Journal Pre-proof

Estimating Soil Organic Carbon Using Field VNIR-SWIR Spectroscopy and Existing Soil Spectral Libraries: Mitigating Heterogeneity, Roughness and Moisture Effects

F Castaldi , B Stenberg , F Liebisch , K Metzger , E Ben-Dor , M Knadel , T Koganti , J Wetterlind , R Barbetti , G Debaene , K Klumpp , M Lippl , R Lorenzetti , C Lozano Fondon , T Sanden , A Schaumberger , D Stajnko

PII: S2772-3755(25)00584-2
DOI: https://doi.org/10.1016/j.atech.2025.101353
Reference: ATECH 101353

To appear in: *Smart Agricultural Technology*

Received date: 10 July 2025
Revised date: 5 August 2025
Accepted date: 21 August 2025

Please cite this article as: F Castaldi , B Stenberg , F Liebisch , K Metzger , E Ben-Dor , M Knadel , T Koganti , J Wetterlind , R Barbetti , G Debaene , K Klumpp , M Lippl , R Lorenzetti , C Lozano Fondon , T Sanden , A Schaumberger , D Stajnko , Estimating Soil Organic Carbon Using Field VNIR-SWIR Spectroscopy and Existing Soil Spectral Libraries: Mitigating Heterogeneity, Roughness and Moisture Effects, *Smart Agricultural Technology* (2025), doi: https://doi.org/10.1016/j.atech.2025.101353

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2025 Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Highlights

- VNIR-SWIR field spectra are mostly impacted by roughness and soil moisture
- Smoothing the soil surface improved stability of spectral measurements
- ML models trained on SSLs satisfactory predicted SOC from field spectra
- Combining ISS and EPO correction yielded the highest SOC prediction accuracy
- Using SSL-based models on field spectra to estimate SOC is a game-changer

Journal Pre-proof

Estimating Soil Organic Carbon Using Field VNIR-SWIR Spectroscopy and Existing Soil Spectral Libraries: Mitigating Heterogeneity, Roughness and Moisture Effects

CASTALDI F.[a], STENBERG B.[b], LIEBISCH F.[c], METZGER K.[d], BEN-DOR E.[e], KNADEL M.[f], KOGANTI T.[f] WETTERLIND J.[b], BARBETTI R.[g], DEBAENE G.[h], KLUMPP K.[i], LIPPL M.[j], LORENZETTI R.[a], LOZANO FONDON C.[k], SANDEN T.[l], SCHAUMBERGER A.[m], STAJNKO D.[n]

[a] Institute of BioEconomy, National Research Council of Italy (CNR), Via Giovanni Caproni 8, 50145 Firenze, Italy

[b] Department of Soil and Environment, Swedish University of Agricultural Sciences (SLU), Gråbrödragatan 19, 532 23 Skara, Sweden.

[c] Agroscope, Research Division Agroecology and Environment, Zürich,Switzerland

[d] Agroscope, Field-Crop Systems and Plant Nutrition, Route de Duillier 60, 1260, Nyon, Switzerland.

[e] Department of Geography, Porter School of Environmental and Earth Science, Tel Aviv University, Tel Aviv-Yafo, Israel

[f] Department of Agroecology, Aarhus University, Blichers Alle 20, 8830 Tjele, Denmark

[g] Research Centre for Forest and Wood, Council for Agricultural Research and Economics, Strada Frassineto, 35, 15033 Casale Monferrato, Alessandria, Italy.

[h] Department of Soil Science and Environmental Analyses, Institute of Soil Science and Plant Cultivation – State Research Institute, ul. Czartoryskich 8, 24-100 Puławy, Poland.

[i] INRAE, VetAgro-Sup, University of Clermont Auvergne, UREP, 5 Chemin de Beaulieu, Clermont Ferrand , France

[j] Austrian Agency for Health and Food Safety (AGES), Spargelfeldstrasse 191.

[k] Research Centre for Agriculture and Environment, Council for Agricultural Research and Economics, via di Lanciola 12/A, 50125 Florence, Italy.

[l] Research Soils and Agroecology, Department for Soil Health and Plant Nutrition, Austrian Agency for Health and Food Safety (AGES), Spargelfeldstrasse 191, 1220 Vienna, Austria.

[m] Agricultural Research and Education Center (AREC) Raumberg-Gumpenstein, Raumberg 38, A-8952 Irdning-Donnersbachtal, Austria.

[n] Faculty of Agriculture and Life Sciences, University of Maribor, Pivola 10, 2311 Hoče, Slovenia.

Journal Pre-proof

ABSTRACT

VNIR-SWIR spectra acquired in the field are inherently affected by uncontrolled conditions, such as variable illumination, surface roughness, and soil moisture. As a result, models trained on soil spectral libraries (SSLs), typically composed of dry, sieved samples analyzed in the lab, often fail when applied directly to field spectra. With this study we propose a routine to succeed with this desirable approach. We collected field spectra from 178 locations across seven countries under heterogeneous field conditions using different spectrometers. At each site, two surface smoothing intensities were compared. Two SSLs, LUCAS topsoil and GEO-CRADLE, were used to train machine learning models for predicting soil organic carbon (SOC), later applied to the field spectra under different correction scenarios: with or without Internal Soil Standard (ISS) harmonization and External Parameter Orthogonalization (EPO) to mitigate the effects of soil moisture. Combining ISS and EPO enables SSL-based models to reliable predict SOC from field-acquired spectra, particularly when using the LUCAS SSL in combination with a spectrally localized approach to reduce training set size (R^2 = 0.70; RPD = 1.66). Model performances are consistent with previous laboratory-based studies despite the diverse field conditions. A refined workflow for SOC estimation using hybrid spectral data is proposed, consisting of three steps: i) Spectral acquisition on highly smoothed surfaces; ii) ISS harmonization to align spectra across from different instruments; iii) EPO correction to reduce non-systematic spectral variability due to masking factors such as moisture, enhancing spectral consistency under variable field conditions.

Keywords: Field Spectroscopy; SOC; LUCAS; EPO; Soil; Machine learning; Soil spectral library, VNIR-SWIR

Introduction

Soil reflectance measurement across the visible–near infrared–shortwave infrared (VNIR–SWIR; 400 – 2500 nm) region is a widely employed technique for both proximal and remote sensing of soil properties [1–4]. The technique has proven effective in predicting various soil properties in the laboratory with air dried and sieved soils, offering faster results compared to traditional wet chemistry methods, and maintaining reliable accuracy [5–7]. Over the years, numerous soil spectral libraries (SSLs) have been generated using spectral and chemical information of soil samples stored in physical archives. These are valuable resources forming a basis for efficient and robust estimation of soil properties. Additionally, VNIR–SWIR spectroscopy has found applications not only in laboratory settings but also in the field. Technically, spectroscopy in this wavelength range is well suited for field applications and offers significant advantages in terms of time and cost, as it eliminates the need for sample preparation steps such as drying, grinding, and sieving [8]. However, soil reflectance measurements in the field are influenced by numerous environmental factors that are avoided in the laboratory.

A standard method for measuring soil reflectance in the field for quantitative estimation of soil composition should be robust, rapid, representative, and as reproducible as possible considering soil environment aspects such as moisture and surface structure. For field spectroscopy, various measurement setups and optical geometries for sample presentation are employed [9]. Collecting reflectance using a bare fiber (BF) relies on solar radiation and thus suffers from challenges such as variations in solar elevation, atmospheric attenuation, bidirectional reflectance distribution function (BRDF) effects, operator skills and a lack of information in certain wavelength regions due to atmospheric attenuation. A contact probe (CP) with an internal light source rules out these challenges and allows stable, pinpointed measurements. On the other hand, a CP is limited by its

Journal Pre-proof

small sampling area, which ranges from a few square centimeters to several hundred cm², over which spectral information is captured either on soil surfaces or along sampling cores. Another limitation is that the sapphire window of the CP must be in full contact with the surface. A rugged and uneven surface may prevent full contact between the CP and the soil surface; in very dry soil conditions, it may be necessary to break soil crusts, while in extremely wet conditions, the CP lens may become covered with mud. Several replicate spectral acquisitions may be required to account for spatial variation. As an alternative to a contact probe the SoilPRO$^{(R)}$ device [10,11] for instance combines a stable light source with a relatively large sample area of 700 cm^2 and screens off ambient light ensuring a consistent geometric acquisition. This is particularly advantageous in studies where an undisturbed soil surface is required regardless of atmospheric and sun illumination conditions, but spectra will be affected by any soil surface contamination.

Soil property estimation models, calibrated using soil spectral libraries (SSL) based on dry and sieved samples in the lab, typically fail when applied to spectra collected in the field [12], mainly due to the differences in environmental factors. Such factors are soil moisture (SM), roughness, surface sealing, light source variations, atmospheric attenuation and sun elevation changes. Additionally, other factors related to the measurement setup, such as user experience, measurement geometry, and measurement procedures must also be considered [10,13]. These factors further introduce uncertainties in data and pose challenges in harmonizing data from field and laboratory. As a result, the accuracy of quantitative estimates from acquired field spectral will be affected and the estimation of soil properties based on SSLs is not straightforward. Since SM and roughness are the two factors with the greatest impact on the spectral responses of soils, SSL and field spectra will not correspond as a default [14] and novel approaches are required to integrate them.

Pronounced and variable soil roughness can cause transmittance, shadows, multiple reflections and scatter, between surface irregularities and particle arrangements. The results in reduced albedo and changed curvatures in spectra [7]. Conversely, a smoother surface generates a better signal-to-noise ratio and more uniform and repeatable measurements. When the aim is to match with laboratory spectra this is desirable. To transform spectra to improve spectral quality is a standard procedure also for laboratory measurements. Common transformations like first and second order derivatives and Standard Normal Variate and De-Trending (SNV-DT) [15] correct for baseline shifts and linear trends in spectra and thereby reduce the effects of scatter [7].

In general, when soil is scanned, electromagnetic radiation travels through a thin layer of particles and is reflected back to the sensor, producing a soil spectrum influenced by both the real and imaginary components of the refractive index, indicating, respectively, the phase speed and the amount of absorption loss when the electromagnetic wave propagates through the material [14]. The complex refractive index of soil results from the combined effects of its mineral and water content; when the soil is dry, the mineral component dominates the imaginary part of the refractive index. In contrast, under moist conditions, the contribution of water to the real part becomes increasingly dominant as moisture levels rise, gradually masking spectral features associated with the mineral fraction. The influence of moisture on the soil spectrum stems from changes in the relative refractivity at particle surfaces caused by the presence of a thin water film in the visible range, and also from the strong absorption features of water in the infrared range [17]. As a result, the spectral features of mineral and organic constituents are diminished or "masked" out by the spectral contribution of water [7,18,19]. Several mathematical approaches exist to mitigate or remove SM effects on spectra such as Direct Standardization (DS), Piecewise Direct Standardization (PDS) and Orthogonal Signal Correction (OSC) (e.g. [14,20]).

According to [14], who reviewed mathematical techniques for reducing moisture effects on the VNIR–SWIR spectra, the External Parameter Orthogonalization (EPO) algorithm is the most effective method, focusing on the spectral variability linked to the effects originating from external factor only. It can significantly reduce errors in estimating SOC, total nitrogen, or clay content [21,22], particularly under varying SM levels. Ackerson et al. [23] successfully tested the EPO algorithm for predicting clay content using a dried and ground SSL on moist soils. However, they collected spectra in the laboratory from moist samples, assuming these conditions were similar to in-situ. Wijewardane et al. [24] and Murad et al. [25] used a U.S. spectral library to predict soil properties from spectral data acquired in the field with a multi-sensing penetrometer system equipped with the same spectroradiometer as the SSL. Both studies successfully applied EPO to improve prediction accuracy, also thanks to the reduction of other disturbing factors during field acquisitions. In fact, the penetrometer system enabled spectral acquisition without the influence of sunlight and with limited effects from roughness and temperature.

Independent to data sources, spectral harmonization between SSLs and the spectra of the samples to be predicted is essential, particularly when data are collected using different spectroradiometers, setups or protocols. For this purpose, the Australian white sand from Lucky Bay (LB) has been successfully used as an internal soil standard (ISS) for spectral normalization across diverse datasets [26–28].

The LUCAS topsoil spectral library includes a large number of spectra collected across Europe, providing an exceptional spectral resource with extensive soil variability [29–31]. However, LUCAS samples were scanned using a benchtop spectroradiometer and did not use the ISS harmonization procedure, making it challenging to align these data to field measurements or to other soil spectral libraries that have been assembled using other spectrometers and setups.

