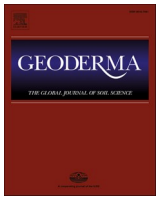
Geoderma 452 (2024) 117116

Contents lists available at ScienceDirect

Geoderma

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



A Moroccan soil spectral library use framework for improving soil property prediction: Evaluating a geostatistical approach

Tadesse Gashaw Asrat [a,*], Timo Breure [a,e], Ruben Sakrabani [a], Ron Corstanje [a], Kirsty L. Hassall [b], Abdellah Hamma [d], Fassil Kebede [c], Stephan M. Haefele [b]

[a] Cranfield University, Cranfield, UK
[b] Rothamsted Research, Sustainable Soils and Crops, Harpenden, UK
[c] Centre for Soil and Fertilizer Research in Africa, College of Agriculture and Environmental Science, Mohammed VI Polytechnic University, Ben Guerir, Morocco
[d] BU Al Moutmir, College of Agriculture and Environmental Sciences, Mohammed VI Polytechnic University, Ben Guerir, Morocco
[e] Wageningen University & Research, Wageningen, Netherlands

ARTICLE INFO

Handling Editor: Jingyi Huang

Keywords:
Soil properties
IR spectroscopy
Spatial non-stationarity
Geostatistics
MBL
Covariate clustering
Morocco

ABSTRACT

A soil spectrum generated by any spectrometer requires a calibration model to estimate soil properties from it. To achieve best results, the assumption is that locally calibrated models offer more accurate predictions. However, achieving this higher accuracy comes with associated costs, complexity, and resource requirements, thus limiting widespread adoption. Furthermore, there is a lack of comprehensive frameworks for developing and utilizing soil spectral libraries (SSLs) to make predictions for specific samples. While calibration samples are necessary, there is the need to optimize SSL development through strategically determining the quantity, location, and timing of these samples based on the quality of the information in the library. This research aimed to develop a spatially optimized SSL and propose a use-framework tailored for predicting soil properties for a specific farmland context. Consequently, the Moroccan SSL (MSSL) was established utilizing a stratified spatially balanced sampling design, using six environmental covariates and FAO soil units. Subsequently, various criteria for calibration sample selection were explored, including a spatial autocorrelation of spectra principal component (PC) scores (spatial calibration sample selection), spectra similarity memory-based learner (MBL), and selection based on environmental covariate clustering. Twelve soil properties were used to evaluate these calibration sample selections to predict soil properties using the near infrared (NIR) and mid infrared (MIR) ranges. Among the methods assessed, we observed distinct precision improvements resulting from spatial sample selection and MBL compared to the use of the entire MSSL. Notably, the Lin's Concordance Correlation Coefficient (CCC) values using the spatial calibration sample selection was improved for Olsen extractable phosphorus (OlsenP) by 41.3% and Mehlich III extractable phosphorus (P_M3) by 8.5% for the MIR spectra and for CEC by 25.6%, pH by 13.0% and total nitrogen (Tot_N) by 10.6% for the NIR spectra in reference to use of the entire MSSL. Utilizing the spatial autocorrelation of the spectra PC scores proved beneficial in identifying appropriate calibration samples for a new sample location, thereby enhancing prediction performance comparable to, or surpassing that of the use of the entire MSSL. This study signifies notable advancement in crafting targeted models tailored for specific samples within a vast and diverse SSL.

1. Introduction

Proximal spectral signatures acquired from soil samples in the near infrared (NIR) to mid infrared (MIR) ranges of the electromagnetic spectrum need to be calibrated, mostly using conventional wet chemistry analysis. A soil spectral library (SSL), which is fundamentally a paired dataset consisting of soil spectral records and conventional wet chemistry data of the same samples, is built to calibrate the predictive models for estimating soil properties for new samples (Knadel et al., 2012; Dematte et al., 2019; Baumann et al., 2017; Summerauer et al., 2021). Based on this approach, soil spectroscopy has been evaluated for diverse applications with the aim of minimising the need for future expensive and time-consuming conventional soil analysis such as for digital soil mapping (Brodský et al., 2011), soil classification and

* Corresponding author.
E-mail address: tadesse-gashaw.asrat@cranfield.ac.uk (T.G. Asrat).

https://doi.org/10.1016/j.geoderma.2024.117116
Received 22 April 2024; Received in revised form 18 November 2024; Accepted 18 November 2024
Available online 24 November 2024

groupings (Viscarra Rossel et al., 2016; Zeng et al., 2017), and precision agriculture (Ng, et al., 2020; Breure et al., 2021; Asrat et al., 2023). However, for the practical application of soil spectroscopy, it is essential to establish a comprehensive reference pattern in the SSL database. Addressing this need involves ensuring that the calibration samples used in the models are the most representative of the criteria such as the target region, the spatial scale, sample conditions, the range of soil variability, and the spectral space in which the models are to be applied (Savvides et al., 2010; Grunwald et al., 2018).

Therefore, it could be possible to achieve the required high resolution soil information easily and more economically (although not equally well for all soil characteristics) using soil spectroscopy to support informed decisions and monitoring of sustainable soil functions and ecological services obtained from soil systems compared to the conventional soil analysis methods. However, to take the approach towards wider application, there are still concerns over the source of uncertainties related to SSL development and selection of appropriate calibration sample entries from a large SSL for new samples to be predicted (Ramirez-Lopez et al., 2013a; Sun et al., 2021). This is because the large SSL might possess non-informative samples (regarding the local spectral and soil variability) to develop a specific predictive model for a specific farmland. The routine evaluation of soil predictions using soil spectroscopy is to split the data into calibration and a validation sets, either randomly or with stratifying algorithms in the spectra space such as k-means clustering (Næs, 1987), Kennard-Stone sampling (Kennard and Stone, 1969; Nawar and Mouazen, 2018), conditioned hypercubic sampling (Minasny and McBratney, 2006; Moloney et al., 2023), a combination of these (Ogen et al., 2018), or by stratification using environmental covariates (Wijewardane et al., 2018; Dotto et al., 2020).

The spectral libraries might cover differing spatial areas, ranging from local (Guerrero et al., 2016) and national (Demattê et al., 2019) to regional (Shepherd and Walsh, 2002), continental (Stevens et al., 2013), or even global scales (Viscarra Rossel et al., 2016). In all these scaled SSLs, apart from variation in the sampling design, number of samples required, and spectra information generated, a spatial component is embedded which the current evaluation of prediction performances may have not considered. Further on, the validation set used to evaluate the predictive models may not be similar in spectral and soil characteristics with the new samples to be analysed (Stenberg et al., 2010). Consequently, this may result in model overfitting which is characterized by smaller calibration errors and large prediction deviation causing a reliability issue (Summerauer et al., 2021).

Considering this, while calibration samples are essential, there is significant potential to optimize the quantity, spatial distribution, and type of samples used to enhance the quality of the spectral library (Shepherd et al., 2022). Such optimised spectral libraries are supposed to minimize and/or avoid further requirement of conventionally analysed soil samples to calibrate a spectroscopic model while predicting soil parameters for new samples. Besides, large datasets might not be computationally effective and can cause problems when using routine calibration methods for regression by increasing noise and leading to model inadequacy (Sáiz-Abajo et al., 2005). By selecting representative (spatially, spectrally, or other improving criteria and combinations of these) and well-distributed calibration sets from abundant SSL data entries, fewer samples could be sufficient to build a robust and reliable model without losing the prediction accuracy.

Thus, the challenge is to determine the best calibration samples from a large and diverse SSL to predict soil properties for any unknown sample which only possess geographic coordinates and soil spectra information. A few approaches have been tried and assessed to downscale large spectral libraries for local scale, avoiding the need of new local calibration sample wet chemistry analysis while improving prediction performance to minimizing the cost of soil information. Among other, geographically weighted partial least square regression Kriging (PLS-GWRK) implements a spatial weighted regression and gives larger weights to the coefficients closest to the new sample location (Song et al., 2021). Another approach is the implementation of a continuous spectra similarity analysis (Memory-based Learner [MBL]) to narrow down calibration sample entries based on the spectra distance to the new samples from a large SSL (Dangal et al., 2019; Summerauer et al., 2021). Both approaches have been evaluated for their contribution to improved soil property prediction for new samples from fewer calibration samples (≪the entire SSL) in geographic and spectra space, respectively.

However, these studies did not consider the spatial autocorrelation of spectra when selecting calibration samples for predicting a new sample. Spatial autocorrelation refers to the principle that spatially close objects are more likely to exhibit similar properties than those further apart. In the context of soil spectra, it means that the spectral signatures of soil samples located near each other are more likely to be similar in most structures. By accounting for the spatial dependency through variogram analysis, it can be more effective in selecting calibration samples that represent the spatial variability within a study area, ultimately improving the predictive accuracy of soil property models from spectra.

Additionally, previous studies limited their evaluation to a small number of soil properties and a single spectral source. In order to obtain a robust assessment of the effectiveness of subsetting entire SSLs, it is essential to examine and model the spatial relationship of spectra within a SSL generated from various soil scanning instruments. In this paper we consider both MIR and NIR spectral sources.

We targeted the Moroccan rainfed wheat growing areas as the research domain. To date, there are only a few samples collected from a Mediterranean climate included in the global (GSSL) (https://explorer.soilspectroscopy.org/) (SoilSpec4GG, 2020) as well as the Geocradle Regional Soil Spectral Libraries (GRSSL) (http://datahub.geocradle.eu/dataset/regional-soil-spectral-library) (Geocradle, 2018), with no soil samples included from Morocco. To evaluate different approaches for selecting calibration samples, the Moroccan soil spectral library (MSSL) was developed. The compared methods included selection based on spectra PC spatial autocorrelation (spatial sample selection), soil spectra similarity (MBL), and environmental covariate clusters, in comparison to the use of the entire MSSL. The specific objectives of this study were: − 1) to develop a representative SSL for the rainfed wheat-growing areas of Morocco, 2) to study the spatial autocorrelation of the spectra PC scores in MIR and NIR spectra, 3) to evaluate the prediction performance of the calibration models built with spatial, spectral and covariate clustering based calibration sample selection for various soil properties and soil spectral range combinations, and 4) to propose a use-framework on how best to use the MSSL for prediction of soil properties for a specific farmland.

2. Material and methods

2.1. Soil sampling location and sampling strategy

This research targeted the rainfed wheat-growing areas of Morocco and the soil samples were collected from regions of agricultural importance and their dominant soil types. Morocco is a country in the Maghreb region of North-western Africa with a Mediterranean climate. Morocco is mostly situated in the arid to semi-arid climate region, and the rainfed agriculture accounts for around 80 % of the utilized agricultural area (Mamassi et al., 2023). Based on the 2017/2018 crop cover map (https://lcviewer.vito.be/2018) (Copernicus:Europe's eyes on Earth, 2020) and the FAO 1974 soil classification (Spaargaren and Batjes, 1995), Calcic Kastanozems (Kk), Calcic Cambisols (Bk) and Chromic Luvisols (Lc) dominate the cropped areas of Morocco in this order of importance (Supp. Doc. Sec. 1).

We selected a subset of 599 soil sampling locations from the Al Moutmir BU (College of Agriculture and Environmental Sciences, Mohammed VI Polytechnic University, Ben Guerir, Morocco) dataset that classified as rainfed wheat farms in the 2019–2020 cropping season. Sampling sites were selected with a stratified balanced coverage

sampling (StrBCS) design to have a balanced representation of the soil types, geographic coordinates, and environmental covariates. Stratified balanced coverage sampling or the doubly balanced sampling eliminates the poor spread of covariates on the axis in their spatial space. Details of the sampling algorithms are provided in the Supp. Doc. Sec. 2. For the stratification, we used the *lcubestratified* function from the *BalancedSampling* package of R (Grafström, 2018).

Mean annual rainfall, mean annual temperature, slope percent and elevation, which are the most influential drivers of soil formation at local scale, were used as balancing covariates, longitude and latitude were used as spreading covariates, and the FAO soil types were used as stratifying variables. Details of the environmental covariate data types and sources are elaborated in the Supp. Doc. Sec. 3. Furthermore, the StrBCS design was used to subset the database into a calibration database (further on named as the Moroccan Soil Spectral Library (MSSL) and new samples (validation set) to evaluate different methodologies of sample selection for calibrating soil spectra to laboratory data (Fig. 1).

