Accessible spectral engineering for calibrated estimation of soil organic carbon using PLSR and a low-cost Vis-NIR sensor

Andrea Stephanie Isla Aguilar[1] Carlos Julian Mestanza Novoa[1]

1 Faculty of Agronomy, Universidad Nacional Agraria La Molina (UNALM), Lima, Peru

Abstract. Quantifying soil organic carbon (SOC) is central to evaluating soil quality, guiding agronomic management, and supporting carbon accounting schemes, but conventional wet-chemistry methods limit high-density measurement because of operational cost, laboratory demand, reagent use, and analytical turnaround time. This study developed and internally evaluated a Vis-NIR/PLSR spectral calibration function for estimating SOC (%) using a compact low-cost LinkSquare 1 spectrometer coupled to a 3D-printed Optical Station for Soils (EOS-3D). The system was designed to stabilize sensor-surface geometry, reduce ambient-light interference, and enable readings on air-dried fine soil presented in glass Petri dishes under a reproducible laboratory configuration. One LinkSquare reading was performed per sample, and each reading generated two spectral signatures that were averaged to obtain one representative spectrum per sample, avoiding artificial duplication of the experimental unit and preserving the sample as the statistical unit. Spectra were preprocessed by treating zero values as missing readings, applying linear interpolation, Standard Normal Variate normalization, and the first derivative of Savitzky-Golay smoothing. The nominal LinkSquare range was considered approximately 400-1000 nm, whereas the exported matrix ranged from 400.89 to 1099.50 nm and the majority-populated spectral interval was 416.79-799.10 nm; therefore, nominal and effective spectral ranges were reported separately. The number of PLSR components was selected by repeated k-fold cross-validation under the minimum RMSECV rule, and 11 components produced the lowest mean RMSECV. The final matrix included 297 independent observations, split into calibration (n = 237) and internal hold-out validation (n = 60). In internal validation, the model achieved R^2 = 0.7741, RMSE = 0.6683% SOC, MAE = 0.5089% SOC, RPD = 2.12, and RPIQ = 2.36. These results indicate moderate to good predictive ability within the evaluated sample and instrumental domain. However, the calibration function should be interpreted as an empirical, locally documented model rather than as external validation or universal transferability to other sensors, sessions, or soil collections.

Keywords. proximal soil sensing; diffuse reflectance; PLSR; chemometrics; spectral calibration; internal validation; precision agriculture.

Introduction

Soil organic carbon (SOC) represents a relatively small fraction of the solid soil matrix, but it controls essential pedogenic processes such as aggregation, water retention, nutrient exchange, biological activity, and functional soil stability. SOC is an integrative component of physical, chemical, and biological soil properties (Weil and Brady 2017), and it supports soil quality through sensitive and comparable indicators (Bünemann et al. 2018). Therefore, repeatable SOC measurements are necessary to evaluate carbon sequestration strategies and climate-change mitigation practices (Minasny et al. 2017).

Precision agriculture has increased the demand for SOC measurements with higher spatial density, lower unit cost, and shorter response time. Wet-oxidation methods remain useful as analytical references, but their scalability is limited when fields, laboratory repositories, or regional gradients must be characterized. Proximal sensors can support soil carbon accounting when calibration is verifiable and uncertainty is controlled (England and Viscarra Rossel 2018). Routine adoption of soil spectroscopy also requires integration between spectral data, wet laboratories, and operational users (Paiva et al. 2022; Poppiel et al. 2022).

Diffuse reflectance spectroscopy in the visible and near-infrared region provides rapid, non-destructive, and environmentally favorable information. Soil spectral response is not controlled only by carbon; it integrates color, mineralogy, texture, residual moisture, physical preparation, and scattering properties. Global spectral libraries have demonstrated the capacity of soil spectra to encode multiple edaphic attributes (Viscarra Rossel et al. 2016), but NIR spectroscopy requires robust empirical calibrations because many signals are indirect and overlapping (Pasquini 2018). Consequently, SOC prediction performance depends on the instrument, measurement environment, and calibration domain (Hutengs et al. 2019).

For compact low-cost sensors, the challenge is not only statistical. Stability of sensor-surface distance, protection against stray light, and sample homogeneity determine whether the model learns edaphic variation or acquisition noise. Physical protocols are important for harmonizing soil reflectance

measurements (Ben-Dor et al. 2024). In addition, operational measurement variables can modify spectral quality (Semella et al. 2022; Xu et al. 2022), indicating that the reproducibility of portable spectra depends on acquisition protocol and uncertainty in the reference method. Nevertheless, limited evidence is still available on accessible optical architectures combining compact sensors, 3D-printed physical control, and local chemometric calibration to estimate SOC in heterogeneous soil collections.

Based on this premise, this study proposes a Vis-NIR/PLSR spectral calibration function to estimate SOC (%) using a compact low-cost sensor coupled to an EOS-3D station. The objective was to develop and evaluate an accessible, reproducible, and chemometrically documented protocol for estimating SOC within an explicit sample domain. The hypothesis was that physical restriction of optical geometry, combined with spectral preprocessing and objective selection of PLSR components, can produce a defensible calibrated estimation model under comparable laboratory conditions.

Materials and methods

Soil samples and analytical domain

Samples were obtained from the repository of the Laboratory of Soil, Plant, Water, and Fertilizer Analysis (LASPAF) at Universidad Nacional Agraria La Molina (UNALM), Lima, Peru. Because samples came from different regions of Peru, the study was not designed as a single-site trial, but rather as a heterogeneous edaphic collection evaluated under controlled laboratory conditions. Samples were prepared as air-dried fine soil, homogenized, and placed in glass Petri dishes to maintain a flat surface and a constant reading depth.

The study used mineral soil material within the analytical domain defined for the model. The instrument exported 594 spectral signatures because each single LinkSquare reading generated two spectral outputs for the same sample. These paired instrumental signatures were averaged to obtain one representative spectrum per sample, resulting in 297 independent sample-level observations. The final matrix used by the PLSR model consisted of 297 independent observations, distributed into 237 calibration samples and 60 internal hold-out validation samples.