To the best of our knowledge, there are no successful attempts in the literature to exploit existing laboratory based VNIR-SWIR SSLs, such as the LUCAS dataset, to predict soil properties from entirely independent spectral data collected under field conditions using different portable spectroradiometers. Within the scope of the @@@@@@ project, we tested different approaches for estimating SOC content through field spectroscopy [32]. Here we report on the effectiveness of two in-field soil surface pre-treatments to reduce spectral noise caused by surface roughness, spectral harmonization performed using the LB sand as an ISS and the effect of mitigating SM in field spectra by applying the EPO algorithm. Spectral transformations were also applied to reduce scatter effects remaining after soil pre-treatments in the field. We hypothesized that this three-step workflow could create a dataset of field spectra that closely resembles the spectra of dry and sieved soils, such as those in existing laboratory-based SSLs. In this approach we may facilitate the development of SOC prediction models from SSLs that can be applied directly to field spectra at various SM conditions. Consequently, the objectives of this study are: i) to evaluate the impact of surface pre-treatments on harmonizing in-field spectral measurements, and ii) to test the effectiveness of combining spectral harmonization with the EPO transformation in improving the accuracy of in-field SOC predictions using machine learning models trained on dried and sieved lab-based VNIR–SWIR SSLs.

Journal Pre-proof

Materials and Methods

Soil spectral libraries

Two large soil spectral libraries (SSLs) were used to train machine learning models for predicting SOC on independent validation spectra collected in the field (Table 1 and Figure 1): Land Use and Coverage Area frame Survey (LUCAS) 2015 European topsoil dataset [33] and GEO-CRADLE regional soil spectral library [34].

Table 1: Main characteristics of the soil spectral libraries used in this work. RM: spectral library composed of re-moistened soil samples; LS: spectral library obtained by field measurement on lightly-smoothed surface ; HS: spectral library obtained by field measurement on Highly-smoothed surface ; DS: spectral library obtained by laboratory measurement on dried and sieved soil samples.

Spectral dataset	*Name*	*N*	*Laboratory measurements*		*Field measurements*
			Dry and sieved	*Sieved and re-wetted*	*Wet*
Soil spectral libraries	LUCAS	12907	X		
	GEO-CRADLE	1754	X		
Re-moistened dataset	RM	81	X	X	
Validation field spectra	LS	178			X
	HS	178			X
Validation lab spectra	DS	178	X		

Journal Pre-proof

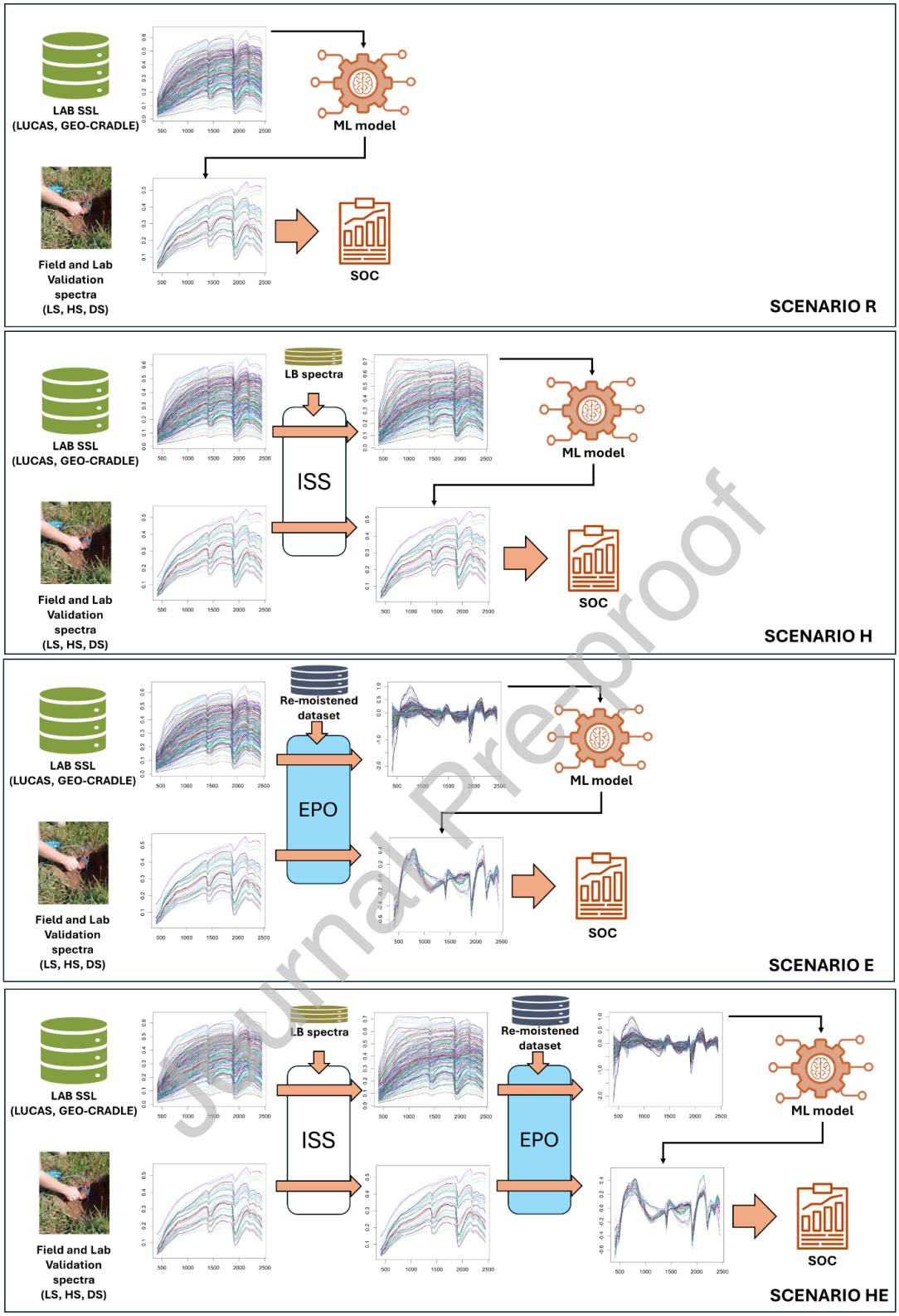

Figure 1: Four workflow scenarios (R, H, E and HE) for estimating soil organic carbon (SOC) using field VNIR-SWIR spectra and soil spectral libraries (SSL) with increasing complexity. LS: spectral library obtained by field measurement on lightly-smoothed surface ; HS: spectral library obtained by field measurement on Highly-smoothed surface ; DS: spectral library obtained by laboratory measurement on dried and sieved soil samples. LB spectra refers to the Lucky Bay sands as internal soil standard.

For calibration, a total of 12907 soil samples from 28 European countries were selected from the LUCAS SSL taking only soil samples collected on agricultural land. From each soil sample, the

SOC content and the VNIR-SWIR spectrum were extracted for our study. Soil spectra in the SSL were obtained using a FOSS XDS rapid content analyzer (FOSS NIR Systems Inc., Laurel, MD, USA) acquiring the spectral range between 400 and 2499 nm with a spectral resolution of 0.5 nm. The spectra were resampled to 1 nm of resolution to homogenize them with those provided by the main spectroradiometers operating in our field study. The SOC values were determined using the dry combustion method (elementary analysis) (ISO 10694:1995) in a central laboratory in Hungary.

The GEO-CRADLE database SSL contains spectral data of 1754 soil samples collected on agricultural lands from 9 Mediterranean and Eastern European countries which were measured for reflectance between 350 and 2500 nm with a spectral resolution of 1 nm. The spectra were acquired using two different spectroradiometers following a standardized protocol: ASD Fieldspec PRO FR (PANanalytical B-V, Boulder, CO, USA) and PSR+ (Spectral Evolution Inc. Lawrance Massachusetts, USA). The spectroradiometers and the protocol are specified in [26]. For this SSL, spectra were harmonized by using the Lucky Bay (LB) sand as internal soil standard (ISS). The LB sand was chosen as ISS due to its spectral stability (being composed almost entirely of quartz), grain size and shape, which are similar to natural soils, and negligible spectral features. The SOC content was not determined in a central laboratory, but each country involved in the SSL measured SOC in their own laboratory following the Walkley-Black method. Only those samples for which both the spectrum and the SOC value were present were kept for subsequent analysis (1621). Due to the limited geographical area and the similar climatic region covered by the GEO-CRADLE dataset, the SOC range and variability are narrower than those observed in LUCAS (Figure 2). On the contrary, the clay range is wider in GEO-CRADLE (1 - 91%) as compared to LUCAS (2 - 62%).

Both the LUCAS and GEO-CRADLE spectra were measured in laboratory conditions on air-dried samples that were gently crushed and then sieved to < 2 mm.

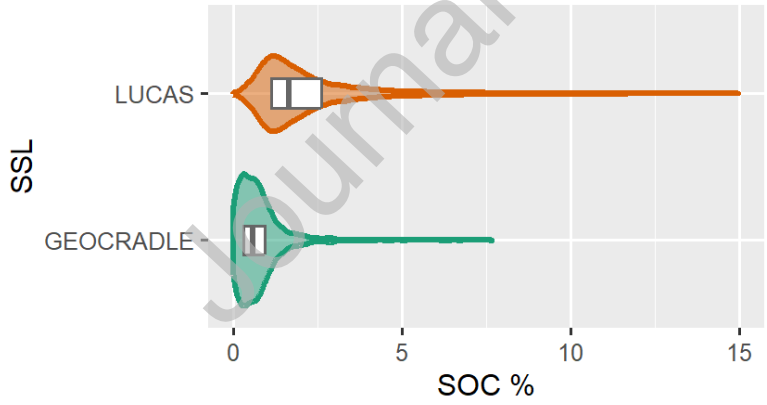

Figure 2: Violin diagrams describing the soil organic carbon (SOC) range of the LUCAS and GEO-CRADLE soil spectral libraries (SSL)

Field spectra collection

Topsoil spectra were acquired directly in the field at 178 sampling locations across seven European countries (Table 2) between 350 and 2500 nm (1 nm spectral resolution) under naturally moist and field roughness conditions using different spectroradiometers (Table 2), which all were equipped with a contact probe including a built-in halogen light source furnished with fiber optic to collect reflected spectra relative to white reference on Spectralon$^{(R)}$. A different brand and model of spectroradiometer was used for each national campaign, with the exception of Switzerland and Poland, where the Spectral Evolution PSR 3500 was employed for both, and the Italian and Slovenian campaigns, which utilized the exact same instrument: a Spectral Evolution RS5400 (Table 2).

A protocol over two surface pre-treatments based on [35] was followed (Figure 3):

Light smooth (LS): lightly smoothing the surface after removing green and dry vegetation, stones and other contaminating materials.

High smooth (HS): after LS, highly smoothing and flattening the surface using a plastic hammer or similar tools.

Journal Pre-proof

*Table 2: Information about the sites where the field spectral campaign was conducted. All the Soil organic content values were determined using the Dry combustion method except for the Switzerland sample for which the Walkley-Black method was used. *Koppen-Geiger climate types (adapted from Peel et al., 2007), ** [36].*

Country	**Region**	**Climate* region**	**Land use**	**Soil type****	**N° soil samples**	**SOC range (%)**	**SM %**	**Instrument**
Italy	Tuscany	Csa	Cropland	Vertisol	23	0.59 – 1.13	9.8 – 14.1	Spectral evolution RS-5400
Sweden	Västra Götaland	Dfb	Cropland	Cambisol and Regosol	12	1.2 – 5.5	19.5 – 33.8	ASD FieldSpec Pro FR
Poland	Lublin Voivodeship	Dfb	Cropland and Meadow	Luvisol, Podzol, Alluvial soils	8	0.63 – 1.91	7.0– 16.8	Spectral Evolution RS 3500
Denmark	Central Jutland	Cfb	Cropland	Cambisol/Phaeozem	8	2.2 – 11.96	13 – 27.9	ASD LabSpec
Switzerland	Cantons of Vaud, Zürich, Thurgau	Cfb	Cropland	Cambisol, calcaric Cambisol, Luvisol	96	1.03 – 4.6	11.1 – 48.5	Spectral Evolution PSR 3500
Austria	Marchfeld	Cfb	Cropland	Chernozem	9	1.67 – 2.05	12.6 – 16.1	ASD FieldSpec PRO FR
Slovenia	Drava	Cfb			22			

			Cropland and Meadow	Cambisol, Eutric; Cambisol, Dystric		1.07 – 2.59	19.1 – 47.7	Spectral evolution RS5400

At each scanning location, a soil sample was collected and brought into the laboratory for SOC measurement by wet chemistry. The soil samples were collected within the same area interested by spectral measurements and at a maximum depth of 20 cm (Figure 3d). Moreover, the soil samples were air dried in the laboratory and then sieved at <2 mm and soil spectra were acquired in laboratory condition using the same instruments as used in the field forming dry spectral (DS) dataset. Thus, for each sampling location we collected spectra according to the LS, HS and DS. According to the protocol followed, five replicates of soil spectra were collected within each sampling location for LS, DS and HS, respectively. The average spectrum was calculated for further analysis (Figure 3). Moreover, for each sampling location *s*, the standard deviation of the reflectance values for the five replicates was computed for each wavelength ($\sigma_{s\lambda}$) and the average standard deviation across the spectral range 400 – 2450 nm was calculated (σ_s). For each spectral measurement protocol (DS, LS, HS) the average standard deviation σ_p was computed according to the following formula

$$\sigma_p = \frac{1}{k}\sum_{s=1}^{k} \sigma_s \quad (1)$$

where k is the number of sampling locations. The σ_p was computed to evaluate the variability of the spectral measurement protocols. Lower σ_p values could indicate greater spectral uniformity, likely resulting from reduced spectral noise due to surface conditions (e.g. soil moisture and roughness) and other influencing factors. The σ_p of the spectral measurement protocols was compared using the Welch's test (p-value < 0.05) [37], a robust and appropriate statistical test for comparing groups of measurements when the homoscedasticity assumption is not satisfied.