The soil samples were collected from the 0–20 cm soil depth with a soil auger, each representing the field of a collaborating farmer. At each location, soil samples were collected in a 'W' pattern across the field, mixed thoroughly and the composite sample was taken for air drying and further sample processing for analysis. All sampling locations had linked data of farming practices, coordinates, and crop productivity. An overview of the sampling sites is given in Fig. 1.

2.2. Baseline soil property analyses

The soil samples were air-dried, crushed, sieved to less than 2 mm, and further processed (milling and weighing) for laboratory analysis as required. Twelve soil properties were considered for this study based on their importance for plant nutrition, soil health and soil productivity. Soil particle size distribution was determined by a laser diffraction on a L-960 particle-size analyser (Horiba scientific Ltd.) in the Dry Spectral Laboratory at Rothamsted Research. The instrument allows for the measurement of particle sizes (volume, %) within the size range of 0.02–2000 μm using a diode laser of 650 nm wavelength and a blue LED light source of 405 nm wavelength in wet mode. The intervals used were clay diameter < 5 μm, silt $5 < d < 63$ μm, sand 63 μm $< d < 2000$ μm (Thomas et al., 2021). Total carbon and nitrogen were measured using the dry combustion method with a Leco analyser (Nelson and Sommers, 1996; Bremner, 1996). Inorganic carbon was determined via wet acidification using a Skalar Primacs AIC 100 (Skalar Analytical BV, Breda, Netherlands). Soil samples were introduced into the reactor, cleared with oxygen, then treated with Phosphoric acid and heated to 135 °C. The released CO2 was measured by an infrared detector. Total soil organic carbon (SOC) was then calculated by subtracting inorganic carbon from total soil carbon.

The ammonium acetate extracts (ISO 13536:1995) of exchangeable potassium (Exch. K) and magnesium (Exch. Mg) were determined by atomic emission and atomic absorbance, respectively, and CEC was determined with the ammonium acetate method (Ciesielski and Sterckeman, 1997a). Soil pH-H_2O was measured in a 1:2.5 soil: water suspension (ISO 10390: 2005) one hour after mixing, using a semi-micro sealed pH electrode from Fisher scientific. Sodium bicarbonate extracted soil phosphorus (OlsenP) was determined following the Olsen method (Olsen, 1954), and Mehlich III extractable phosphorus (P_M3) was determined with an inductively-coupled plasma membrane analyser (Mehlich, 1984). Available zinc (DTPA Zn) and manganese (DTPA Mn) were determined by atomic absorption spectrophotometer after complexing with DTPA (Lindsay and Norvell, 1972).

2.3. Soil spectral collection and spectra preprocessing

Absorbance data in the MIR range (2500 − 16,666 nm) were obtained in two replicates, and the average of these was used for subsequent spectral modelling. The spectral signatures were acquired from finely milled samples (< 50 μm following overnight drying at 40 °C. Measurements in the NIR range (1350–2600 nm) were obtained in three replicates per sample. Each replicate was obtained from the same Petri dish by shifting the scanning spot across the sample holder. These samples were air-dried and sieved through 2 mm stainless steel sieve. For more detailed information on each instrument's specifications, including the labelling used in subsequent graphs and discussions, please refer to the Supplementary Documentation Section (4). Spectra collected by each instrument is depicted in Fig. 2.

The reflectance (R) measurements recoded by Neospectra was transformed to the logarithmic apparent absorbance using $A = \log(1/R)$ after which spectra beyond 2450 nm were trimmed off due to low signal-

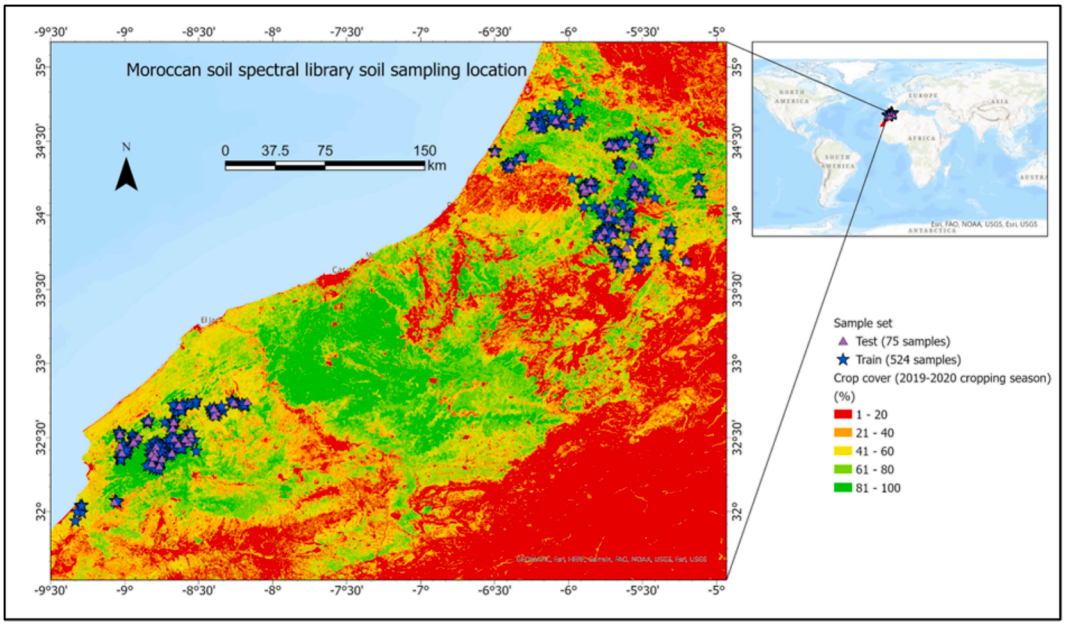


Fig. 1. Soil sample locations with crop cover map of the 2019–2020 cropping season and sample subset. There were two geographic distinct sampling regions situated in the north and south of Morocco. Note that soil sampling for the cropping area between the two geographic distinct sampling regions is planned.

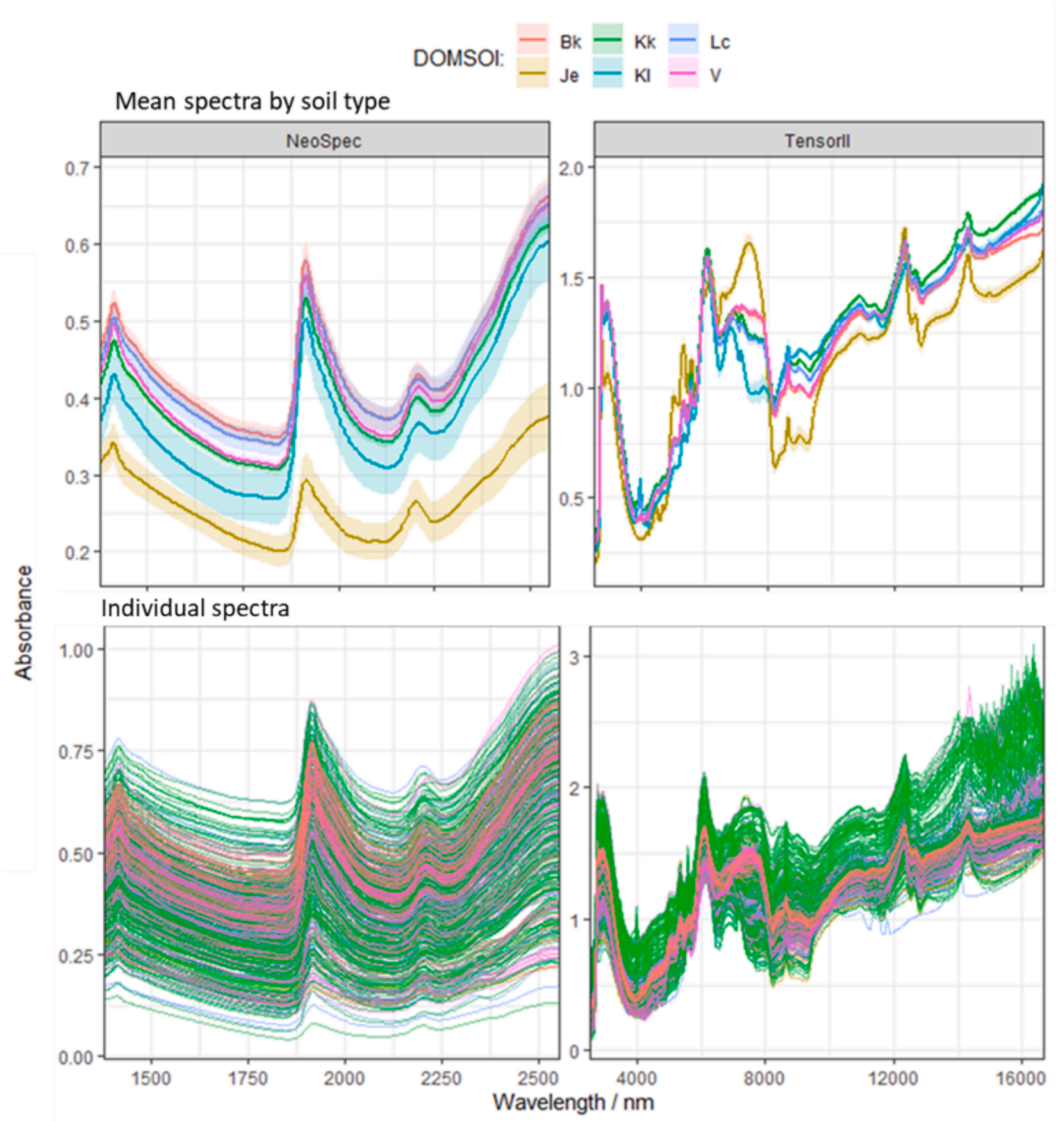


Fig. 2. Mean with standard error (top row) and raw (bottom) absorbance spectra of the dominant soil types in the NIR and MIR region. Bk = Calcic Cambisols; Kk = Calcic Kastanozems; Lc = Chromic Luvisols; Je = Eutric Gleysols; Kl = Luvic Kastanozems; V = Vertisols.

to-noise ratio often caused by light scattering effects of quartz sand or instrument drift. For both spectra signatures, a combination of spectral pre-processing techniques were applied that improved absorption features by reducing or eliminating noise and thus enhancing the correlation with soil properties (Vestergaard et al., 2021). A correlation of these pre-processing algorithms with the conventional soil properties' attribute was performed. The smoothing was done with a window size of 11 for MIR and 3 for NIR and a polynomial order of 1. Afterwards, the spectral data was cleaned from water absorption regions (H_2O band 1 between 1350 and 1460 nm; H_2O band 2 between 1790 and 1960 nm) in the NIR, and CO_2 peaks in the MIR absorption regions (between 4274 and 4464 nm). Spectral pre-processing was implemented in R with the prospectr package (Stevens and Lopez, 2022).

2.4. Frameworks for calibration sample selection

We evaluated four different methodologies for calibration sample selections and analysed their impact on the prediction performance, following the workflow illustrated in Fig. 3. These were use of 1) the entire MSSL sample collection (MSSL), 2) subset by first four PC scores (weighted by their variance explained) variogram (Spatial selection), 3) based on spectra similarity (MBL) and 4) based on clustering using environmental covariates and FAO soil units (covariate clustering). The details of each methodology are described in section 2.5, 2.6, 2.7, and 2.8, respectively. We used 524 calibration sample entries for the library to develop the predictive model and 75 validation set samples. This was done for two spectral sources, i) MIR and ii) NIR. We also considered prediction of soil properties using the reflectance spectra in the NIR using the entire MSSL dataset.