Reference soil organic carbon

Reference SOC values were determined by wet oxidation using the Walkley-Black method, following procedures described in the Soil Survey Field and Laboratory Methods Manual (Soil Survey Staff 2014). This method estimates oxidizable organic carbon and not necessarily total organic carbon determined by dry combustion. Accordingly, the reference value has method-specific analytical uncertainty that must be considered when interpreting model error (De Vos et al. 2007; Ellison and Williams 2012; Soil Survey Staff 2014).

Analytical uncertainty in the reference method is transferred to the spectral model. Therefore, RMSE should not be interpreted only as sensor error, but as an integrated measure including optical error, physical sample heterogeneity, spectral preprocessing effects, analytical uncertainty, and bias inherited from the laboratory reference method.

Vis-NIR acquisition prototype and EOS-3D station

Spectral response was acquired using a compact low-cost Vis-NIR LinkSquare-type sensor coupled to a 3D-printed Optical Station for Soils (EOS-3D). According to the manufacturer specification, the LinkSquare 1 spectrometer (Stratio Inc., San Jose, CA, USA) operates by diffuse reflectance spectroscopy and records light approximately between 400 and 1000 nm, with optical resolution varying from 1 to 44 nm depending on the spectral region (Stratio Inc. n.d.). In the exported dataset used in this study, spectral columns ranged from 400.89 to 1099.50 nm, whereas the interval with non-zero signal in more than 50% of samples was 416.79-799.10 nm. Therefore, the nominal instrument range and the effective modeling range were reported separately. This interval covers mainly visible and short near-infrared information; thus, the predictive signal associated with SOC was interpreted primarily as indirect information linked to soil color, darkening, iron oxides, texture, and organo-mineral covariation, rather than as fundamental absorptions of specific organic functional groups.

The EOS-3D station was fabricated using a dark material to reduce lateral reflection and maintain the sensor in an orthogonal position above the Petri dish surface. The design aimed to stabilize sensor-surface geometry and reduce exposure to ambient light during reading. One reflectance reading was taken per sample under the controlled EOS-3D configuration. The soil was placed in a glass Petri dish, and the physical station was used to reduce variability associated with distance, angle, and measurement position.

No user-accessible white/dark calibration was applied because the low-cost spectrometer used in this study did not provide an integrated white/dark calibration procedure available to the user. This limitation was recognized as part of the low-cost instrumental approach rather than treated as an uncontrolled omission. An operating distance of 0.5 mm was adopted after comparing direct contact, 0.5 mm, and 2.0 mm. Direct contact preserved signal intensity but increased the risk of contaminating the optical window; a 2.0-mm separation reduced intensity and increased exposure to stray light. The 0.5-mm separation provided an operational compromise between stability, intensity, and optical-window protection (Fig. 1).

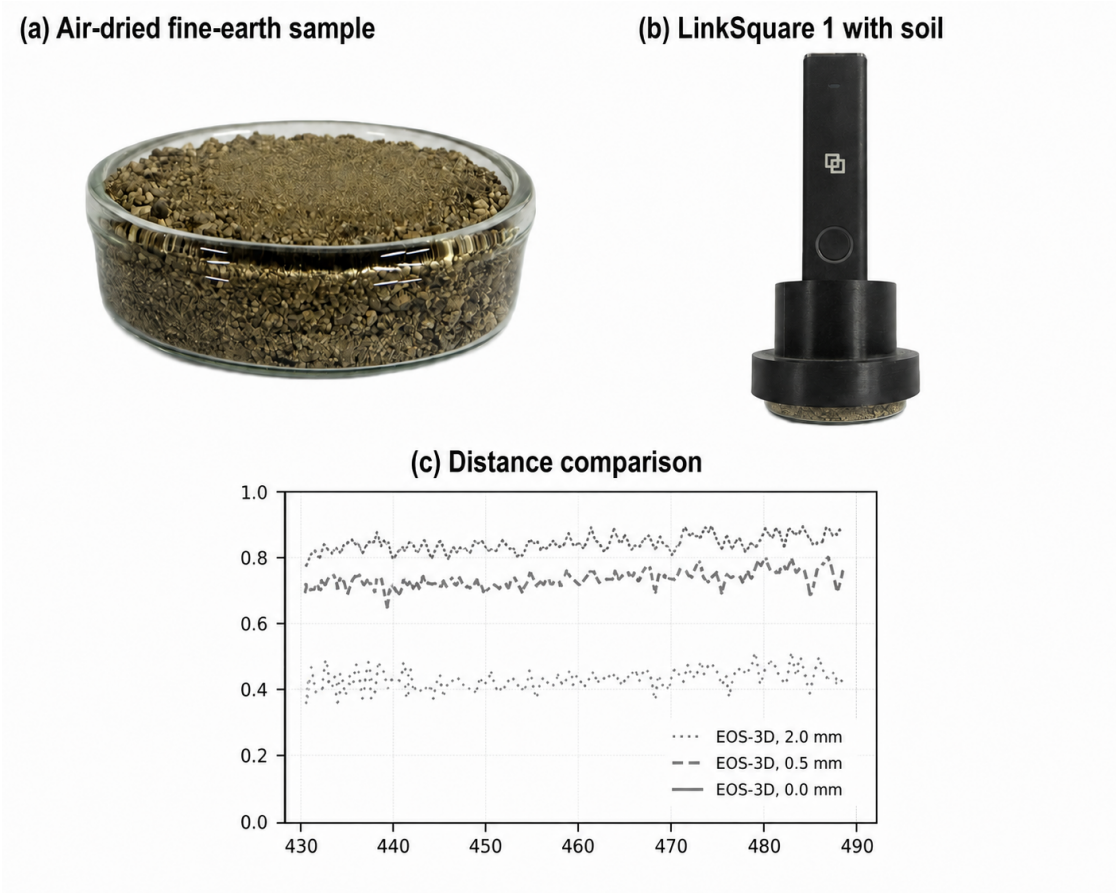

Fig. 1 Vis-NIR acquisition system and comparison of sensor-surface distance. The 0.5-mm configuration was adopted as the operating condition because it balanced signal intensity, optical-window protection, and reduced stray light

Repeated spectral signatures and construction of the modeling matrix

Raw spectral data included two instrumental signatures per LinkSquare reading. Because both signatures belonged to the same physical soil sample and the same measurement event, they were not treated as independent observations. The two signatures associated with each sample were averaged to obtain one representative sample-level spectrum. This procedure converted 594 instrumental signatures into 297 independent observations and avoided artificial duplication of the data.

