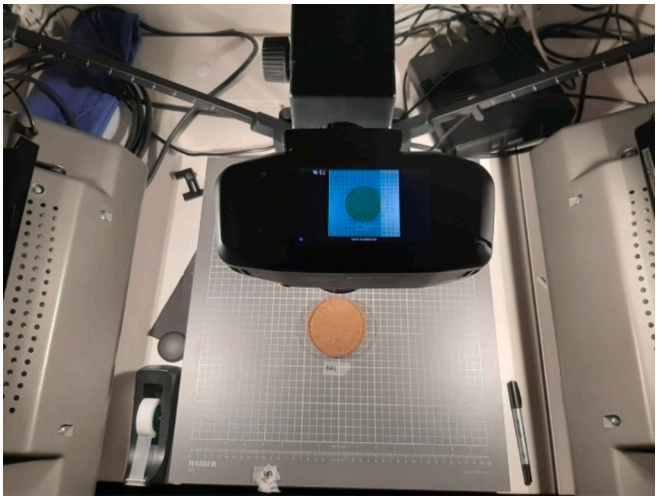

Fig. 4. Imaging setup with the Specim IQ camera fixed with a stand. Two external halogen lamps are illuminating the soil sample at a 45° angle. [Photo by Niklas Sädekoski, 2020].

process the SIG files. Noisy spectral data were eliminated between 350–399 nm and 2400–2500 nm, reducing the number of spectral bands from 1024 to 900. The spectra from the overlapping sensors at 990 nm and 1900 nm regions were matched and spliced together utilizing the method proposed by Meireles et al. (2023) to remove spectral inconsistency. The three replicate measurements from each sample were averaged.

We employed HoughCircles, an image-recognition algorithm to automatically alter the petri dish border and background from each Specim IQ square image, which resulted in a maximum 154 × 154-pixel square area that exclusively shows the soil (see Pellikka et al., 2023 for details). For further analysis, a total of 182 bands (440–972 nm) were retained after omitting 22 noisy bands from the blue and NIR regions. The spectral dataset was derived by taking the mean pixel values from the 154 × 154-pixel image of whole dish sample. Finally, all spectra from both instruments were standard normal variates (SNV) corrected using equation (1). SNV transformation minimizes the multiplicative effect caused by scattering and particle size and centers each spectrum (X_i) by subtracting its average ($\overline{X}_i$) and dividing by its standard deviation (σ_i) (Barnes et al., 1989).

$$X_i^{snv} = \frac{X_i - \overline{X}}{\sigma_i} \tag{1}$$

2.6. Prediction algorithms

Considering a comparatively small dataset, we selected PLSR, LASSO, and GPR after assessing model performance in SOC and SN prediction by Pellikka et al. (2023). PLSR (Martens & Næs, 1992) has demonstrated efficacy in managing spectral data by simplifying highly correlated predictors into fewer latent variables and maximizing response correlation (Serbin et al., 2019; Wang et al., 2020). Thus, PLSR aims to maximize covariance between the two matrices of latent variables and to build a regression model that offers the highest R^2 and the lowest RMSE values, and the optimal count of components (Viscarra Rossel et al., 2006b). This algorithm has been widely used by many scientists for soil properties (SOC and SN) analysis from spectral information (Cao et al., 2020; Pellikka et al., 2023; Rossel & Behrens, 2010).

LASSO (Tibshirani, 1996) can handle linear and nonlinear scenarios by choosing variables (important wavelength bands) from the sample data through a penalty process. The smaller coefficients are forced to zero; hence, only variables with non-zero coefficients are retained (Yang & Bao, 2017). The study from Yang & Bao (2017) showed that the accuracy of SOC prediction with hyperspectral data was improved by utilizing the band-selection ability from LASSO. The potential for LASSO in predicting SOC was also demonstrated by Pellikka et al. (2023), where LASSO outperformed the PLSR algorithm with an excellent ratio of performance to deviation (RPD) indicator. The LASSO model contains a single hyperparameter (denoted as α), which can be tuned to control the level of regularization.

GPR is a non-parametric method based on a Bayesian approach that employs different choices of a covariance (i.e. kernel function) (Rasmussen, 2004) to adapt to different types of data and modeling requirements. Rather than adopting a fixed parametric form, it assigns a functional prior directly on the regression function f(x), which maps predictors to the response. In geo-statistics, GPR is a technique widely applied in studies related to pedometric approaches (Williams & Rasmussen, 1995). As described by Williams & Rasmussen (1995), GPR is a powerful algorithm for approximating functions in spaces with high dimensionality. Ramirez-Lopez et al. (2013) utilized GPR to model SOC, clay content (CC), and exchangeable calcium (Ca^{++}) from VIS-NIR spectra.

2.7. Model performance metrics

The performance of the models was assessed based on root mean squared error (RMSE), coefficient of determination (R^2 score), the RPD and ratio of performance to inter-quartile distance (RPIQ) using equations (2) to (5), respectively. A higher R^2 (close to 1.0) and smaller RMSE (close to zero) indicates a good fit for the model (Yang Yang, 2015). RMSE has the same units as the response variables and consistently produces non-negative values. If RPD < 1.0, the estimation capacity is very poor; RPD values between 1 and 1.4 indicates approximate prediction of low and high value; RPD between 1.4 and 1.8 indicates fair model; RPD lies between 1.8 and 2.0 indicates good model; RPD > 2.0 indicates the model's estimation capability is very good. According to Xu et al. (2023), RPIQ values between 2.02 and 2.70 indicates fair estimation, 2.70 to 3.37 suggests approximate estimation, 3.37 to 4.05 can be characterized by good accuracy.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3}$$

$$RPD = \frac{SD}{RMSE} = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}{RMSE} \tag{4}$$

$$RPIQ = \frac{Q3 - Q1}{RMSE} \tag{5}$$

Where y is measured response values (SOC and SN), $\overline{y}$ is the average of response values, $\hat{y}$ is predicted response values, n is the total sample number, SD is the standard deviation, and Q1 and Q3 are the first quartile (25 %) and third quartile (75 %) of the total dataset, respectively.

2.8. Cross-validation and model optimization

Model calibration and validation were performed using k-fold cross validation (CV) (Vapnik, 2000). In the k-fold cross-validation process, the data set is split into k sections or folds. A classifier is trained using k-1 of these folds, and the excluded fold is used to calculate an error value by testing the classifier (Rodriguez et al., 2010). We performed 10-fold cross validation as suggested by Bouckaert (2003) and Witten et al.

(2011). The average RMSE, R^2, RPD, and RPIQ were reported over all the folds. In each fold, further nested cross-validation was employed to determine the optimum values for the hyperparameters of every model. The dataset was further divided into five inner folds, forming one fold as the validation and the other four folds as the inner training set. On the inner training set, we trained the models across a range of hyperparameter selections, their performance metrics were assessed on the validation set, choosing the hyperparameters for the final model that offered the highest average accuracy over the five folds. However, for GPR, we optimized the hyperparameters directly by maximizing the marginal likelihood. We selected hyperparameter values within inner folds using Bayesian optimization (Snoek et al., 2012), implemented through scikit-optimize (Pellikka et al., 2023).

