International Journal of Design & Nature and Ecodynamics
Vol. 19, No. 4, August, 2024, pp. 1405-1413
Journal homepage: http://iieta.org/journals/ijdne

Near Infrared Spectroscopy for Rapid Prediction of Soil Organic Carbon Content in Agricultural Soils

Darusman Darusman[1,2], Agus Arip Munawar[2,3*], Zulfahrizal Zulfahrizal[3]

[1] Department of Soil Science, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[2] Center for Biochar and Sustainable Tropical Forest, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[3] Department of Agricultural Engineering, PUSMEPTAN Universitas Syiah Kuala, Banda Aceh 23111, Indonesia

Corresponding Author Email: aamunawar@usk.ac.id

https://doi.org/10.18280/ijdne.190433

Received: 17 October 2023
Revised: 1 February 2024
Accepted: 19 March 2024
Available online: 28 August 2024

Keywords:
NIRS, model, PLSR, soil classification, SOC prediction

ABSTRACT

Soil organic carbon (SOC) or C-organic is a key component of soil quality that affects the properties of organic materials and soil mixtures. It also holds practical value and importance in agriculture. Traditionally, determining SOC has involved expensive and time-consuming procedures that require the use of chemicals and may cause pollution. Therefore, there is a need for an alternative method that is fast, environmentally friendly, and cost-effective, as they are key important factors in precision agriculture practices the near infrared reflectance spectroscopy (NIRS) technique can be considered as a suitable option since it is non-destructive, requires simple preparation, and does not cause pollution. The main objective of this study is to apply the NIRS technique to predict SOC levels and classify soils based on geographical and soil land-use characteristics. Soil samples were collected from four different locations, and their spectra data were acquired in the wavenumbers range of 4000-10,000 cm^{-1}. A prediction model for SOC was developed using the NIR spectra data and the partial least squares regression (PLSR) method, followed by k-fold cross-validation. The results demonstrated that the NIRS technique successfully predicted SOC levels, with a maximum correlation coefficient (r) of 0.96 and a residual predictive deviation (RPD) index of 4.05, indicating excellent model performance. In conclusion, the NIRS technique can be applied as a rapid and non-destructive method for predicting C-organic levels and classifying soil characteristics.

1. INTRODUCTION

Soil serves as the primary medium for plant growth, including food crops and plantations. Healthy soil, characterized by suitable physical and chemical properties, is essential for optimal plant growth. Visual indicators of healthy soil include texture, structure, and moisture content [1]. The chemical properties of soil, particularly nutrient levels, vary throughout different stages of plant growth. Imbalances in soil nutrients can negatively impact plant growth, leading to slower growth rates and increased susceptibility to disease. Additionally, excessive nutrient levels can have adverse effects on plant growth and the environment [2].

In precision agriculture, the excessive use of fertilizers can result in environmental pollution by creating unnecessary nutrient deposits. Soil organic carbon (SOC), also known as organic carbon content, plays a vital role in precision agriculture [3, 4]. Understanding the levels of SOC and soil fraction is crucial in determining appropriate fertilization dosages and fertilizer types to ensure adequate nutrient availability in the soil. These factors also influence soil structure, which affects plant growth and water storage capacity.

SOC is a main component of soil organic matter (SOM) and serves as a reservoir of nutrients for plants, including nitrogen, phosphorus, and sulfur.

As microorganisms break down organic matter, these nutrients are released in a form that plants can absorb and utilize for growth, a process that is essential for the productivity of agricultural systems. Organic carbon is vital for soil structure. It helps bind together soil particles into aggregates, which improves soil structure and stability. This aggregation allows for better air and water movement through the soil, reduces erosion, and promotes root growth. Good soil structure is crucial for agricultural soils to sustain and improve crop yields.

Soils rich in organic carbon typically have improved water retention capabilities. SOC can influence the soil's ability to absorb and hold water, which is beneficial for plant growth, especially in drier regions or during periods of drought. Moreover, soils with higher organic carbon levels have better infiltration rates, reducing runoff and the likelihood of flooding. SOC is fundamental to maintaining a diverse and active soil biota, including bacteria, fungi, and soil fauna. These organisms are key for decomposing organic matter, cycling nutrients, and they even help suppress soil-borne diseases. The biological activity driven by SOC is a critical factor for maintaining and enhancing soil fertility.

From a global environmental perspective, SOC is an important carbon sink. Through the process of photosynthesis, plants fix carbon from the atmosphere and ultimately deposit some of this carbon in the soil as organic matter. By storing carbon in soils, it is sequestered from the atmosphere, which helps mitigate climate change by reducing greenhouse gas concentrations.

Soils with higher organic carbon levels often exhibit better structure and greater resistance to erosion. Through the bolstering of aggregate stability, SOC lessens the susceptibility of soil to being washed or blown away, which is critical for preserving topsoil and maintaining land productivity. Soils with high levels of SOC tend to be more resilient, they recover more quickly from disturbances such as drought or compaction. This resilience is key in a changing climate, where extreme weather events are becoming more common. The presence of adequate SOC buffers soils against such changes and helps sustain agricultural productivity. SOC also influences soil pH by acting as a buffer against changes. This is important for maintaining a pH range that is conducive to crop growth and beneficial soil microbial activity.

Accurate knowledge of SOC and soil fraction aids in decision-making processes for precision agriculture, such as plant selection and suitable fertilization and irrigation practices [4, 5]. However, predicting organic carbon content and soil fractions in real-time poses challenges. Laboratory testing is time-consuming and complicated, often requiring the use of chemicals that can contribute to environmental pollution. Therefore, there is a need for alternative methods that are fast, reliable, non-destructive, and environmentally friendly to determine soil quality parameters, including organic carbon content.

Near-infrared reflectance spectroscopy (NIRS) is one such method that has been successfully applied in various sectors, including agriculture and soil science [6]. NIRS works by measuring the interaction between electromagnetic radiation and biological objects. It offers advantages such as simple sample preparation, non-destructiveness, absence of chemical waste, and high-speed analysis [7, 8].