The field spectral campaign was conducted in locations across different climatic regions: a hot-summer Mediterranean climate (Csa) in Italy, a temperate oceanic climate (Cfb) in Denmark, Switzerland, Austria, and Slovenia, and a warm-summer humid continental climate (Dfb) in Sweden and Poland (Table 2).

Journal Pre-proof

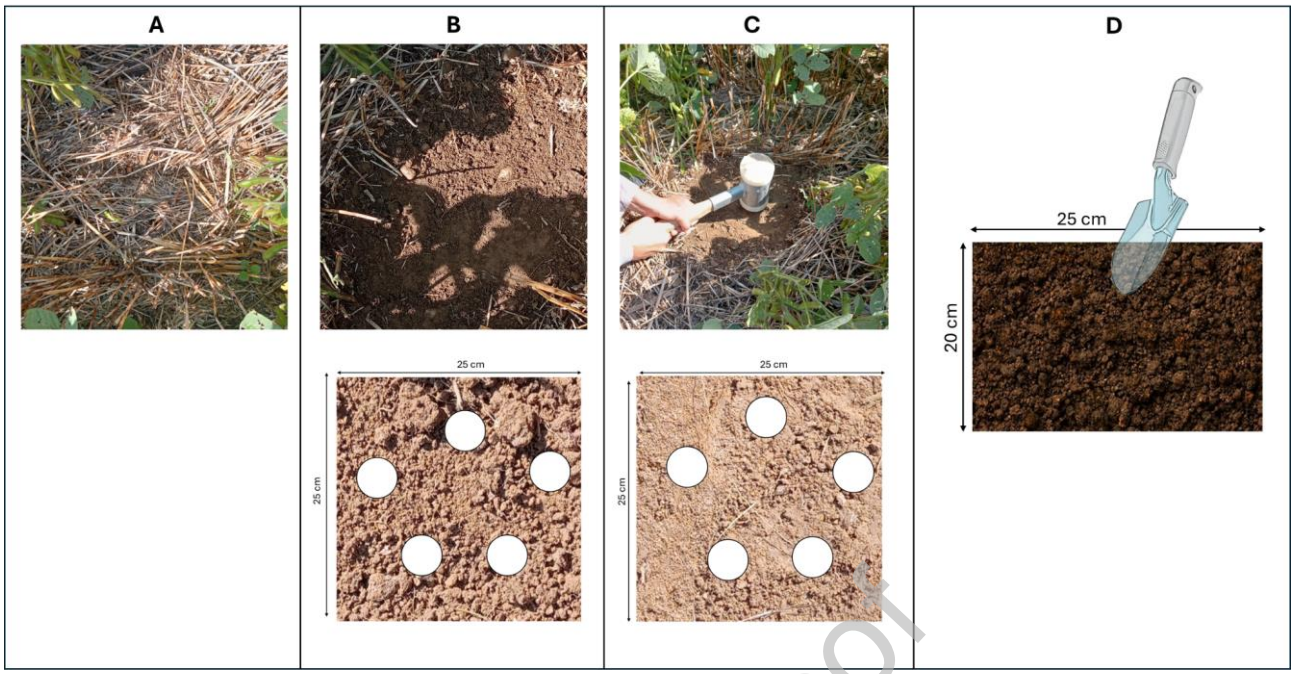

Figure 3: Example of the surface soil pre-treatments: before surface pre-treatment (A), after lightly smoothed (LS) (B), after highly smoothed (HS) surface (C) and the soil sampling collection (D). The white circles approximately indicate the five areas where spectral measurements were taken (5 replicates).

Re-moistened spectral dataset

For calibrating the moisture correction model described in section 2.5, a dataset of 81 soil samples was collected from French, Italian, Polish, Swedish, and Swiss soil archives (Table 3), where SOC and texture measurements were available. The selection aimed to include different soil types and ensure a wide variability in SOC and soil texture. This dataset, henceforth referred to as RM, is completely independent of the field spectral dataset described in section 2.2, therefore these 81 soil samples were not collected in the same sampling locations as the field spectral measurements. The soil samples were air dried and sieved at 2 mm. The SOC ranged between 0.2 and 9.2 %, and clay content between 4 and 55.4%. Soil spectra were collected in the country of origin using different spectroradiometers (Table 3). Before collecting the soil spectra, each lab scanned the ISS LB sands. Each soil sample was split into five subsamples and placed in five petri dishes: one was left dry, while a quantity of water was added to the other four subsamples to reach 5, 10, 20 and 30% of gravimetric soil moisture. Spectral measurements were performed in laboratory dark rooms using a contact probe and at constant temperature and humidity for all 5 subsamples and 3 replicate soil spectra were collected for each subsample and the average spectrum was computed.

Table 3: Information about the soil spectral measurements conducted to create the re-moistened spectral dataset in the laboratory.

Country	N	Instrument	SOC range (%)
France	8	ASD QualitySpec Trek	1.2 - 5.9
Italy	26	Spectral evolution RS5400	0.6 - 3.0
Poland	17	Spectral Evolution PSR 3500	0.2 - 2.3

Sweden	20	ASD FieldSpec Pro FR	1.4 - 9.2
Switzerland	10	Spectral Evolution PSR 3500	0.7 - 2.9

SOC and Spectral data harmonization

Since SOC content was measured using different methods (Dry combustion and Walkley-Black), for the various SSLs used in this work, a data harmonization was carried out by multiplying the SOC values determined by the Walkley-Black by a correction factor according to [38].

To harmonize the spectral data acquired using different instruments, the LB sand was used as an ISS [26] for all the spectral datasets used for this work, that is both large SSLs, field and lab spectral data.

The reflectance spectrum of the ISS (relative to a white reference; WR panel) was scanned for each instrument *i* used to scan the soils of all spectral datasets described so far (Table 1; Figure 4). A correction factor for each wavelength λ was then computed according to the following formula

$$CF_{\lambda i} = 1 - \frac{S_{\lambda i} - M_{\lambda i}}{S_{\lambda i}} \quad (2)$$

Where $S_{\lambda i}$ is the LB reflectance measured at the user's setup and $M_{\lambda i}$ is the soil benchmark ISS reference measured at the CSIRO laboratory. Each spectrum collected using instrument *i* was then multiplied by the CF_i vector: $CF_i=[CF_{(\lambda 1)},CF_{(\lambda 2)},\ldots,CF_{(\lambda N)}]$.

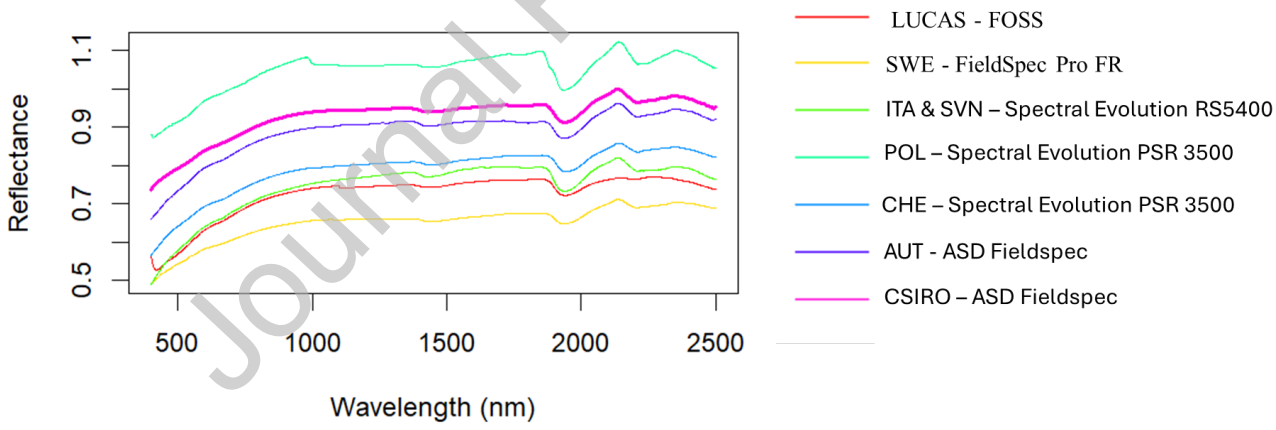

Figure 4: Lucky Bay (LB) spectrum used as internal soil standard (ISS) acquired by using different spectroradiometers and by different labs.

Reducing/removing soil moisture effects

From the RM dataset, a spectral matrix *D* was computed as a difference between the dry spectra and those collected at different soil moisture levels. The matrix *D* was used for the EPO algorithm. This algorithm aims to identify the spectral variability due to the external parameter, that is the water content in this case, therefore projecting the spectral regions orthogonal to the space where the effects of external parameters are dominant [18,39]. The EPO algorithm aims to remove the

Journal Pre-proof

parasitic information, i.e. that affected by the external parameter, decomposing the spectra matrix (X) into three main components:

$$X = XP + XQ + N \quad (3)$$

Where XP is the informative component that we would like to isolate for extracting only the information related to the target variables, XQ is the parasitic component and N is the spectral noise. To obtain XP, the correction matrix P is calculated, obtaining the projection matrix of XQ as a singular value decomposition of D, and multiplied by X. The Wilks' Λ method was used to determine the optimal number of EPO components [39]. The Wilks' Λ value is obtained as the ratio between the trace of the inter-group variance-covariance matrix of EPO-transformed spectra averaged across all the five moisture levels for each sample and the trace of the variance-covariance matrix of the EPO-transformed spectra, therefore Λ values close to 1 indicate that the inter-sample variation is close to the intra-sample variation and, consequently, the EPO worked properly and there is a good separation of samples in the spectral space [40]. Several combinations of spectral pre-treatment were tested for the optimization of the EPO's performance based on the maximization of the Wilks' Λ value. The spectral transformations were carried out using the *prospectr* package in R [41]. Moreover, to evaluate whether the EPO algorithm effectively reduces spectral variability caused solely by changes in soil moisture content, we calculated the standard deviation (SD) of the spectral measurements before (ρ^{raw}) and after EPO transformation (ρ^{epo}), taken at four different moisture levels (m), for each wavelength (λk) and each of the 81 samples (n) of the re-moistened spectral dataset.

$$SD_{raw}(\lambda k) = sd(\rho_{1,m}^{raw}, \rho_{2,m}^{raw}, \rho_{3,m}^{raw}, \rho_{4,m}^{raw}) \quad (4)$$

$$SD_{epo}(\lambda k) = sd(\rho_{1,m}^{epo}, \rho_{2,m}^{epo}, \rho_{3,m}^{epo}, \rho_{4,m}^{epo}) \quad (5)$$

We then compared the average standard deviation before and after applying the EPO transformation. This comparison allowed us to test whether the differences were statistically significant, that is, whether the EPO algorithm leads to a meaningful reduction in spectral variability.

Spectral transformation

All spectral datasets in both SSLs (LUCAS and GEO-CRADLE), LS, HS and DS (Table 1 and Figure 1) were tested in three formats:

R: Raw spectra without any transformation

H: harmonized spectra using the ISS

E: the EPO correction matrix was applied to R spectra

HE: the EPO correction matrix was applied to H spectra

The only exception is the GEO-CRADLE SSL for which the original reflectance data (R) were not available and consequently, it was not possible to obtain the E processing stage

For the field spectra, eight combinations of surface pre-treatments and spectral transformations were obtained: LS_R, LS_H, LS_E, LS_HE, HS_R, HS_H, HS_E, HS_HE where the letters before the underscore indicate the surface pre-treatment (see 2.1) and the letter after the underscore the spectral transformation (Table 4).

Table 4: Description of the three surface pre-treatments and four spectral transformations/scenarios for the laboratory and field spectral libraries.