Spectral preprocessing, operational floor, and PLSR modeling

Spectral columns were converted to numeric format and sorted by wavelength. Zero values were treated as possible point-reading failures and replaced by missing values. Linear interpolation was then applied along the spectral axis, followed by spectrum-wise Standard Normal Variate (SNV) normalization and the first derivative of Savitzky-Golay smoothing using an 11-point window and a second-order polynomial. This workflow was intended to reduce multiplicative variation, baseline shifts, and high-frequency noise without changing the empirical interpretation of the model.

An operational floor of 0.10% SOC was applied as a quality-control criterion before modeling. This minimum threshold was intended to reduce the influence of values close to the lower analytical limit, where relative uncertainty in the reference method may be proportionally greater and may affect multivariate fitting in an unbalanced manner. Therefore, the operational floor was not interpreted as a general soil property, but as a prudent methodological decision to define the analytical domain of the model.

Computational processing was implemented in Python using scikit-learn (Pedregosa et al. 2011) and SciPy (Virtanen et al. 2020). Partial least squares regression (PLSR) was used because spectral channels are naturally multicollinear and because the Vis-NIR matrix must be projected into latent variables associated with spectrum-SOC covariance. The number of latent components was objectively selected by repeated k-fold cross-validation within the calibration set, evaluating 1 to 30 components with 10 folds and 20 repetitions. The official selection rule was minimum RMSECV. For each configuration, RMSECV, MAECV, and cross-validated R^2 were calculated; the internal hold-out validation set was not used for hyperparameter selection and was reserved for final assessment of the fitted model.

Selecting components through RMSECV is appropriate in PLSR because the number of latent variables controls the balance between extracting useful spectral information and avoiding overfitting. Bai et al. (2022) identify the number of latent variables as a critical PLSR parameter for SOC and optimize it by cross-validation. Reyna et al. (2017) warn that models without adequate cross-validation can appear strong but generalize poorly. Tu et al. (2021) also emphasize that redundant latent variables may increase overfitting risk. In this study, the final number of components was selected by examining the RMSECV curve under expert judgment and retaining the configuration that reached the minimum mean RMSECV without using the hold-out validation set for model tuning. The one-standard-error (1-SE) criterion was used only as an exploratory parsimony reference; it identifies the smallest number of components whose mean RMSECV is within one standard error of the minimum.

Model performance was summarized using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), bias, the ratio of performance to deviation (RPD), the ratio of performance to interquartile distance (RPIQ), and the regression slope and intercept between predicted and observed SOC. Bias was defined as predicted − observed, whereas residuals were defined as observed − predicted. RPD was calculated as the standard deviation of observed SOC divided by RMSE, and RPIQ was calculated as the interquartile range of observed SOC divided by RMSE.

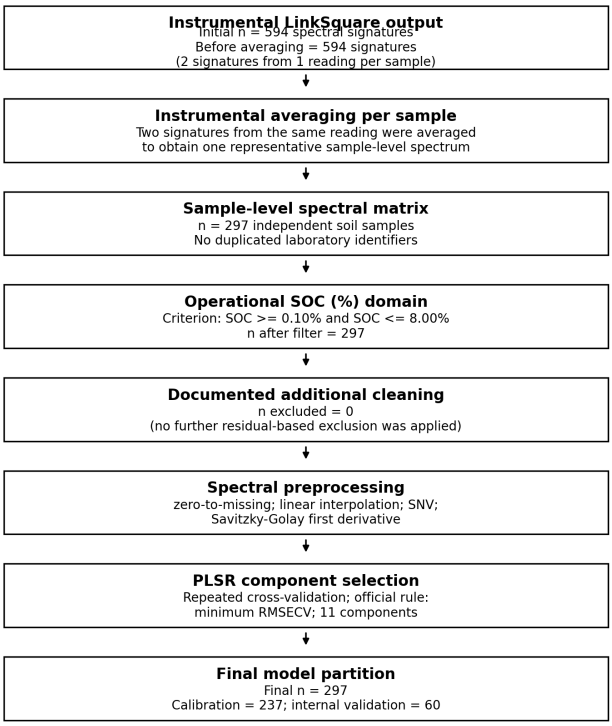

Fig. 2 Experimental and computational workflow for calibrated SOC estimation using a Vis-NIR sensor, EOS-3D station, sample-level averaging, chemometric preprocessing, RMSECV-based component selection, and internal validation

Spectral equation and numerical verification

The model equation was reconstructed from the PLSR spectral coefficients and the effective intercept. The formal published expression was $SOC_{predicted}\ (\%) = b0 + \Sigma\ \beta\lambda\ X_{proc,\lambda}$, where $X_{proc,\lambda}$ is the spectral signal after the complete preprocessing workflow. The direct internal intercept from scikit-learn was not used for the expanded equation because it did not exactly reproduce predictions. In contrast, the effective intercept b0 = 1.342397 reproduced model.predict() within numerical tolerance in both calibration and internal validation. Because the full βλ vector is extensive, the main text reports the verified intercept and

formal model expression rather than tabulating all coefficients in the proceedings paper.

Predictive performance was evaluated using an internal 20% hold-out partition reserved for samples not used during final model fitting. This set is referred to as internal validation, not external validation, because it originated from the same experimental matrix and acquisition protocol.