Numerous studies have shown that NIRS is a viable tool for predicting quality attributes in agriculture, with strong prediction model performance indicated by correlation coefficients ranging from 0.93 to 0.99 and residual predictive deviation (RPD) indices categorized as coarse, sufficient, or excellent [9-11]. Building on these advantages, we aim to apply the NIRS method to predict soil quality parameters, specifically C-organic or SOC. Prediction models will be established using soil spectra from near-infrared data and the partial least square regression (PLSR) method. The results will be compared to actual SOC measurements obtained through standard laboratory procedures.

One of the primary challenges is the requirement for extensive calibration. NIRS must be calibrated against reference laboratory data, which can be a time-consuming process and requires a diverse set of soil samples to develop accurate models. Additionally, due to soil's heterogeneous nature comprising various organic and inorganic components, obtaining consistent NIRS readings can be difficult. Such variability necessitates a broad calibration covering a wide range of soil types.

The moisture content of soil also significantly affects NIR measurements, as water has strong NIR absorption bands. Calibrations must, therefore, either control for soil moisture or directly include it in the measurement process to ensure accuracy. Moreover, NIRS has lower sensitivity to components present in trace amounts, such as micronutrients and certain pollutants, which limits its ability to detect these elements related to soil quality.

Physical characteristics, like the roughness of the soil surface and the presence of debris, can interfere with NIR reflectance readings. This calls for careful sample preparation to ensure that such factors do not compromise the analysis's accuracy. There's also the matter of spectral interferences, where organic matter and other soil constituents can cause overlapping spectral features that make distinguishing between different compounds challenging.

Lastly, the accuracy of NIRS measurements can be influenced by environmental factors like ambient light, temperature, and humidity, which need to be controlled or accounted for during analysis. Despite these challenges, advancements in technology are continuously improving NIRS applications in soil quality assessment.

2. MATERIALS AND METHODS

2.1 Soil samples

Soil samples are collected from *Aceh Besar* district to capture the spatial variability of the research area. These are often mixed to create a representative composite sample. The collected samples are then air-dried at room temperature; heat is avoided as it can alter the soil's organic components. The drying process is followed by sieving the soil through a < 2mm mesh to remove larger particles such as stones and roots, which could interfere with the NIRS results. Subsequently, the sieved soil is finely ground to ensure a consistent particle size, as NIRS is sensitive to variations in particle size which can affect the scattering of NIR light.

Homogenization of the finely ground soil is then conducted to ensure uniformity throughout the sample. This step is crucial as heterogeneity within the sample can impact the spectral analysis.

2.2 Spectral data of soil samples

Soil samples were analyzed using an infrared instrument to collect near-infrared spectral data in the form of diffuse reflectance spectrum. For the actual NIRS scanning, the homogenized soil is placed into a suitable, non-reflective petri dish consistency in container size and shape is maintained to negate any spectral variations that might arise from these factors. The collected infrared spectra covered a range of wavenumbers from 4000 to 10,000 cm^{-1} with 32 scans averaged [12]. The resulting spectra data were stored in three different file formats: *.SPA, *.JDX, and *.CSV. respectively.

2.3 Actual SOC measurement

After collecting the spectra, the soil samples were immediately measured for soil organic carbon (SOC) using an elemental analyzer and thermal conductivity detector. These measurements were done in triplicate and averaged [4]. The soil organic carbon data were expressed as a percentage of SOC.

2.4 Spectral data corrections

To ensure accurate and reliable prediction results, the near-infrared spectral data of the soil samples were enhanced and corrected using de-trending (DT), multiplicative scatter correction (MSC), and a combination of both (DT+MSC).

DT is particularly important in removing linear trends found within the spectral data, which often originate from instrumental effects or baseline drift that occur over time. These trends, if left uncorrected, can mask the true spectral features that are indicative of the sample's properties.

On the other hand, MSC is employed to mitigate scattering effects that arise due to variations in particle size, surface reflectance properties, and differences in the optical path of the NIR light across different samples. These variations can lead to multiplicative errors in the recorded spectra. MSC works by standardizing the spectral data against a reference spectrum, attenuating these scatter-related discrepancies.

This normalization process helps ensure that differences in the spectral readings are due to genuine chemical variations among samples rather than physical inhomogeneities, thus bolstering the accuracy of multivariate analysis models used in the interpretation of NIR spectra.

When DT and MSC are used in combination (DT+MSC), they provide a synergistic benefit. This combination aims to harness the strengths of both methods to achieve a high-quality representation of the spectral data. MSC is capable of addressing multiplicative errors, while DT concurrently corrects any additive or linear trends not accounted for by MSC exclusively.

2.5 Prediction models

Prediction models were then established to predict the SOC of the soil samples using the raw spectrum data and the enhanced spectrum data (DT, MSC, and DT+MSC). Partial least square regression (PLSR) was used as the regression method for developing these prediction models as illustrated in Figure 1.

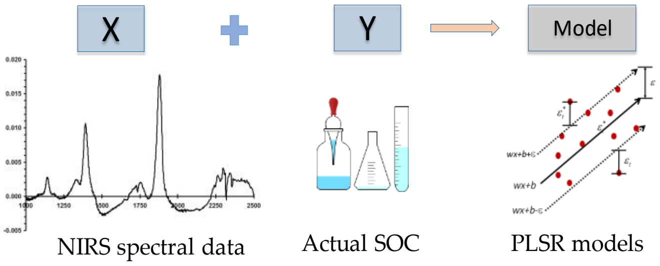


Figure 1. Building NIRS prediction model using PLSR algorithm to determine SOC

The performance of the prediction models was evaluated based on statistical indicators including the coefficient of determination (R^2), correlation coefficient (r), root mean square error (RMSE), and the residual predictive deviation (RPD). A higher RPD indicates a stronger and more accurate model for predicting the soil organic carbon of the samples. Ideally, a good model should have high R^2 and r coefficients, low RMSE, and a low number of latent variables in PLSR. An excellent prediction model should have a high R2 and r coefficient (at or above 0.8), a high RPD index (above 3), and a low RMSE [13-16].