Dataset code	**Surface pre-treatment**
DS	Spectra collected in the laboratory condition
LS	Spectra collected in the field after lightly smoothing the ground surface
HS	Spectra collected in the field after highly smoothing the ground surface
	Spectral transformation/Scenario
R	No spectral transformation applied
H	Spectral harmonization using ISS correction
E	Spectral correction using EPO transformation
HE	Spectral harmonization using ISS correction + spectral correction using EPO transformation

SOC prediction models from SSLs

Due to the different numerosity and SOC range (Figure 2) of the two SSLs, two different and independent SOC prediction strategies were followed: for the GEO-CRADLE dataset we used all the data for training a global Random Forest model [42], while for LUCAS we adopted a local-spectral approach to reduce the size of the training dataset, similar to that proposed by [43], selecting a different training dataset for each validation spectrum according to the spectral similarities based on the Mahalanobis distance. A principal component analysis (PCA) was previously applied to the spectral matrix to reduce the dimensionality of the spectral dataset to a number of components for which the eigenvalues were greater than 1. Therefore, the LUCAS and GEO-CRADLE SSLs were always used separately and were never combined to build a single training dataset.

For the local-spectral approach we tested models using 100, 200, 500 and 1000 samples for the training dataset, in order to evaluate the influence of the number of the calibration data on SOC prediction accuracy.

The SOC prediction models generated from SSLs were tested on the spectra acquired in the field (LS and HS) and the estimation accuracy was assessed by computing the root mean square error (RMSE; Eq. 6), ratio of performance to deviation (RPD; Eq.7) and the coefficient of determination (R^2; Eq.8)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$RPD = \frac{SD}{RMSE} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (8)$$

Where y_i are the observed values, $\hat{y}_i$ are the predicted values, $\bar{y}_i$ is the average observed value, n and SD respectively the number and the standard deviation of the observed values.

Furthermore, considering the heterogeneity of the soil samples being assessed, the normalized error (NE; Eq.9) was calculated as the percentage error estimate for each sample *i*.

$$NE = \frac{y_i - \hat{y}_i}{y_i} \, x \, 100 \quad (9)$$

All the models were trained and validated using all the three spectral formats indicated in section 2.6 (Table 4) separately, to assess the effectiveness of the ISS and EPO transformation according to the four scenarios described in Figure 1.

Results

Field spectral library

The dataset exhibits a wide variability in soil types, although Cambisols are the most prevalent, particularly in Denmark, Sweden, Switzerland, and Slovenia. Most spectral measurements were performed in croplands, with the exception of a few locations in Poland and Slovenia, where meadows were surveyed. The SOC content in the samples ranges from 0.59% in Italy to 11.9% in Denmark. Clay content ranges between 1.87% in Poland and 55.46 in Switzerland, while soil moisture content at the time of spectral measurement also varied greatly, from dry conditions in Italy (average of 10%) and Poland (12%) to very high moisture levels in Slovenia (47%). Significant SM variability was observed even within individual national campaigns, driven by differences in soil types, land management practices or rain events prior to the measurements.

Uniformity of field spectroscopy procedures

The σ_p value of DS measurements was the lowest (0.020); however, it was not significantly different from HS (0.022) according to the analysis of variance and Welch's test (Figure 5). Both DS and HS values are significantly lower than LS (light smooth) measurements (0.043). In other words, LS resulted in the least uniform spectral measurements, whereas the HS (high smooth) surface pre-treatment achieved a spectral uniformity comparable to that obtained in the DS lab.

The σ_p values did not show a significant correlation with SM and SOC content, therefore neither of the two variables had an effect on measurement uniformity.

Journal Pre-proof

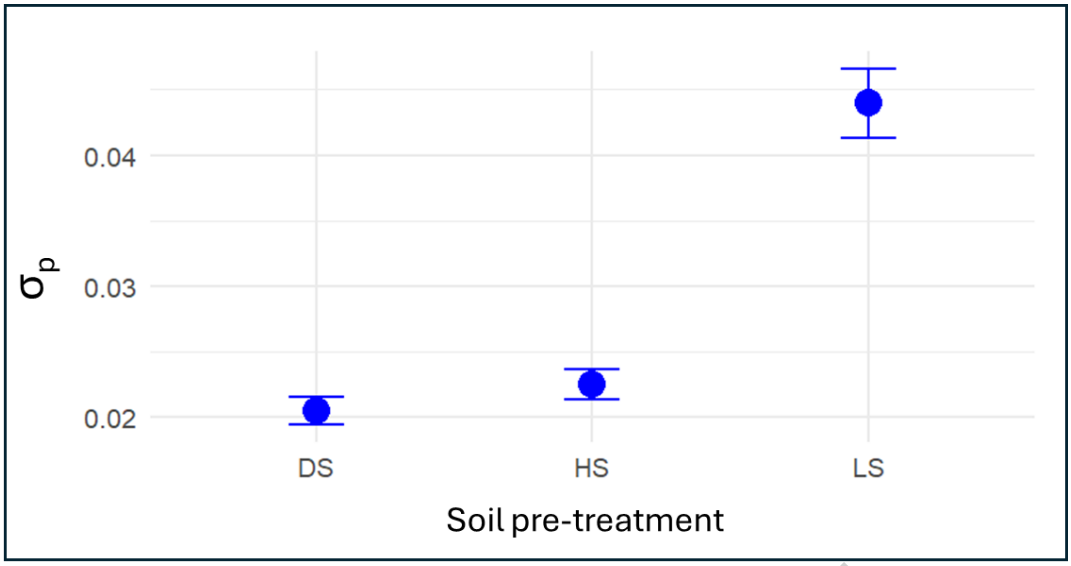

Figure 5: Mean standard deviation of the spectral measurements for each group of soil pretreatment: DS (lab spectra acquired on dried and sieved soils), LS (field spectra acquired on lightly smoothed surfaces), HS (field spectra acquired on highly smoothed surfaces)

EPO transformation

We tested the efficiency of the EPO transformation by computing the Wilk's Λ values for a number of EPO components ranging from 1 to 15 and using different spectral pre-treatments and spectral transformations. The combination of Savitzky-Golay filter and detrend transformation using three EPO components provided the highest Wilk's Λ value: 0.94. Consequently, before applying the EPO algorithm, all the spectral data were smoothed by the Savitzky-Golay filter (window size of 11 nm), detrended applying a standard normal variate transformation, and finally a second order linear model was fitted returning the fitted residuals (Figure 6).

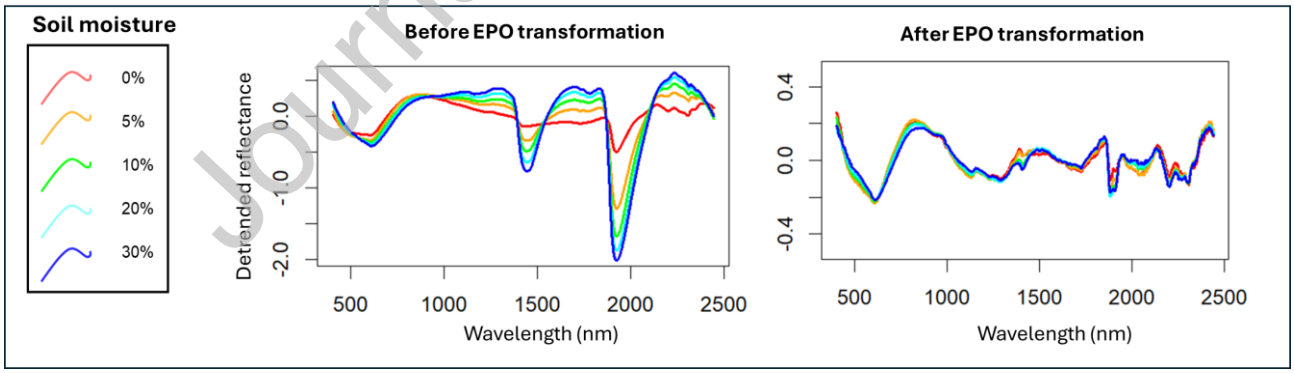

Figure 6: Example of spectra acquired at different soil moisture content on the same soil sample before and after the external parameter orthogonalization (EPO) transformation.

The Wilcoxon test demonstrated the average value of the standard deviation after EPO transformation ($\overline{SD}_{post}$= 0.03) is significant lower than the average value measured before the correction ($\overline{SD}_{pre}$= 0.16) (p<0.01).

Journal Pre-proof

SOC prediction

The Random Forest models trained on the LUCAS and GEO-CRADLE spectral libraries (under DS conditions) were applied to in-field spectra collected under LS and HS pre-treatments and those scanned under laboratory conditions (DS). The statistics presented in Table 5 for Scenario R, which uses raw spectra without any spectral transformation, revealed poor performance (with RPD close to 1). Scenario H, where spectra were harmonized using ISS, generally showed slightly poorer accuracy compared to Scenario R, while significant improvements were observed with the application of the EPO transformation (Scenario E). The best prediction performance was achieved in Scenario HE using the LUCAS SSL dataset, DS spectra, and 100 samples for calibration (RPD = 1.65; Table 5). The same level of accuracy (RPD = 1.66) was also obtained under Scenario HE with HS spectra and 500 samples for generating the training model. For the GEO-CRADLE models, Scenario HE delivered the highest accuracy as well, with RPD values of 1.35 for LS spectra and 1.33 for HS spectra (Table 6), but with over three times more calibration samples used to train the model as compared to LUCAS models.

Table 5: Validation results of the soil organic carbon (SOC) prediction models using the LUCAS spectral library. The statistics used are: R^2 = coefficient of determination; RMSE = root mean square error; RPD = ratio of performance to deviation. SSL = soil spectral libraries. DS = lab spectra acquired on dried and sieved soils, LS = field spectra acquired on lightly smoothed surfaces, HS = field spectra acquired on highly smoothed surfaces. Scenario R = Raw spectra without any transformation, Scenario H = harmonized spectra using the ISS, Scenario E: the EPO correction matrix was applied to R spectra, Scenario HE = the EPO correction matrix was applied to H spectra. N° Cal is the number of LUCAS soil samples used to train the SOC prediction model.

Scenario	**SSL**	**Soil pre-treatment**	**N° Cal**	**RMSE%**	**R^2**	**RPD**
R	LUCAS	DS	100	1.41	0.25	1.11
			200	1.38	0.28	1.13
			500	1.37	0.29	1.15
			1000	1.37	0.28	1.15
R	LUCAS	LS	100	4.05	0.05	0.39
			200	3.44	0.14	0.46
			500	3.50	0.26	0.45
			1000	3.73	0.18	0.42
R	LUCAS	HS	100	4.52	0.04	0.35
			200	3.60	0.10	0.44
			500	2.85	0.34	0.55
			1000	2.92	0.31	0.54
H	LUCAS	DS	100	1.67	0.17	0.94
			200	1.52	0.23	1.03

Journal Pre-proof

			500	1.52	0.23	1.03
			1000	1.50	0.24	1.04
H	LUCAS	LS	100	3.49	0.08	0.45
			200	3.51	0.16	0.45
			500	4.19	0.21	0.38
			1000	4.43	0.16	0.36
H	LUCAS	HS	100	5.00	0.02	0.31
			200	4.09	0.05	0.38
			500	3.52	0.26	0.44
			1000	3.58	0.25	0.44
E	LUCAS	DS	100	1.27	0.35	1.24
			200	1.21	0.43	1.3
			500	1.18	0.46	1.33
			1000	1.21	0.41	1.29
E	LUCAS	LS	100	1.55	0.24	1.01
			200	1.52	0.27	1.04
			500	1.5	0.34	1.05
			1000	1.51	0.34	1.04
E	LUCAS	HS	100	1.15	0.54	1.37
			200	1.16	0.56	1.35
			500	1.22	0.56	1.28
			1000	1.25	0.59	1.25
HE	LUCAS	DS	100	0.95	0.67	1.65
			200	0.96	0.66	1.62
			500	1.00	0.64	1.56
			1000	0.98	0.64	1.60
HE	LUCAS	LS	100	1.31	0.45	1.20
			200	1.33	0.42	1.18
			500	1.33	0.44	1.18
			1000	1.34	0.47	1.18
HE	LUCAS	HS	100	1.04	0.63	1.51

Journal Pre-proof

200	0.98	0.67	1.59
500	**0.94**	**0.70**	**1.66**
1000	0.97	0.71	1.62

Table 6: Validation results of the soil organic carbon (SOC) prediction models using the GEO-CRADLE spectral library. The statistics used are: R^2 = coefficient of determination; RMSE = root mean square error; RPD: ratio of performance to deviation. SSL = soil spectral libraries. DS =lab spectra acquired on dried and sieved soils, LS = field spectra acquired on lightly smoothed surfaces, HS = field spectra acquired on highly smoothed surfaces. Scenario H =: harmonized spectra using the ISS, Scenario HE = the EPO correction matrix was applied to H spectra. N° Cal is the number of GEO-CRADLE soil samples used to train the SOC prediction model.