Results

Spectral response and geometric control

Comparison of sensor-surface distances showed that small changes in reading geometry modified relative signal intensity. The 0.5-mm condition produced a more operationally stable response than direct contact and the 2.0-mm separation. The grouped spectral response of the soil set showed a broad envelope, compatible with the edaphic heterogeneity of the collection and the evaluated SOC range (Fig. 3).

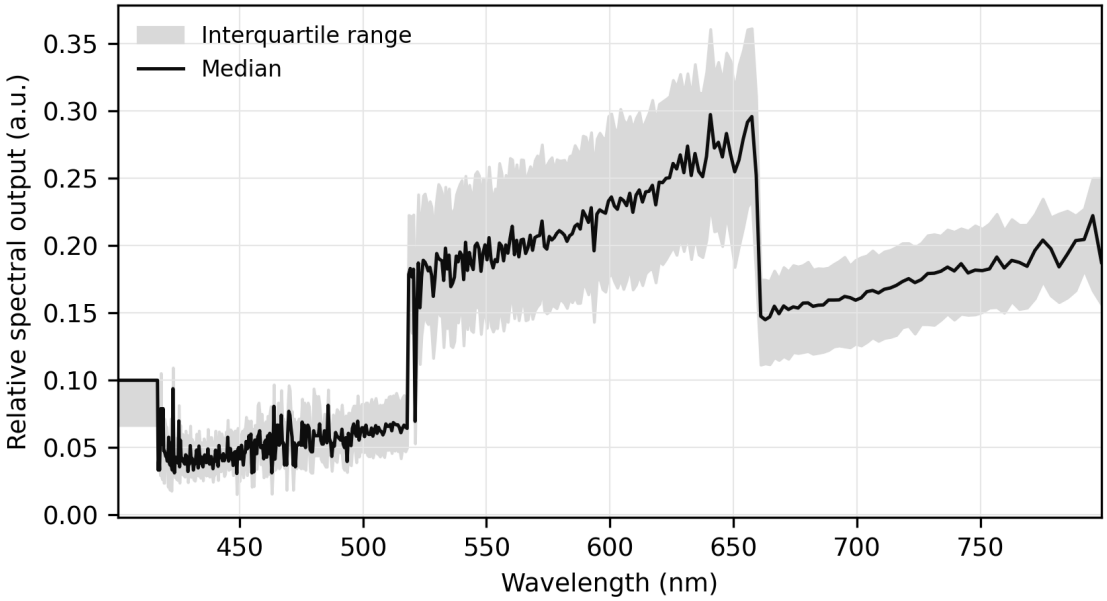

Fig. 3 Relative spectral output of the soil set after grouping by laboratory identifier. The line represents the median and the envelope represents the interquartile range

Selection of the number of PLSR components

Repeated cross-validation showed a progressive decrease in RMSECV until the minimum value was reached with 11 latent components (mean RMSECV = 0.7757% SOC). The 1-SE criterion, used only as an exploratory parsimony reference, indicated that the 10-component model was the simplest option within one standard error of the minimum. The 12-component model showed a similar RMSECV and remained comparable, but it did not produce the lowest mean error. Therefore, after inspecting the RMSECV curve and applying the official minimum-RMSECV rule, the final model was fitted with 11 components. This decision avoided arbitrary post-hoc selection and reduced the risk of choosing a model based on the internal hold-out validation set (Fig. 4; Table 1).

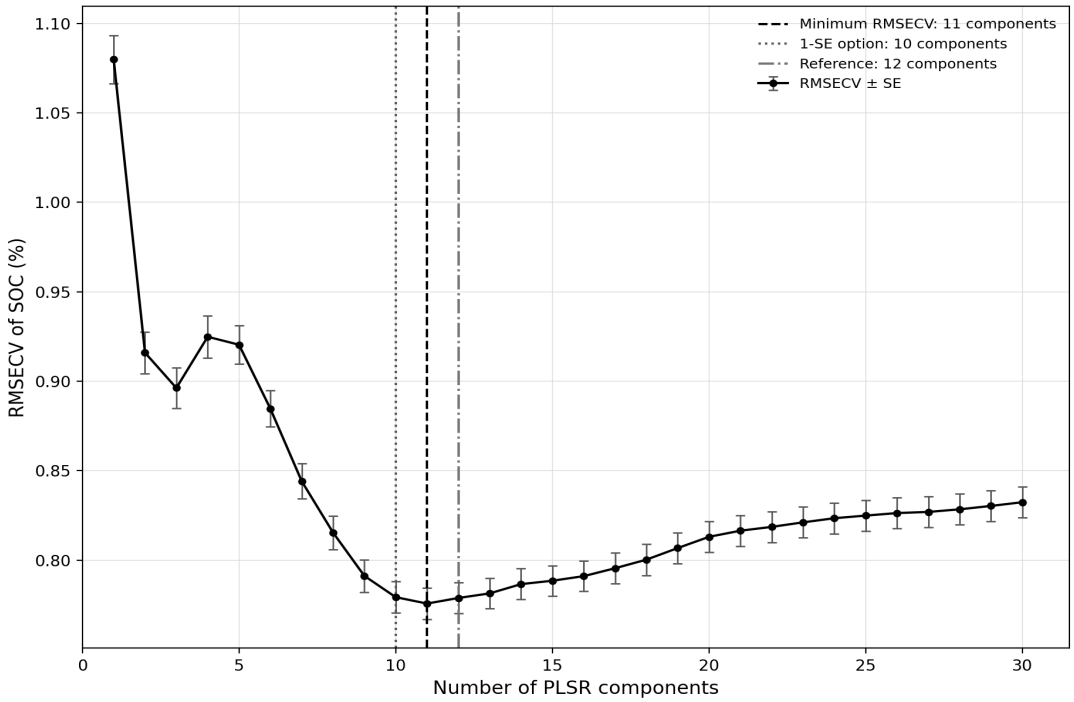

Fig. 4 Selection of the number of PLSR components by repeated k-fold cross-validation. The lowest mean RMSECV was obtained with 11 components; 1-SE denotes the one-standard-error parsimony reference, and the 12-component model was comparable but not optimal

Table 1. Comparison of PLSR components relevant to the manuscript.