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

$$RPD = \frac{SD_{ref}}{RMSE} \quad (2)$$

where, $\hat{y}_i$ is the predicted value of the i^{th} observation, y_i is the measured value of the i^{th} observation from soil organic carbon, n is the number of observations in the calibration, validation or prediction set, and y_m is the mean value of the calibration or validation data set. The root mean square error prediction (RMSEP) is an estimate of total prediction error for an independent validation dataset.

3. RESULTS AND DISCUSSION

3.1 Spectra features

Diffuse reflectance spectrum of the soil samples is presented in Figure 2, which displays peaks corresponding to molecular bond vibrations of C-C, O-H, N-H, C-H-O, and C-H. Raw spectral data, without any processing, retain all the original spectral features, including both the valuable chemical information and the undesirable noise or variability. The raw spectrum often includes: baseline drift consisted of variations in the baseline of the spectrum due to instrumental or environmental factors. Random variations that can be due to electronic noise and interference in the detector, photon noise, or environmental noise.

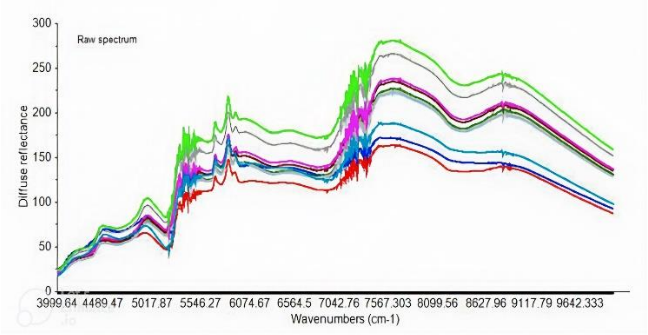


Figure 2. Spectra features in near infrared region of soil samples before correction

Variability due to the physical properties of the sample, such as particle size or surface roughness, which can affect the scatter of the light in the sample. However, the original spectra data before correction were affected by interference from noise caused by light scattering. In spectroscopic analysis, raw spectra can be significantly affected by systematic variability that is not related to the chemical properties of interest. This variability can be caused by factors such as light scattering, instrument variation, sample particle size, or surface irregularities.

To address this issue, various pre-treatment methods, including de-trending (DT), multiplicative scatter correction (MSC), and a combination of both (DT+MSC), were employed. As depicted in Figure 3, the DT correction method significantly improved the appearance of the spectra and effectively eliminated some of the noise resulting from light scattering. The application of DT rectifies the baseline, thereby elevating the signal-to-noise ratio, which directly contributes to the fidelity of the calibration models. This rectification allows for a more precise correlation of spectral features with the specific properties of interest, such as the soil organic carbon content.

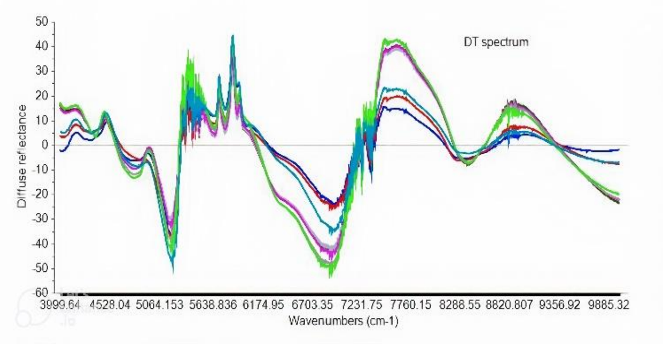


Figure 3. Near infrared spectra after corrections using DT algorithm

The spectral data obtained from the near infrared instrument often contain background information and noise, which can interfere with the desired soil quality information, specifically SOC. Various factors such as light scattering, variations in path length, and random noise caused by sample properties or instrumental effects can introduce interference in the spectra.

To ensure the accuracy, reliability, and stability of the calibration models used for predicting SOC, it is crucial to remove or minimize these interfering spectral parameters. Therefore, preprocessing of the spectral data before the development of prediction models for SOC is necessary.

The principle of de-trending spectra correction involves removing systematic trends or baseline variations present in the spectra. These trends can arise from various factors such as instrumental variations, sample handling, or environmental conditions.

To perform de-trending, a polynomial function is fitted to the baseline of the spectrum. This baseline represents the overall trend of the spectrum without the presence of absorption peaks. By subtracting this polynomial fit from the original spectrum, the systematic trends are removed, leaving behind the desired spectral features related to the composition of the sample.

De-trending is a process used to remove trends in the spectral data that can obscure the true signal, reduced baseline drift. DT helps in minimizing fluctuations in the baseline, which may not be related to the analyte of interest. By reducing the trend, the variability of the spectrum due to the actual chemical composition becomes more obvious, clear and prominent, improving the interpretability of the data. Hence, calibration models become more robust and less sensitive to the variability in the data that is not associated with concentration levels of the analyte.

By applying these corrections, spectral data become more reliable for quantitative analysis, allowing chemometric models to better predict properties of interest, such as soil organic carbon content. These preprocessing steps are essential in many analytical applications to ensure that the resultant spectra reflect the true chemical information with minimal interference from other sources of variation.

The de-trending process helps to eliminate unwanted variations and enhance the specific spectral information related to the targeted analyte. It improves the accuracy and reliability of the subsequent analysis by reducing the influence of non-analyte-related factors on the spectra.

Besides using DT algorithm, spectral data were also corrected and enhanced using multiplicative scatter correction (MSC) approach as shown in Figure 4. The principle of multiplicative scatter correction (MSC) aims to correct for unwanted scattering effects in spectral data. When light interacts with a sample, it can undergo scattering, leading to distortions in the recorded spectrum.