2.5. The entire MSSL entries with partial least square regression (PLSR)

The first methodology is the reference method in soil spectroscopy, generalising and combining features from principal component analysis and multiple regression to develop the 'global' PLSR predictive model. From a systematic review and *meta*-analysis, Ahmadi et al., (2021) reported that the most employed machine learning methods was the PLSR family which accounted for around 70 % of the studies in soil spectroscopy. It establishes a linear relationship between a set of latent variables extracted from the spectral matrix (X) while maximizing the covariance with the soil property (y). As such, it eliminates the multicollinearity problem that would arise by regressing y on X. For more

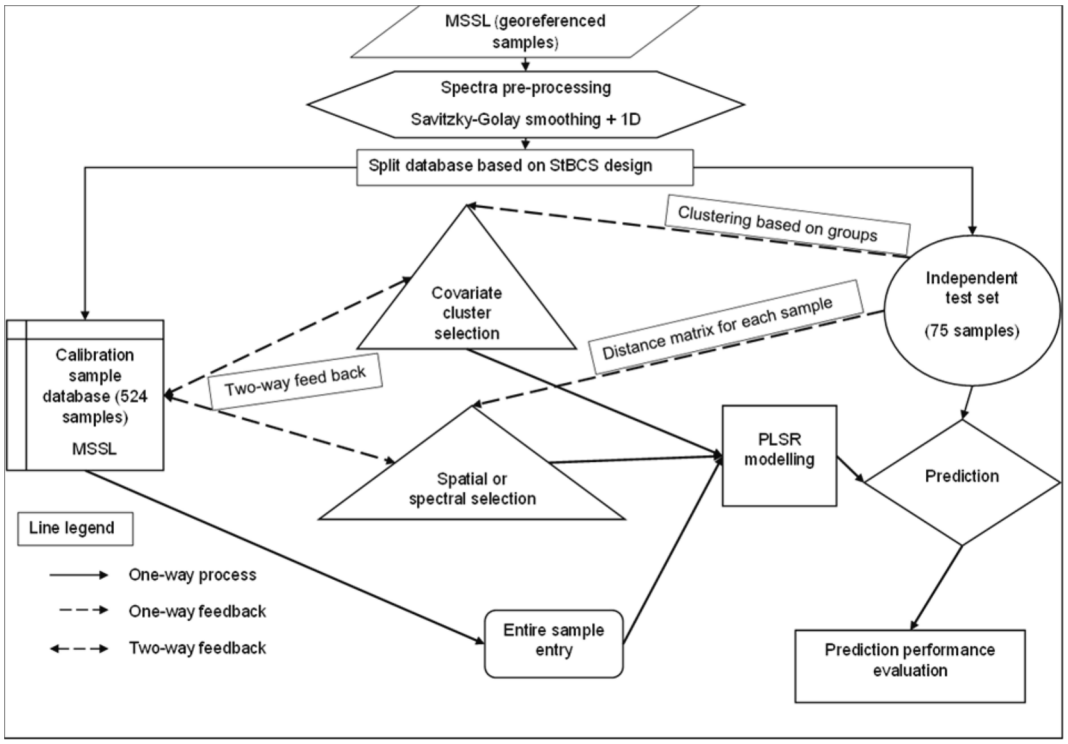


Fig. 3. Schematic demonstrating the sample selection processes for soil property prediction from a MIR and NIR soil spectral library. Tensor-II −> Bruker Tensor 27; Neospec = Neospectra. StBCS −> Stratified Spatially Balanced Coverage Sampling.

details see Forina et al. (2007). We selected the number of latent variables, or components to retain, by the minimum RMSE from a 10-fold cross-validation and considered a maximum of 15 components. We used the plsr function in the pls R package (Liland et al., 2023). This process generates a single model (per soil property) that was used to predict each sample in the validation set.

2.6. Spatial autocorrelation of spectra PC scores

The second methodology selected calibration samples under the assumption that soil variation is spatially autocorrelated. A principal component analysis (PCA) was computed on the soil spectra. A single vector representing the majority of the variation in the spectra space was calculated by weighting the first four principal components according to the variance they explained. We calculated the experimental variograms by the method of moments with Eq. (1), considering a cutoff distance of 120 km and a lag distance of 10 km. Additionally, we computed the experimental variograms for a longer cut-off distance (200 km) to investigate any further variogram characteristics that might be present in the MSSL dataset. Elevation, slope percentage, mean rainfall and temperature were used as covariates to account for long-distance trends.

$$\widehat{\gamma}(h) = \frac{1}{2\mathrm{m}(h)} \sum_{\mathrm{i}=1}^{\mathrm{m}(h)} [\mathrm{Z}(\mathrm{s_i}) - \mathrm{Z}(\mathrm{s_i} + h)]^2 \quad (1)$$

where $\widehat{\gamma}(\mathrm{h})$ is the empirical semivariogram value at lag h, $m(\mathrm{h})$ is the number of paired comparisons separated by h, $Z(s_i)$ and $Z(s_i + \mathrm{h})$ are the values of the weighted PC vectors at locations s_i and $s_i + \mathrm{h}$, respectively.

After evaluating the residuals of various covariance models, including Exponential and Matern, we fitted a spherical covariance model ($\gamma(\mathrm{h})$) for characterising spatial autocorrelation and parameter estimation using Eq. (2).

$$\gamma(h) = \begin{cases} 0 if h = 0 \\ \mathrm{c}_0 + c\left[\frac{3\mathrm{h}}{2\mathrm{a}} - \left(\frac{\mathrm{h}^3}{2\mathrm{a}^3}\right)\right], if 0 < h \le a \\ \mathrm{c}_0 + c, if h > a \end{cases} \quad (2)$$

Where c_0 is the nugget, c is the sill, and a is the effective range.

Restricted maximum likelihood fitting led to non-unique solutions depending on the initial range parameter. Hence, we used the '*fit.variogram*' function from the gstat package in R (Pebesma and Graeler, 2023), which uses weights based on the number of point pairs and the lag distance, i.e. smaller lag distances and bins with more data get more weight. Thus, the spatial dependency analysis determined the distance for which the weighted PC was autocorrelated (the effective range). We employed subsampling with spatial constraints to select a set of the MSSL dataset for each prediction location, based on a local neighbourhood determined by the effective range estimate (Pebesma and Wesseling, 1998).

Following this, the predictions of soil properties $\widehat{z}(s_n)$ from spectral dataset obtained by a specific soil scanning instrument j for new sample at any given location s_n were derived using the following equation (Eq. (3)):

$$\widehat{\mathrm{z}}(\mathrm{s_n})_{\mathrm{j}} = \mathrm{PLSR}\left\{\mathrm{z}(\mathrm{s_i})_{\mathrm{j}}, \gamma(\mathrm{h})_{\mathrm{j}}, \mathrm{s_n} \mid \mathrm{s} \in \mathbb{A}\right\} \quad (3)$$

where the PLSR denotes the Partial Least Square Regression function, $z(s_i)$ are the input point data values within area $\mathbb{A}$, $\gamma(h)$ is the semivariogram model defining the spatial autocorrelation of the weighted PC vector at a defined cut-off distance, s_n is the location of the new sample, $s \in \mathbb{A}$ indicates that the points are within the spatial area of $\mathbb{A}$.

This process generates a separate PLSR model (using a specific subset of calibration samples) for every sample in the validation set.

2.7. MBL spectra similarity analysis

In the third methodology, we followed the approach of Summerauer et al., (2021) with slight modification to enhance calibration sample selection based on spectral similarity. In this approach, calibration samples were selected for each target sample based on spectral similarity. The process involved the following steps:

Moving Windows Correlation Dissimilarity: We calculated the dissimilarity between spectra using a moving window correlation method. This was done for a range of window sizes from 11 to 151 in steps of 10. For each window size *w*, the correlation dissimilarity matrix was computed using Eq. *(4)*.

$$\text{dissimilarity}_{ij} = 1 - \text{cor}\left(X_i^w, X_j^w\right) \tag{4}$$

Where $cor\left(X_i^w, X_j^w\right)$ is the correlation between the spectral windows of size *w* for sample *i* and *j*. The optimal window size was determined by calculating the root mean squared errors (RMSE) between the nearest neighbours for each target soil property, selecting the window size that yielded the lowest RMSE (Suppl. Fig. S2) using Eq. *(5)*.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

Where y_i is the actual value, and $\hat{y}_i$ is the predicted value.

Selection of Nearest Neighbours and weighted Average PLSR Model: For each target sample, we selected the nearest neighbours (NN_i) with a dissimilarity below a threshold of 0.1 with Eq. (6).

$$\text{NN}_i = \left\{j : \text{dissimilarity}_{ij} < 0.1\right\} \tag{6}$$

And a weighted average PLSR model was fitted (Shenk et al., 1997) for each target sample. The weights for each component *k* of the PLSR model were calculated based on the reconstruction error E_k of the PLS loadings for the target samples using Eq. *(7)*.

$$w_k = \frac{1}{E_k} \tag{7}$$

These weights were in turn used to average all PLSR models across the range of components considered (minimum 5 and maximum of 15 components) using Eq. *(8)*.

$$\hat{y} = \sum_{k=1}^{k} w_k \hat{y}_k \tag{8}$$

Where $\hat{y}_k$ is the prediction from the $k-th$ component.

Optimum Number of Neighbours and Execution of MBL Routine: Lastly, the final number of nearest neighbours was determined by nearest neighbour cross-validation (Ramirez-Lopez et al., 2013b) testing a range from 20 to 100 in interval of 10. The number that minimized the cross-validation RMSE was selected. To execute the MBL (Moving Block Local) routine, the 'resemble' package in R (Ramirez-Lopez et al., 2022) was used.

2.8. Cluster grouping using environmental covariates and FAO soil groups

For the fourth method, calibration samples were selected based on the similarity in environmental conditions and FAO soil types. We used five environmental variables (mean annual rainfall, mean annual temperature, slope percent, elevation, and FAO soil types) to cluster the entire dataset (n = 599). All numerical variables were scaled to unit variance, and the FAO soil types were converted into binary variables using one-hot encoding. The one-hot encoding process converted each category of the FAO soil types into a separate binary (0 or 1) variable. If there are *n* unique categories $\{c_1, c_2 \cdots c_n\}$ in the FAO soil types (FAO_{Soil}), the one-hot encoding process created *n* new binary variables (FAO_{c_i}) as follow (Eq. *(9)*,

$$\text{FAO}_{c_i} = \begin{cases} 1 if \text{FAO}_{\text{Soil}} = c_i \\ 0 otherwise \end{cases} \quad \text{for i} = 1, 2, \cdots, n \tag{9}$$

After converting the FAO soil types, the dataset was expanded to include these binary variables. We checked for multicollinearity among the numberical variables using a correlation matrix and identify highly correlated variables (correlation > 0.9), which were removed to avoid multicollinearity issues in the clustering process.

We then applied the k-prototypes clustering algorithm to the scaled dataset with the original numerical variables and newly created binary variables. The k-prototype algorithm, which is suitable for datasets with mixed numerical and categorical variables, was employed to minimize the total within-cluster variation. The total within-cluster variation (total within-cluster sum of squares) was calculated from the sum of squared Euclidean distances between variables and the corresponding centroid following Eq. (10):

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{10}$$

where x_i is a data point belonging to the cluster C_k, and μ_k is the mean value of the points assigned to the cluster C_k.

To determine the optimum number of clusters, we ran the *k*-prototype algorithm for a range of clusters and repeated this process 10 times for each cluster count to ensure robustness. The choice of the optimum number of clusters was based on evaluating the total within-cluster sum squares, selecting the number of clusters that minimised this measure while maintaining the minimum number of calibration sample for a PLSR model. Each observation x_i was then assigned to the cluster such that the sum of squares (SS) distance of the observation to the cluster centers μ_k was a minimum. The total within-cluster sum of squares, which measures the compactness (i.e. goodness) of the clustering for the entire dataset, is given by Eq. (11):

$$\text{tot.withinss} = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{11}$$

where *tot.withinss* is the sum of the within-cluster variations for all clusters.

We used the *'kproto'* function from the '*clusteMixType*' R package and considered tuning the lambda parameter to adjust the weighting between numerical and categorical distances as needed. In addition, we evaluated the stability of the clusters using silhouette scores to confirm the robustness of the chosen cluster number. The clustering resulted in 5 distinct clusters, and training and validation sets were assigned accordingly (Supp. Doc. Table 2). Calibration regression was then based on cluster associations and the PLSR as described above (2.5). For each validation sample, one of 5 possible PLSR models was used according to the cluster assignment.

2.9. Accuracy assessment

Ten-fold-leave-group out cross validation was used to optimise PLSR model. The models' performance was further assessed using the measures a) Lin's concordance correlation coefficient (LCCC) (Lin, 1989), b) the coefficient of determination (R^2), c) the root mean square error of prediction, d) the ratio of performance to the interquartile range (RPIQ) (Bellon-Maurel et al., 2010), e) the ratio of performance to deviation and f) bias according to the equations below using the test set:

$$CCC = \frac{2 r s_x s_y}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})} \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(x_i - y_i)^2} \quad (14)$$

$$RPIQ = \frac{(Q3 - Q1)}{RMSE} \quad (15)$$

$$Bias = \frac{1}{n}\sum_{i}^{n}(x_i - y_i) \quad (16)$$

Where x_i and y_i are the observed and predicted value at point i; $\bar{x}$ and $\bar{y}$ are the mean of the observed and predicted values; r is the correlation coefficient between the observed and predicted values, s_x and s_y are the variance of observed and predicted values, n is the number of samples, Q_3 and Q_1 are the 75th and 25th percentiles, respectively.