Components	Mean RMSECV	SE	Global RMSECV	Global MAECV	Global R^2CV	Interpretation
10	0.7793	0.0088	0.7894	0.6126	0.7217	Parsimonious exploratory model according to 1-SE
11	0.7757	0.0087	0.7856	0.6103	0.7245	Final model by lowest mean RMSECV
12	0.7788	0.0087	0.7887	0.6127	0.7222	Comparable within 1-SE; not optimal

Note. PLSR = partial least squares regression; RMSECV = root mean square error of cross-validation; SE = standard error; MAECV = mean absolute error of cross-validation; R^2CV = cross-validated coefficient of determination; 1-SE = one-standard-error criterion.

Predictive performance of the final model

The final 11-component model showed a high calibration fit; however, the main interpretation must be based on internal validation. In this hold-out set, the model achieved R^2 = 0.7741, RMSE = 0.6683% SOC, MAE = 0.5089% SOC, RPD = 2.12, and RPIQ = 2.36. These metrics indicate moderate to good predictive ability within the evaluated sample domain, but they do not constitute external validation. The difference between calibration and internal validation confirms that calibration metrics should not be used as the main indicator of predictive performance (Table 2; Fig. 5).

Table 2. Metrics of the final 11-component PLSR model.

Set	n	R^2	RMSE (% SOC)	MAE (% SOC)	Bias predicted − observed	RPD	RPIQ	Slope pred~obs	Intercept pred~obs
Calibration	237	0.9821	0.2001	0.1571	0.0000	7.49	8.00	0.9821	0.0285
Internal validation	60	0.7741	0.6683	0.5089	0.0590	2.12	2.36	0.8382	0.3161

Note. RMSE = root mean square error; MAE = mean absolute error; RPD = ratio of performance to deviation; RPIQ = ratio of performance to interquartile distance; bias = predicted − observed; pred~obs = regression of predicted SOC on observed SOC.

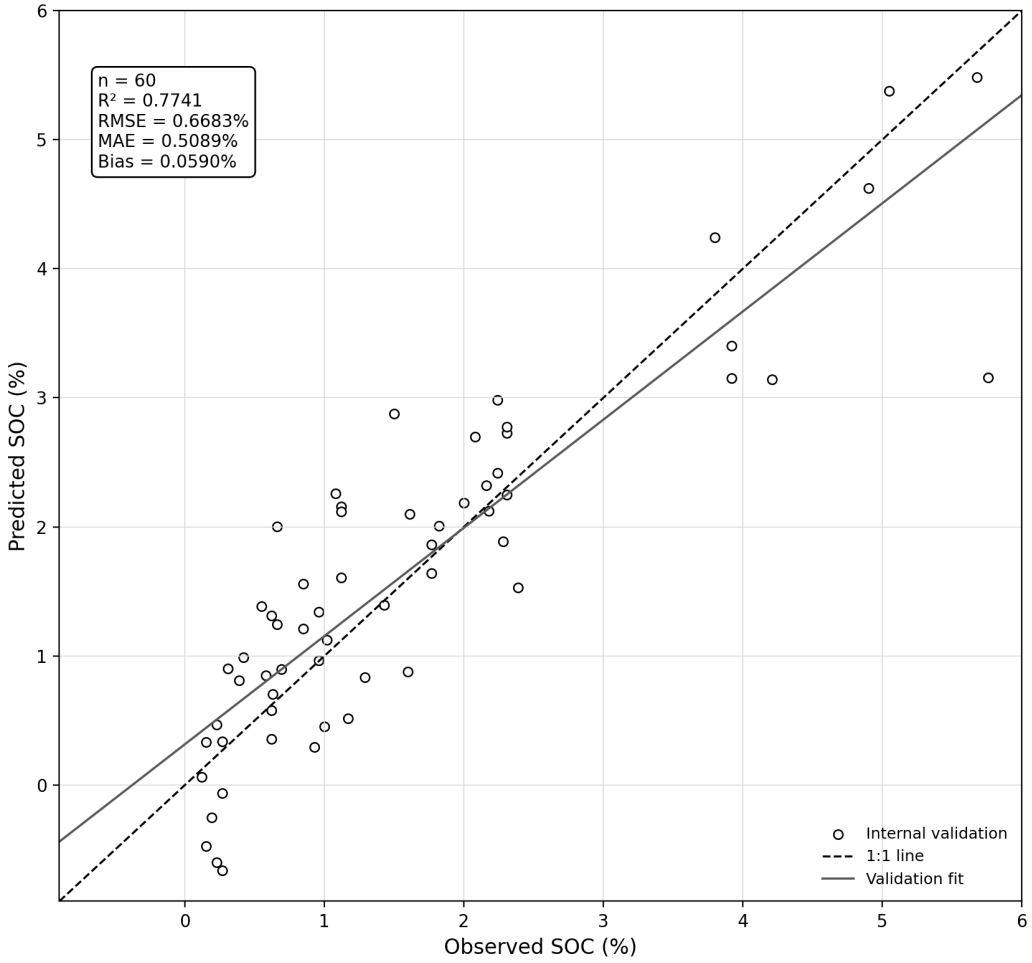

Fig. 5 Observed versus predicted SOC (%) for the complete internal validation set. The dashed line represents the 1:1 relationship and the solid line represents the fitted validation trend

Verified manual equation

The expanded PLSR equation was numerically verified against model.predict(). The direct expression using the internal scikit-learn intercept did not exactly reproduce predictions. In contrast, the equation using the effective intercept reproduced model predictions within numerical tolerance. Therefore, the model was expressed as:

$$SOC\_predicted\ (\%) = 1.342397 + \Sigma\ \beta\lambda\ X\_proc,\lambda \qquad (1)$$

where X_proc,λ is the spectral signal after the complete preprocessing workflow.