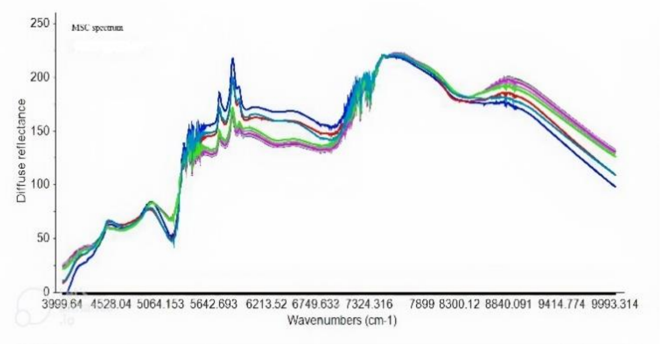


Figure 4. Near infrared spectra after corrections using MSC algorithm

MSC addresses this issue by dividing the original spectrum by a reference spectrum, which is typically a smooth or noise-free spectrum [17]. This division operation normalizes the spectral intensities and reduces the influence of scattering. The reference spectrum represents the average scattering effects present in the sample set and is used as a scaling factor.

By dividing the original spectrum by the reference spectrum, the spectral intensities are adjusted to compensate for the scattering variations. This correction allows for clearer and more accurate spectral analysis, as it separates the true spectral features from the scattering effects.

Moreover, MSC is a technique used to correct for multiplicative interferences in the spectral data, like scatter effects. The MSC adjusts the spectra to correct for light scattering caused by particles of different sizes, shapes, or densities. This makes the spectra more comparable and reduces the variability caused by physical characteristics of the sample.

The spectral data are normalized, making the comparison between different spectra more straightforward, as they are on a similar scale. Important spectral features that are related to the composition of the sample can be more easily discerned after MSC application, which can enhance the quality of predictive models based on the spectra.

MSC is particularly useful when dealing with samples that exhibit significant scattering, as it minimizes the impact of scattering variations across different samples. It helps to improve the comparability and interpretability of spectral data, enabling more reliable analysis and prediction of targeted sample properties.

Moreover, we attempted to combines the method of de-trending and multiplicative scatter correction (DT+MSC) as presented in Figure 5, to enhance spectral data and minimize unwanted variations. First, the de-trending method is applied to remove systematic trends or baseline variations in the spectra. A polynomial function is fitted to the baseline, and subtracting this polynomial fit from the original spectrum eliminates the systematic variations.

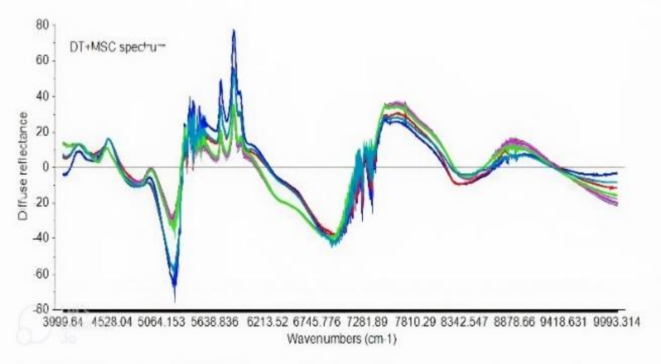


Figure 5. Near infrared spectra after corrections using combination of DT and MSC algorithms

Next, the multiplicative scatter correction is performed by dividing the de-trended spectrum by a reference spectrum. This division operation normalizes the intensity values and corrects for scattering effects caused by variations in sample composition or instrumental conditions.

By combining these two techniques, the DT+MSC method leverages their individual strengths to effectively reduce noise, correct for scattering, and remove baseline drifts. This combined approach enhances the spectral appearance by preserving the relevant spectral features associated with the analyte of interest while minimizing interferences or variations caused by scattering or other factors.

The DT+MSC method is particularly useful in scenarios where both baseline drift and scattering effects are present in the spectral data. By applying these complementary correction techniques, the resulting spectra are optimized for subsequent analysis and prediction models, ensuring accurate and robust estimations of the targeted sample properties.

Employing both techniques together allows for a more thorough correction of the spectral data. The resulting improvements in data quality can significantly enhance the predictive accuracy of NIRS analyses since both potential sources of error scatter and baseline shifts are duly corrected. This preprocessing makes the data more conducive to the development of precise and reliable quantitative prediction models, crucial for various applications in spectroscopic analysis.

3.2 SOC prediction models

Once the pre-processing of the spectra was concluded, we proceeded to develop prediction models for estimating C-organic levels in soil samples. These models were based on partial least squares regression (PLSR) and utilized both untreated and treated spectra from the soil sample dataset encompassing wavenumbers ranging from 4,000 to 10,000 cm^{-1}. In evaluating the models, we compared the correlation coefficient (r), standard error of prediction (RMSE), and residual predictive deviation (RPD) index.

Table 1. PLSR prediction model's performance of different NIR spectral data, before and after corrections

Spectrum	Statistical Indicator			
	R^2	r	RMSE	RPD
Raw	0.75	0.87	0.26	1.73
DT	0.80	0.90	0.21	2.70
MSC	0.86	0.93	0.18	3.15
DT+MSC	0.93	0.96	0.14	4.05

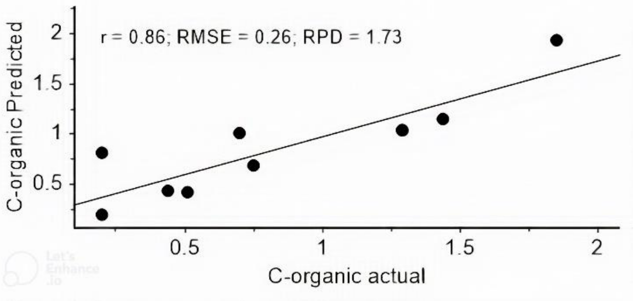
The SOC prediction model captures the relationship between the observed response variable, represented by the SOC content (Y-variables), and the independent variable, which comprises the diffuse near infrared reflectance spectrum (X-variables). Through the interaction of near-infrared radiation with the biological object, significant information pertaining to its physical, optical, and chemical properties can be extracted. The prediction results for SOC are presented in Table 1.