3. Results

3.1. Baseline soil properties and their correlations with spectra pre-processing

Descriptive statistics are presented in Table 1 for the entire MSSL soil samples for both the training and validation sets. The mean and median values of SOC across all samples are moderate, though the lowest and highest values fell into extremely low and very high categories, respectively (Agvise Laboratories, 2019). Higher SOC is typically linked to better soil health and nutrient availability. Most of the soil samples exhibited a mildly to moderately alkaline pH, with some reaching the strongly alkaline range (Hazelton and Murphy, 2016), a characteristics often found in semiarid climates with limited leaching. The CEC of the soils was generally high, indicating good availability of exchangeable cations, which helps regulate soil pH and its response to inputs (Agvise Laboratories, 2019). In contrast, low CEC is usually associated with lower resistance to change in soil chemistry. Similarly, exchangeable K and Mg were found in moderate to high availability categories, likely due to the high CEC of soils and low precipitation levels. The distribution of clay content across all samples followed a normal distribution in the soil textural diagram, according to the International Union of Soil Sciences (IUSS) soil textural classification system (Suppl. Fig. 2).

The distribution of soil properties was largely similar between the MSSL and validation sets, though the range was generally higher for the former. While Olsen P, P_M3, Exch.K, DTPA.Zn, DTPA.Mn and Tot_IC showed right-skewed distribution, CEC and pH were left-skewed. In contrast, Exch.Mg, TotN and SOC exhibited normal distributions (Suppl. Doc. Figs. 4 and 5). Data transformations were applied due to the skewed distribution of the data for some of the soil properties. Log transformation was applied for those positively skewed datasets viz., OlsenP, DTPA.Zn, DTPA.Mn, P-m3, and Exch.Mg, while a square root transformation was used for Exch.K. For negatively skewed data of pH and CEC, reflection and log transformation were carried out.

The correlation boxplot of spectral pre-processing methods against soil property attributes (Fig. 4; Wang et al., 2018) highlighted the critical role of spectral pre-processing in refining the selection of explanatory variables for modelling via dimension reduction algorithms. The box plots effectively illustrated how different pre-processing techniques influence the identification of relevant spectral variables, which are shown at the tails of the plots, while irrelevant variables were concentrated around the centre, closer to zero correlation values.

The use of pre-processing methods such as SG + 1D and SG + gap 1D pre-processing demonstrated superior performance in generating clearer correlations between the spectral data and soil properties across both soil scanning instruments. This suggested that these pre-processing techniques enhanced the ability to distinguish relevant spectral features for modelling purposes. However, the analysis also revealed that certain soil properties, such as Exch.K exhibited the least correlation values on either side of the boxplot, indicating that such soil properties might be less responsive to spectral pre-processing or that its spectral signature is more challenging to capture accurately. This insight is crucial for understanding the limitations and strengths of different pre-processing methods in relation to specific soil properties and can guide the selection of appropriate spectral pre-processing techniques to improve model accuracy and predictive capability.

3.2. Spatial autocorrelation and scale dependency of spectra PC scores

The first four PC components, used as a PC vector, explained 80 % and 99 % of the variation in the spectra for the MIR and NIR spectra, respectively (Suppl. Fig. 6 & 7). We assumed that these PC components could reliably explain most of the spatial autocorrelation for the spectra generated. The number of PC components explaining 99 % of the

Table 1
Descriptive statistics of soil properties in the training and validation sets.

Soil property	Sample set	N	Min	Median	Mean	Max	Std	Skewness
CEC (cmol (+) kg-1 of soil)	Validation	75	4.20	31.10	30.19	41.80	6.51	−1.50
	Train	524	0.60	31.00	29.46	41.90	7.60	−1.24
Clay (%)	Validation	75	8.46	42.35	40.62	65.62	14.63	−0.30
	Train	524	2.61	40.91	39.90	71.69	15.11	−0.19
DTPA.Mn (mg kg-1)	Validation	75	0.50	5.50	8.92	87.20	11.85	4.29
	Train	524	0.50	5.20	8.74	87.70	10.15	3.12
DTPA.Zn (mg kg-1)	Validation	75	0.10	0.70	0.75	2.60	0.52	1.18
	Train	524	0.10	0.60	0.78	7.20	0.74	3.38
Exch.K (cmol kg-1 of soil)	Validation	75	0.38	0.94	1.13	6.86	0.88	4.15
	Train	524	0.04	0.95	1.12	10.23	0.73	4.83
Exch.Mg (cmol kg-1 soil)	Validation	75	0.63	8.99	10.25	49.87	8.09	2.27
	Train	524	0.30	9.05	9.30	56.28	6.51	2.17
OlsenP (ppm)	Validation	75	0.72	9.37	14.41	96.22	14.55	2.93
	Train	524	0.52	10.08	15.49	130.27	16.93	3.23
P_M3 (ppm)	Validation	75	3.00	29.00	47.91	301.00	54.37	2.19
	Train	524	1.00	27.00	49.30	381.00	60.33	2.41
pH	Validation	75	6.01	8.06	7.98	8.68	0.42	−2.15
	Train	524	5.45	8.08	7.97	8.94	0.48	−1.74
SOC (%)	Validation	75	0.35	1.19	1.30	2.74	0.48	0.86
	Train	524	0.09	1.16	1.22	4.29	0.46	1.25
Tot_IC (%)	Validation	75	0.00	0.50	0.82	3.87	0.96	1.12
	Train	524	0.00	0.21	0.99	9.16	1.39	1.97
Tot_N (%)	Validation	75	0.03	0.13	0.14	0.29	0.05	0.71
	Train	524	0.03	0.13	0.13	0.42	0.05	0.86

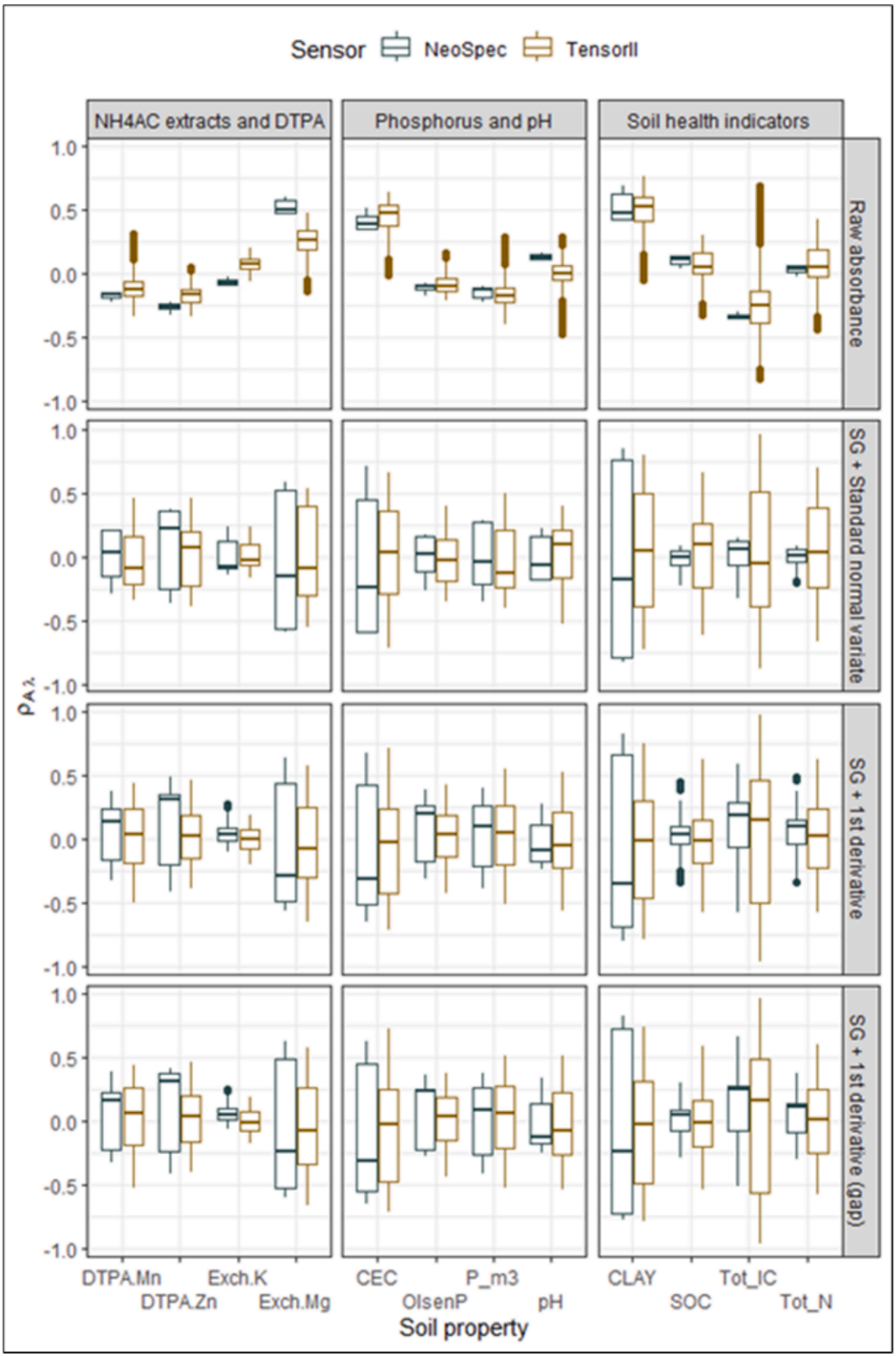


Fig. 4. Correlation boxplot of each soil properties with spectra pre-processing techniques (SG −> Savitzsky-Golay smoothing) at each wavelength by the scanning instrumentation (Neospec −> Neospectra, Tensor-II −> Bruker Tensor 27). The lower and upper limits of each box are the 25th and 75th percentiles, the tails show the correlation range, the dots represent outliers and the horizontal line in each box represents the median.

variation for the MIR were eleven. The spatial autocorrelation of the PC vector is depicted in Fig. 5 with cut-off distances of 120 and 200 km. The distances were chosen to reflect the width and length of the distinct sampling areas (Fig. 1). The separation distance between the two distinct sampling areas was around 300 km.

The variograms exhibited distinct patterns of spatial autocorrelation among the PC vectors of the NIR and MIR spectral ranges. However, the general pattern of the variograms for each spectral range tended to be similar between the cut-off distances used. The nugget effect, random variation which could be caused by mainly undetectable experimental error and field variation within the minimum sampling space (Sun et al., 2003), remained similar among these cut-off distances. The distinct variations notably were 1) a shorter cut-off distance of 120 km yielded better fit for spherical variograms and 2) the longer cut-off distance of 200 km brought up a new developing variogram characteristics for both NIR and MIR spectra. This divergence could stem from the additional spatial autocorrelation of the PC vectors, given that soil spectra constitute a multiresolution dataset (Lark and Webster, 1999; Song et al., 2021) which is influenced by various soil chromophores exhibiting different spatial scales (Savvides et al., 2010).

With the shorter cut-off distance, the spatial autocorrelation for the PC vectors reached a sill parameter at a range of 117 and 67 km for the MIR and NIR, respectively. These distance values for each spectra source indicated the bounding spatial extent within which the first four spectra PC scores of the soil samples were spatially autocorrelated. The spherical variogram fit showed the presence of a strong spatial autocorrelation for both NIR and MIR spectra PC vectors with a steep rise to the sill. This means that the soil sample spectra are strongly correlated over the effective range.

3.3. Prediction performance among sample selection processes

3.3.1. Use of entire MSSL sample (MSSL)

The PLSR models, created for each soil property using the entire MSSL dataset (524 samples), was assessed based on their effectiveness in predicting the respective soil properties from the NIR and MIR spectra. The models demonstrated excellent predictive accuracy for clay and Tot_IC, as evidenced by high CCC (>0.8) and high RPIQ (>2) values for both NIR and MIR spectra (Fig. 6). This suggested that these soil properties have strong, consistent spectral signatures across the samples of the entire MSSL library, making them well-suited for prediction using the entire dataset with a PLSR models.