Residual diagnostics and retained outlier

Residual diagnostics were acceptable in calibration, with mean close to zero, approximately normal distribution, and no extreme residuals with |z| > 3. In internal validation, the Shapiro-Wilk test indicated non-normal residuals (p = 0.0134), mainly associated with a high-SOC sample underestimated by the model. In addition, the correlation between residuals and observed SOC was significant (r = 0.3417; p = 0.0075), suggesting range-dependent bias, especially at high SOC values (Table 4).

Table 4. Summary of residual diagnostics.

Set	n	Mean residual (obs − pred)	RMSE	MAE	Shapiro p	r residual vs. observed	p	|z| > 3
Calibration	237	-0.0000	0.2001	0.1571	0.5767	0.1337	0.0397	0
Internal validation	60	-0.0590	0.6683	0.5089	0.0134	0.3417	0.0075	1

Note. Residuals are defined as observed − predicted. RMSE = root mean square error; MAE = mean absolute error.

Sample lab 640 had observed SOC = 5.76%, predicted SOC = 3.1551%, and residual = 2.6049% SOC. This sample was retained in the official analysis because no analytical or instrumental error was confirmed. Its presence allows a conservative assessment of the model and prevents artificial overestimation of predictive performance (Table 5; Fig. 6).

Table 5. Sample with extreme residual retained in internal validation.

Lab	Observed SOC (%)	Predicted SOC (%)	Residual (obs − pred)	Absolute error	z residual	Decision
640	5.76	3.1551	2.6049	2.6049	3.8804	Retained; model limitation

Note. The sample was retained because no analytical, instrumental, or data-entry error was confirmed.

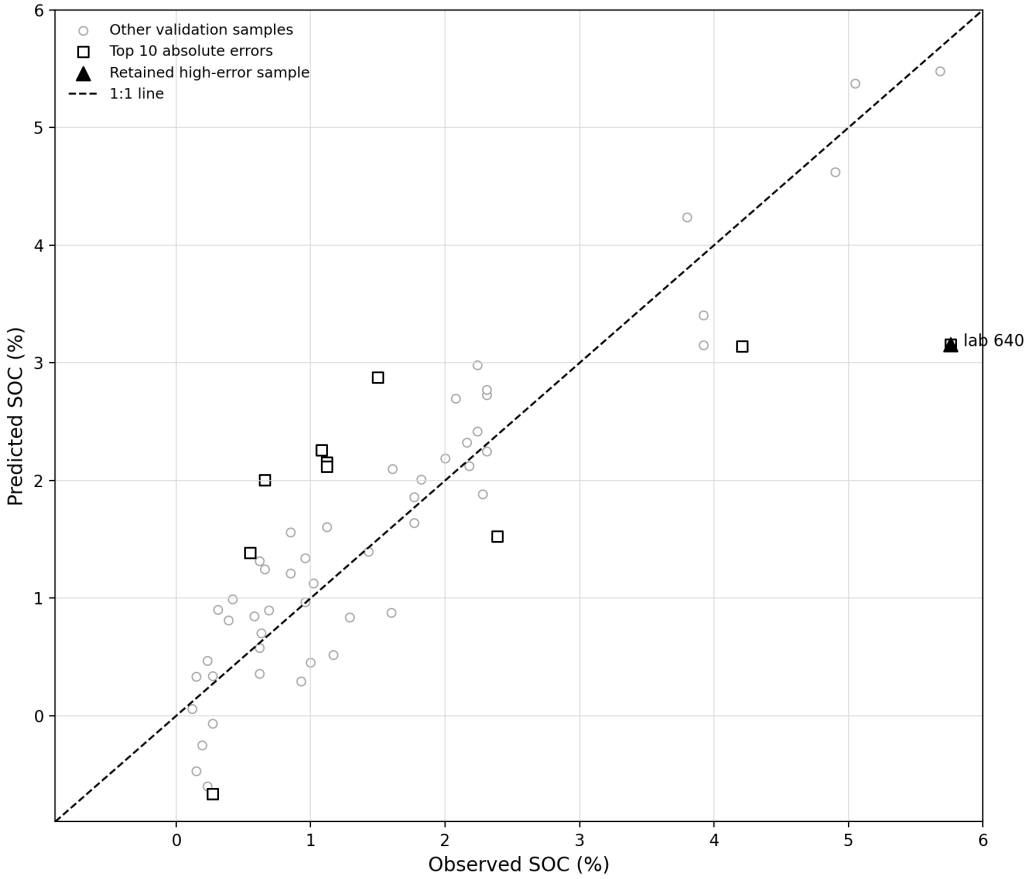

Fig. 6 Observed-predicted relationship for the ten samples with the highest absolute error in internal validation. Sample lab 640 was retained according to a conservative methodological criterion

Discussion

The spectral calibration developed in this study responds to a practical methodological need: estimating SOC at lower operational cost and with faster turnaround than conventional wet chemistry, while maintaining traceability to a laboratory reference. Repeatable SOC measurement is central to carbon-sequestration assessment (Minasny et al. 2017), and proximal sensing can support carbon accounting when calibration and uncertainty are explicitly controlled (England and Viscarra Rossel 2018). In this context, the main contribution of the study is not only the validation R^2, but the integration of a low-cost physical prototype, a local reference matrix, and an auditable chemometric workflow consistent with the routine-implementation perspective described by Poppiel et al. (2022).

Geometric stabilization through EOS-3D was a critical component of the measurement design. In compact sensors, small variations in distance, angle, and stray light can modify relative intensity and shift the spectral baseline, generating variation unrelated to SOC. This agrees with the need for physical harmonization in soil reflectance measurements reported by Ben-Dor et al. (2024) and with evidence that laboratory VNIR-SWIR spectra are sensitive to protocol variables (Xu et al. 2022). The procedure also addresses the measurement-representativeness problem discussed by Fonseca et al. (2022), because the two spectral outputs generated by one LinkSquare reading were averaged before modeling instead of being treated as independent samples. Thus, EOS-3D should be interpreted as part of the measurement protocol, not merely as an accessory.