Initially, a prediction model for SOC was developed using the raw, untreated spectra data. This model achieved a correlation coefficient of 0.87, with an RMSE value of 0.28 and an RPD index of 1.73. The raw spectrum data yields an R^2 of 0.75, which demonstrates that 75% of the variability in the dependent variable can be explained by the model based on the raw spectral data. This is a reasonably good fit, indicating the raw data does have a significant predictive power. The correlation coefficient (r) of 0.87 is quite strong, illustrating a robust positive linear relationship between observed and predicted values. However, with an RMSE of 0.26, there's room for improvement as this number indicates that the predictions deviate from the actual observations by some margin. The RPD of 1.73 solidifies the conclusion that while useful, the predictive accuracy of the raw data model could be enhanced with further data processing.

Subsequently, by employing de-trending (DT) on the spectra data, the correlation coefficient improved to 0.90, the RMSE for prediction errors decreased to 0.21, and the RPD index showed noticeable improvement compared to the previous model. Furthermore, the accuracy and robustness of the C-organic prediction model were significantly enhanced when the spectra data underwent multiplicative scatter correction (MSC). This model yielded an r value of 0.93, an RPD of 3.15, and a reduced prediction error of 0.18.

When comparing DT and MSC corrected spectra to the raw spectrum, the main differences are: the corrected spectra have a reduced influence of instrument and sample-related noise and variability. They exhibit a more uniform baseline, and their features are more attributable to the chemical composition of the sample. The spectra have undergone transformations to mitigate effects that overshadow the signal of interest, thus they are typically smoother and offer clearer information for the development of calibration models. Both DT and MSC, either used separately or combined, aid in enhancing the accuracy of subsequent quantitative and qualitative spectral analyses.

Ultimately, a combination of DT and MSC was employed to establish the C-organic prediction model, resulting in the most accurate and robust prediction outcome. This model achieved a maximum correlation coefficient of 0.96, an RPD index of 4.05, and the lowest RMSE error of 0.14. Figure 6 presents a scatter plot depicting the actual C-organic values versus the predicted C-organic values.



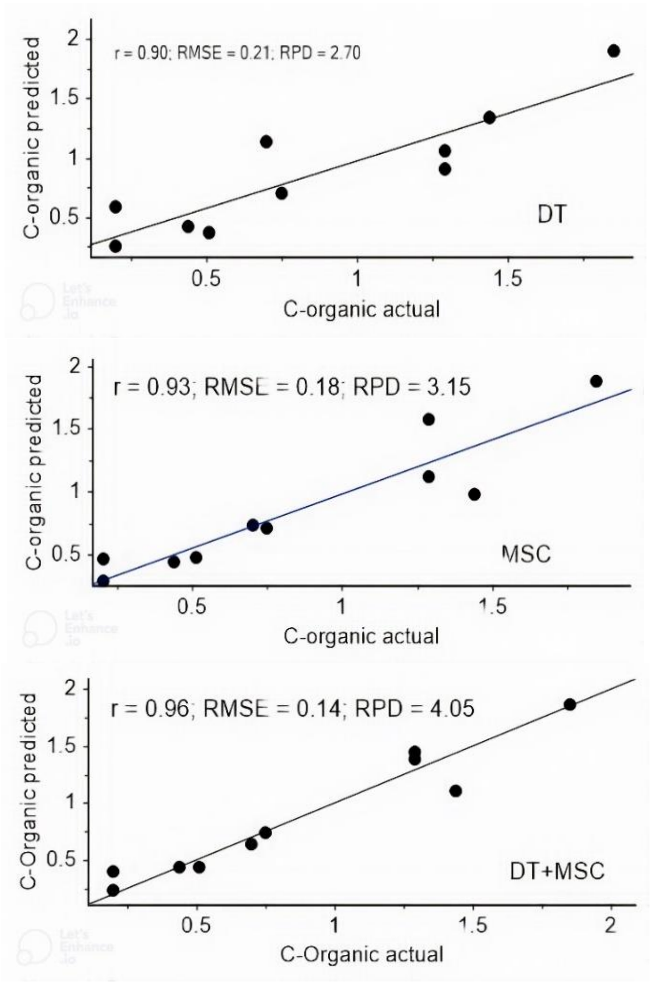


Figure 6. Scatter plot between SOC predicted and SOC measured using four different NIR spectrum coupled with PLSR algorithm

Prediction models involves establishing a relationship between the spectral data (X-variable) and the corresponding property or analyte of interest (Y-variable). To build a prediction model, a representative dataset comprising both spectral data and reference values for the property of interest is required. This dataset is typically split into two subsets: a calibration set and a validation set. The calibration set is used to develop the model, while the validation set is used to assess the model's performance.

The calibration process involves applying a regression algorithm, such as partial least squares regression (PLSR), to the calibration set. PLSR identifies the underlying relationships between the spectral data and the reference values to establish a predictive model. Partial least squares regression is a widely used method for developing calibrations based on near-infrared (NIR) spectral data to determine Soil Organic Carbon (SOC) content. PLS is a powerful multivariate modeling technique that is particularly well-suited to handle the high-dimensional nature of spectral data and to establish robust relationships between the spectral information and the target property, such as SOC.

The PLS algorithm constructs a linear regression model describing the relationship between the predictor variables (NIR spectra) and the response variable (SOC content). The main steps involved in PLS modeling are includes latent variable selection, model training and optimization. PLS regression reduces the dimensionality of the spectral data by creating a set of new latent variables (also known as PLS components) that are linear combinations of the original NIR variables. The number of latent variables is chosen to maximize the explained variance in the SOC content while avoiding overfitting.