For SOC and Tot_N, the MIR spectra provided excellent predictive agreement (CCC = 0.86 and 0.91, respectively) while the NIR spectra showed good but lower agreement (CCC = 0.69 and 0.73, respectively). The discrepancy between the MIR and NIR performance could be attributed to the absence of critical absorbance ranges (500–1350 nm) in the NIR spectra, which includes important C-H and O-H absorption regions (Stenberg, Rossel, Mouazen, Wetterlind, *et al.*, 2010). These regions are essential for accurately capturing the spectral characteristics associated with organic carbon and nitrogen, thus explaining the superior performance of MIR in these cases. We also included the prediction performance using the reflectance spectra of NIR for further comparison, applying SG-1D preprocessing (see Suppl. Doc. Fig. 7). However, the results showed similar predictive accuracy to those obtained with absorbance spectra in most soil properties.

The predictive performance for soil properties like CEC, DTPA.Zn, Exch.Mg, and P_M3 ranged from moderate to satisfactory when using both MIR and NIR spectra. However, the models showed low prediction accuracy for OlsenP and Exch.K across both spectral regions, as well as DTPA.Mn and pH specifically in the NIR spectra, with CCC values below 0.5 and RPIQ values under 1. Notably, NIR spectra outperformed MIR in

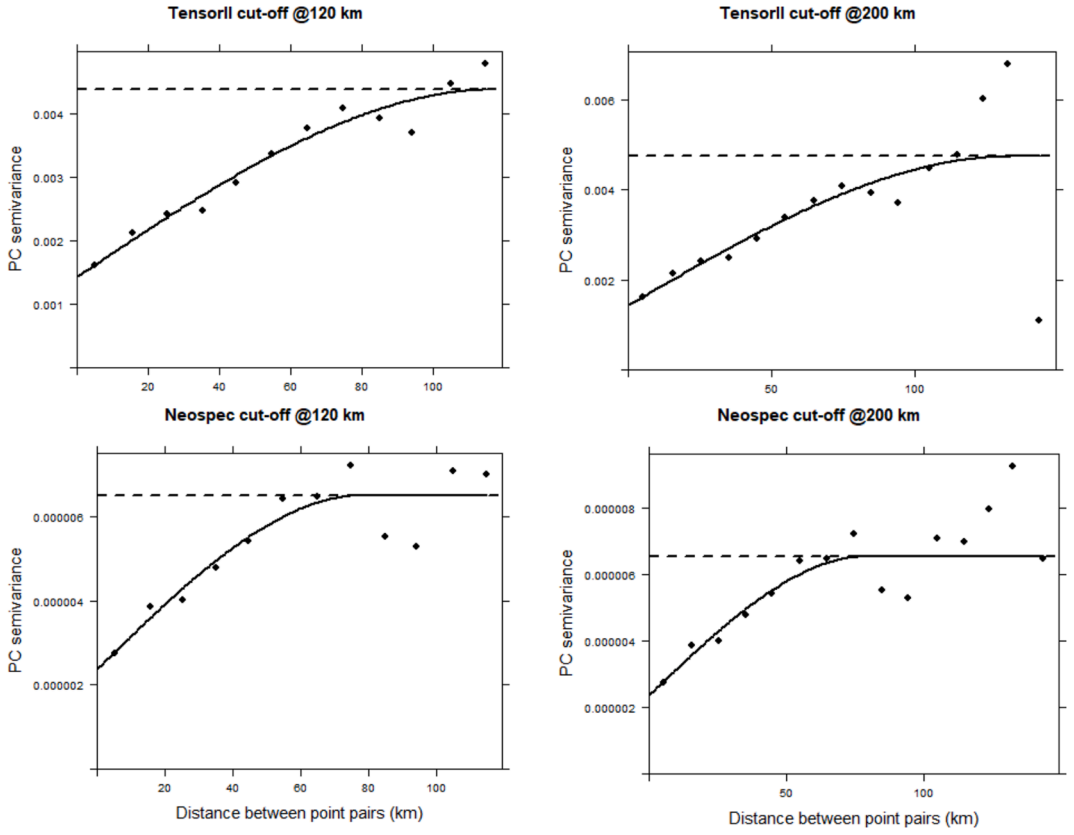


Fig. 5. Semi-variograms of PC vectors fitted with a spherical model (solid line) from the respective soil scanning instruments (Neospec −> Neospectra for NIR spectra; Tensor-II −> Bruker Tensor 27 for MIR spectra) with cut-off distances of 120 and 200 km. The dashed line is the sill.

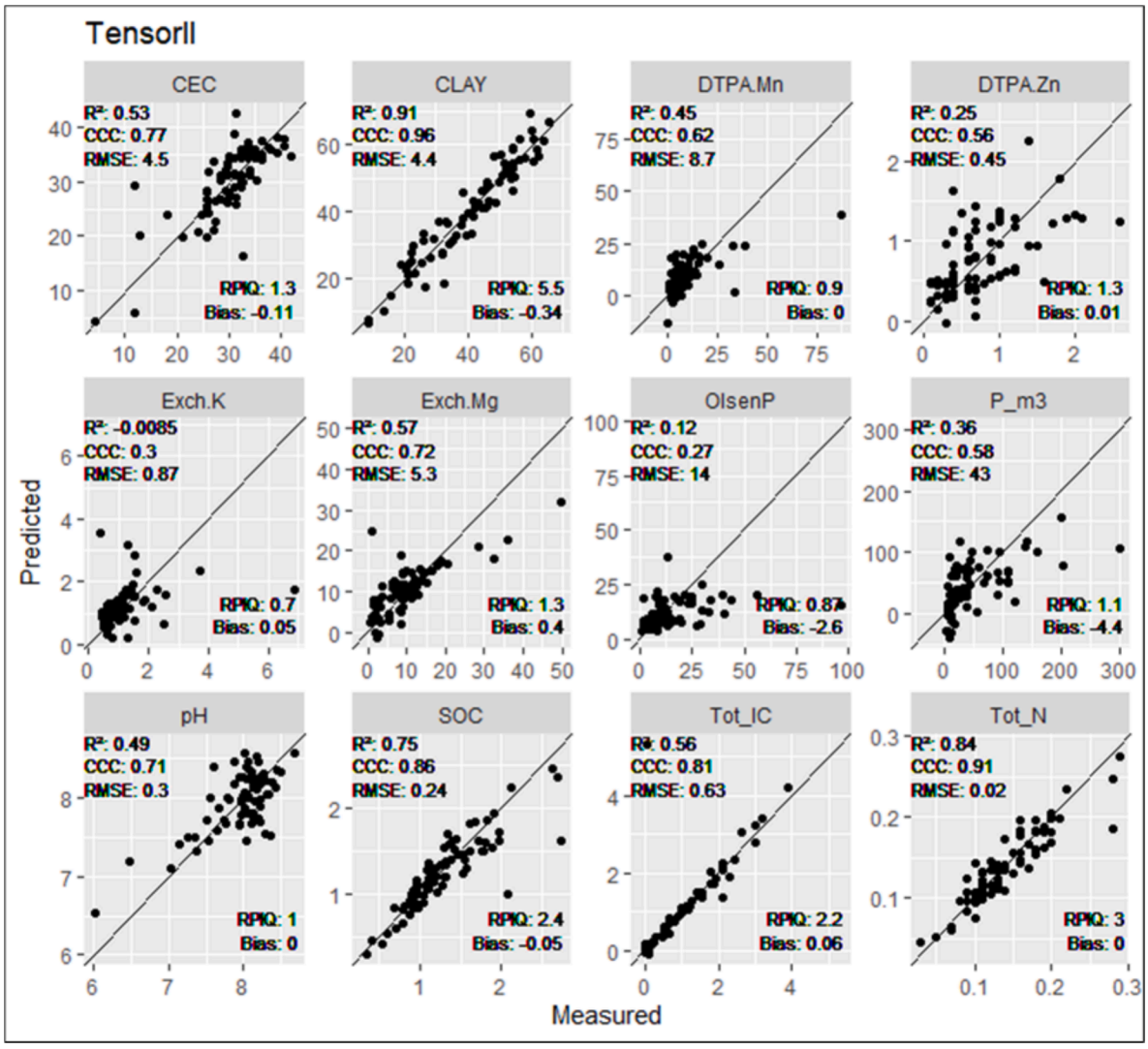


Fig. 6. Prediction performance for the soil properties of the validation set and soil scanning instruments (NeoSpec –> Neospectra for NIR spectra; TensorII –> Bruker Tensor 27 for MIR spectra) using the entire MSSL. The statistics shown are the R^2, concordance correlation coefficient (CCC), the root mean square error (RMSE), Bias, and the ratio of performance to interquartile range (RPIQ).

predicting OlsenP and Exch.K. These low performance indicators suggest that these properties may have weaker or more variable spectral signatures, limiting accurate prediction with PLSR models (Dangal et al., 2019).

Interestingly, the poor prediction performance for certain soil properties correlated with the characteristics observed in the correlation boxplot between spectral pre-processing and conventional soil analysis values (Fig. 4). The boxplot, which showed the distribution of correlations between spectral data and soil properties, could serve as a qualitative indicator of prediction performance. Specifically, the variability in box and tail length for properties like CEC, DTPA.Mn, pH, Exch.K, and OlsenP highlights the challenges in predicting these parameters accurately across different soil spectra ranges using the MSSL dataset.

Despite these challenges, the study demonstrated that PLSR models, particularly those using MIR spectra, can effectively predict a range of soil properties. The higher RPIQ values (>1.4) observed for Clay, SOC, Tot_IC and Tot_N in the MIR region underscores the potential of both spectra regions in delivering more accurate predictions of these soil properties. The stronger performance of MIR compared to NIR for certain soil properties can be attributed to the richer spectral information in the MIR range. These findings are consistent with existing literature, which highlights the superior predictive ability of MIR (Johnson et al., 2019).

3.3.2. Performance comparison between PLSR model calibration sample selection methods and the entire MSSL

The selection of calibration samples for the PLSR model produced varying prediction performances across different soil property-spectral region combinations, with both positive and negative impacts (Table 2). The scatter plots illustrating the spread of predicted values against the 1:1 line for each sample selection methods are also provided in the Supp.doc. Figs. 8–10. For the MIR spectra, the spatial calibration sample selection improved the CCC values notably for OlsenP (by 41.3 %) and P_M3 (by 8.5 %). Similarly, for the NIR spectra, improvements were obtained in CEC (by 25.6 %), pH (by 13.0 %), Tot_N (by 10.6 %), Exch.Mg (by 9.6 %) and SOC (5.6 %) compared to using the entire MSSL dataset. However, the CCC values decreased for DTPA.Mn (−21.4), Exch.K (−13.7 %), pH (by −8.3 %) and DTPA.Zn (−7.1)for the MIR spectra, and for Exch.K (by –23.2 %), OlsenP (by −13.4 %), DTPA.Mn (by −9.5 %) DTPA.Zn (by −7.3 %) and P_m3 (−6.7 %) for the NIR spectra. Despite these variations, the spatial calibration sample selection generally provided comparable prediction performance for other soil

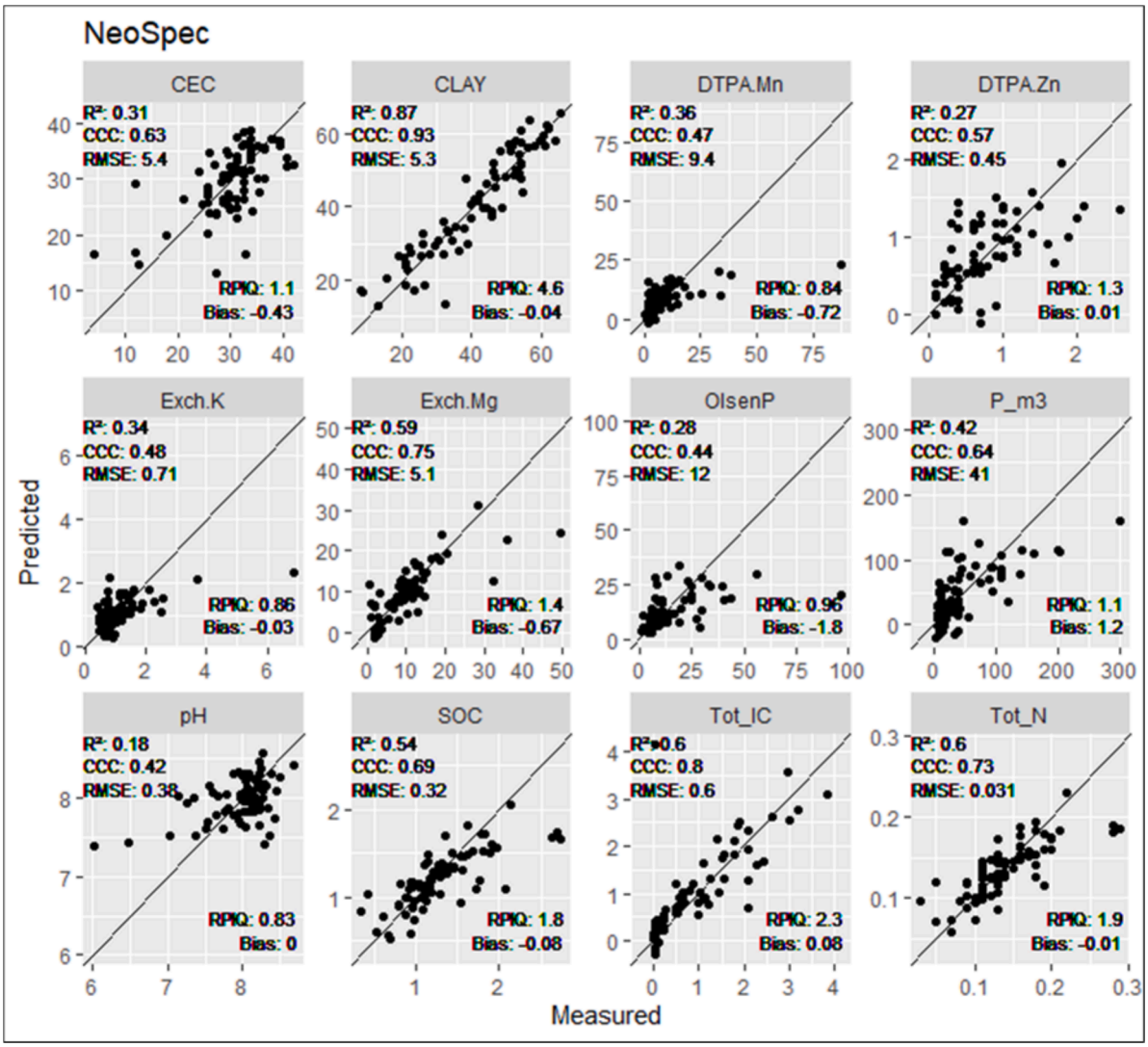


Fig. 6. (*continued*).

property-spectral range combinations, even with smaller sample size (<175 samples) compared to the entire MSSL dataset.

On the other hand, the MBL approach enhanced the CCC value for OlsenP (by 85.2 %), P_m3 (by 21.2 %) pH (by 12.2 %), and Exch.Mg (by 7.7 %) in the MIR spectra, and for pH (by 18.7 %), Tot_N (by 13.5 %), OlsenP (by 9.9 %), SOC (by 8.5 %), DTPA.Zn (by 8.1 %), and Exch.Mg (by 6.3 %) in the NIR spectra. However, it also resulted in decreased CCC values for DTPA.Zn (by −15.1 %), Exch.K (by −10.9 %), CEC (by −7.6 %), and DTPA.Mn (by −7.0 %) in the MIR spectra, as well as for Exch.K (by −18.9 %) and DTPA.Mn (by −6.2 %) in the NIR spectra. Overall, the spectra similarity selection demonstrated an improvement in prediction performance compared to using the entire MSSL dataset aligning with findings from Gholizadeh et al. (2018) and Moloney et al. (2023). Conversely, the cluster-based calibration sample selection improved the CCC values only for OlsenP (by 9.1 %); but led to reduced prediction performance for DTPA.Mn (by −27.5 %), DTPA.Zn (by −21.9 %), Exch.K (−20.9 %), pH (by −15.8 %), CEC (by −7.0), and P_m3 (by −6.3 %) in the MIR spectra. Similarly, in the NIR spectra, it improved prediction performance for CEC (by 28.2 %) and Exch.Mg (by 7.3 %) while reducing it for DTPA.Mn (by −34.4 %), OlsenP (by −29.1 %), Exch.K (by −19.0 %), pH (by −17.7 %) and P_m3 (by −5.5 %). These findings partially concur with those of Wijewardane et al. (2018), who reported that clustering with environmental covariates did not improve prediction performance for Exch.K and P_m3 using MIR spectra. Nonetheless, the cluster-based selection yielded prediction outcomes comparable to using the entire MSSL dataset for other soil property – spectral range combinations with a PLSR model developed with only 48 samples.

4. Discussion

4.1. Robustness of the MSSL for soil property predication and its framework

The Moroccan rainfed wheat belt soils mostly had calcic diagnostic characteristics with a pH in alkaline to strongly alkaline ranges which is mostly due to the low annual precipitation, ≪600 mm per annum, for the majority of this region. A further decrease in precipitation amount of 20 % by 2050 and of 40 % by 2080 is projected because of the expected climate change which could further the severity of recurring and severe drought combined with increasing demand for water (Mokhtar et al., 2022). Hence, soil input applications and soil management for agricultural productivity requires informed decision making with the support of reliable tools and methods. The database established comprised 599 soil samples collected across an area spanning approximately 12,200 km^2, with 524 of these samples constituting the entire calibration set. This equated to an average of about 23 km^2 per sample.

Our sampling strategy and sampling points were determined and selected to acquire representative samples to develop the MSSL for soil

Table 2
Prediction performance for the soil properties using calibration sample selection approaches in the NIR and MIR spectral ranges.

Soil properties	Wavelength Region	Spatial					MBL					Cluster				
		CCC	RMSE	RPIQ	Bias	*CCC deviation (%)	CCC	RMSE	RPIQ	Bias	*CCC deviation (%)	CCC	RMSE	RPIQ	Bias	*CCC deviation (%)
CEC	MIR	0.77	4.50	1.30	0.0	0.8 %	0.71	5.00	1.20	−0.9	−7.6 %	0.71	4.90	1.20	0.1	−7.0 %
	NIR	0.79	4.10	1.50	0.1	25.6 %	0.61	5.90	1.00	−1.2	−3.1 %	0.80	3.90	1.50	0.0	28.2 %
Clay	MIR	0.95	4.60	5.30	−0.1	−0.6 %	0.95	4.70	5.20	0.2	−0.5 %	0.94	5.20	4.70	−0.2	−1.8 %
	NIR	0.94	5.10	4.80	0.2	0.8 %	0.91	6.10	4.00	−0.4	−2.1 %	0.92	5.70	4.30	−0.4	−1.2 %
Mn	MIR	0.49	9.90	0.80	−0.1	−21.4 %	0.58	9.50	0.83	−1.1	−7.0 %	0.45	10.00	0.76	−0.8	−27.5 %
	NIR	0.43	9.90	0.80	−0.8	−9.5 %	0.44	11.00	0.75	−0.3	−6.2 %	0.31	11.00	0.72	−0.9	−34.4 %
Zn	MIR	0.52	0.49	1.20	0.0	−7.1 %	0.48	0.56	1.10	0.1	−15.1 %	0.44	0.53	1.10	0.0	−21.9 %
	NIR	0.53	0.46	1.30	−0.1	−7.3 %	0.62	0.52	1.10	0.1	8.1 %	0.59	0.45	1.30	0.0	3.2 %
K	MIR	0.26	0.87	0.70	0.1	−13.7 %	0.27	0.88	0.69	0.0	−10.9 %	0.24	0.89	0.69	0.0	−20.9 %
	NIR	0.37	0.79	0.77	−0.1	−23.2 %	0.39	0.79	0.77	0.0	−18.9 %	0.39	0.75	0.81	0.0	−19.0 %
Mg	MIR	0.73	5.40	1.30	0.5	1.3 %	0.77	4.70	1.50	0.0	7.7 %	0.74	5.20	1.30	0.2	2.9 %
	NIR	0.82	4.60	1.50	−0.4	9.6 %	0.79	4.70	1.50	−0.5	6.3 %	0.80	4.60	1.50	−0.8	7.3 %
OlsenP	MIR	0.39	13.00	0.93	−2.1	41.3 %	0.51	12.00	1.00	−1.5	85.2 %	0.30	16.00	0.76	−1.5	9.1 %
	NIR	0.38	13.00	0.92	−1.8	−13.4 %	0.48	13.00	0.93	−1.2	9.9 %	0.31	14.00	0.84	−1.9	−29.1 %
P_m3	MIR	0.63	42.00	1.10	1.1	8.5 %	0.70	38.00	1.20	1.3	21.2 %	0.54	48.00	0.94	−1.6	−6.3 %
	NIR	0.59	43.00	1.10	2.8	−6.7 %	0.62	45.00	1.00	1.1	−2.7 %	0.60	43.00	1.10	−4.3	−5.5 %
pH	MIR	0.65	0.33	0.96	0.0	−8.3 %	0.80	0.26	1.20	0.0	12.2 %	0.60	0.35	0.90	0.0	−15.8 %
	NIR	0.47	0.38	0.83	0.0	13.0 %	0.49	0.39	0.82	0.0	18.7 %	0.34	0.41	0.76	0.0	−17.7 %
SOC	MIR	0.87	0.24	2.40	−0.1	0.7 %	0.88	0.22	2.60	0.0	2.6 %	0.84	0.26	2.20	−0.1	−2.3 %
	NIR	0.72	0.31	1.80	−0.1	5.6 %	0.74	0.31	1.80	−0.1	8.5 %	0.68	0.33	1.70	0.0	−1.3 %
Tot_IC	MIR	0.81	0.63	2.30	0.1	0.1 %	0.82	0.61	2.30	0.1	1.3 %	0.79	0.66	2.20	0.1	−2.4 %
	NIR	0.80	0.61	2.30	0.0	0.0 %	0.80	0.64	2.20	0.0	−0.9 %	0.80	0.62	2.30	0.1	−0.6 %
Tot_N	MIR	0.92	0.02	3.20	0.0	0.8 %	0.91	0.02	3.00	0.0	−0.1 %	0.88	0.02	2.70	0.0	−3.0 %
	NIR	0.81	0.03	2.20	0.0	10.6 %	0.83	0.03	2.10	0.0	13.5 %	0.74	0.03	1.90	0.0	1.0 %

*Percent CCC deviation for the respective sample entry selection methods in reference to the entire MSSL sample entry.

property prediction in rainfed wheat growing regions of Morocco, optimized by a stratified balanced coverage sampling design. This approach enabled prioritization for those soil types which had the largest area coverage while addressing the spatial distribution and soil variation, and hence optimized the number of samples required for soil property prediction using soil spectroscopy (Potash et al., 2023). Consequently, we targeted the relevant geographic coverage of rainfed wheat productivity and spread the sampling locations according to the spatial variation expected from the environmental covariates (Janik et al., 1998).

Using the MSSL established within the outlined framework, the most significant agricultural soil properties were accurately predicted with satisfactory to excellent agreement, accompanied by low RMSEs for the respective soil variables. However, certain soil properties, which are spectrally inactive, exhibited unsatisfactory predictions in both the NIR and MIR spectral regions (Shepherd et al., 2022). Sarathjith et al. (2014) classified SOC and clay as spectrally active, while pH, P, K, and Zn fall under spectrally inactive soil properties. These associations were also explained for conventional extraction procedures by e.g., Ciesielski and Sterckeman (1997b) who noted that the ammonium acetate determination method, a percolation extraction method buffering the pH of the extracts, led to significant variations in the proportion of negatively charged sites and particularly those bonded to organic matter. For instance, the relatively poor prediction performance for Exch.K in this study mirrored findings by Hu et al. (2013) for Missouri soils, and attributed to low SOC levels. Conversely, Jin et al. (2020) reported better performance for Anhui soils in China, although detailed soil types and property information were lacking for comparison. This may also indicate that variations in prediction performance for certain soil properties can emanate from differing conventional extraction methods followed, besides general soil variations. Notably, Exch.K determined with cobalt hexamine (Cohex) trichloride showed excellent agreement (CCC of 0.81–0.94) with soil spectra in both vis-NIR and MIR for East African soils (Asrat et al., 2023). These disparities in prediction performance were also observed for phosphorus, i.e., for OlsenP versus P_m3, reinforcing the notion that the choice of conventional soil extraction methods substantially affect the estimation of specific soil property attributes using soil spectroscopy methods. Therefore, it is hypothesized that discrepancies in quantifying soil properties contribute to variation in prediction performance when utilizing spectroscopic methods, underscoring the importance of selecting the appropriate conventional analysis method for specific purposes and research objectives for a defined soil system.

A general observation was that the prediction performance within this database also depended on the two spectral regions used. The MIR spectra demonstrated better prediction accuracies for soil properties associated with organic matter (SOC, Tot_N and CEC), while the NIR spectra yielded superior predictions for soil properties related to geological materials and/or land use (DTPA extracts of Zn & Mn, Exch.K, Exch.Mg, OlsenP and P_M3). Hence, this study and the database created will support access to spatially explicit soil information easily and in time to support the soil information gap for soil spectroscopic methods, especially for the studied region and represented soil types.

4.2. *Suitability of spatial autocorrelation of spectra for calibration sample selection*

The scales of spatial variation is dependent on the sampling design and sampling intensity in a given area for specific soil properties, and each of the soil properties can be influenced at varying extent by the combination of soil forming factors and soil forming processes in space and time (Bogunovic et al., 2017). With that in mind, the parameter selection in variogram modelling and fitting can influence the analysis and interpretation of the outcome (Vasu et al., 2017). This may impose challenges to compare results across such research fields as the data used might have varying spatial density (Ye et al., 2017). To better fit a

variogram in any geographic extent, a minimum of about a 100 sampling points is required with a set interval for the scale of variation needed (Iqbal et al., 2005; Kerry and Oliver, 2007). This study used the spectral PC scores to understand the overall spatial autocorrelation as it was noted that the general soil characteristics were encoded in the soil spectral signature (Viscarra Rossel et al., 2016; Demattê et al., 2019). Hence, the spatial variability will dictate the extent of the general soil variation within the sampling extent and scale (Hengl, 2009). We anticipated that the spatial dependency analysis in the soil spectra space can be performed to inform calibration sample entry selection with consideration of the databases' spatial extent and subsequent representation of the domain variability within it. Consequently, the semivariogram of the spectra PC vectors delineated the spatial variability and extent of the spatial autocorrelation across both spectral ranges. Longer spatial autocorrelation was observed for the MIR region (range = 117 km) than the NIR region (range = 67 km) which supports the findings reported by Qiu et al. (2013) as the relatively longer wavelengths tended to result in higher spatial variability. In parallel, the spatial autocorrelation of the PC vectors for both spectral ranges were strong which indicates the geostatistical approach could reliably be used to inform calibration sample entry selection for local sample prediction using a soil spectral database.

The spatial autocorrelations of the spectra PC scores were useful in identifying proper and reliable calibration samples per target sample location, improving prediction performances or equivalent to using the entire MSSL. The observed spectral spatial autocorrelation likely reflects the influence of factors such as parent material, agroclimatic, soil management practices, and crop production patterns, which contribute to the spatial variability in specific soil property across the study region. Using the spatial selection approach, mainly the prediction performance was improved for soil property − spectral region combinations which had poor prediction results when using the entire MSSL samples, suggesting that this approach was more reliable. We therefore recommend further research on spatial dependency analysis of the specific responsible wavelength regions for a soil property and its spatial variation in any geographic extent.

4.3. Spectra similarity and clustering performance for calibration sample selection

Spectra similarity analysis using MBL had been evaluated and anticipated to be a better way to subset calibration sample entries for local scales with samples within the domain area of a large SSL. Summerauer et al. (2021) found that despite the spectral similarity of a new sample to those in the large SSL database, reasonable predictions of soil properties were not guaranteed. Their analysis revealed that prediction performance using MBL was subpar when calibration sample entries were selected for the new samples which were geographically distant from the SSL database location. Conversely, when the new samples fell within the geographic coverage of the SSL database, the MBL yielded improved prediction performance. Our findings and others (i.e., Gholizadeh et al., 2016) suggested that the MBL approach could be preferred for improved soil property prediction of new samples when a calibration model was developed within the geographic area covered by the SSL, as compared with the use of the entire database.

On the other hand, it was suggested that predictions of soil properties based on spectra might show enhanced performance within a database organized into homogeneous units using environmental covariates or readily accessible soil characteristics (Ogen et al., 2018; Angelopoulou et al., 2020). Such approaches have been applied in other studies such as digital soil mapping (Dunkl and Ließ, 2022). We implemented clustering of the entire MSSL entries using five environmental covariates, as elaborated in section 2.4.4, into their respective homogeneous groupings. The prediction performances were only better or similar to the use of the entire MSSL for soil properties predicted satisfactorily or better with the use of the entire MSSL. It is vital to note that a wide range of soil property attributes within the homogenous clusters that address the variability within the local and/or homogeneous clusters might be required for a better outcome (Morais et al., 2018). Hence, the question which needs to be addressed in future research is to optimize homogenous clusters with the range and distribution of soil property attributes.

5. Conclusion

Ordinary predictive models, such as PLSR, developed from a large soil spectral library might have high uncertainty because the models learn the general variation to explain the target soil property from the diverse spectral database. Hence, appropriate sub-setting methods could help to use such huge and important databases for the purpose of generating soil information in real-time and at the required spatial resolution from the proximal spectra information acquired. As soil properties vary in space due to varying combination and extent of soil forming factors, their multiresolution spectral signature will vary in space parallelly. We characterized this spatial variation of soil spectral PC vectors and used it for sub-setting calibration samples from a large SSL to predict soil properties at farm level in Morocco. We compared this approach with the use of the entire MSSL, spectral similarity sub-setting (MBL) and clustering with environmental covariates for twelve agricultural important soil properties and two soil scanning instruments. The spatial autocorrelation of the PC vectors varied among soil spectral ranges and gave rise to varying lag distances. Our findings suggested that the spatial and spectra similarity-based calibration sample entry selection improved prediction for those soil properties (OlsenP, P_m3, Exch.K and DTPA.Mn) with low prediction performances when using the entire MSSL database. Otherwise, these sub-setting methods resulted in equivalent prediction performance for soil properties predicted well with the entire MSSL samples. In general, the spatial selection and MBL approaches demonstrated superior performance in predicting most soil properties that were inadequately predicted by using the entire MSSL, indicating the effectiveness of these targeted calibration sample selections methods. This improvement likely stems from the fine-tuning of calibration samples, which effectively addresses local variations arising from the site-specific soil-forming factors and land management practices. The observed spatial variation in soil property and spectral characteristics may be influenced by the scale of observation, landscape heterogeneity and other contextual factors. On the other hand, this study suggests that utilizing clustering based on environmental covariates and soil units can be a viable strategy for sub-setting a large SSL for predicting local samples, particularly for soil properties that achieve satisfactory or better predictions with the entire MSSL. Hence, this study brought up use-case frameworks for large SSLs to predict soil properties per location which have geographic coordinates and spectral recordings. Our study employed PCA to explore the general spatial autocorrelation of spectra data, providing a comprehensive overview of spatial variability across samples. This may generalize the specific variation per soil property whereby PLSR latent variables could offer more targeted understanding of spatial autocorrelations related to specific soil attributes. We anticipated future research investigating the application of PLSR latent variables to enhance the spatial analysis of soil properties, addressing this gap and potentially improving the precision of soil property predictions.

Author contributions

T.G.A., S.M.H, R.S., F.K., R.C. and K.H. conceptualized the study, and acquired and administered project funding. A.H, and A.K carried out soil sampling, processing, and data generation processes. T.G.A., K.H. and T. B. contributed to data analyses, including method development, data visualizations and data interpretation. T.G.A wrote the primary draft of the paper with editing and reviewing inputs from all other authors.

CRediT authorship contribution statement

Tadesse Gashaw Asrat: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Timo Breure:** Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Data curation. **Ruben Sakrabani:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Ron Corstanje:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Kirsty L. Hassall:** Visualization, Formal analysis, Data curation. **Abdellah Hamma:** Writing – review & editing, Methodology, Investigation, Data curation. **Fassil Kebede:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition. **Stephan M. Haefele:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgment

The authors want to thank Mohammed VI Polytechnic University (UM6P) and OCP group, for the technical and financial support, respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2024.117116.

Data availability

Data will be made available upon contacting the OCP Group.

References

Ahmadi, A., Emami, M., Daccache, A., He, L., 2021. Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: A systematic review and meta-analysis. Agronomy 11 (3). https://doi.org/10.3390/agronomy11030433.

Angelopoulou, T., Balafoutis, A., Zalidis, G., Bochtis, D., 2020. From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation — A Review. Sustainability 12 (2), 1–24.

Asrat, T.G., Sakrabani, R., Corstanje, R., Breure, T., Hassall, K.L., Kebede, F., Haefele, S. M., 2023. Spectral soil analysis for fertilizer recommendations by coupling with QUEFTS for maize in East Africa: A sensitivity analysis. Geoderma, Elsevier b.v. 432 (February), 116397.

Baumann, P., Helfenstein, A., Gubler, A., Keller, A., Meuli, R.G., Wächter, D., Lee, J., et al., 2017. Developing the Swiss mid-infrared soil spectral library for local estimation and monitoring. Eur. J. Soil Sci. 68 (6), 840–852.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC - Trends in Analytical Chemistry, Elsevier Ltd 29 (9), 1073–1081.

Bogunovic, I., Trevisani, S., Seput, M., Juzbasic, D., Durdevic, B., 2017. Short-range and regional spatial variability of soil chemical properties in an agro-ecosystem in eastern Croatia. Catena, Elsevier b.v. 154, 50–62.

Bremner, J.M. (1996), "Nitrogen Total", Methods of Soil Analysis: Part 3 Chemical Methods, pp. 1085–1121.

Breure, T.S., Milne, A.E., Webster, R., Haefele, S.M., Hannam, J.A., Moreno-Rojas, S., Corstanje, R., 2021. Predicting the growth of lettuce from soil infrared reflectance spectra: the potential for crop management. Precision Agriculture, Springer US 22 (1), 226–248.

Brodský, L., Klement, A., Penížek, V., Kodešová, R., Borůvka, L., 2011. Building Soil Spectral Library of the Czech Soils for Quantitative Digital Soil Mapping. Soil Water Res. 6 (4), 165–172.

Ciesielski, H., Sterckeman, T., 1997a. A comparison between three methods for the determination of cation exchange capacity and exchangeable cations in soils. Agronomie 17 (1), 9–16.

Ciesielski, H., Sterckeman, T., 1997b. Determination of cation exchange capacity and exchangeable cations in soils by means of cobalt hexamine trichloride. Agronomie 17 (1), 1–7.

Copernicus:Europe's eyes on Earth. (2020), "Land Cover Classification:Global Land Cover", Land Monitoring Service, available at: https://lcviewer.vito.be/2018.

Dangal, S.R.S., Sanderman, J., Wills, S. and Ramirez-lopez, L. (2019), "Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library", pp. 1–23.

Dematte, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M.V., Dalmolin, R.S.D., de Araújo, M. do S.B., da Silva, E.B., et al., 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. Geoderma, Elsevier 354 (October), 113793.

Dotto, A.C., Dematte, J.A.M., Viscarra Rossel, R.A., Rizzo, R., 2020. Soil environment grouping system based on spectral, climate, and terrain data: a quantitative branch of soil series. Soil 6 (1), 163–177.

Dunkl, I., Ließ, M., 2022. On the benefits of clustering approaches in digital soil mapping: an application example concerning soil texture regionalization. Soil 8 (2), 541–558.

Forina, M., Lanteri, S., Casale, M., 2007. Multivariate calibration. J. Chromatogr. A 1158 (1–2), 61–93.

Geocradle. (2018), "Regional Soil Spectral Library", PILOT 2: Improved Food Security – Water Extremes Management (IFS), available at: http://datahub.geocradle.eu/dataset/regional-soil-spectral-library.

Gholizadeh, A., Borůvka, L., Saberioon, M., Vašát, R., 2016. A memory-based learning approach as compared to other data mining algorithms for the prediction of soil texture using diffuse reflectance spectra. Remote Sens. (Basel) 8 No. 4. https://doi.org/10.3390/rs8040341.

Gholizadeh, A., Saberioon, M., Carmon, N., Boruvka, L., Ben-Dor, E., 2018. Examining the performance of PARACUDA-II data-mining engine versus selected techniques to model soil carbon from reflectance spectra. Remote Sens. (Basel) 10 No. 8. https://doi.org/10.3390/rs10081172.

Grafström, A.L., 2018. Balanced and Spatially Balanced Sampling. R Package Version 1 (5), 4. Http://Www.Antongrafstrom.Se/Balancedsampling.

Grunwald, S., Yu, C., Xiong, X., 2018. Transferability and Scalability of Soil Total Carbon Prediction Models in Florida, USA. Pedosphere, Soil Science Society of China 28 (6), 856–872.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-Sinoga, J.D., Zornoza, R., et al., 2016. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil and Tillage Research, Elsevier b.v. 155, 501–509.

Hazelton, P. and Murphy, B. (2016), Interpreting Soil Test Results: What Do All the Numbers Mean?., CSIRO Publishing, available at:https://doi.org/10.1071/9781486303977.

Hengl, T., 2009. A Practical Guide to the Superintendent of Documents Classification System, Government Publications Review Vol. 13. https://doi.org/10.1016/0277-9390(86)90082-8.

Hu, G., Sudduth, K.A., Myers, D.B., Agriscience, C. and Nathan, M. (2013), "Factors Affecting Soil Phosphorus and Potassium Estimation by Reflectance Spectroscopy", American Society of Agricultural and Biological Engineers, Vol. In 2013 Ka No. April 2016.

Iqbal, J., Thomasson, J.A., Jenkins, J.N., Owens, P.R., Whisler, F.D., 2005. Spatial Variability Analysis of Soil Physical Properties of Alluvial Soils. Soil Sci. Soc. Am. J. 69 (4), 1338–1350.

Janik, L.J., Merry, R. and Skjemstad, J.O. (1998), "Can mid infrared diffuse reflectance analysis replace soil extractions?", Vol. 38 No. January, pp. 681–696.

Jin, X., Li, S., Zhang, W., Zhu, J. and Sun, J. (2020), "Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms", Applied Sciences (Switzerland), Vol. 10 No. 4, available at:https://doi.org/10.3390/app10041520.

Johnson, J.M., Vandamme, E., Senthilkumar, K., Sila, A., Shepherd, K.D., Saito, K., 2019. Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-Saharan Africa. Geoderma 354 (May). https://doi.org/10.1016/j.geoderma.2019.06.043.

Kennard, R.W., Stone, L.A., 1969. Computer Aided Design of Experiments. Technometrics 11 (1), 137–148.

Kerry, R., Oliver, M.A., 2007. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma 140 (4), 383–396.

Knadel, M., Deng, F., Thomsen, A., Greve, M.H., 2012. Development of a Danish national vis-NIR soil spectral library for soil organic carbon determination. Digital Soil Assessments and beyond 403–408.

Laboratories, A., 2019. Interpreting a Soil Test Report. available at: https://www.agvise.com/wp-content/uploads/2020/09/guide-soil-interpretation-report-2019.pdf.

Lark, R.M., Webster, R., 1999. Analysis and elucidation of soil variation using wavelets. Eur. J. Soil Sci. 50 (2), 185–206.

Liland, H., Mevik, B.-H., Wehrens, R. and Hiemstra, P. (2023), Package 'Pls'.

Lin, L.-I.-K., 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45 (1), 255.

Lindsay, W.L., Norvell, W.A., 1972. Development of a DTPA soil test for zinc, iron, manganse and copper. Soil Sci. Soc. Am. Proc. 36, 778–783.

Mamassi, A., Balaghi, R., Devkota, K.P., Bouras, H., El-Gharous, M., Tychon, B., 2023. Modeling genotype × environment × management interactions for a sustainable intensification under rainfed wheat cropping system in Morocco. Agriculture and Food Security, BioMed Central 12 (1), 1–23.

Mehlich, A., 1984. Mehlich 3 soil test extractant: A modification of Mehlich 2 extractant. Commun. Soil Sci. Plant Anal. 15 (12), 1409–1416.
Mokhtar, M.A.E., Laouane, R.B., Anli, M., Boutasknit, A., Fakhech, A., Wahbi, S., 2022. "Climate change and its impacts on oases ecosystem in Morocco", Research Anthology on Environmental and Societal Impacts of. Clim. Change 1103–1131.
Moloney, J.P., Malone, B.P., Karunaratne, S., Stockmann, U., 2023. Leveraging large soil spectral libraries for sensor-agnostic field condition predictions of several agronomically important soil properties. Geoderma, Elsevier b.v. 439 (March), 116651.
Morais, P., Inácio, E., Filho, F. and Rocha, M. (2018), "Digital Soil Mapping of Soil Properties in the ' Mar de Morros ' Environment Using Spectral Data", pp. 1–19.
Næs, T., 1987. The design of calibration in near infra-red reflectance analysis by clustering. J. Chemometr.
Nawar, S., Mouazen, A.M., 2018. Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. Computers and Electronics in Agriculture, Elsevier 151 (June), 469–477.
Nelson, D.W. and Sommers, L.E. (1996), "Total Carbon, Organic Carbon, and Organic Matter", Methods of Soil Analysis: Part 3 Chemical Methods, pp. 961–1010.
Ng, W., Husnain, A.L., Siregar, A.F., Hartatik, W., Sulaeman, Y., Jones, E., et al., 2020. Developing a soil spectral library using a low-cost NIR spectrometer for precision fertilization in Indonesia. Geoderma Regional, Elsevier b.v. 22, e00319.
Ogen, Y., Zaluda, J., Francos, N., Goldshleger, N., Ben-Dor, E., 2019. Cluster-based spectral models for a robust assessment of soil properties". Geoderma, Elsevier 340 (August 2018), 175–184.
Olsen, S., 1954. Estimation of available phosphorus in soils by extraction with sodium bicarbonate. US Department of Agriculture., Vol. 939, pp. 1–20.
Pebesma, E. and Graeler, B. (2023), Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation: Package 'Gstat', available at: https://github.com/r-spatial/gstat/issues/.
Pebesma, E.J., Wesseling, C.G., 1998. Gstat: A program for geostatistical modelling, prediction and simulation. Comput. Geosci. 24 (1), 17–31.
Potash, E., Guan, K., Margenot, A.J., Lee, D.K., Boe, A., Douglass, M., Heaton, E., et al., 2023. Multi-site evaluation of stratified and balanced sampling of soil organic carbon stocks in agricultural fields. Geoderma, Elsevier b.v. 438 (May), 116587.
Qiu, B., Zeng, C., Chen, C., 2013. Comparative spatio-spectral heterogeneity analysis using multispectral and hyperspectral airborne images. Geo-spatial Inf. Sci. 16 (2), 83–90.
Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R.A.V., Demattê, J.A.M., Scholten, T., 2013a. Distance and similarity-search metrics for use with soil vis-NIR spectra. Geoderma, Elsevier b.v. 199, 43–53.
Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013b. The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets. Geoderma 195–196, 268–279.
Ramirez-lopez, L., Stevens, A., Orellano, C., Viscarra, R., Lobsey, C., Wadoux, A., 2022. resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics. R Package Version 1.2.2. 1 (2), 1–47.
Sáiz-Abajo, M.J., Mevik, B.H., Segtnan, V.H., Næs, T., 2005. Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. Anal. Chim. Acta 533 (2), 147–159.
Sarathjith, M.C., Das, B.S., Wani, S.P. and Sahrawat, K.L. (2014), "Dependency Measures for Assessing the Covariation of Spectrally Active and Inactive Soil Properties in Diffuse Reflectance Spectroscopy", available at:https://doi.org/10.2136/sssaj2014.04.0173.
Savvides, A., Corstanje, R., Baxter, S.J., Rawlins, B.G., Lark, R.M., 2010. The relationship between diffuse spectral reflectance of the soil and its cation exchange capacity is scale-dependent. Geoderma, Elsevier b.v. 154 (3–4), 353–358.
Shenk, J.S., Westerhaus, M.O., Berzaghi, P., 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. J. Near Infrared Spectrosc. 5 (4), 223–232.
Shepherd, K.D., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J., Ge, Y., 2022. A global soil spectral calibration library and estimation service. Soil Security, Elsevier Ltd 7 (April), 100061.
Shepherd, K.D., Walsh, M.G., 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. Soil Sci. Soc. Am. J. 66 (3), 988–998.
SoilSpec4GG. (2020), "Open Soil Spectral Library (OSSL)", available at:https://doi.org/10.7560/320624-010.
Song, Y., Shen, Z., Wu, P. and Rossel, R.A.V. (2021), "Wavelet geographically weighted regression for spectroscopic modelling of soil properties", Scientific Reports, Nature Publishing Group UK, No. 0123456789, pp. 1–11.
Spaargaren, O.C., Batjes, N.H., 1995. Report on the Clsification Into Fao-Unesco Soil Units of Profiles Selected From the. Nrcs Pedon Data Base for Igbp-Dis.
Stenberg, B., Rossel, R.A.V., Mouazen, A.M., Wetterlind, J., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and Near Infrared Spectroscopy in Soil Science. Adv. Agron. 107 (C), 163–215.
Stevens, A. and Lopez, L.R. (2022), An Introduction to the Prospectr Package.
Stevens, A., Nocita, M., Tóth, G., Montanarella, L. and van Wesemael, B. (2013), "Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy", PLoS ONE, Vol. 8 No. 6, available at:https://doi.org/10.1371/journal.pone.0066409.
Summerauer, L., Baumann, P., Ramirez-Lopez, L., Barthel, M., Bauters, M., Bukombe, B., Reichenbach, M., et al., 2021. The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis. Soil 7 (2), 693–715.
Sun, Y., Yuan, M., Liu, X., Su, M., Wang, L., Zeng, Y., Zang, H., et al., 2021. A sample selection method specific to unknown test samples for calibration and validation sets based on spectra similarity. Spectrochimica Acta - Part a: Molecular and Biomolecular Spectroscopy, Elsevier b.v. 258 (119870), 1–8.
Sun, B., Zhou, S., Zhao, Q., 2003. Evaluation of spatial and temporal changes of soil quality based on geostatistical analysis in the hill region of subtropical China. Geoderma 115 (1–2), 85–99.
Thomas, C.L., Hernandez-Allica, J., Dunham, S.J., McGrath, S.P., Haefele, S.M., 2021. A comparison of soil texture measurements using mid-infrared spectroscopy (MIRS) and laser diffraction analysis (LDA) in diverse soils. Scientific Reports, Nature Publishing Group UK 11 (1), 1–12.
Vasu, D., Singh, S.K., Sahu, N., Tiwary, P., Chandran, P., Duraisami, V.P., Ramamurthy, V., et al., 2017. Assessment of spatial variability of soil properties using geospatial techniques for farm level nutrient management. Soil and Tillage Research, Elsevier b.v. 169, 25–34.
Vestergaard, R.-J., Vasava, H.B., Aspinall, D., Chen, S., Gillespie, A., Adamchuk, V., Biswas, A., 2021. Evaluation of Optimized Preprocessing and Modeling. Sensors 21 (20), 6745.
Viscarra Rossel, R.A., Behrens, T., Ben-dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., et al., 2016. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230.
Wang, G., Wang, W., Fang, Q., Jiang, H., Xin, Q., Xue, B., 2018. The application of discrete wavelet transform with improved partial least-squares method for the estimation of soil properties with visible and near-infrared spectral data. Remote Sens. (Basel) 10 (6). https://doi.org/10.3390/rs10060867.
Wijewardane, N.k., Ge, Y., Wills, S., Libohova, Z., 2018. Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. Soil Sci. Soc. Am. J. 82 (3), 722–731.
Ye, H., Huang, W., Huang, S., Huang, Y., Zhang, S., Dong, Y., Chen, P., 2017. Effects of different sampling densities on geographically weighted regression kriging for predicting soil organic carbon. Spatial Statistics, Elsevier b.v. 20, 76–91.
Zeng, R., Rossiter, D.G., Yang, F., Li, D.C., Zhao, Y.G., Zhang, G.L., 2017. How accurately can soil classes be allocated based on spectrally predicted physio-chemical properties? Geoderma, Elsevier 303 (March), 78–84.