Scenario	**SSL**	**Soil pre-treatment**	**N° Cal**	**RMSE%**	**R^2**	**RPD**
H	GEO-CRADLE	DS	1621	1.69	0.44	0.92
H	GEO-CRADLE	LS	1621	2.88	0.35	0.54
H	GEO-CRADLE	HS	1621	1.76	0.55	0.89
HE	GEO-CRADLE	DS	1621	1.16	0.64	1.34
HE	GEO-CRADLE	LS	1621	1.16	0.68	1.35
HE	GEO-CRADLE	HS	1621	1.17	0.66	1.31

In Figure 7, we show the scatterplots related to the best performing prediction models using LUCAS and GEO-CRADLE SSLs. The figure highlights the better correlation between observed and predicted SOC values obtained through LUCAS SSL (Figure 7a and b) as compared to those achieved using GEO-CRADLE (Figure 7c and d), in particular for OC-rich soils (SOC >3%) and for those collected in Switzerland.

Journal Pre-proof

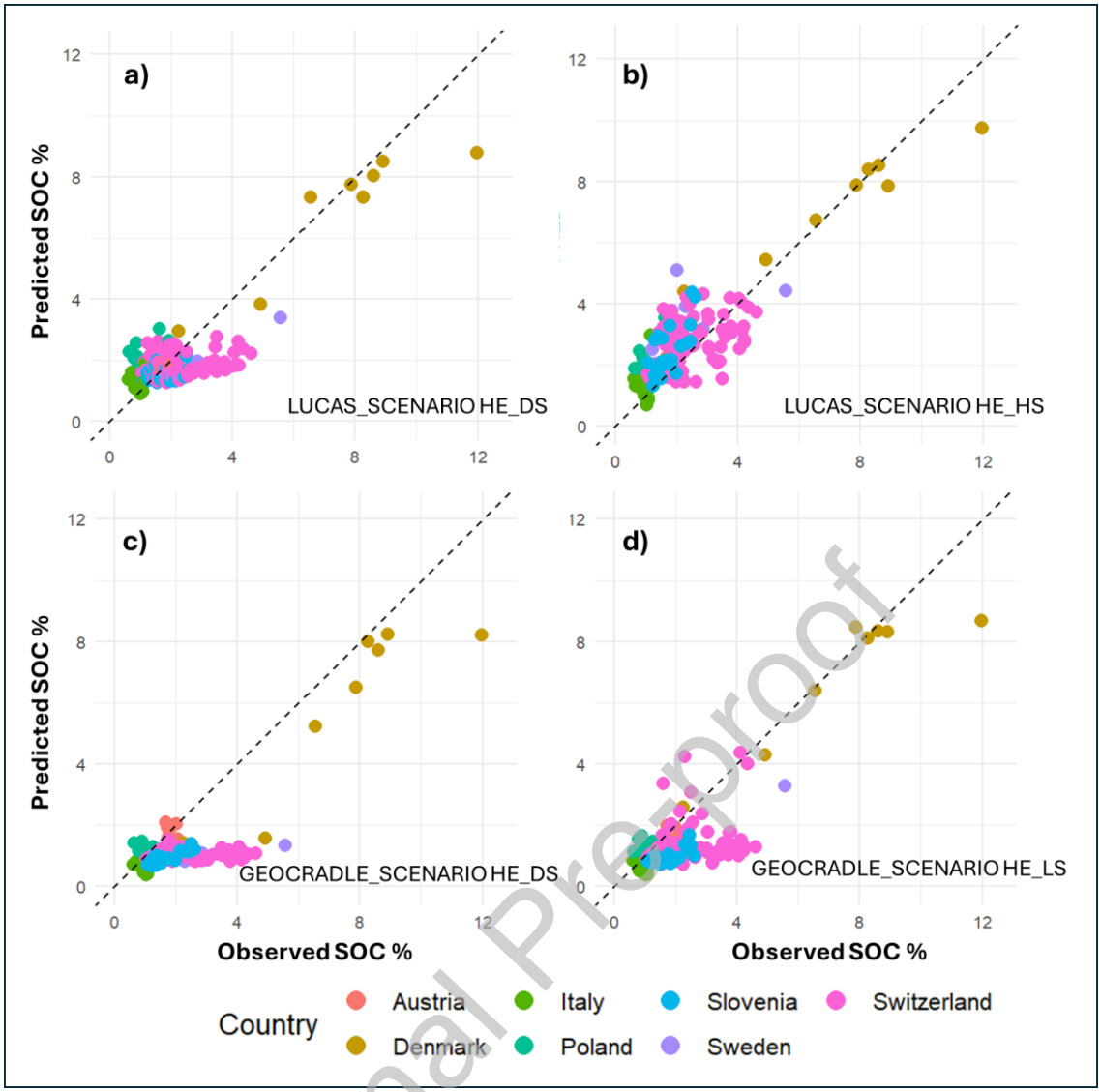

Figure 7: Scatterplots of the validation results using different soil spectral libraries and acquisition scenarios. ***HE_DS****: spectra acquired under laboratory conditions on dried and sieved samples (a, c), then harmonized using ISS correction and adjusted with the EPO transformation.* ***HE_HS****: spectra acquired in the field after highly smoothing the soil surface, then harmonized using ISS correction and adjusted with the EPO transformation (b).* ***HE_LS****: spectra acquired in the field after lightly smoothing the soil surface, then harmonized using ISS correction and adjusted with the EPO transformation (d).*

Journal Pre-proof

The relationship between the normalized error (NE) and the observed SOC values is positive and significant ($p<0.05$ according to the Spearman's rank correlation coefficient) and showed a similar logarithmical trend for all the models (Figure 8): we observed high and negative values (overestimation) for the lowest observed values, that correspond to Italian and Polish samples, and increasing NE values up to around 5% of SOC, after that the NE is generally very close to 0. Comparing the difference between LUCAS and GEO-CRADLE models in terms of NE values we observed the negative values were lower for GEO-CRADLE models, however within the SOC range where we can find most of the soil samples (1 – 3 %), the NE is closer to 0 for the LUCAS models (Figure 8a and b).

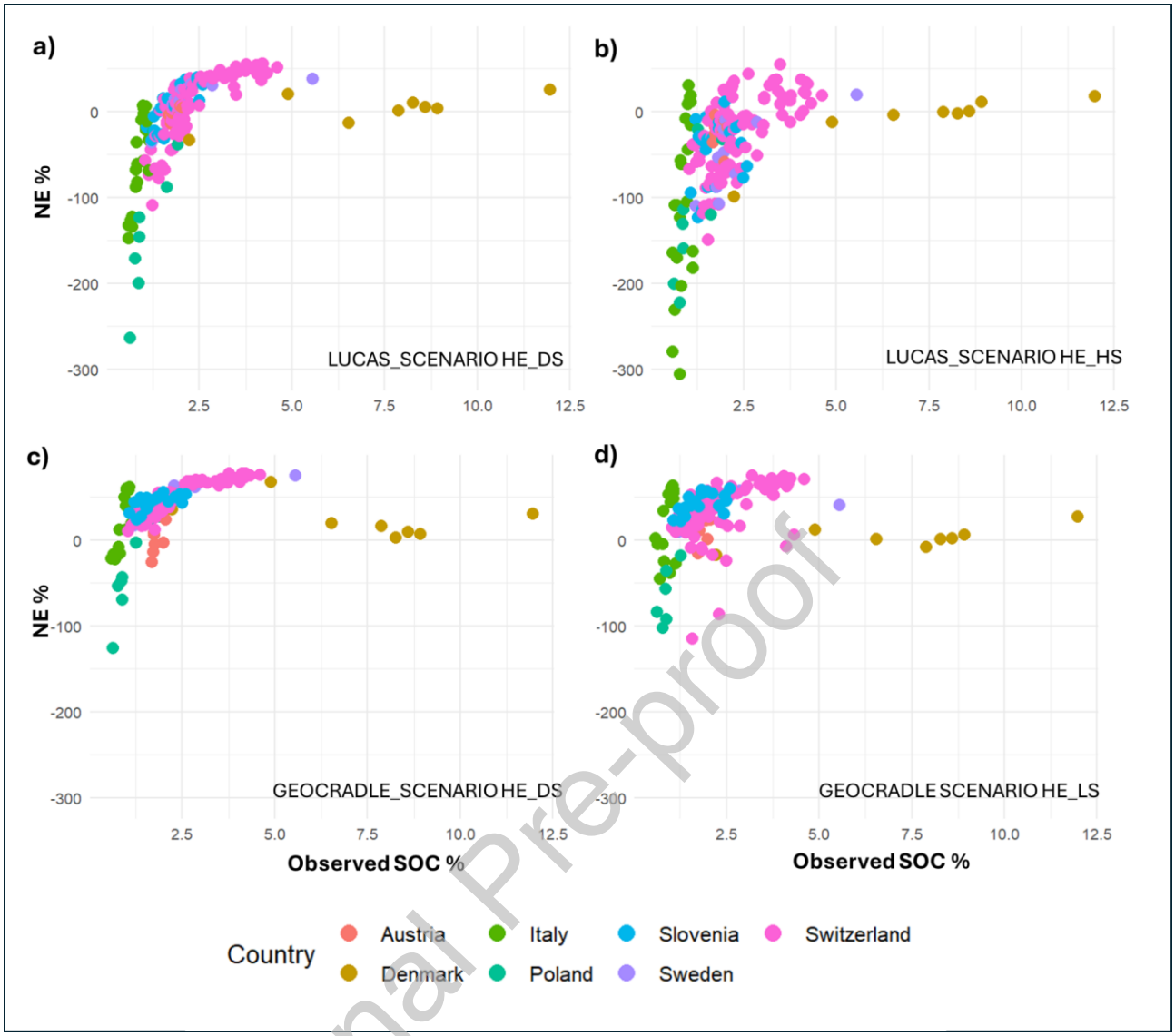

Figure 8: Normalized error (NE) as related to the observed soil organic carbon value for each soil sample according to different scenarios.

The absolute value of NE ($|NE|$) has a weak, though significant, inverse correlation with SM (-0.40; p<0.05). Considering the average $|NE|$ for country, the only significant differences (p<0.05) exists between the Polish dataset (mean $|NE| = 124\%$) and Danish (mean $|NE|$ = 18%), Austrian (mean $|NE|$ = 21%) and Swiss (mean $|NE|$ = 38%) datasets according to Dunn's test and Bonferroni method to adjust p values for multiple comparisons.

Discussion

4.1 Data harmonization and mitigation of moisture effects

In general, robust prediction models require that the training populations are representative and share key characteristics with the prediction samples, such as particle and aggregate size distribution, sample preparation methods and spectrometer type. When a training dataset consists of laboratory-prepared samples, but the prediction targets are field samples, discrepancies arise due to differences in measurement and sample conditions. This mismatch can make field predictions

difficult or even impossible using laboratory-based SSLs for calibration. This challenge was evident in this study, as the R-scenario consistently underperformed (Table 5 and 6; Figure 9).

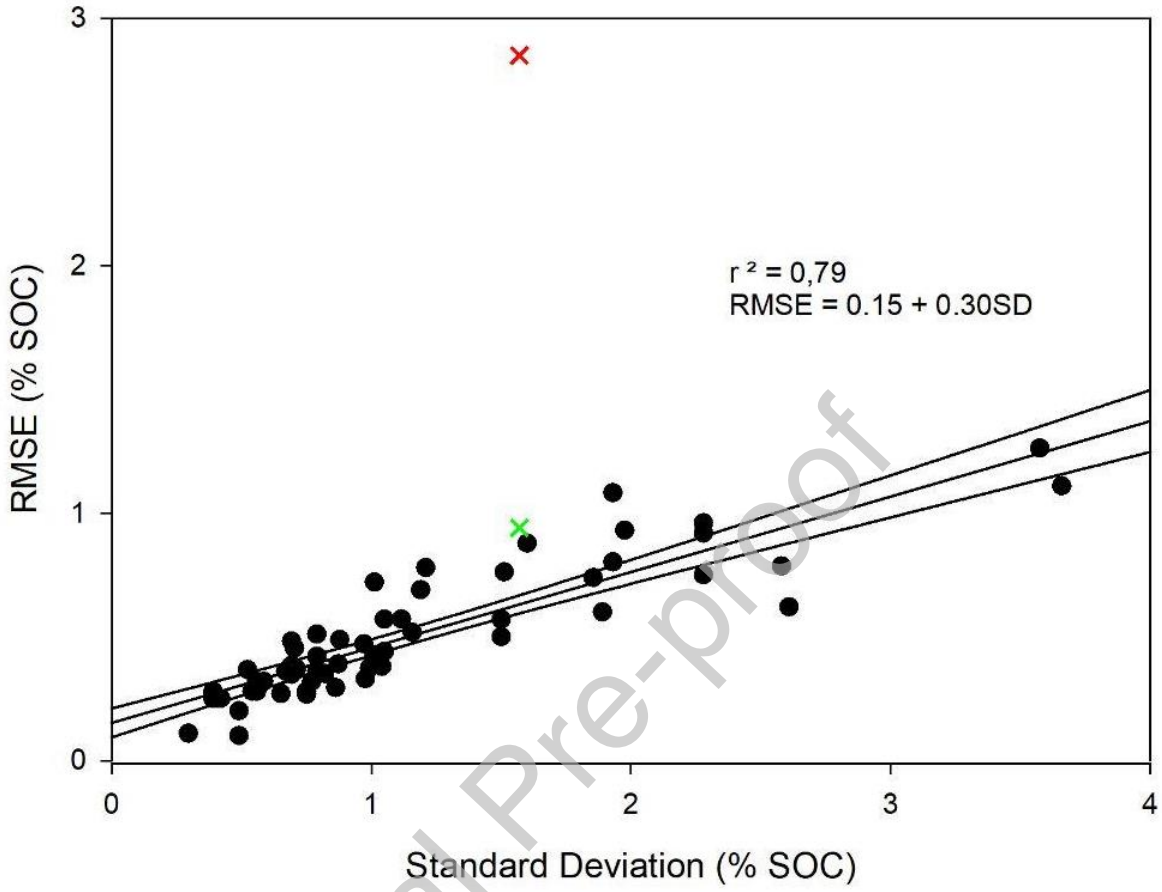

*Figure 9: The **R_HS** (red cross) and **HE_HS** (green cross) results from Table 5 projected on the relationship between RMSE and SD in 58 observations from 38 studies in the last 25 years. The regression line and 95 % confidence interval are indicated. The calibration methods used are PLS and PCR, Machine learning, and spectrally localized selections. Only studies from regional scale and up, without the influence of organic soils and where both RMSE and SD could be extracted are included [2,7,29,44–77]*