In our analysis, we incorporated three spectral ranges from FS, specifically the FWR (400–2400 nm), VNIR (400–1000 nm), and SWIR (1000–2400 nm) and one spectral range (VNIR: 400–1000 nm) from IS, treated as four independent datasets. In addition, we employed PLSR, LASSO, and GPR on these datasets for modeling SOC and SN, resulting in 12 groups of cross-validation results. To test if spectral ranges and machine learning methods differ significantly, we utilized paired student's t-tests to compare the performance metrics (R^2*andRMSE*) of different groups (algorithms and spectral ranges) obtained through 10-fold cross-validation (Dietterich, 1998). We used the same k-fold cross-validation split of the data (e.g., the same random seed to split the data in each case) and calculated a score for each split.

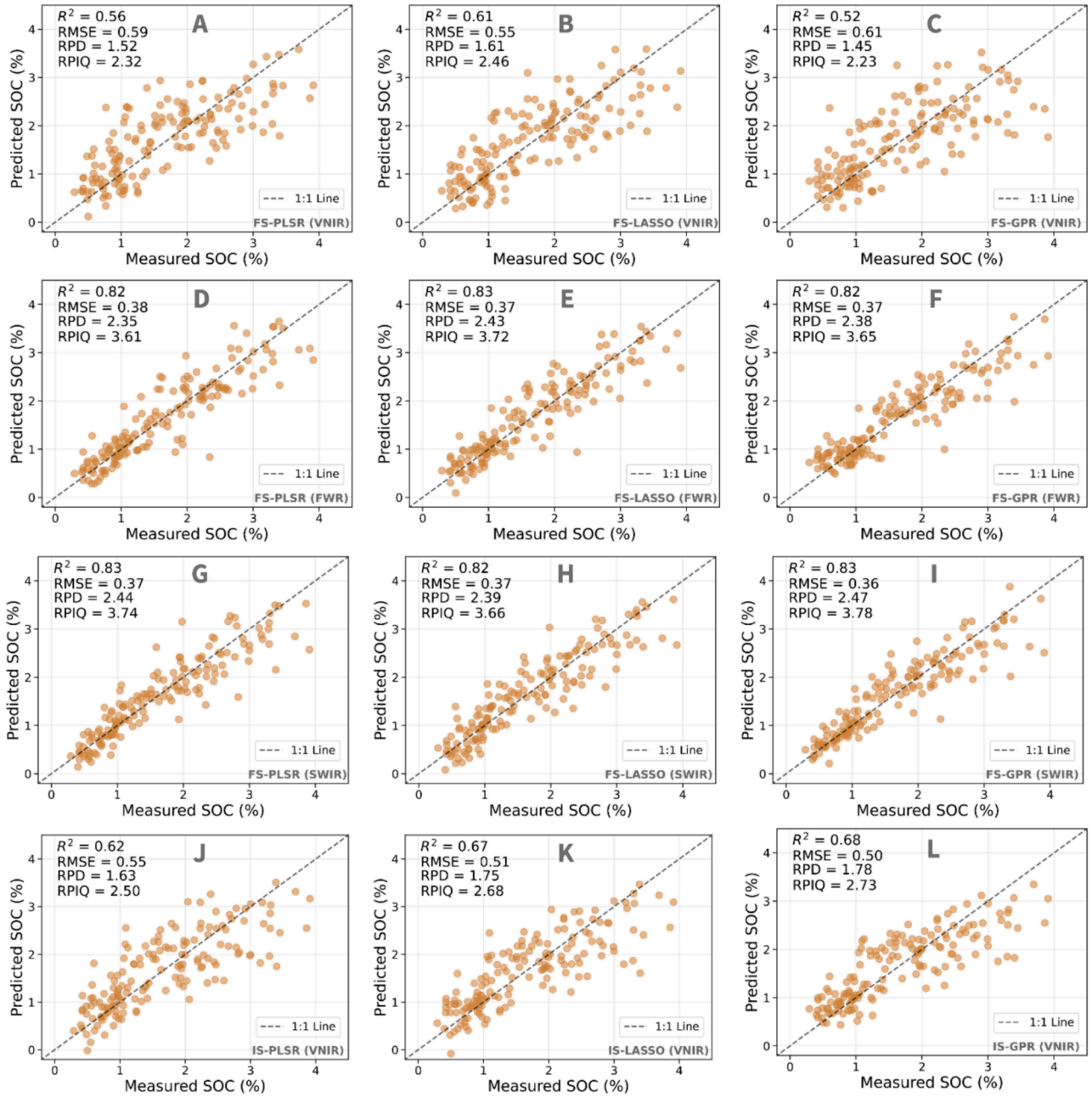

Fig. 5. Laboratory measured against the model-predicted soil organic carbon (SOC) with corresponding model performance metrics for FS (panels A-I) and IS (panels J-L). The plots in columns depict SOC estimation from three machine-learning models (PLSR, LASSO and GPR) and rows illustrate the ranges of wavelength bands (VNIR: 400–1000 nm, FWR: 400–2400 nm, SWIR: 1000–2400 nm) used as model predictors.

3. Results

3.1. Soil carbon estimation

Fig. 5 (A-I) shows the SOC prediction results and model performance metrics for different wavelength ranges using FS. The Supplementary Tables 1 and 2 provide a detailed account of the hyperparameters and program library versions employed for SOC modeling. Overall, all the models (PLSR, LASSO, and GPR) showed better performance in the SWIR range and with the GPR model, having the lowest RMSE (0.36 %) and the highest R^2 (0.83) (Fig. 5I). In contrast, SOC predictions in the VNIR range had poor accuracy across all models, having the lowest RMSE (0.55 %) and the highest R^2 (0.61) for the LASSO model (Fig. 5B). The integration of the SWIR range as a predictor significantly enhanced the accuracy of the models, leading to a significant reduction in the RMSE (to 0.37 %) and achieving the highest R^2 (0.83) for LASSO (Fig. 5E) in the FWR. All three models achieved good SOC predictive accuracy in FWR and SWIR wavelengths regions, with R^2 between 0.83 and 0.82 and RMSE between 0.36 % and 0.38 %.

The SOC prediction results for IS are shown in Fig. 5 (J-L). In general, the SOC prediction with IS showed less accuracy, reaching the lowest RMSE of 0.50 % and the highest R^2 of 0.68 with GPR model. Among the models tested for SOC estimation using the IS, the PLSR model (Fig. 5J) exhibited lower accuracy (RMSE = 0.55 %, R^2= 0.62) compared with both GPR and LASSO.

The results of hypothesis testing for accuracy metrics of SOC modeling are provided in Supplementary Tables 3 and 4. In most cases, R^2 and RMSE differ significantly ($p < 0.05$) between the 12 groups. Non-significant differences occur mainly between modeling methods in the same wavelength ranges (e.g. LASSO VNIR, PLSR VNIR) but sometimes also between wavelength ranges in the same algorithms (e.g. GPR SWIR, GPR FWR).