The selection of 11 components by the minimum RMSECV rule strengthens the methodological defensibility of the model. Although models with a nearby number of components were comparable during exploration, the objective criterion used in this study was the lowest cross-validation error, which occurred with 11 components. This avoids a post-hoc choice based on the hold-out set and reduces the risk of overfitting. The decision is consistent with the use of cross-validation to control latent-variable complexity in PLSR soil models (Bai et al. 2022; Reyna et al. 2017; Tu et al. 2021).

The difference between calibration and internal validation must be interpreted cautiously. Calibration R^2

was high (0.9821), but the relevant performance for assessing internal generalization was R^2 = 0.7741 and RMSE = 0.6683% SOC in the hold-out set. This reduction suggests that the model strongly captured the structure of the calibration set but performed less well when applied to samples not used during fitting. Therefore, the result should be presented as internal validation within the sample domain and not as external validation. This interpretation is consistent with the warning by Reyna et al. (2017) regarding apparently strong PLSR models without sufficient validation control and with the need for explicit application domains in soil spectral libraries (Viscarra Rossel et al. 2016).

The performance obtained is coherent with the behavior reported for compact sensors and local models, although comparisons must consider differences in instrument, matrix, preparation, and scale. Fonseca et al. (2023) reported RMSEP values from 5.2 to 6.3 g kg^{-1}, equivalent to 0.52-0.63% SOC, with R^2 between 0.522 and 0.645 for NeoSpectra under independent validation and with emphasis on sample preparation and local modeling. Goodwin et al. (2022) reported an average R^2 of 0.54 and approximate accuracy of +/- 0.3% SOC for a portable reflectometer, but noted lower accuracy in samples above 2.0% SOC; when reflectance was combined with remotely sensed variables, performance improved to R^2 = 0.75 and RMSE = 0.447% SOC. Hutengs et al. (2019) also show that performance varies according to instrument, in situ or laboratory condition, and calibration domain. In this study, R^2 = 0.7741 and RMSE = 0.6683% SOC in internal validation are defensible within the EOS-3D protocol, but they should not be extrapolated to other collections, sessions, or devices without recalibration or independent verification.

The outlier identified in internal validation was not removed because no analytical, instrumental, or data-entry error was verified. This conservative decision avoids artificially improving the reported performance. The underestimated high-SOC sample reveals a relevant model limitation: predictions may shrink toward the center of the calibration distribution when high-SOC spectral patterns are poorly represented. Its retention is consistent with robust anomaly-detection principles, where exclusion requires technical justification rather than statistical convenience (Rousseeuw and Hubert 2018; Thennadil et al. 2018).

Uncertainty in the Walkley-Black method constrains the performance ceiling of the spectral equation. This procedure estimates oxidizable organic carbon and not total carbon by dry combustion; therefore, the reported RMSE integrates optical error, physical sample variability, preprocessing effects, analytical uncertainty, and bias inherited from the laboratory reference method. De Vos et al. (2007) document recovery, limitations, and uncertainty of Walkley-Black analysis in soils. Ellison and Williams (2012) establish that every analytical measurement should be interpreted with associated uncertainty. This interpretation prevents attributing all observed error to the sensor and strengthens the methodological reading of the result.

The LinkSquare spectral range must be reported with methodological precision. The nominal range of the equipment is approximately 400-1000 nm, whereas the exported spectral matrix ranged from 400.89 to 1099.50 nm and the majority-populated signal interval was 416.79-799.10 nm. Because this effective information is located mainly in the visible and short near-infrared region, the model signal may depend on soil properties correlated with SOC, such as color, darkening, iron oxides, texture, and mineralogy. Pasquini (2018) describes the indirect and overlapping nature of many NIR signals, and Viscarra Rossel et al. (2016) show that soil spectra encode multiple edaphic attributes.

Conclusions

A protocol for spectral measurement of soil organic carbon was developed using a Vis-NIR LinkSquare sensor coupled to the EOS-3D station, integrating physical control of acquisition geometry, sample-level averaging, homogeneous sample presentation, and reproducible chemometric preprocessing.

Each LinkSquare reading generated two spectral signatures that were averaged to form one representative spectrum per sample. This procedure converted 594 instrumental signatures into 297 independent sample-level observations and avoided pseudoreplication.

The nominal range of the LinkSquare sensor was distinguished from the spectral matrix effectively available for modeling. The exported matrix ranged from 400.89 to 1099.50 nm, whereas the majority-populated signal interval was 416.79-799.10 nm.

The number of components in the final model was selected objectively by repeated k-fold cross-validation under the minimum RMSECV rule. The lowest mean RMSECV was obtained with 11 components; therefore, the final reported model uses 11 PLSR components.

The final model achieved R^2 = 0.7741, RMSE = 0.6683% SOC, MAE = 0.5089% SOC, RPD = 2.12, and RPIQ = 2.36 in internal validation. These results indicate moderate to good predictive ability within the

evaluated sample and instrumental domain, but they do not constitute external validation.

The manual equation was numerically verified against model.predict(). The expanded expression requires the effective intercept b0 = 1.342397 and the same preprocessed spectral variables used during model fitting.

Residual diagnostics identified one high-SOC sample underestimated by the model. Because no analytical or instrumental error was confirmed, this sample was retained in the official analysis. Its presence reveals a real model limitation for high SOC values with spectral patterns poorly represented in the calibration domain.

The final RMSE integrates sensor uncertainty, physical sample heterogeneity, preprocessing effects, and uncertainty inherited from the Walkley-Black reference method. Therefore, the calibration function must be externally validated before application to new sessions, sensors, or soil collections.

Acknowledgments

The authors thank the Laboratory of Soil, Plant, Water, and Fertilizer Analysis of Universidad Nacional Agraria La Molina for technical support and availability of analytical information for this study. They also thank Bruno Gamarra for his assistance with laboratory activities. This work was supported by the Vice-Rectorate for Research of Universidad Nacional Agraria La Molina (UNALM)

References

Bai, Z., Xie, M., Hu, B., Luo, D., Wan, C., Peng, J., et al. (2022). Estimation of soil organic carbon using Vis-NIR spectral data and spectral feature bands selection in Southern Xinjiang, China. *Sensors*, 22(16), 6124. doi:10.3390/s22166124

Ben-Dor, E., Efrati, B., Amir, O., Francos, N., Shepherd, J., Khosravi, V., et al. (2024). A standard and protocol for in-situ measurement of surface soil reflectance. *Geoderma*, 447, 116920. doi:10.1016/j.geoderma.2024.116920

Bünemann, E. K., Bongiorno, G., Bai, Z., Creamer, R. E., De Deyn, G., de Goede, R., et al. (2018). Soil quality: A critical review. *Soil Biology and Biochemistry*, 120, 105-125. doi:10.1016/j.soilbio.2018.01.030

De Vos, B., Lettens, S., Muys, B., & Deckers, J. A. (2007). Walkley-Black analysis of forest soil organic carbon: Recovery, limitations and uncertainty. *Soil Use and Management*, 23(3), 221-229. doi:10.1111/j.1475-2743.2007.00084.x

Ellison, S. L. R., & Williams, A. (Eds.). (2012). *Quantifying uncertainty in analytical measurement* (3rd ed.). Eurachem/CITAC.

England, J. R., & Viscarra Rossel, R. A. (2018). Proximal sensing for soil carbon accounting. *SOIL*, 4, 101-122. doi:10.5194/soil-4-101-2018

Fonseca, A. A., Pasquini, C., Costa, D. C., & Soares, E. M. B. (2022). Effect of the sample measurement representativeness on soil carbon determination using near-infrared compact spectrophotometers. *Geoderma*, 409, 115636. doi:10.1016/j.geoderma.2021.115636

Fonseca, A. A., Pasquini, C., & Soares, E. M. B. (2023). Large-scale measurement of soil organic carbon using compact near-infrared spectrophotometers: Effect of soil sample preparation and the use of local modelling. *Environmental Science: Advances*, 2, 1372-1381. doi:10.1039/D3VA00046J

Goodwin, D. J., Kane, D. A., Dhakal, K., Covey, K. R., Bettigole, C., Hanle, J., et al. (2022). Can low-cost, handheld spectroscopy tools coupled with remote sensing accurately estimate soil organic carbon in semi-arid grazing lands? *Soil Systems*, 6(2), 38. doi:10.3390/soilsystems6020038

Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., & Vohland, M. (2019). In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. *Geoderma*, 355, 113900. doi:10.1016/j.geoderma.2019.113900

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., et al. (2017). Soil carbon 4 per mille. *Geoderma*, 292, 59-86. doi:10.1016/j.geoderma.2017.01.002

Paiva, A. F. S., Poppiel, R. R., Rosin, N. A., Greschuk, L. T., Rosas, J. T. F., & Demattê, J. A. M. (2022). The Brazilian Program of soil analysis via spectroscopy (ProBASE): Combining spectroscopy and wet laboratories to understand new technologies. *Geoderma*, 421, 115905. doi:10.1016/j.geoderma.2022.115905

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Poppiel, R. R., Paiva, A. F. S., & Demattê, J. A. M. (2022). Bridging the gap between soil spectroscopy and traditional laboratory: Insights for routine implementation. *Geoderma*, 425, 116029. doi:10.1016/j.geoderma.2022.116029

Reyna, L., Dube, F., Barrera, J. A., & Zagal, E. (2017). Potential model overfitting in predicting soil carbon content by visible and near-

infrared spectroscopy. *Applied Sciences*, 7(7), 708. doi:10.3390/app7070708

Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2), e1236. doi:10.1002/widm.1236

Semella, S., Hutengs, C., Seidel, M., Ulrich, M., Schneider, B., Ortner, M., et al. (2022). Accuracy and reproducibility of laboratory diffuse reflectance measurements with portable VNIR and MIR spectrometers for predictive soil organic carbon modeling. *Sensors*, 22(7), 2749. doi:10.3390/s22072749

Soil Survey Staff. (2014). *Soil Survey field and laboratory methods manual* (Soil Survey Investigations Report No. 51, Version 2.0). Lincoln, NE, USA: U.S. Department of Agriculture, Natural Resources Conservation Service.

Stratio Inc. (n.d.). LinkSquare product information. https://linksquare.io/product-linksquare. Accessed 14 June 2026.

Thennadil, S. N., Dewar, M., Herdsman, C., Nordon, A., & Becker, E. (2018). Automated weighted outlier detection technique for multivariate data. *Control Engineering Practice*, 70, 40-49. doi:10.1016/j.conengprac.2017.09.006

Tu, Y., Zou, B., Feng, H., Zhou, M., Yang, Z., & Xiong, Y. (2021). A near standard soil samples spectra enhanced modeling strategy for Cd concentration prediction. *Remote Sensing*, 13(14), 2657. doi:10.3390/rs13142657

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, 261-272. doi:10.1038/s41592-019-0686-2

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., et al. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, 155, 198-230. doi:10.1016/j.earscirev.2016.01.012

Weil, R. R., & Brady, N. C. (2017). *The nature and properties of soils* (15th ed.). Harlow, United Kingdom: Pearson.

Xu, Z., Chen, S., Lu, P., Wang, Z., Li, A., Zeng, Q., et al. (2022). Optimizing a standard spectral measurement protocol to enhance the quality of soil spectra: Exploration of key variables in lab-based VNIR-SWIR spectral measurement. *Remote Sensing*, 14(7), 1558. doi:10.3390/rs14071558