Journal Pre-proof

However, the integration of both ISS and EPO on field spectra from highly smoothed soil surfaces (HS) in the field yielded the highest accuracy and thus represents best practice for estimating organic carbon from laboratory-based calibration using both LUCAS and GEO-CRADLE SSLs (Table 5 and 6). This dual correction approach not only improved SOC prediction accuracy to match that of the corresponding DS procedure but also aligned model performance with expectations arising from a wide range of previous laboratory studies (Figure 9) or studies using stratification approaches to align different databases representing different methods and data quality [31]. This is despite our spectral acquisition being performed under diverse field conditions with diverse instrumentation that in turn was different from that used for SSL development. Notably, studies summarized in Figure 9 are all entirely laboratory based, and with one exception [2] based on spectra originating from the same instrument and laboratory conditions with data sets split into calibration and validation sets. This highlights the efficiency of the ISS+EPO protocol (HE

scenario) in mitigating discrepancies from disparate measurement systems and the combination of heterogeneous data sets. The heterogeneous scenario represented by the present study resembles what could be expected from in practice implementation. Additionally, the study included disperse reference methods for SOC which, from the perspective of wet chemical analysis, is known to involve considerable uncertainty and divergence [38].

Globally, interest in using SSLs is increasing and various initiatives are put forward for developing open-access SSLs [78]. In our study the LUCAS SSL performed slightly better than GEO-CRADLE for the ISS+EPO corrected spectra. This is not surprising as the LUCAS SSL is substantially larger, more diverse and represents the entire Europe and thus geographically also the origin of our field samples. GEO-CRADLE is dominated by samples from the Mediterranean and Eastern Europe. Interestingly, GEO-CRADLE did not perform substantially worse than LUCAS, despite none of the sample locations in the present study being geographically represented. Instead, SOC-content appears to be a regulatory factor (Figure 8). Similar findings were reported by [79], who found that geographical origin, or climate, or soil type, are not necessarily the drivers for successful training sample selection, but rather the consistency in the relationship between studied soil property and soil spectra.

The most effective methods for identifying spectral similarities related to soil properties are based on Mahalanobis distance [63], therefore in this study we implemented a training set selection approach for the LUCAS SSL based entirely on spectral similarity using Mahalanobis distance that allowed for a reduction of the dataset size without compromising the model's accuracy. In fact, the localized selection of 100 samples showed very similar statistics to those obtained with 200-1000 samples (Table 5). Using 1000 samples never showed the best result indicating the importance of selecting a representative training set. Thus, the selection of a spectrally local modeling set for reducing the required number of training samples is effective for both processing time and for extracting relevant information out from the SSL. These insights support that global machine learning models trained on large SSLs are rarely optimal in relation to local models [80].

In theory, ISS is limited to correcting systematic biases, such as those arising from differences in spectrometer types or calibration protocols, and alone, it cannot correct for differences in moisture content or aggregation stages in the field [27], while EPO corrects for non-systematic effects like soil moisture and particle size variation [14]. In our study, predictions from field spectra without any correction, and predictions with only ISS correction failed, while spectra from samples dried and sieved (DS) in the laboratory reached RPDs slightly over 1 without any corrections. Unexpectedly, ISS correction did not improve predictions from DS spectra either, despite moisture and particle size effects should be largely reduced by drying and sieving in the laboratory. EPO correction, on the other hand, did improve prediction results on its own from both laboratory and field spectra, especially from highly smoothed surface field spectra. Interestingly, ISS combined with EPO improved results further and both laboratory and highly smoothed surface spectra reached similar accuracy (Table 5 and 6). These results suggest that ISS is sensitive to disturbances it cannot correct for. However, when examining the results obtained using the GEO-CRADLE dataset, the differences in SOC prediction accuracy between scenarios H and HE (Table 6) appear less pronounced than those observed with the LUCAS dataset (Table 5). This suggests that the ISS algorithm may have performed more effectively for GEO-CRADLE than for LUCAS. A possible explanation lies in the type of instruments used to build the two spectral libraries. For GEO-CRADLE, field spectra were collected using two portable spectrometers operating in direct reflectance mode, similar to the method used for field measurements in our study. In contrast,

LUCAS samples were scanned using a benchtop instrument (FOSS XDS) that acquires spectra in diffuse reflectance using an integrated dome (S1). This fundamental difference in spectral acquisition methods likely introduced a greater heterogeneity between the LUCAS SSL and our field dataset, thereby reducing the effectiveness of ISS in this case. It also shows that EPO corrected for other disturbances than the moisture variability the EPO calibration was set up for. Presumably those related to instrument differences, laboratory conditions, and surface characteristics.

Unlike previous studies that reported reduced prediction accuracy in wetter soils (e.g. [40,81]), our application of EPO removed any positive relation between soil moisture and prediction errors. In fact, a slight negative correlation was observed. This suggests that EPO effectively mitigates the typical decline in accuracy associated with higher moisture levels.

The EPO matrix used in this study was developed using an independent spectral dataset. Therefore, we propose that it could be effectively applied to remove the influence of soil moisture from other field-acquired spectral datasets, provided they have similar soil organic carbon (SOC) content, moisture ranges, and soil types to those used in this study. In particular, the harmonization sequence developed in this study (ISS+EPO) could be applied to other soil attributes analyzed using wet laboratory methods with established consensus protocols and known uncertainties, broadening the applicability of spectroscopy in soil analysis.

4.2 Field-spectroscopy procedures and mitigation of roughness effects

The uncertainty of the spectral measurements, calculated as the average standard deviation (σ_p) of repeated scans at the same sampling point, did not show significant differences between spectra acquired in the laboratory (DS) and those collected in the field after highly smoothing the soil surface (HS; figure 5). This indicates that the mechanical smoothing procedure applied just before the field scans was effective in reducing noise and artefacts not related to soil properties. The lightly smoothing procedure (LS), resulted in a significantly higher σ_p compared to the HS procedure. Presumably this difference was primarily due to surface roughness as soil moisture conditions were similar. This demonstrates that the additional flattening step, involving soil compression, effectively reduced surface irregularities and allowed a more homogenous sample structure resulting in higher quality spectra. The HS procedure in combination with ISS+EPO correction led to an increase in SOC prediction performances with the LUCAS SSL (Table 5). The positive effect of this procedure in the field was not surprising as EPO was not systematically calibrated for the type of structure variability that can be expected in the field. The influence of structure on spectral quality is well established [82] as a coarse structure causes both scatter and shadow effects, and decreased signal to noise ratios [83].

4.3 Future Directions

VNIR-SWIR spectroscopy is on track to become a standard method for in-field agricultural soil analysis [9,84]. This study suggests a potential to reach an accuracy for field spectra in line with what is achieved in studies purely based on laboratory spectra. In addition, this study largely resembles real case situations, with a range of operators sampling in the field and, although using a common protocol, with a diversity of spectrophotometers and field accessories, resulting in data sets totally independent from each other in every aspect. This should be a strong incentive to focus future research and development efforts to refine and develop approaches that can ensure reliable and robust results from field spectroscopy and still take advantage of the efficiency of using existing laboratory-based large and robust SSLs. We also find it important that aspects emerging

from the heterogeneity and independence of real case scenarios are accounted for in future studies. This should include field sampling procedures. The apparent requirement of some degree of soil preparation in the field is a draw-back that presumably can be partly mitigated by automated sampling platforms. Removing non-soil contamination of the surface can hardly be avoided, but we should not exclude the possibility to manage structure issues mathematically or by improved probe geometry.

To harmonize the reference method for SOC will not change predictions but can be assumed to improve validation results and prevent undermining reliability. It could be expected that the harmonization of instrumentation would further improve results in addition to the ISS correction.

However, the widespread use of a universal SSL like LUCAS is unlikely to be feasible, as field instrumentation, like in this study, will inevitably vary between operators and evolve over time. Therefore, procedures and algorithms for aligning datasets and instruments will remain crucial.

In future research, the instance-based transfer learning approaches suggested by [79] can be tested for localization in combination with ISS + EPO. This would be especially feasible for monitoring small scale variation, like in fields, farms or watersheds. This involves measuring the studied variable, such as SOC, for a few samples in the field along with field spectra to extract the most relevant information from the large SSLs based on the relationship structure between spectra and the studied soil variable.

We systematically calibrated EPO for moisture effects, but found that it also corrected for other but not fully identified artefacts, as well as for dried and sieved soil. Possibly there is room for further improvements of the EPO calibration by establishing mechanisms behind these corrections.

Conclusions

Our approach demonstrates that in-field spectroscopy following a standardized protocol for reducing surface roughness and contamination, such as described in [35], can aid in the collection of good enough spectra and facilitate a rapid and low-cost assessment of SOC. Together with an internal soil standard for alignment of spectral data of dispersed origin, and EPO correction for artefacts like the influence of moisture, field spectra can be used together with existing laboratory based soil spectral libraries. This enables efficient large-scale employment of in field soil spectroscopy for a number of applications, such as soil monitoring and high resolution soil mapping for adoption in the study and management of agriculture and environmental systems.

This study further demonstrates that highly heterogeneous data collected by a range of operators using different instrumentation in terms of spectrophotometers and field accessories, and varying sample conditions in both field and laboratory environments can be successfully harmonized. EPO in addition to ISS was to a very large extent able to overcome this heterogeneity.

For our heterogenous data the ISS had no effect on its own, for laboratory or field spectra alike, but results improved when EPO was also applied. This suggests that EPO has the potential to correct for a diversity of factors simultaneously and should be considered for heterogeneous data sets also when of altogether laboratory origin. Our approach opens up broad application potential, and could potentially be used for the estimation of other soil properties that exhibit spectral features within the VNIR-SWIR range and derived information. That calibrations can be conducted or adapted from existing SSLs, and that building geographically local calibration sets or extensive field spectral

SSL's is found superfluous is a game-changing insight that should make the difference for widespread adoption of field spectroscopy in agriculture.

References

[1] M. Nocita, A. Stevens, B. van Wesemael, M. Aitkenhead, M. Bachmann, B. Barthès, E. Ben Dor, D.J. Brown, M. Clairotte, A. Csorba, P. Dardenne, J.A.M. Demattê, V. Genot, C. Guerrero, M. Knadel, L. Montanarella, C. Noon, L. Ramirez-Lopez, J. Robertson, H. Sakai, J.M. Soriano-Disla, K.D. Shepherd, B. Stenberg, E.K. Towett, R. Vargas, J. Wetterlind, Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring, Advances in Agronomy 132 (2015) 139–159. https://doi.org/10.1016/BS.AGRON.2015.02.002.

[2] R.A. Viscarra Rossel, T. Behrens, E. Ben-Dor, D.J. Brown, J.A.M. Demattê, K.D. Shepherd, Z. Shi, B. Stenberg, A. Stevens, V. Adamchuk, H. Aïchi, B.G. Barthès, H.M. Bartholomeus, A.D. Bayer, M. Bernoux, K. Böttcher, L. Brodský, C.W. Du, A. Chappell, Y. Fouad, V. Genot, C. Gomez, S. Grunwald, A. Gubler, C. Guerrero, C.B. Hedley, M. Knadel, H.J.M. Morrás, M. Nocita, L. Ramirez-Lopez, P. Roudier, E.M.R. Campos, P. Sanborn, V.M. Sellitto, K.A. Sudduth, B.G. Rawlins, C. Walter, L.A. Winowiecki, S.Y. Hong, W. Ji, A global spectral library to characterize the world's soil, Earth Sci Rev 155 (2016) 198–230. https://doi.org/10.1016/J.EARSCIREV.2016.01.012.