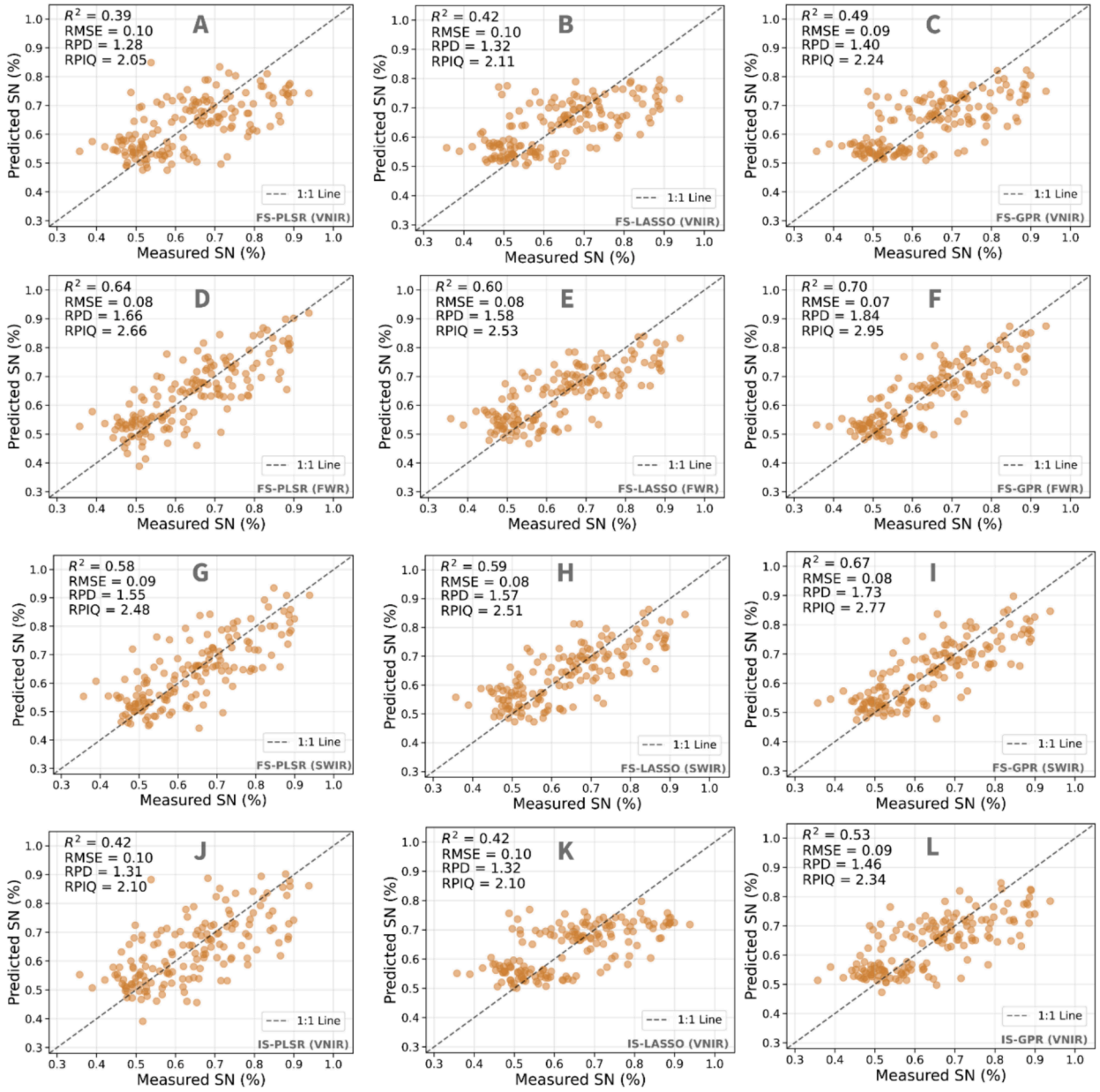

Fig. 6. Laboratory measured against model-predicted soil nitrogen (SN) with corresponding model performance metrics from FS (panels A-I) and IS (panels J-L). The plots in columns depict SN estimation from three machine-learning models (PLSR, LASSO, and GPR) and rows illustrate the ranges of wavelengths bands (VNIR: 400–1000 nm, FWR: 400–2400 nm, SWIR: 1000–2400 nm) used as model predictors.

3.2. Soil nitrogen estimation

Fig. 6 (A-I) shows the results of SN predictions and model performance metrics across different wavelength ranges using FS. The FWR range yielded optimal SN results for all tested models (PLSR, LASSO, and GPR) where GPR outperformed the other two models with the lowest RMSE of 0.07 % and the highest R^2 of 0.70 (Fig. 6 D-F). Conversely, the accuracy of SN predictions in the VNIR spectrum for all models were the lowest, with R^2 between 0.49 and 0.39 and RMSE of 0.10 % (Fig. 6 A-C). All three models in SWIR range alone achieved accuracy close to FWR, with R^2 ranging from 0.67 to 0.58 and RMSE between 0.08 % and 0.09 % (Fig. 6 G-I).

For IS, the maximum SN prediction accuracy was achieved by the GPR model, with the highest R^2 of 0.53 and the lowest RMSE of 0.09 (Fig. 6L). PLSR and LASSO (Fig. 6 J and K) showed similar accuracy (R^2 = 0.42 and RMSE = 0.10) but was lower than GPR. Overall, the patterns of the SN prediction model accuracy in different wavelength ranges illustrate similar patterns as SOC prediction, showing better predictive accuracy in FWR and SWIR than VNIR.

Hypothesis testing showed less significant differences ($p < 0.05$) for R^2 and RMSE between the 12 groups for SN compared to groups for SOC modeling (Supplement Tables 5 and 6). Non-significant differences occur mainly between modeling methods (e.g. LASSO FWR, GPR FWR) but sometimes also between algorithms (e.g. GPR SWIR, GPR FWR) and sensors (e.g. GPR IS, GPR VNIR).

3.3. Informative band selection

Finally, we studied which bands are most informative for SOC and SN modeling using FWR FS data and LASSO (Fig. 7). Most of the informative bands for SOC and SN were in NIR to SWIR wavelength range (980–2400 nm). There was a consistent band selection around 980 nm and 1360 nm for both SOC and SN modeling. Selected bands were intensified in the SWIR region over all folds starting from 2150 nm to 2400 nm. Most of the selected bands were common for both SOC and SN, showing good correlation with SOC and SN. There were clear inconsistencies and less spectral bands selected over the folds in VIS region for both SOC and SN prediction. Overall linear coefficient values were higher for SOC than SN in all selected bands.

4. Discussion

4.1. Model performance

4.1.1. Soil organic carbon

All the SOC models using FWR and SWIR as predictors had $R^2 > 0.82$ and had a "good prediction" result as classified by Williams et al. (2019). In contrast, prediction using only VNIR can discriminate between low

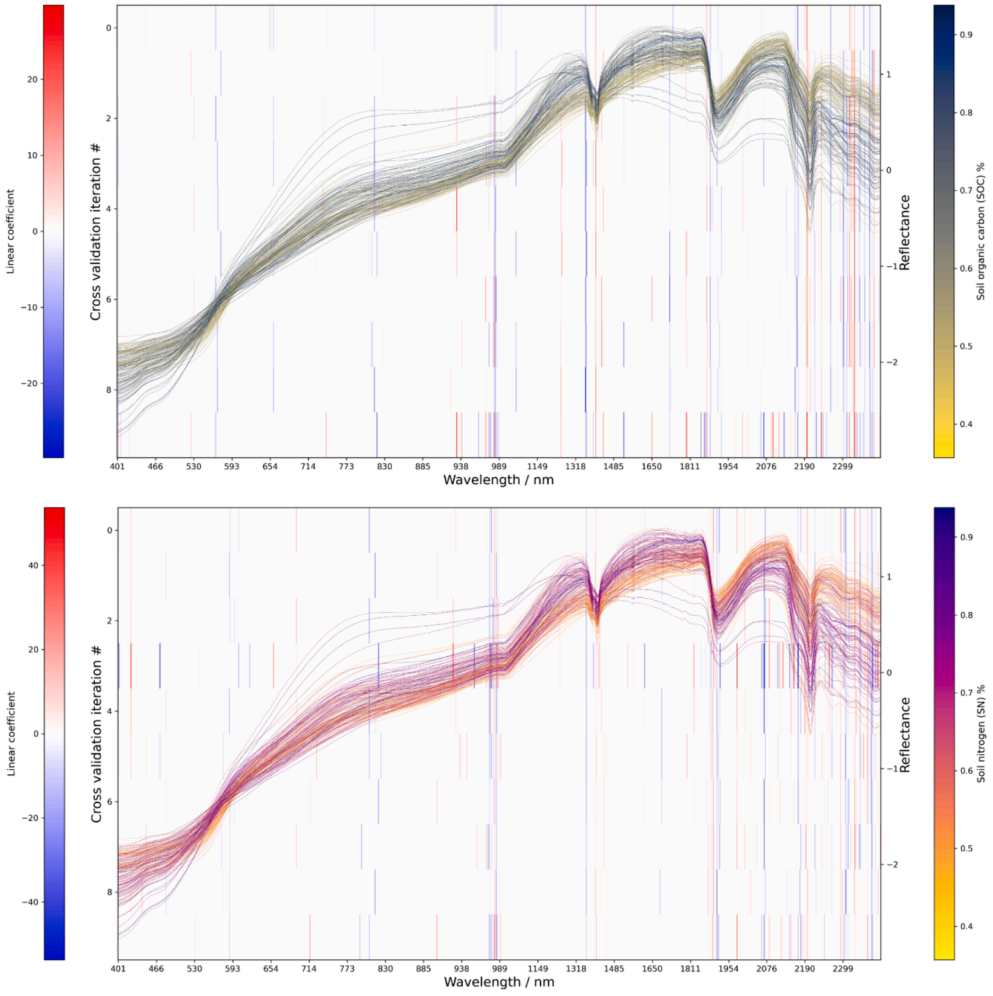

Fig. 7. Band selection by LASSO over 10 folds in SOC and SN prediction using full wavelength range (FWR) field spectrometer data. The strength of the correlation with spectra and SOC/SN is marked with the red to blue color palette. The white areas show bands with zero correlation coefficient and bands were not selected by LASSO. Mean normalized spectra are shown with line color indicating SOC concentration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and high values, as R^2 was between 0.50 and 0.68. Furthermore, we achieved RPD between 2.0 and 2.5 in FWR and SWIR, which indicates "very good quantitative prediction", whereas VNIR showed only "fair model" accuracy (RPD = 1.4–1.8) according to Viscarra Rossel et al., (2006b). Finally, RPIQ range of SOC prediction using FWR and SWIR can be characterized as "good prediction" accuracy (RPIQ = 3.37–4.05) according to Xu et al. (2023).

The performance of the PLSR model for SOC prediction in the FWR showed good correspondence in accuracy with previous studies (Chen et al., 2019; Lazaar et al., 2020; Miloš & Bensa, 2017; Morellos et al., 2016; Wijewardane et al., 2016). Moreover, the accuracy of our PLSR model in VNIR and SWIR aligns with the earlier findings (Miloš & Bensa, 2017). dos Santos et al. (2023), Guo et al. (2022), and Pellikka et al. (2023) concluded that employing LASSO for important band selection improved overall accuracy in all tested models in SOC prediction. This is in good agreement with improved accuracy of our LASSO model, when considering all wavelength bands for SOC prediction. We noticed that the predictive capacity of GPR was highest in SWIR among our all-tested models and spectral ranges. Similarly, Dotto et al. (2018) demonstrated the potential of the GPR model, which utilized a faster learning process with FWR spectral data in SOC prediction.

4.1.2. Soil nitrogen

The accuracy of the GPR model of SN in FWR and SWIR indicates "approximate quantification" with R^2 between 0.66 and 0.81 (Williams et al., 2019). All the models in VNIR achieved $R^2 < 0.50$, which indicates incapability of the model in predicting SN. SN prediction with all tested models with FWR and SWIR as predictor achieved "fair" prediction (RPD = 1.4–1.8) accuracy as indicated in Viscarra Rossel et al., (2006a). In contrast, the accuracy of all three models in VNIR is poor, where only low and high SN can be discriminated with RPD between 1 and 1.4. SN with GPR model in FWR and SWIR yielded approximate model accuracy (RPIQ = 2.70–3.37) where LASSO and PLSR achieved "fair" SN prediction (RPIQ range 2.02 to 2.70). All models managed to achieve "fair" SN estimation when predicted with VNIR range with RPIQ ranging from 2.02 to 2.70.

Our PLSR model underperformed in predicting SN in comparison to He et al. (2005), Vibhute et al. (2020), and Xiao & He (2019). However, results from Erler et al. (2020), Madari et al. (2006), and Martin et al. (2002) show good correspondence with our results, with the PLSR model having lower R^2 (range 0.51 to 0.68) with different sample sizes and range of SN concentrations over the FWR. The LASSO model performed poorly in SN prediction using IS compared with the results obtained by Pellikka et al. (2023), as we excluded forest samples with higher SN. More investigation is required in testing LASSO across a wider range of SN concentrations over FWR using FS. Martínez-España et al. (2019) concluded that GPR is the most reliable and the best performing method in SN prediction among all employed models including PLSR, which aligns with our GPR model in reaching the highest accuracy in SN prediction over FWR using FS. Erler et al. (2020) utilized PLSR, LASSO, and GPR models for SN estimation but observed lower accuracy compared with our results.

4.2. Potential of VNIR and SWIR spectral regions for soil properties modeling

Studies from Lazaar et al. (2020) and Miloš & Bensa (2017) validate our findings, indicating that the SWIR spectrum is better in predicting SOC compared with VNIR spectrum. Moreover, Miloš & Bensa, (2017) found very low accuracy in visible (400–700 nm) with R^2 of 0.42 and RPD of 1.13. This strongly supports our results in VNIR (400–1000 nm) for both IS and FS in showing the lowest accuracy with all tested models ($R^2 < 0.67$ and RPD < 1.75). However, using VNIR as predictor from IS, Pellikka et al. (2023) achieved R^2 up to 0.78 and RPD of 2.25 with the soil sample including higher SOC range from forest, which we did not incorporate in our SOC modeling. Therefore, this confirms that VNIR does not have enough variance to explain the variation in low SOC range (Table 1). However, all the models can achieve good predictive accuracy in low-range SOC estimation using SWIR as predictors, either independently or pairing it with VNIR.

Our model better estimated SN when SWIR range was included as a predictor, which is consistent with the findings by Madari et al. (2006), Mouazen et al. (2006), and Nawar & Mouazen, (2017). Mouazen et al. (2006) achieved R^2 of 0.69 over 350–2500 nm and a much lower R^2 of 0.56 over 300–1700 nm. This suggests the potential of SWIR range for the accurate estimation of nitrogen. Similar to SOC prediction (section 4.1.1), both FS and IS performed poorly in estimating SN by employing only VNIR wavelength range as predictor. However, the slightly better results in IS compared with FS can be explained by better sampling of the spectra over the samples captured by IS.

The visible-short wave NIR region (400–1000 nm) primarily exhibits absorption features attributed to iron-oxides while in the NIR-SWIR (1000–2500 nm) region, absorption feature may arise from water, clay minerals, and organic matter (Viscarra Rossel & Behrens, 2010). LASSO results show that for both SOC and SN, the most important spectral bands were in the SWIR region, with very few bands selected from VNIR range when using FWR data from FS. In our study, selected bands were mostly from overtone bands between 2000–2400 nm where C–H, C = O, N–H, O–H and S-H show sensitivity to soil reflectance (Dalal & Henry, 1986; Viscarra Rossel & Behrens, 2010). Bands located around 2200 and 2300 nm are found to be important for SOC calibration by many researchers (Ben-Dor & Banin, 1995a; Dalal & Henry, 1986; Stenberg et al., 2010; Viscarra Rossel & Behrens, 2010). This highlights the importance of SWIR region and explains why VNIR models performed poorer in both cases.

4.3. The potential of proximal sensing techniques for regenerative agriculture initiatives

RA initiatives require rapid but reliable estimates of SOC and SN, possibly in remote locations with limited access to well-equipped laboratories. Therefore, transportation of the field-collected samples for analysis involves a significant cost and can limit sample size. For example, ML-based remote-sensing models for mapping and monitoring soil characteristics are preferably based on a large number of samples (Angelopoulou et al., 2019).

We compared two sensors, SVC HR1024i FS and Specim IQ IS, for modeling SOC and SN to reduce the amount of time and resources needed for laboratory analysis and transportation of the field samples to remote laboratories. Relatively high-cost FS collects FWR spectra, whereas relatively low-cost IS covers only VNIR range. Only FS with SWIR range performed very well for SOC modeling in accuracy in the considered low range of SOC concentrations. The models were not as good for SN, or when using only VNIR range and IS. This suggests that an instrument with SWIR range is preferable for the given application, which is supported by earlier studies (Lazaar et al., 2020; Miloš & Bensa, 2017). However, in the VNIR range, IS, which captures images instead of point-wise measurements of FS, performed slightly better. Therefore, image-wise measurements may be preferable in some cases. Increasing the number of point-wise measurements may also reduce the difference between FS and IS, although this remains to be studied.

Good modeling results based on FS demonstrate that the measurement protocol using the leaf clip and contact probe (LC-RP PRO) can achieve adequate modeling results. One of the benefits with LC-RP PRO is that it has an internal light source. This provides the benefits of conducting spectral measurements in any light condition with FS coupled with LC-RP PRO. An internal light source makes use of FS easier in field laboratories; therefore, separate external light sources are not needed. Moreover, FS generates simple outputs (basically text files) with small file size, while IS creates larger image files, which requires computation resources and development of more demanding processing pipelines.

Considering the performance of all the tested models, FS has the ability to measure SOC and SN at an adequate accuracy utilizing the SWIR range for undertaking RA initiatives. In our study, we compared VNIR spectral range among IS and FS, as IS sensor (Specim IQ) does not cover the SWIR spectral range, which is one of the limitations of this study. Moreover, although our study covers a range of altitudinal zones within the Taita Hills, it is still limited to this one region and we plan to assess the transferability of the models to different agricultural areas in the future studies.

5. Conclusion

We assessed the compatibility of FS and IS over different wavelength ranges (VNIR, FWR, and SWIR) in modeling SOC and SN to complement expensive and time-demanding laboratory measurements for low-range SOC and SN concentration across different land covers. We developed a controlled soil spectral measurement protocol for FS coupled with a contact probe. We tested three ML models (PLSR, LASSO, and GPR) in predicting SOC and SN. We conclude the following based on our results:

- SOC estimation accuracy over FWR and SWIR with FS is competitive according to literature-defined accuracy standards;
- SOC and SN estimation models yielded high accuracy over SWIR independently or combined with VNIR;
- GPR achieved the highest model accuracy in SOC and SN prediction;
- All tested models show slightly better SOC estimation with IS than FS in VNIR range.

For future studies, our models should be assessed for transferability to soil samples collected from a different area. Detailed soil type classification of the soil samples may provide insight on how soil spectral signatures are attributed to soil class and SOC and SN concentration.

CRediT authorship contribution statement

Ashfak Mahmud: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Markku Luotamo:** Writing – review & editing, Visualization, Data curation. **Kristiina Karhu:** Writing – review & editing. **Petri Pellikka:** Writing – review & editing. **Juuso Tuure:** Writing – review & editing. **Janne Heiskanen:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We are immensely thankful to Niklas Sädekoski for his essential role in soil sample collection and analysis with SpecimIQ. We acknowledge the help in field work from the Taita Research Station staff in Kenya, especially Mr. Mwadime Mjomba. Additionally, we extend our appreciation to Matti Räsänen for compiling the meteorological data obtained from the weather stations at Taita Research Station. We acknowledge the Academy of Finland for funding REACT (Regenerative agricultural systems for climate resilience in agroecological gradient in East Africa) [decision number: 351087, 2022]. We also acknowledge the National Commission for Science, Technology, and Innovation (NACOSTI) for granting research permits (no. P/23/23238 and P/23/23239) for Kenya.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.catena.2024.108180.

References

Abdalla, M., Hastings, A., Cheng, K., Yue, Q., Chadwick, D., Espenberg, M., Truu, J., Rees, R.M., Smith, P., 2019. A critical review of the impacts of cover crops on nitrogen leaching, net greenhouse gas balance and crop productivity. Glob. Chang. Biol. 25 (8), 2530–2543. https://doi.org/10.1111/gcb.14644.

Abera, T.A., Vuorinne, I., Munyao, M., Pellikka, P.K.E., Heiskanen, J., 2022. Land Cover Map for Multifunctional Landscapes of Taita Taveta County, Kenya, Based on Sentinel-1 Radar, Sentinel-2 optical, and topoclimatic data. Data 7 (3), Article 3. https://doi.org/10.3390/data7030036.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: a review. Remote Sensing 2019, Vol. 11, Page 676, 11(6), 676. doi: 10.3390/RS11060676.

Autio, A., Johansson, T., Motaroki, L., Minoia, P., Pellikka, P., 2021. Constraints for adopting climate-smart agricultural practices among smallholder farmers in Southeast Kenya. Agr. Syst. 194, 103284 https://doi.org/10.1016/J.AGSY.2021.103284.

Bao, N., Wu, L., Ye, B., Yang, K., Zhou, W., 2017. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. Geoderma 288, 47–55. https://doi.org/10.1016/j.geoderma.2016.10.033.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43 (5), 772–777. https://doi.org/10.1366/0003702894202201.

Behmann, J., Acebron, K., Emin, D., Bennertz, S., Matsubara, S., Thomas, S., Bohnenkamp, D., Kuska, M.T., Jussila, J., Salo, H., Mahlein, A.-K., Rascher, U., 2018. Specim IQ: evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. Sensors 18 (2), Article 2. https://doi.org/10.3390/s18020441.

Ben Dor, E., Ong, C., Lau, I.C., 2015. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma 245–246, 112–124. https://doi.org/10.1016/j.geoderma.2015.01.002.

Ben-Dor, E., Banin, A., 1995a. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Sci. Soc. Am. J. 59 (2), 364–372. https://doi.org/10.2136/sssaj1995.03615995005900020014x.

Ben-Dor, E., Banin, A., 1995b. Quantitative analysis of convolved Thematic Mapper spectra of soils in the visible near-infrared and shortwave-infrared spectral regions (0·4–2·5 μm). Int. J. Remote Sens. 16 (18), 3509–3528. https://doi.org/10.1080/01431169508954643.

Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Taylor, G.R., Hill, J., Whiting, M.L., Sommer, S., 2009. Using Imaging Spectroscopy to study soil properties. Remote Sens. Environ. 113, S38–S55. https://doi.org/10.1016/j.rse.2008.09.019.

Bouckaert, R.R., 2003. Choosing between two learning algorithms based on calibrated tests. In: Proceedings of the Twentieth International Conference on International Conference on Machine Learning, pp. 51–58.

Bricklemyer, R.S., Brown, D.J., Turk, P.J., Clegg, S., 2018. Comparing vis–NIRS, LIBS, and Combined vis–NIRS-LIBS for Intact Soil Core Soil Carbon Measurement. Soil Sci. Soc. Am. J. 82 (6), 1482–1496. https://doi.org/10.2136/sssaj2017.09.0332.

Cao, Y., Bao, N., Liu, S., Zhao, W., Li, S., 2020. Reducing moisture effects on soil organic carbon content prediction in visible and near-infrared spectra with an external parameter othogonalization algorithm. Can. J. Soil Sci. 100 (3), 253–262. https://doi.org/10.1139/cjss-2020-0009.

Chen, H., Pan, T., Chen, J., Lu, Q., 2011. Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods. Chemom. Intel. Lab. Syst. 107 (1), 139–146. https://doi.org/10.1016/j.chemolab.2011.02.008.

Chen, Y., Wang, J., Liu, G., Yang, Y., Liu, Z., Deng, H., 2019. Hyperspectral estimation model of forest soil organic matter in northwest Yunnan Province. China. *Forests* 10 (3), Article 3. https://doi.org/10.3390/f10030217.

Conant, R.T., Ogle, S.M., Paul, E.A., Paustian, K., 2011. Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. Front. Ecol. Environ. 9 (3), 169–173. https://doi.org/10.1890/090153.

Condit, H., 1970. The spectral reflectance of American Soils. Photogramm. Eng.

Dalal, R.C., Henry, R.J., 1986. Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. Soil Sci. Soc. Am. J. 50 (1), 120–123. https://doi.org/10.2136/sssaj1986.03615995005000010023x.

Demattê, J.A.M., Ramirez-Lopez, L., Marques, K.P.P., Rodella, A.A., 2017. Chemometric soil analysis on the determination of specific bands for the detection of magnesium and potassium by spectroscopy. Geoderma 288, 8–22. https://doi.org/10.1016/j.geoderma.2016.11.013.

Dick, W.A., Blevins, R.L., Frye, W.W., Peters, S.E., Christenson, D.R., Pierce, F.J., Vitosh, M.L., 1998. Impacts of agricultural management practices on C sequestration in forest-derived soils of the eastern Corn Belt. Soil Tillage Res. 47 (3–4), 235–244. https://doi.org/10.1016/S0167-1987(98)00112-3.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923. https://doi.org/10.1162/089976698300017197.

dos Santos, E. P., Moreira, M. C., Fernandes-Filho, E. I., Demattê, J. A. M., Santos, U. J. dos, da Silva, D. D., Cruz, R. R. P., Moura-Bueno, J. M., Santos, I. C., & Sampaio, E. V.

de S. B. (2023). Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data. *Ecological Informatics*, *77*, 102240. doi: 10.1016/j.ecoinf.2023.102240.

Dotto, A.C., Dalmolin, R.S.D., 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, 262–274. https://doi.org/10.1016/j.geoderma.2017.11.006.

Dyar, M.D., Carmosino, M.L., Breves, E.A., Ozanne, M.V., Clegg, S.M., Wiens, R.C., 2012. Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples. Spectrochim. Acta B At. Spectrosc. 70, 51–67. https://doi.org/10.1016/j.sab.2012.04.011.

Ehrenfeld, J.G., 2013. Plant-Soil Interactions. In: Levin, S.A. (Ed.), Encyclopedia of Biodiversity, Second Edition. Academic Press, pp. 109–128. https://doi.org/10.1016/B978-0-12-384719-5.00179-9.

Erler, A., Riebe, D., Beitz, T., Löhmannsröben, H.-G., Gebbers, R., 2020. Soil nutrient detection for precision agriculture using handheld laser-induced breakdown spectroscopy (LIBS) and Multivariate Regression Methods (PLSR, Lasso and GPR). Sensors 20 (2), Article 2. https://doi.org/10.3390/s20020418.

Gholizadeh, A., Neumann, C., Chabrillat, S., van Wesemael, B., Castaldi, F., Borůvka, L., Sanderman, J., Klement, A., Hohmann, C., 2021. Soil organic carbon estimation using VNIR–SWIR spectroscopy: The effect of multiple sensors and scanning conditions. Soil Tillage Res. 211, 105017 https://doi.org/10.1016/j.still.2021.105017.

Guo, H., Zhang, R., Dai, W., Zhou, X., Zhang, D., Yang, Y., Cui, J., 2022. Mapping Soil Organic Matter Content Based on Feature Band Selection with ZY1-02D Hyperspectral Satellite Data in the Agricultural Region. Agronomy 12 (9), Article 9. https://doi.org/10.3390/agronomy12092111.

Haynes, R. J. (2005). Labile Organic Matter Fractions as Central Components of the Quality of Agricultural Soils: An Overview. In *Advances in Agronomy* (Vol. 85, pp. 221–268). Academic Press. doi: 10.1016/S0065-2113(04)85005-3.

He, Y., Song, H., Pereira, A.G., Gómez, A.H., 2005. Measurement and analysis of soil nitrogen and organic matter content using near-infrared spectroscopy techniques. J. Zhejiang Univ. Sci. B 6 (11), 1081–1086. https://doi.org/10.1631/jzus.2005.B1081.

Hong, Y., Chen, S., Zhang, Y., Chen, Y., Yu, L., Liu, Y., Liu, Y., Cheng, H., Liu, Y., 2018. Rapid identification of soil organic matter level via visible and near-infrared spectroscopy: Effects of two-dimensional correlation coefficient and extreme learning machine. Sci. Total Environ. 644, 1232–1243. https://doi.org/10.1016/j.scitotenv.2018.06.319.

Jaetzold, R., Schmidt, H., 1983. *Farm Management Handbook of Kenya. East Kenya* (pp. 245–285). Kenya Ministry of Agriculture. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1747826.

Jaetzold, R., Schmidt, H., & Shisanya, C. (2012). *Coast Province: Taita-Taveta County, in: Farm Management Handbook of Kenya VOL.II*. Ministry of Agriculture, Nairobi.

Jakab, G., Rieder, Á., Vancsik, A.V., Szalai, Z., 2018. Soil organic matter characterisation by photometric indices or photon correlation spectroscopy: Are they comparable? *Hungarian Geographical Bulletin*, *67*(2). Article 2. https://doi.org/10.15201/hungeobull.67.2.1.

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M. F., Bampa, F., van Wesemael, B., Harrison, R. B., Guerrini, I. A., Richter, D. deB, Rustad, L., Lorenz, K., Chabbi, A., & Miglietta, F. (2014). Current status, uncertainty and future needs in soil organic carbon monitoring. *Science of The Total Environment*, *468–469*, 376–383. doi: 10.1016/j.scitotenv.2013.08.026.

Johnson, C.E., Ruiz-Méndez, J.J., Lawrence, G.B., 2000. Forest Soil Chemistry and Terrain Attributes in a Catskills Watershed. Soil Sci. Soc. Am. J. 64 (5), 1804–1814. https://doi.org/10.2136/sssaj2000.6451804x.

Keesstra, S., Pereira, P., Novara, A., Brevik, E.C., Azorin-Molina, C., Parras-Alcántara, L., Jordán, A., Cerdà, A., 2016. Effects of soil management techniques on soil water erosion in apricot orchards. Sci. Total Environ. 551–552, 357–366. https://doi.org/10.1016/j.scitotenv.2016.01.182.

Lazaar, A., Mouazen, A.M., EL Hammouti, K., Fullen, M., Pradhan, B., Memon, M.S., Andich, K., & Monir, A. (2020). The application of proximal visible and near-infrared spectroscopy to estimate soil organic matter on the Triffa Plain of Morocco. *International Soil and Water Conservation Research*, *8*(2), 195–204. doi: 10.1016/J.ISWCR.2020.04.005.

Li, Z., Deng, C., Zhao, B., Tian, Y., & Huang, Y. (2019). Hyperspectral inversion for soil moisture and temperature based on Gaussian process regression. In: 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), 1–4. doi: 10.1109/ICSIDP47821.2019.9172823.

Loisel, J., Casellas Connors, J.P., Hugelius, G., Harden, J.W., Morgan, C.L., 2019. Soils can help mitigate CO2 emissions, despite the challenges. Proc. Natl. Acad. Sci. 116 (21), 10211–10212. https://doi.org/10.1073/pnas.1900444116.

Madari, B.E., Reeves, J.B., Machado, P.L.O.A., Guimarães, C.M., Torres, E., McCarty, G. W., 2006. Mid- and near-infrared spectroscopic assessment of soil compositional parameters and structural indices in two Ferralsols. Geoderma 136 (1), 245–259. https://doi.org/10.1016/j.geoderma.2006.03.026.

Martens, H., Næs, T., 1992. Multivariate Calibration. John Wiley & Sons.

Martin, P.D., Malley, D.F., Manning, G., Fuller, L., 2002. Determination of soil organic carbon and nitrogen at the field level using near-infrared spectroscopy. Can. J. Soil Sci. 82 (4), 413–422. https://doi.org/10.4141/S01-054.

Martínez-España, R., Bueno-Crespo, A., Soto, J., Janik, L.J., Soriano-Disla, J.M., 2019. Developing an intelligent system for the prediction of soil properties with a portable mid-infrared instrument. Biosyst. Eng. 177, 101–108. https://doi.org/10.1016/j.biosystemseng.2018.09.013.

Meireles, J.E., Schweiger, A.K., Cavender-Bares, J., 2023. *spectrolab: Class and Methods for Spectral Data* (0.0.18). [Computer software].

Miloš, B., Bensa, A., 2017. Prediction of soil organic carbon using VIS-NIR spectroscopy: Application to Red Mediterranean soils from Croatia. *Eurasian*. J. Soil Sci. 6 (4), Article 4. https://doi.org/10.18393/ejss.319208.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., Mouazen, A.M., 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst. Eng. 152, 104–116. https://doi.org/10.1016/j.biosystemseng.2016.04.018.

Mouazen, A.M., De Baerdemaeker, J., Ramon, H., 2005. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. Soil Tillage Res. 80 (1), 171–183. https://doi.org/10.1016/j.still.2004.03.022.

Mouazen, A.M., De Baerdemaeker, J., Ramon, H., 2006. Effect of wavelength range on the measurement accuracy of some selected soil constituents using visual-near infrared spectroscopy. J. Near Infrared Spectrosc. 14 (3), 189–199. https://doi.org/10.1255/jnirs.614.

Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Arrouays, D., 2016. GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. Sci. Total Environ. 573, 1352–1369. https://doi.org/10.1016/j.scitotenv.2016.07.066.

Nawar, S., Mouazen, A.M., 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. Catena 151, 118–129. https://doi.org/10.1016/j.catena.2016.12.014.

Nelson, D. w., & Sommers, L. e. (1983). Total Carbon, Organic Carbon, and Organic Matter. In *Methods of Soil Analysis* (pp. 539–579). John Wiley & Sons, Ltd. doi: 10.2134/agronmonogr9.2.2ed.c29.

Njeru, C.M., Ekesi, S., Mohamed, S.A., Kinyamario, J.I., Kiboi, S., Maeda, E.E., 2017. Assessing stock and thresholds detection of soil organic carbon and nitrogen along an altitude gradient in an east Africa mountain ecosystem. Geoderma Reg. 10, 29–38. https://doi.org/10.1016/j.geodrs.2017.04.002.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E.B., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Wetterlind, J., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. Adv. Agron. 132, 139–159. https://doi.org/10.1016/BS.AGRON.2015.02.002.

O' Rourke, S. M., & Holden, N. M. (2011). Optical sensing and chemometric analysis of soil organic carbon – a cost effective alternative to conventional laboratory methods? *Soil Use and Management*, *27*(2), 143–155. doi: 10.1111/j.1475-2743.2011.00337.x.

Omran, E.S.E., 2017. Rapid prediction of soil mineralogy using imaging spectroscopy. Eurasian Soil Sci. 50 (5), 597–612. https://doi.org/10.1134/S106422931705012X.

Ontl, T., Schulte, L.A., 2012. Soil Carbon Storage. Nature Education Knowledge 3 (10), 35.

Parton, W.J., Schimel, D.S., Cole, C.V., Ojima, D.S., 1987. Analysis of factors controlling soil organic matter levels in Great Plains Grasslands. Soil Sci. Soc. Am. J. 51 (5), 1173–1179. https://doi.org/10.2136/sssaj1987.03615995005100050015x.

Pellikka, P., Clark, B.J.F., Gosa, A.G., Himberg, N., Hurskainen, P., Maeda, E., Mwang'ombe, J., Omoro, L.M.A., & Siljander, M. (2013). Chapter 13—Agricultural Expansion and Its Consequences in the Taita Hills, Kenya. In P. Paron, D. O. Olago, & C. T. Omuto (Eds.), *Developments in Earth Surface Processes* (Vol. 16, pp. 165–179). Elsevier. doi: 10.1016/B978-0-444-59559-1.00013-X.

Pellikka, P.K.E., Heikinheimo, V., Hietanen, J., Schäfer, E., Siljander, M., Heiskanen, J., 2018. Impact of land cover change on aboveground carbon stocks in Afromontane landscape in Kenya. Appl. Geogr. 94, 178–189. https://doi.org/10.1016/j.apgeog.2018.03.017.

Pellikka, P., Luotamo, M., Sädekoski, N., Hietanen, J., Vuorinne, I., Räsänen, M., Heiskanen, J., Siljander, M., Karhu, K., Klami, A., 2023. Tropical altitudinal gradient soil organic carbon and nitrogen estimation using Specim IQ portable imaging spectrometer. Sci. Total Environ. 883, 163677 https://doi.org/10.1016/j.scitotenv.2023.163677.

Poeplau, C., Don, A., 2015. Carbon sequestration in agricultural soils via cultivation of cover crops – A meta-analysis. Agr Ecosyst Environ 200, 33–41. https://doi.org/10.1016/j.agee.2014.10.024.

Powlson, D.S., Brookes, P.C., Whitmore, A.P., Goulding, K.W.T., Hopkins, D.W., 2011. Soil Organic Matters. Eur. J. Soil Sci. 62 (1), 1–4. https://doi.org/10.1111/j.1365-2389.2010.01338.x.

Prescott, C.E., Rui, Y., Cotrufo, M.F., Grayston, S.J., 2021. Managing plant surplus carbon to generate soil organic matter in regenerative agriculture. J. Soil Water Conserv. 76 (6), 99A–104A. https://doi.org/10.2489/jswc.2021.0920A.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing* [Computer software]. https://www.R-project.org/.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013. The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma 195–196, 268–279. https://doi.org/10.1016/j.geoderma.2012.12.014.

Rasmussen, C.E., 2004. Gaussian Processes in Machine Learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (Eds.), Advanced Lectures on Machine Learning: ML Summer Schools 2003. Springer, Canberra, Australia, pp. 63–71. https://doi.org/10.1007/978-3-540-28650-9_4. February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised Lectures.

Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32 (3), 569–575. https://doi.org/10.1109/TPAMI.2009.187.

Sáez-Plaza, P., Navas, M.J., Wybraniec, S., Michałowski, T., Asuero, A.G., 2013. An Overview of the Kjeldahl Method of Nitrogen Determination. Part II. Sample

Preparation, Working Scale, Instrumental Finish, and Quality Control. Crit. Rev. Anal. Chem. 43 (4), 224–272. https://doi.org/10.1080/10408347.2012.751787.
Serbin, S.P., Wu, J., Ely, K.S., Kruger, E.L., Townsend, P.A., Meng, R., Wolfe, B.T., Chlus, A., Wang, Z., Rogers, A., 2019. From the Arctic to the tropics: Multibiome prediction of leaf mass per area using leaf reflectance. New Phytol. 224 (4), 1557–1568. https://doi.org/10.1111/nph.16123.
Sinfield, J.V., Fagerman, D., Colic, O., 2010. Evaluation of sensing technologies for on-the-go detection of macro-nutrients in cultivated soils. Comput. Electron. Agric. 70 (1), 1–18. https://doi.org/10.1016/j.compag.2009.09.017.
Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical Bayesian Optimization of Machine Learning Algorithms. Adv. Neural Inf. Proces. Syst. 25 https://doi.org/10.48550/arXiv.1206.2944.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter Five—Visible and Near Infrared Spectroscopy in Soil Science. In: Sparks, D.L. (Ed.), Advances in Agronomy, Vol. 107. Academic Press, pp. 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.
Stoner, E.R., Baumgardner, M.F., 1981. Characteristic Variations in Reflectance of Surface Soils. Soil Sci.Society of America Journal 45 (6), 1161–1165. https://doi.org/10.2136/sssaj1981.03615995004500060031x.
Stuart, B.H., 2004. Infrared Spectroscopy: Fundamentals and Applications Vol. 8. https://doi.org/10.1002/0470011149.
Svc, 2020. *SVC LC-RP Pro Manual 2.4* (User Mannual Revision 2.4. Spectra Vista Corporation, pp. 1–29.
Svc, 2021. *SVC HR i Series User Manual Revision 1.17* (User Mannual Revision 1.17. Spectra Vista Corporation, pp. 1–147.
Tan, B., You, W., Tian, S., Xiao, T., Wang, M., Zheng, B., & Luo, L. (2022). Soil Nitrogen Content Detection Based on Near-Infrared Spectroscopy. *Sensors*, *22*(20), Article 20. doi: 10.3390/s22208013.
Tan, S.S.X., Kuebbing, S.E., 2023. A synthesis of the effect of regenerative agriculture on soil carbon sequestration in Southeast Asian croplands. Agr Ecosyst Environ 349, 108450. https://doi.org/10.1016/j.agee.2023.108450.
Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer. https://doi.org/10.1007/978-1-4757-3264-1.
Vibhute, A.D., Kale, K.V., Gaikwad, S.V., Dhumal, R.K., 2020. Estimation of soil nitrogen in agricultural regions by VNIR reflectance spectroscopy. SN Applied Sciences 2 (9), 1523. https://doi.org/10.1007/s42452-020-03322-9.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1), 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.
Viscarra Rossel, R.A., McGlynn, R.N., McBratney, A.B., 2006a. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. Geoderma 137 (1), 70–82. https://doi.org/10.1016/j.geoderma.2006.07.004.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006b. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1), 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.
Vuorinne, I., Heiskanen, J., Maghenda, M., Mwangala, L., Muukkonen, P., Pellikka, P.K.E., 2021. Allometric models for estimating leaf biomass of sisal in a semi-arid environment in Kenya. Biomass Bioenergy 155, 106294. https://doi.org/10.1016/j.biombioe.2021.106294.
Wachiye, S., Merbold, L., Vesala, T., Rinne, J., Leitner, S., Räsänen, M., Vuorinne, I., Heiskanen, J., Pellikka, P., 2021. Soil greenhouse gas emissions from a sisal chronosequence in Kenya. Agric. For. Meteorol. 307, 108465 https://doi.org/10.1016/j.agrformet.2021.108465.
Walkley, A., 1935. An Examination of Methods for Determining Organic Carbon and Nitrogen in Soils. J. Agric. Sci. 25 (4), 598–609. https://doi.org/10.1017/S0021859600019687.
Wang, Z., Chlus, A., Geygan, R., Ye, Z., Zheng, T., Singh, A., Couture, J.J., Cavender-Bares, J., Kruger, E.L., Townsend, P.A., 2020. Foliar functional traits from imaging spectroscopy across biomes in eastern North America. New Phytol. 228 (2), 494–511. https://doi.org/10.1111/nph.16711.
Wijewardane, N.K., Ge, Y., Wills, S., Loecke, T., 2016. Prediction of Soil Carbon in the Conterminous United States: visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. Soil Sci. Soc. Am. J. 80 (4), 973–982. https://doi.org/10.2136/sssaj2016.02.0052.
Williams, C., Rasmussen, C., 1995. Gaussian processes for regression. Adv. Neural Inform. Process. Syst., 8. https://papers.nips.cc/paper_files/paper/1995/hash/7cce53cf90577442771720a370c3c723-Abstract.html.
Williams, P., Manley, M., Antoniszyn, J., 2019. Near Infrared Technology: Getting the best out of light. AFRICAN SUN MeDIA.
Witten, I.H., Frank, E., Hall, M.A., 2011. *Data mining: Practical machine learning tools and techniques* (3rd ed). Morgan Kaufmann.
Xiao, S., He, Y., 2019. Application of near-infrared spectroscopy and multiple spectral algorithms to explore the effect of soil particle sizes on soil nitrogen detection. Molecules 24 (13), Article 13. https://doi.org/10.3390/molecules24132486.
Xu, X., Chen, Y., Dai, X., Lei, T., Wang, S., Li, K., 2023. An Improved Vis-NIR estimation model of soil organic matter through the artificial samples enhanced calibration set. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 4626–4637. https://doi.org/10.1109/JSTARS.2023.3275745.
Yang, D., Bao, W., 2017. Group lasso-based band selection for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 14 (12), 2438–2442. https://doi.org/10.1109/LGRS.2017.2768074.
Yang Yang, G.X., 2015. Hyperspectral retrieval of soil organic matter for different soil types in the three-river headwaters region. Remote Sensing Technology and Application 30 (1), 186–198. https://doi.org/10.11873/j.issn.1004-0323.2015.1.0186.