The PLS algorithm uses the calibration set to find the optimum linear relationship between the latent variables and the reference SOC values, taking into account the covariance between the two sets of variables. PLS models are optimized using techniques to prevent overfitting, such as cross-validation or other regularization methods, to ensure good predictive performance on new samples.

During the calibration process, the model is developed by finding the optimal number of latent variables (factors) that explain the maximum variance in the spectral data while minimizing the prediction error. This optimization process ensures that the model extracts the most relevant information from the spectral dataset.

Determining the optimal number of latent variables is a critical aspect of model optimization in the realm of spectroscopy and multivariate statistics. Latent variables are created features from observed data, crucial for capturing the underlying structure while reducing its dimensionality. Selecting the right number of latent variables is essential, as too few may lead to under fitting, while too many could result in overfitting, capturing noise instead of relevant signal. Here's a general overview of how the optimal number of latent variables is determined.

One common approach is through cross-validation, where the dataset is split into training and validation sets. The model is then built on the training set using an increasing number of latent variables, and its performance is evaluated on the validation set. Key performance metrics such as root-mean-square error of cross-validation (RMSECV), root-mean-square error of prediction (RMSEP), or the coefficient of determination (R^2) are observed as the number of latent variables increases. The optimal number is typically determined by identifying the point where the error metrics reach a minimum before starting to increase again or stabilize, finding the balance between bias and variance.

With detrending, we see noticeable improvements. The R^2 value increases to 0.80, suggesting that after removing trends from the data, the model can explain 80% of the variance. This is a good leap forward towards predictive accuracy. The improvement in the correlation coefficient to 0.90 underscores a stronger linear relationship post-detrending. A lower RMSE of 0.21 is a direct indicator of more accurate predictions compared to the raw data model. The average error has decreased. An RPD of 2.70 is significant, as it suggests that the detrending process has substantially enhanced the model's ability to use the spectral data for making predictions.

MSC applies another layer of refinement. With an R^2 of 0.86, we see an additional increment in the variance explained by the model. The correlation coefficient is now at a very high 0.93, indicating an even stronger relationship between observed and predicted values, the linear association is becoming tighter with each data processing step. The RMSE drops further to 0.18, which implies that the MSC technique is successful in reducing the prediction errors. The model is becoming reliably precise. The RPD sees a more notable increase to 3.15, reinforcing that MSC is effectively standardizing the data, leading to improved calibration models.

Once the model is developed, it is evaluated using the validation set. The model's performance is assessed using statistical indicators such as the coefficient of determination (R^2), correlation coefficient (r), root mean square error (RMSE), and residual predictive deviation (RPD). These indicators determine how well the model predicts the property of interest based on the spectral data.

According to the literatures [3, 18, 19], the RPD index ranging from 1.0 to 1.5 suggests that the prediction performance is coarse and requires improvement, particularly in terms of spectral data correction. An RPD value between 1.5 and 2.5 indicates that the prediction performance can be categorized as sufficient. Furthermore, an RPD value between 2.5 and 3 suggests a good prediction performance, while an RPD above 3 indicates an excellent level of accuracy in predictions.

Significant improvements in prediction performance were observed when SOC prediction models were developed using treated and enhanced spectral data. All spectral correction approaches demonstrated higher accuracy and robustness indices compared to the raw, uncorrected spectra. Notably, the combination of enhanced spectra data using DT and MSC correction methods proved to be the most effective, yielding a maximum correlation coefficient between reference SOC and predicted SOC. Therefore, it is evident that enhancing the spectra data leads to improved prediction accuracy and robustness.

Combining the two preprocessing techniques results in the best outcomes across the board. The R^2 shoots up to 0.93, meaning the model can now explain 93% of the variance, which is indicative of high predictive accuracy. The correlation coefficient reaches a near perfect 0.96, showcasing an extremely strong positive linear relationship in the data. The RMSE at its lowest, 0.14, confirms that the combination approach lowers the prediction error significantly. Lastly, the RPD at 4.05 is the highest among all models, implying that this combined model has excellent predictive performance and can reliably be used to estimate the values of the dependent variable.

Based on the prediction performance obtained from all spectra data, it can be concluded that infrared technology has the potential to predict SOC content in soil samples with a maximum correlation coefficient of 0.96 and an RPD index of 4.05. Moreover, this technology can deliver more accurate and reliable prediction results when the spectra data are properly corrected and enhanced.

The trend from raw data to combined DT+MSC preprocessing demonstrates a consistent improvement across all statistical metrics, indicating that the proper preprocessing of spectral data is crucial for enhancing model accuracy. Each successive technique, detrending and MSC, contributes to refining the model's predictive capabilities, and their combined usage seems to offer synergistic benefits. Such insights could be invaluable for those working in fields where spectral analysis is pivotal, like analytical chemistry, remote sensing, or agriculture, guiding them toward the best practices for preprocessing their data.

General remarks of NIRS application for predicting SOC of agricultural soil samples: NIRS allows for the non-destructive analysis of soil samples, enabling researchers and practitioners to gather spectral data without altering or compromising the integrity of the soil, which is crucial for follow-up physical and chemical analyses. NIRS analysis is rapid and high-throughput, enabling swift data collection and large-scale assessment of SOC levels in soils. This can significantly expedite the pace of soil monitoring efforts.

Compared to traditional wet-chemistry methods, the NIRS approach typically offers cost efficiencies in terms of time, labor, and equipment, while still providing reliable predictions of SOC content. NIRS can provide simultaneous analysis of multiple soil properties beyond SOC, such as nutrient levels, texture, pH, and soil structure, offering holistic insights into soil health and fertility.

The analysis clearly indicates a significant enhancement of model accuracy with each step of data preprocessing applied to spectral data. The raw data shows initial trends, but the execution DT and MSC preprocessing techniques refines these trends substantially. This sequential improvement across all statistical metrics highlights a key takeaway: the choice and execution of appropriate preprocessing techniques are critical. They play a synergistic role in enhancing the predictive prowess of models [20-22]. In fields heavily reliant on spectral analysis, such as analytical chemistry, remote sensing, or agriculture, understanding and utilizing these preprocessing strategies can dramatically improve data reliability and model outcomes. This sets the stage for more accurate, efficient, and meaningful data interpretation and application.

Each preprocessing step contributes uniquely: detrending removes systematic trends that could bias the baseline, while MSC adjusts for multiplicative scatter effects due to particle size variations or other inconsistencies. Their combined effect results in a data set that is much cleaner and more representative of true underlying chemical or physical properties [23-25]. Researchers and practitioners should, therefore, consider an iterative approach to preprocessing, adjusting techniques as needed to optimize their models for the specific characteristics of their spectral data.

Moreover, NIRS enables data-driven decision making in agriculture, land management, and environmental assessments by providing accurate and timely information on soil properties, thereby facilitating informed choices for sustainable and efficient agricultural practices.

Accurate and efficient prediction of SOC through NIRS allows farmers and land managers to implement precision agriculture and tailored soil management practices, which can optimize crop productivity, conserve resources, and minimize environmental impact. NIRS-based soil analysis supports informed land use planning and conservation efforts by providing insights into soil carbon dynamics, helping to identify areas for reforestation, habitat restoration, or sustainable land use that promote carbon sequestration and biodiversity conservation.

An accurate SOC prediction through NIRS aids in assessing the potential role of soils as carbon sinks and sources, informing climate change mitigation strategies, and contributing to accounting for carbon credits or offsets. By facilitating efficient soil monitoring and management, NIRS contributes to global food security efforts by promoting sustainable agricultural practices that maintain soil productivity, preserve natural ecosystems, and safeguard agricultural yields for a growing global population. NIRS applications in predicting SOC content offer an avenue for advancing scientific research and capacity building in soil science, fostering collaborations among researchers, and equipping soil scientists and agricultural practitioners with valuable tools for sustainable development.

Effective monitoring of soil carbon content through NIRS aids in the efficient use of agricultural inputs, such as fertilizers and irrigation, thereby conserving resources and minimizing environmental impacts associated with agricultural activities. In summary, the application of NIRS for predicting SOC content in soil samples has wide-ranging benefits for environmental sustainability, agriculture, and global development. It enables informed decision making, promotes sustainable land management, and contributes to global efforts related to climate change mitigation and food security, ultimately fostering a more resilient and sustainable global ecosystem.

4. CONCLUSIONS

The obtained results suggest that near infrared spectroscopy (NIRS) coupled with partial least square regression (PLSR) has the potential to be utilized as a quick and eco-friendly approach for predicting soil organic carbon (C-organic). The application of De-trending (DT) and multiplicative scatter correction (MSC) for spectral correction significantly enhanced the accuracy and reliability of C-organic prediction. This was evident in the increased correlation coefficient between actual and predicted SOC values, higher RPD index, and reduced RMSE error prediction. Among the different correction methods, DT+MSC exhibited the best performance in accurately and robustly predicting SOC, with a correlation coefficient of 0.96 and RPD of 4.05.

The developed method can be integrated into current soil testing practices in a manner that offers profound implications for precision agriculture. NIRS potentially coupled with the described optimization techniques, harbors several practical applications and benefits that could revolutionize soil management.

The rapid and nondestructive nature of NIRS analysis allows for real time assessment of soil properties, potentially leading to the development of on site testing devices. This advancement would enable farmers to make quick, informed decisions about their soil management practices without having to wait for lengthy laboratory results.

The cost-effectiveness and sustainability of this method are also noteworthy. By providing instant, accurate feedback, NIRS could notably reduce the reliance on traditional, often expensive, and time-consuming laboratory analysis, benefiting both large-scale agricultural operations and smaller, resource-constrained farms.

ACKNOWLEDGMENT

This work is supported by the directorate of research and community service LPPM *Universitas Syiah Kuala* through *Penelitian Professor* scheme 2023.

REFERENCES

[1] Afriyie, E., Verdoodt, A., Mouazen, A.M. (2022). Potential of visible-near infrared spectroscopy for the determination of three soil aggregate stability indices. Soil and Tillage Research, 215: 105218. https://doi.org/10.1016/j.still.2021.105218

[2] Reda, R., Saffaj, T., Itqiq, S.E., Bouzida, I., Saidi, O., Yaakoubi, K., Lakssir, B., El Mernissi, N., El Mernissi, N. (2020). Predicting soil phosphorus and studying the effect of texture on the prediction accuracy using machine learning combined with near-infrared spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 242: 118736. https://doi.org/10.1016/j.saa.2020.118736

[3] Wang, Q., Zhang, H., Li, F., Gu, C., Qiao, Y., Huang, S. (2021). Assessment of calibration methods for nitrogen estimation in wet and dry soil samples with different wavelength ranges using near-infrared spectroscopy. Computers and Electronics in Agriculture, 186: 106181. https://doi.org/10.1016/j.compag.2021.106181

[4] Reda, R., Saffaj, T., Ilham, B., Saidi, O., Issam, K., Brahim, L. (2019). A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. Chemometrics and Intelligent Laboratory Systems, 195: 103873. https://doi.org/10.1016/j.chemolab.2019.103873

[5] Cai, H.T., Liu, J., Chen, J.Y., Zhou, K.H., Pi, J., Xia, L.R. (2021). Soil nutrient information extraction model based on transfer learning and near infrared spectroscopy. Alexandria Engineering Journal, 60(3): 2741-2746. https://doi.org/10.1016/j.aej.2021.01.014

[6] Darusman, D., Juwita, I.R., Munawar, A.A., Zainabun, Z., Zulfahrizal, Z. (2021). Rapid determination of mixed soil and biochar properties using a shortwave near infrared spectroscopy approach. IOP Conference Series: Earth and Environmental Science, 667(1): 012003. https://doi.org/10.1088/1755-1315/667/1/012003

[7] Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives–A review. Analytica Chimica Acta, 1026: 8-36. https://doi.org/10.1016/j.aca.2018.04.004

[8] Munawar, A.A., von Hörsten, D., Wegener, J.K., Pawelzik, E., Mörlein, D. (2016). Rapid and non-destructive prediction of mango quality attributes using Fourier transform near infrared spectroscopy and chemometrics. Engineering in Agriculture, Environment and Food, 9(3): 208-215. https://doi.org/10.1016/j.eaef.2015.12.004

[9] Allo, M., Todoroff, P., Jameux, M., Stern, M., Paulin, L., Albrecht, A. (2020). Prediction of tropical volcanic soil organic carbon stocks by visible-near-and mid-infrared spectroscopy. Catena, 189: 104452. https://doi.org/10.1016/j.catena.2020.104452

[10] Yang, R.M. (2020). Characterization of the salt marsh soils and visible-near-infrared spectroscopy along a chronosequence of Spartina alterniflora invasion in a coastal wetland of eastern China. Geoderma, 362: 114138. https://doi.org/10.1016/j.geoderma.2019.114138

[11] dos Santos, U.J., de Melo Dematte, J.A., Menezes, R.S.C., Dotto, A.C., Guimarães, C.C.B., Alves, B.J.R., Primo, D.C., Sampaio, E.V.S.B. (2020). Predicting carbon and nitrogen by visible near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil. Geoderma Regional, 23: e00333. https://doi.org/10.1016/j.geodrs.2020.e00333

[12] Munawar, A.A., Yunus, Y., Satriyo, P. (2020). Calibration models database of near infrared spectroscopy to predict agricultural soil fertility properties. Data in Brief, 30: 105469. https://doi.org/10.1016/j.dib.2020.105469

[13] Afriyie, E., Verdoodt, A., Mouazen, A.M. (2021). Data fusion of visible near-infrared and mid-infrared spectroscopy for rapid estimation of soil aggregate

stability indices. Computers and Electronics in Agriculture, 187: 106229. https://doi.org/10.1016/j.compag.2021.106229

[14] Sun, W., Zhang, X., Sun, X., Sun, Y., Cen, Y. (2018). Predicting nickel concentration in soil using reflectance spectroscopy associated with organic matter and clay minerals. Geoderma, 327: 25-35. https://doi.org/10.1016/j.geoderma.2018.04.019

[15] Jiang, Q., Chen, Y., Guo, L., Fei, T., Qi, K. (2016). Estimating soil organic carbon of cropland soil at different levels of soil moisture using VIS-NIR spectroscopy. Remote Sensing, 8(9): 755. https://doi.org/10.3390/rs8090755

[16] Benedet, L., Faria, W.M., Silva, S.H.G., Mancini, M., Guilherme, L.R.G., Demattê, J.A.M., Curi, N. (2020). Soil subgroup prediction via portable X-ray fluorescence and visible near-infrared spectroscopy. Geoderma, 365: 114212. https://doi.org/10.1016/j.geoderma.2020.114212

[17] Munawar, A.A., Meilina, H., Pawelzik, E. (2022). Near infrared spectroscopy as a fast and non-destructive technique for total acidity prediction of intact mango: Comparison among regression approaches. Computers and Electronics in Agriculture, 193: 106657. https://doi.org/10.1016/j.compag.2021.106657

[18] Lazaar, A., Mouazen, A.M., Hammouti, K.E., 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. https://doi.org/10.1016/j.iswcr.2020.04.005

[19] van der Meer, F. (2018). Near-infrared laboratory spectroscopy of mineral chemistry: A review. International Journal of Applied Earth Observation and Geoinformation, 65: 71-78. https://doi.org/10.1016/j.jag.2017.10.004

[20] Zhou, W., Wang, Q., Chen, S., Chen, F., Lv, H., Li, J., Chen, Q., Zhou, J.B., Liang, B. (2024). Nitrate leaching is the main driving factor of soil calcium and magnesium leaching loss in intensive plastic-shed vegetable production systems. Agricultural Water Management, 293: 108708. https://doi.org/10.1016/j.agwat.2024.108708

[21] Lei, J., Yin, J., Chen, S., Fenton, O., Liu, R., Chen, Q., Fan, B.Q., Zhang, S. (2024). Understanding phosphorus mobilization mechanisms in acidic soil amended with calcium-silicon-magnesium-potassium fertilizer. Science of The Total Environment, 916: 170294. https://doi.org/10.1016/j.scitotenv.2024.170294

[22] Zhang, Q., Hu, Z., Xu, Z., Zhang, P., Jiang, Y., Fu, D., Chen, Y. (2024). Quantitative determination of TVB-N content for different types of refrigerated grass carp fillets using near-infrared spectroscopy combined with machine learning. Journal of Food Composition and Analysis, 126: 105871. https://doi.org/10.1016/j.jfca.2023.105871

[23] Xu, S., Zhao, Y., Wang, Y. (2024). Optimizing machine learning models for predicting soil pH and total P in intact soil profiles with visible and near-infrared reflectance (VNIR) spectroscopy. Computers and Electronics in Agriculture, 218: 108643. https://doi.org/10.1016/j.compag.2024.108643

[24] Díaz, E.O., Iino, H., Koyama, K., Kawamura, S., Koseki, S., Lyu, S. (2023). Non-destructive quality classification of rice taste properties based on near-infrared spectroscopy and machine learning algorithms. Food Chemistry, 429: 136907. https://doi.org/10.1016/j.foodchem.2023.136907

[25] Ong, P., Jian, J., Li, X., Zou, C., Yin, J., Ma, G. (2023). New approach for sugarcane disease recognition through visible and near-infrared spectroscopy and a modified wavelength selection method using machine learning models. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 302: 123037. https://doi.org/10.1016/j.saa.2023.123037