[3] S.K. Behera, V.I. Adamchuk, A.K. Shukla, P.S. Pandey, P. Kumar, V. Shukla, C. Thiyagarajan, H.K. Rai, S. Hadole, A.K. Sachan, P. Singh, V. Trivedi, A. Mishra, N.P. Butail, P. Kumar, R. Prajapati, K. Tiwari, D. Suri, M. Sharma, The Scope for Using Proximal Soil Sensing by the Farmers of India, Sustainability 2022, Vol. 14, Page 8561 14 (2022) 8561. https://doi.org/10.3390/SU14148561.

[4] E. Vaudour, A. Gholizadeh, F. Castaldi, M. Saberioon, L. Borůvka, D. Urbina-Salazar, Y. Fouad, D. Arrouays, A.C. Richer-De-forges, J. Biney, J. Wetterlind, B. Van Wesemael, Satellite Imagery to Map Topsoil Organic Carbon Content over Cultivated Areas: An Overview, Remote Sensing 2022, Vol. 14, Page 2917 14 (2022) 2917. https://doi.org/10.3390/RS14122917.

[5] E. Ben-Dor, A. Banin, Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties, Soil Science Society of America Journal 59 (1995) 364–372. https://doi.org/10.2136/SSSAJ1995.03615995005900020014X.

[6] R.A. Viscarra Rossel, D.J.J. Walvoort, A.B. McBratney, L.J. Janik, J.O. Skjemstad, Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties, Geoderma 131 (2006) 59–75. https://doi.org/10.1016/J.GEODERMA.2005.03.007.

[7] B. Stenberg, Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon, Geoderma 158 (2010) 15–22. https://doi.org/10.1016/J.GEODERMA.2010.04.008.

[8] J. Liu, J. Xie, J. Han, H. Wang, J. Sun, R. Li, S. Li, Visible and near-infrared spectroscopy with chemometrics are able to predict soil physical and chemical properties, J Soils Sediments 20 (2020) 2749–2760. https://doi.org/10.1007/S11368-020-02623-1/METRICS.

[9] C. Piccini, K. Metzger, G. Debaene, B. Stenberg, S. Götzinger, L. Borůvka, T. Sandén, L. Bragazza, F. Liebisch, In-field soil spectroscopy in Vis–NIR range for fast and reliable soil analysis: A review, Eur J Soil Sci 75 (2024) e13481. https://doi.org/10.1111/EJSS.13481.

[10] E. Ben Dor, B. Efrati, O. Amir, N. Francos, J. Shepherd, V. Khosravi, A. Gholizadeh, A. Klement, L. Borůvka, A standard and protocol for in-situ measurement of surface soil reflectance, Geoderma 447 (2024) 116920. https://doi.org/10.1016/J.GEODERMA.2024.116920.

[11] E. Ben Dor, A. Granot, R. Wallach, N. Francos, D. Heller Pearlstein, B. Efrati, L. Borůvka, A. Gholizadeh, T. Schmid, Exploitation of the SoilPRO® (SP) apparatus to measure soil surface reflectance in the field: Five case studies, Geoderma 438 (2023) 116636. https://doi.org/10.1016/J.GEODERMA.2023.116636.

[12] N. Francos, E. Ben-Dor, A transfer function to predict soil surface reflectance from laboratory soil spectral libraries, Geoderma 405 (2022). https://doi.org/10.1016/J.GEODERMA.2021.115432.

[13] E.J. Milton, M.E. Schaepman, K. Anderson, M. Kneubühler, N. Fox, Progress in field spectroscopy, Remote Sens Environ 113 (2009) S92–S109. https://doi.org/10.1016/J.RSE.2007.08.001.

[14] M. Knadel, F. Castaldi, R. Barbetti, E. Ben-Dor, A. Gholizadeh, R. Lorenzetti, Mathematical techniques to remove moisture effects from visible–near-infrared–shortwave-infrared soil spectra—review, Appl Spectrosc Rev 58 (2022) 629–662. https://doi.org/10.1080/05704928.2022.2128365.

[15] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Correction to the Description of Standard Normal Variate (SNV) and De-Trend (DT) Transformations in Practical Spectroscopy with Applications in Food and Beverage Analysis—2nd Edition, J Near Infrared Spectrosc 1 (1993) 185–186. https://doi.org/10.1255/JNIRS.21.

[16] E. Ben Dor, N. Francos, Y. Ogen, A. Banin, Aggregate size distribution of arid and semiarid laboratory soils (<2 mm) as predicted by VIS-NIR-SWIR spectroscopy, Geoderma 416 (2022) 115819. https://doi.org/10.1016/J.GEODERMA.2022.115819.

[17] D.B. Lobell, G.P. Asner, Moisture Effects on Soil Reflectance, Soil Science Society of America Journal 66 (2002) 722–727. https://doi.org/10.2136/SSSAJ2002.7220.

[18] B. Minasny, A.B. McBratney, V. Bellon-Maurel, J.M. Roger, A. Gobrecht, L. Ferrand, S. Joalland, Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon, Geoderma 167–168 (2011) 118–124. https://doi.org/10.1016/J.GEODERMA.2011.09.008.

[19] F. Castaldi, A. Palombo, S. Pascucci, S. Pignatti, F. Santini, R. Casa, Reducing the influence of soil moisture on the estimation of clay from hyperspectral data: A case study using simulated PRISMA data, Remote Sens (Basel) 7 (2015). https://doi.org/10.3390/rs71115561.

Journal Pre-proof

[20] Y. Ogen, S. Faigenbaum-Golovin, A. Granot, Y. Shkolnisky, N. Goldshleger, E. Ben-Dor, Removing Moisture Effect on Soil Reflectance Properties: A Case Study of Clay Content Prediction, Pedosphere 29 (2019) 421–431. https://doi.org/10.1016/S1002-0160(19)60811-8.

[21] S. Chakraborty, B. Li, D.C. Weindorf, C.L.S. Morgan, External parameter orthogonalisation of Eastern European VisNIR-DRS soil spectra, Geoderma 337 (2019) 65–75. https://doi.org/10.1016/J.GEODERMA.2018.09.015.

[22] K. Metzger, F. Liebisch, J.M. Herrera, T. Guillaume, L. Bragazza, Prediction Accuracy of Soil Chemical Parameters by Field- and Laboratory-Obtained vis-NIR Spectra after External Parameter Orthogonalization, Sensors 2024, Vol. 24, Page 3556 24 (2024) 3556. https://doi.org/10.3390/S24113556.

[23] J.P. Ackerson, J.A.M. Demattê, C.L.S. Morgan, Predicting clay content on field-moist intact tropical soils using a dried, ground VisNIR library with external parameter orthogonalization, Geoderma 259–260 (2015) 196–204. https://doi.org/10.1016/J.GEODERMA.2015.06.002.

[24] N.K. Wijewardane, S. Hetrick, J. Ackerson, C.L.S. Morgan, Y. Ge, VisNIR integrated multi-sensing penetrometer for in situ high-resolution vertical soil sensing, Soil Tillage Res 199 (2020) 104604. https://doi.org/10.1016/J.STILL.2020.104604.

[25] M.O.F. Murad, E.J. Jones, B. Minasny, A.B. McBratney, N. Wijewardane, Y. Ge, Assessing a VisNIR penetrometer system for in-situ estimation of soil organic carbon under variable soil moisture conditions, Biosyst Eng 224 (2022) 197–212. https://doi.org/10.1016/J.BIOSYSTEMSENG.2022.10.011.

[26] E. Ben Dor, C. Ong, I.C. Lau, Reflectance measurements of soils in the laboratory: Standards and protocols, Geoderma 245–246 (2015) 112–124. https://doi.org/10.1016/J.GEODERMA.2015.01.002.

[27] V. Kopačková, E. Ben-Dor, Normalizing reflectance from different spectrometers and protocols with an internal soil standard, Int J Remote Sens 37 (2016) 1276–1290. https://doi.org/10.1080/01431161.2016.1148291.

[28] G. Crucil, F. Castaldi, E. Aldana-Jague, B. van Wesemael, A. Macdonald, K. Oost, Assessing the performance of UAS-Compatible multispectral and hyperspectral sensors for soil organic carbon prediction, Sustainability (Switzerland) 11 (2019). https://doi.org/10.3390/su11071889.

[29] M. Nocita, A. Stevens, G. Toth, P. Panagos, B. van Wesemael, L. Montanarella, Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach, Soil Biol Biochem 68 (2014) 337–347. https://doi.org/10.1016/J.SOILBIO.2013.10.022.

[30] F. Castaldi, S. Chabrillat, C. Chartin, V. Genot, A.R. Jones, B. van Wesemael, Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database, Eur J Soil Sci 69 (2018). https://doi.org/10.1111/ejss.12553.

[31] O. Yuzugullu, N. Fajraoui, A. Don, F. Liebisch, Satellite-based soil organic carbon mapping on European soils using available datasets and support sampling, Science of Remote Sensing 9 (2024) 100118. https://doi.org/10.1016/J.SRS.2024.100118.

[32] R. Lorenzetti, F. Castaldi, C.L. Fondon, L. Boruvka, K. Metzger, E. Ben-Dor, F. Van Egmond, R. Barbetti, M. Fantappie, G. Debaene, K. Klumpp, F. Liebisch, A. Gholizadeh, B. Stenberg, M. Knadel, The contribution of the European Project Probefield to in-field use of proximal soil sensors, 2023 IEEE International Workshop on Metrology for Agriculture and Forestry, MetroAgriFor 2023 - Proceedings (2023) 727–731. https://doi.org/10.1109/METROAGRIFOR58484.2023.10424081.

[33] A. Jones, O. Fernandez-Ugalde, S. Scarpa. LUCAS 2015 Topsoil Survey. Presentation of dataset and results, EUR 30332 EN, Publications Office of the European Union: Luxembourg. 2020, ISBN 978-92-76-21080-1, doi:10.2760/616084, JRC121325

[34] A. Granot, E. Granot, Soil spectral measurements in the field: problems and solutions in light of the GEO-CARDEL project, Https://Doi.Org/10.1117/12.2279661 10444 (2017) 340–346. https://doi.org/10.1117/12.2279661.

[35] B. Stenberg, T. Koganti, F. Castaldi, K. Metzger, G. Buttafuoco, F. van Egmond, J.A. Cayuela, L. Borůvka, G. Debaene, F. Liebisch, T. Sandén, D5.1 ProbeField: Best Practice Protocol for Field Spectroscopy and Assessment by Soil Spectral Library Based Calibrations, (n.d.). https://doi.org/10.5281/ZENODO.14150972.

[36] IUSS Working Group W.R.B. World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports No. 106, Rome: FAO, 2015.

[37] B.L. Welch, The Generalization of `Student's' Problem when Several Different Population Variances are Involved, Biometrika 34 (1947) 28. https://doi.org/10.2307/2332510.

[38] J. Meersmans, B. Van Wesemael, M. Van Molle, Determining soil organic carbon for agricultural soils: a comparison between the Walkley & Black and the dry combustion methods (north Belgium), Soil Use Manag 25 (2009) 346–353. https://doi.org/10.1111/J.1475-2743.2009.00242.X.

[39] J.M. Roger, F. Chauchard, V. Bellon-Maurel, EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits, Chemometrics and Intelligent Laboratory Systems 66 (2003) 191–204. https://doi.org/10.1016/S0169-7439(03)00051-0.

[40] N.K. Wijewardane, Y. Ge, C.L.S. Morgan, Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization, Geoderma 267 (2016) 92–101. https://doi.org/10.1016/J.GEODERMA.2015.12.014.

[41] A. Stevens, L. Ramirez-Lopez. *An introduction to the prospectr package*. R package version 0.2.8. (2025)

[42] L. Breiman, Random forests, Mach Learn 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324/METRICS.

[43] K.J. Ward, S. Chabrillat, M. Brell, F. Castaldi, D. Spengler, S. Foerster, Mapping soil organic carbon for airborne and simulated enmap imagery using the lucas soil database and a local plsr, Remote Sens (Basel) 12 (2020). https://doi.org/10.3390/rs12203451.

Journal Pre-proof

[44] S.R. Araújo, J. Wetterlind, J.A.M. Demattê, B. Stenberg, Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques, Eur J Soil Sci 65 (2014) 718–729. https://doi.org/10.1111/EJSS.12165.

[45] N. Asgari, S. Ayoubi, J.A.M. Demattê, A.C. Dotto, Carbonates and organic matter in soils characterized by reflected energy from 350–25000 nm wavelength, J Mt Sci 17 (2020) 1636–1651. https://doi.org/10.1007/S11629-019-5789-9/METRICS.

[46] C.W. Chang, D.A. Laird, Near-infrared reflectance spectroscopic analysis of soil C and N, Soil Sci 167 (2002) 110–116. https://doi.org/10.1097/00010694-200202000-00003.

[47] C.-W. Chang, D. Laird, C. Hurburgh, Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties, (2005). https://dr.lib.iastate.edu/handle/20.500.12876/1596 (accessed July 1, 2025).

[48] M. Clairotte, C. Grinand, E. Kouakoua, A. Thébault, N.P.A. Saby, M. Bernoux, B.G. Barthès, National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy, Geoderma 276 (2016) 41–52. https://doi.org/10.1016/J.GEODERMA.2016.04.021.

[49] J.A.M. Demattê, A.C. Dotto, A.F.S. Paiva, M. V. Sato, R.S.D. Dalmolin, M. do S.B. de Araújo, E.B. da Silva, M.R. Nanni, A. ten Caten, N.C. Noronha, M.P.C. Lacerda, J.C. de Araújo Filho, R. Rizzo, H. Bellinaso, M.R. Francelino, C.E.G.R. Schaefer, L.E. Vicente, U.J. dos Santos, E. V. de Sá Barretto Sampaio, R.S.C. Menezes, J.J.L.L. de Souza, W.A.P. Abrahão, R.M. Coelho, C.R. Grego, J.L. Lani, A.R. Fernandes, D.A.M. Gonçalves, S.H.G. Silva, M.D. de Menezes, N. Curi, E.G. Couto, L.H.C. dos Anjos, M.B. Ceddia, E.F.M. Pinheiro, S. Grunwald, G.M. Vasques, J. Marques Júnior, A.J. da Silva, M.C. de V. Barreto, G.N. Nóbrega, M.Z. da Silva, S.F. de Souza, G.S. Valladares, J.H.M. Viana, F. da Silva Terra, I. Horák-Terra, P.R. Fiorio, R.C. da Silva, E.F. Frade Júnior, R.H.C. Lima, J.M.F. Alba, V.S. de Souza Junior, M.D.L.M.S. Brefin, M.D.L.P. Ruivo, T.O. Ferreira, M.A. Brait, N.R. Caetano, I. Bringhenti, W. de Sousa Mendes, J.L. Safanelli, C.C.B. Guimarães, R.R. Poppiel, A.B. e Souza, C.A. Quesada, H.T.Z. do Couto, The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges, Geoderma 354 (2019) 113793. https://doi.org/10.1016/J.GEODERMA.2019.05.043.

[50] B.W. Dunn, H.G. Beecher, G.D. Batten, S. Ciavarella, The potential of near-infrared reflectance spectroscopy for soil analysis - A case study from the Riverine Plain of south-eastern Australia, Aust J Exp Agric 42 (2002) 607–614. https://doi.org/10.1071/EA01172.

[51] G. Fystro, The prediction of C and N content and their potential mineralisation in heterogeneous soil samples using Vis-NIR spectroscopy and comparative methods, Plant Soil 246 (2002) 139–149. https://doi.org/10.1023/A:1020612319014/METRICS.

[52] F. Gogé, R. Joffre, C. Jolivet, I. Ross, L. Ranjard, Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database, Chemometrics and Intelligent Laboratory Systems 110 (2012) 168–176. https://doi.org/10.1016/J.CHEMOLAB.2011.11.003.

[53] K. Islam, B. Singh, A. McBratney, Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy, Australian Journal of Soil Research 41 (2003) 1101–1114. https://doi.org/10.1071/SR02137.

[54] M. Knadel, B. Stenberg, F. Deng, A. Thomsen, M.H. Greve, Comparing predictive abilities of three visible-near infrared spectrophotometers for soil organic carbon and clay determination, J Near Infrared Spectrosc 21 (2013) 67–80. https://doi.org/10.1255/JNIRS.1035.

[55] A. Lazaar, A.M. Mouazen, K. EL Hammouti, M. Fullen, B. Pradhan, M.S. Memon, K. Andich, A. Monir, 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 (2020) 195–204. https://doi.org/10.1016/J.ISWCR.2020.04.005.

[56] Y. Liu, Z. Shi, G. Zhang, Y. Chen, S. Li, Y. Hong, T. Shi, J. Wang, Y. Liu, Application of Spectrally Derived Soil Type as Ancillary Data to Improve the Estimation of Soil Organic Carbon by Using the Chinese Soil Vis-NIR Spectral Library, Remote Sensing 2018, Vol. 10, Page 1747 10 (2018) 1747. https://doi.org/10.3390/RS10111747.

[57] D.F. Malley, P.D. Martin, L.M. McClintock, L. Yesmin, R.G. Eilers, P. Haluschak. Feasibility of analysing archived Canadian prairie agricultural soils by near infrared reflectance spectroscopy. In: A.M.C. Davies, R. Giangiacomo (Eds.), Near Infrared Spectroscopy: Proceedings of the 9th International Conference. NIR Publications, Chichester, UK, (2000) 579-585.

[58] T. Miao, W. Ji, B. Li, X. Zhu, J. Yin, J. Yang, Y. Huang, Y. Cao, D. Yao, X. Kong, Advanced Soil Organic Matter Prediction with a Regional Soil NIR Spectral Library Using Long Short-Term Memory–Convolutional Neural Networks: A Case Study, Remote Sensing 2024, Vol. 16, Page 1256 16 (2024) 1256. https://doi.org/10.3390/RS16071256.

[59] A. Morón, D. Cozzolino, Application of near Infrared Reflectance Spectroscopy for the Analysis of Organic C, Total N and pH in Soils of Uruguay, Journal of Near Infrared Spectroscopy, Vol. 10, Issue 3, Pp. 215-221 10 (2002) 215–221. https://opg.optica.org/abstract.cfm?uri=jnirs-10-3-215 (accessed July 1, 2025).

[60] M. Nocita, L. Kooistra, M. Bachmann, A. Müller, M. Powell, S. Weel, Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa, Geoderma 167–168 (2011) 295–302. https://doi.org/10.1016/J.GEODERMA.2011.09.018.

[61] M. Nocita, A. Stevens, C. Noon, B. Van Wesemael, Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy, Geoderma 199 (2013) 37–42. https://doi.org/10.1016/J.GEODERMA.2012.07.020.

[62] É.F.M. Pinheiro, M.B. Ceddia, C.M. Clingensmith, S. Grunwald, G.M. Vasques, Prediction of Soil Physical and Chemical Properties by Visible and Near-Infrared Diffuse Reflectance Spectroscopy in the Central Amazon, Remote Sensing 2017, Vol. 9, Page 293 9 (2017) 293. https://doi.org/10.3390/RS9040293.

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

[64] J.B. Sankey, D.J. Brown, M.L. Bernard, R.L. Lawrence, Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C, Geoderma 148 (2008) 149–158. https://doi.org/10.1016/J.GEODERMA.2008.09.019.

[65] Z. Shi, W. Ji, R.A. Viscarra Rossel, S. Chen, Y. Zhou, Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis-NIR spectral library, Eur J Soil Sci 66 (2015) 679–687. https://doi.org/10.1111/EJSS.12272;JOURNAL:JOURNAL:13652389A;PAGE:STRING:ARTICLE/CHAPTER.

[66] L.K. Sørensen, S. Dalsgaard, Determination of Clay and Other Soil Properties by Near Infrared Spectroscopy, Soil Science Society of America Journal 69 (2005) 159–167. https://doi.org/10.2136/SSSAJ2005.0159.

[67] B. Stenberg, A. Jonsson, A., T. Börjesson, Near infrared technology for soil analysis with implications for precision agriculture. In: A. Davies, R. Cho (eds.), Near Infrared Spectroscopy: Proceedings of the 10th International Conference. NIR publications, Chichester, UK. (2002) 279-284. https://www.researchgate.net/publication/290839451 (accessed July 2, 2025).

[68] A. Stevens, M. Nocita, G. Tóth, L. Montanarella, B. van Wesemael, Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy, PLoS One 8 (2013) e66409. https://doi.org/10.1371/JOURNAL.PONE.0066409.

[69] Y. Tekin, Z. Tumsavas, A.M. Mouazen, Effect of Moisture Content on Prediction of Organic Carbon and pH Using Visible and Near-Infrared Spectroscopy, Soil Science Society of America Journal 76 (2012) 188–198. https://doi.org/10.2136/SSSAJ2011.0021.

[70] M. Todorova, S. Atanassova, R. Ilieva. Determination of soil organic carbon using near-infrared spectroscopy. Agricultural Science and Technology 1(2) (2009) 45-50.

[71] T. Udelhoven, C. Emmerling, T. Jarmer, Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study, Plant Soil 251 (2003) 319–329. https://doi.org/10.1023/A:1023008322682/METRICS.

[72] R.A.V. Rossel, T. Behrens, Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma 158 (2010) 46–54. https://doi.org/10.1016/J.GEODERMA.2009.12.025.

[73] R.A. Viscarra Rossel, W.S. Hicks, Soil organic carbon and its fractions estimated by visible-near infrared transfer functions, Eur J Soil Sci 66 (2015) 438–450. https://doi.org/10.1111/EJSS.12237;WGROUP:STRING:PUBLICATION.

[74] M. Vohland, J. Besold, J. Hill, H.C. Fründ, Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy, Geoderma 166 (2011) 198–205. https://doi.org/10.1016/J.GEODERMA.2011.08.001.

[75] Z. Wang, S. Chen, R. Lu, X. Zhang, Y. Ma, Z. Shi, Non-linear memory-based learning for predicting soil properties using a regional vis-NIR spectral library, Geoderma 441 (2024) 116752. https://doi.org/10.1016/J.GEODERMA.2023.116752.

[76] Z. Wang, J. Ding, Z. Zhang, Estimation of Soil Organic Matter in Arid Zones with Coupled Environmental Variables and Spectral Features, Sensors 2022, Vol. 22, Page 1194 22 (2022) 1194. https://doi.org/10.3390/S22031194.

[77] M. Yang, D. Xu, S. Chen, H. Li, Z. Shi, Evaluation of Machine Learning Approaches to Predict Soil Organic Matter and pH Using vis-NIR Spectra, Sensors 2019, Vol. 19, Page 263 19 (2019) 263. https://doi.org/10.3390/S19020263.

[78] J.L. Safanelli Id, T. Hengl Id, L.L. Parente, R. Minarik, D.E. Bloom Id, K. Todd-Brown Id, A. Gholizadeh, W. De Sousa, M. Id, J. Sanderman, Open Soil Spectral Library (OSSL): Building reproducible soil calibration models through open development and community engagement, PLoS One 20 (2025) e0296545. https://doi.org/10.1371/JOURNAL.PONE.0296545.

[79] R.A. Viscarra Rossel, Z. Shen, L. Ramirez Lopez, T. Behrens, Z. Shi, J. Wetterlind, K.A. Sudduth, B. Stenberg, C. Guerrero, A. Gholizadeh, E. Ben-Dor, M. St Luce, C. Orellano, An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning, Earth Sci Rev 254 (2024) 104797. https://doi.org/10.1016/J.EARSCIREV.2024.104797.

[80] A.L. Kock, P.D. Ramphisa-Nghondzweni, G. Van Zijl, Development of soil spectroscopy models for the Western Highveld region, South Africa: Why do we need local data?, Eur J Soil Sci 75 (2024) e70014. https://doi.org/10.1111/EJSS.70014.

[81] B. Minasny, A.B. McBratney, L. Pichon, W. Sun, M.G. Short, Evaluating near infrared spectroscopy for field prediction of soil properties, Soil Research 47 (2009) 664–673. https://doi.org/10.1071/SR09005.

[82] H. Croft, K. Anderson, N.J. Kuhn, Characterizing soil surface roughness using a combined structural and spectral approach, Eur J Soil Sci 60 (2009) 431–442. https://doi.org/10.1111/J.1365-2389.2009.01129.X;PAGE:STRING:ARTICLE/CHAPTER.

[83] D. Dahm, K. Dahm, Interpreting Diffuse Reflectance and Transmittance: A Theoretical Introduction to Absorption Spectroscopy of Scattering Materials, Interpreting Diffuse Reflectance and Transmittance: A Theoretical Introduction to Absorption Spectroscopy of Scattering Materials (2007). https://doi.org/10.1255/978-1-901019-05-6.

[84] E. Najdenko, F. Lorenz, K. Dittert, H.W. Olfs, Rapid in-field soil analysis of plant-available nutrients and pH for precision agriculture—a review, Precision Agriculture 2024 25:6 25 (2024) 3189–3218. https://doi.org/10.1007/S11119-024-10181-6.

Ethical Statement

The tests and procedures described in this manuscript did not involve human participants or animals, and therefore did not require ethical approval. No ethical issues are associated with the methods used, and no specific permits or authorizations were necessary for the activities described.

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof