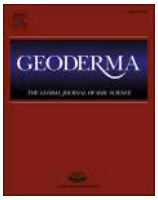
Geoderma 434 (2023) 116491

Contents lists available at ScienceDirect

Geoderma

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



A systematic approach to predicting and mapping soil particle size distribution from unknown samples using large mid-infrared spectral libraries covering large-scale heterogeneous areas

Felipe B. de Santana [a,*], Rebecca.L. Hall [a], Victoria Lowe [b], Margaret A. Browne [b], Eric C. Grunsky [b], Mairéad M. Fitzsimons [b], Vincent Gallagher [b], Karen Daly [a]

[a] Environment, Soils and Land Use Department, Teagasc, Johnstown Castle Research Centre, Wexford Y35 TC97, Ireland
[b] Geological Survey Ireland, Block 1, Booterstown Hall, Booterstown, Blackrock, Co Dublin, A94 N2R6, Ireland

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Soil particle size
MIR
Soil texture
Machine learning
Control chart
Chemometrics

ABSTRACT

Soil spectroscopy has the potential to replace wet chemical methods. However, high prediction accuracy of unknown samples is necessary to confidently map predicted values across large spatial scales. This study developed a systematic approach using MIR spectroscopy to predict soil particle size of 9,009 unknown samples, representing an area of approximately 35,716 km^2 in Ireland. A systematic approach for MIR spectroscopy combined with chemometrics was augmented by including spectral control charts to identify unrepresentative spectra in predicting of unknown samples. Moreover, ~2% of the predicted values (external validation) were selected, covering a wide range of clay, sand and silt, and analysing them using the classical reference method. This step assesses the accuracy of predicted values and provided confidence when projecting predicted values onto soil particle size maps at multiple regional scales.

The MIR models were based on the support vector regression algorithm, and were built using a calibration dataset of ~1,000 mineral and organo-mineral (MOM) soils, i.e., non-peat soils, analysed using the pipette method. The model was used to predict soil particle size from 9009 unknown samples described as MOM soils. Spectral control charts correctly identified 97.59% of the peat soils (3078 out of 3154 samples) as 'out of control' samples, which the spectral model should not predict, and about 90% of the MOM soils (5254 out of 5855 samples) were classified as 'under control' samples, which the spectra models can predict. The accuracy calculated for the ~2% of samples selected from the MOM unknown samples (n = 5254) was similar to the accuracy in the internal validation set (n = 280), with R^2_{val} values of 0.90, 0.84 and 0.71, RMSEP values of 2.21, 4.31 and 3.91%; $RPIQ_{val}$ values of 4.97, 3.71 and 2.30, for clay, sand and silt prediction, respectively. This systematic approach can predict soil attributes using large spectral libraries, providing confidence for building regional and national scale soil maps.

1. Introduction

Soil particle size distribution is a key physical characteristic that regulates soil function and defines soil texture, which is the composition of mineral particles <2 mm in diameter (fine earth fraction). According to the ADAS-UK classification system, particles <2 mm can be classified as clay (<0.002 mm), silt (0.002–0.63 mm) and sand (0.063–2 mm) (Hatten and Liles, 2019). The soil texture is stable and cannot be easily altered directly or indirectly by management practices (Zhang et al., 2022). It is an important variable in environmental and agronomic models due to its influence on soil hydraulic characteristics, compaction, erosion susceptibility, soil moisture movement, nutrient retention, contaminant transport, microbiological activity and, in mineral soils, soil organic carbon accumulation (Dunne et al., 2020; Polakowski et al., 2021; Zhang et al., 2022).

High-resolution information of soil's physical and chemical attributes is mandatory for regional and national digital soil mappings; and the migration to Agriculture 4.0, which incorporates the evolution of precision agriculture with the use of several sensors, and automated machines to collect precise and accurate data used to make decisions

* Corresponding author.
E-mail address: Felipe.bachiondesanta@teagasc.ie (F.B. de Santana).

https://doi.org/10.1016/j.geoderma.2023.116491
Received 17 November 2022; Received in revised form 21 April 2023; Accepted 23 April 2023
Available online 5 May 2023

(Valle and Kienzle, 2020). This, impetus results in high demand for soil analysis, including soil particle size analysis. However, traditional analytical methods used to determine soil particle size are time-consuming and expensive, requiring significant investments from farmers or the government to build regional or national/semi-national maps (large-scale) of high-resolution soil attributes, i.e. <5 km for national surveys (O'Rourke et al., 2016; Valle and Kienzle, 2020). Soil particle size can be considered one of the most problematic soil attributes for large-scale studies or high spatial resolution monitoring. Traditional techniques can require up to two to three months to analyse 100 samples depending on laboratory infrastructure. This lengthy analysis time and associated cost (typically €35.00 per sample excluding taxes) can impede the development of high-resolution mapping of particle size and texture. Consequently, soil texture maps at national and regional scales are typically based on low-resolution sampling, extrapolated across larger scales (de Santana and Daly, 2022; Creamer et al., 2016).

To overcome the high costs and time associated with analysis, vibrational spectroscopy techniques, such as visible and near infrared (vis–NIR, 350–2500 nm, 28,571–4000 cm^{-1}) and mid infrared (MIR, 2500–16,667 nm, 4000–600 cm^{-1}) have been used to quantify or semi-quantify the soil particle size and several other soil attributes from a single spectrum collected in few seconds/minutes (de Santana et al., 2021; Metzger et al., 2020; O'Rourke et al., 2016; Stenberg et al., 2010). These spectroscopy techniques combined with chemometrics/machine learning have the potential to increase sample density without incurring substantial additional costs (de Santana et al., 2019; O'Rourke et al., 2016).

The acquisition of reliable high-resolution soil particle size and texture information at large spatial scales is desirable in agri-environmental modelling. Although spectroscopy can predict this parameter, studies that use spectral models to predict unknown samples using large spectral libraries over a wide spatial area are scarce. In general, most spectroscopy studies for soil analysis build the regression models using the spectra and the reference values (calibration step) and then validate the model using an internal validation set (spectra and reference values). There are also several studies that compare the accuracy of regression models using different approaches or chemometric algorithms, such as partial least squares regression (PLSR), back propagation neural network (BPNN), multiple linear regression (MLR), multivariate adaptive regression splines (MARS), support vector regression (SVR), cubist, random forest (RF) boosted trees (BT), and artificial neural networks (ANN), among others (Brodský et al., 2013; de Santana et al., 2018, 2021). These studies provided strategies and algorithm approaches to improve the model's accuracy. However, to widen the application of spectroscopy for predicting soil particle size, there is a need for the development of protocols/methodologies that can reliably predict values from spectral information of external samples, i.e., samples that were external to the calibration model and not submitted for reference analysis, also known as 'blind samples' or 'unknown samples'.

Ramirez-Lopez et al. (2019) created soil particle size maps at farm scales using only information from the unknown samples' spectra, that is, samples without reference values. In this study the authors used the vis-NIR spectra (350–2500 nm) combined with memory-based learning (MBL) algorithm to determine the soil particle size of 910 surface (0–0.20 m) and subsurface (0.80–1.00 m) samples collected from 458 pits. From these samples, about 345 samples from each depth were selected to build the spectra models (calibration samples) and the remaining samples, ~115 samples from each depth, were used as external/unknown samples. The vis-NIR spectra was collected from dried and sieved 2 mm soil samples and the soil particle size reference values of the calibration samples were obtained using the hydrometer method. In order to restrict the sum of the individual predictions (sand, silt and clay) to 100%, the authors used the additive log-ratio (*alr*) transformation before build the regression models (Aitchison, 1986).

After predicting the unknown samples using the spectroscopy models, these samples were also analysed using the hydrometer method. These external samples were predicted with root mean squared error (RMSEP) values of 7.20, 2.85 and 6.31%; R^2_{val} of 0.88, 0.77 and 0.87 for sand, silt and clay determinations, respectively. The results obtained by Ramirez-Lopez et al. (2019) show that vis-NIR spectroscopy combined with chemometrics can be used to predicted soil particle size with good accuracy at farm scales, reducing costs and time required to map these soil attributes.

Most of the current literature uses NIR spectroscopy to build spectral libraries. However, the results obtained using MIR spectroscopy are generally more accurate, making MIR more appropriate for routine laboratories which require high accuracy. Compared to NIR spectroscopy, the MIR region is more affected by inadequate soil preparation, i.e. moisture in the samples. Furthermore, MIR spectroscopy presents a smaller radiation spot and more hindrances in customisation, i.e. connecting bifurcated optical fibre probes (de Santana and Daly, 2022). In this sense, the best option must be evaluated for each application.

The use of NIR or MIR spectroscopy to predict unknown samples are not common in the literature, and studies that use one of these techniques to predict unknown samples at large scales (regional or national) for digital soil maps are scarce in the literature. According to Poppiel et al. (2022) the shortage of these kind of studies, and routine, robust spectral methods to predict unknown samples can be attributed mainly to (1) the lack of standards and protocols, (2) the costs associated with building large spectral libraries, especially at large scales, (3) lack of professionals with expertise in both chemometrics and analytical chemistry and (4) the lack of capacity of traditional soil laboratories in spectral methods. Furthermore, for soil spectroscopy to replace wet chemical methods for some attributes, there is a need to achieve high prediction accuracy of unknown samples but also to identify unrepresentative samples during the process. This will increase the confidence in the values predicted by the spectral models (MIR or NIR) and encourage the transition from wet chemical methods.

To address this gap, the objective of this study was to develop and validate a methodology for MIR spectral models that ensures a high level of accuracy in predicting soil particle size from large spectral libraries over wide spatial areas. To meet this objective, this study developed a soil particle size regression method using a small subset (~10%) of the Tellus spectral library (TSL), which comprises ~10,000 shallow topsoil samples collected by the Tellus survey. The Tellus survey is a national survey programme that aims to gather geochemical and geophysical data across Ireland and examine the chemical and physical properties of Irish soil, rocks and water.

The regression model will only predict representative (under control) samples among the TSL unknown samples (n = 9,009); for this, we implemented the control charts based on PCA proposed by Laursen et al. (2011) to identify unrepresentative (out of control) samples. The authors used these control charts to monitor the quality of a specific pharmaceutical drug using reverse-phase high-performance liquid Chromatography with UV detector. They highlighted the conventional univariate control charts focussing on particular peaks, forcing the analyst to inspect a control chart for each peak. Besides that, in specific situations, the relationship between the peaks can be affected and not identified by the univariate control charts. In contrast, the control charts based on PCA could simultaneously identify all the peaks and their relationship, besides identifying impurity in ranges not evaluated from the univariate control chart.

By adding this step, we improved the confidence of the predicted results obtained using spectroscopy methods. To the best of our knowledge, this is the first study that used MIR spectroscopy to predict a large number of unknown samples to build soil particle size maps at a large scale with high resolution and a high degree of accuracy.

2. Materials and methods

2.1. Tellus survey soil sampling

A total of 9,993 shallow topsoils (5 – 20 cm) samples were collected from agricultural and non-agricultural lands in the Border, Midlands, West and Periurban regions comprising approximately 35,716 km^2 representing the northern half of Ireland as part of Geological Survey Ireland's Tellus programme (GSI, 2021). Regional and periurban samples were collected on a 4 km^2 and 1 km^2 grid, respectively. The predominant Irish Great Soil Groups used in this study are Peat (n = 3181), Grey Brown Podzolic (n = 2433), Gley (n = 1156), Brown Earth (n = 754), Brown Podzolic (n = 317), Complex (n = 307), Podzol (n = 231) and Brown Earth L.B.S (n = 231) (Gardiner and Radford, 1980).

This large variability is attributed to the underlying subsoils and the bedrock from which the soils are derived, such as granite, gneiss, schist, limestone, sandstone and shale. In addition, there are large areas underlain by blanket bog in upland areas and raised bog in the midlands (Gardiner and Radford, 1980).

2.2. Soil particle size analysis

2.2.1. Gravimetric pipette method

The soil particle size reference analysis of the calibration samples (n = 984) was performed in an analytical laboratory accredited to ISO/IEC 17025:2005. This laboratory followed the standard ADAS-UK pipette method procedure (ISO 11277), using ~20 g of oven-dried soil samples sieved to <2 mm (fine earth fraction) for the analysis. Hydrogen peroxide was used to oxidise the organic matter (OM), after clay aggregates were dispersed using a sodium hexametaphosphate solution and the dispersed material was sieved using a 63 μm sieve to retain the sand fraction. The remaining fraction containing silt and clay was determined through sedimentation for 7 h. If free carbonates were present in the soil, after the OM oxidation, 10% hydrochloric acid was added until effervescence ceased (Massey et al., 2013).

2.2.2. PARIO method

The PARIO method adopts the same sample preparation methodology used for the Pipette method, differing only in the sedimentation step. This system is based on the integral suspension pressure (ISP) method, which measures the suspension pressure continuously at a given depth without any mechanical interference, not requiring the use of a pipette to get aliquots from the soil solution that require further drying and weighing. Using the system pressure, an inverse simulation is used by the PARIO method to calculate the soil particle size distribution (Durner et al., 2017; Durner and Iden, 2021). Ninety-six mineral and organo-mineral (MOM) unknown samples from TSL classified as under control samples by the control chart, were selected and analysed using the PARIO method for laboratory determination of the soil particle size.

2.3. Tellus spectral library (TSL)

2.3.1. MIR spectra acquisition

A large spectral library of 9,993 topsoil samples from the Tellus programme was developed in the MIR spectrometer using the following method. The fine earth fraction was ball-milled using a planetary ball mill to a particle size nominally 95 % <0.032 mm. Ball-milled soils were scanned in the MIR range in the diffuse reflectance mode using the FTIR Spectrometer INVENIO S® (Bruker, Massachusetts - USA). The ball-milled soils were filled into aluminium micro-cups of 7 mm internal diameter and its surface was smoothed with a blade before the spectral acquisition. A globar (u-shaped silicon carbide - SiC) was used as a radiation source and a mercury cadmium telluride (MCT) detector was used for the diffuse reflection measurements in the MIR (4000–600 cm^{-1}). The FTIR spectrometer was coupled with an autosampler composed of a microplate reader extension for high-throughput screening (HTS-XT), also from Bruker (Bruker, Massachusetts - USA), allowing analyse of 60 samples per hour.

Using this system the radiation beam strikes the soil samples at an angle of 90° in the centre of the micro-cups with a diameter of ~2.5 mm. At the beginning of each batch of analysis of 23 samples, the background was obtained using a gold reference disk, and the radiance of each soil spectra was collected using 32 scans at a spectral resolution of 1.43 cm^{-1}, totalling 2378 variables per spectra.

2.3.2. Spectral pre-processing

Prior to modelling, the reflectance MIR spectra were converted to pseudo absorbance, log (1/Reflectance). Then some pre-processing algorithms were tested, such as smoothing, first and second derivative Savitzky–Golay smoothing with different windows sizes, standard normal variate (SNV) and multiplicative signal correction (MSC) (Rinnan et al., 2009). For this, a regression model was built for each pre-processing, and the pre-processing with the lowest error in the cross-validation was the first derivative Savitzky–Golay smoothing, with a window size of 11 points (~14 cm^{-1}) and second order polynomial (Savitzky and Golay, 1964).

2.4. Sample selection

Three subsets of soil samples from the Tellus archive were selected for traditional laboratory analysis of soil particle size. The aims of this were (1) to develop a reference database for soil particle size for model building, (2) to compare the gravimetric pipette and PARIO methods for soil particle size determination and (3) to calculate the accuracy of unknown samples in the external validation set using a subsampling strategy. A flow chart indicating how these subsets were set is shown in Fig. 1, and their selection are described here.

2.4.1. Laboratory reference database

A laboratory reference database on soil particle size was developed using samples from the Tellus sample archive. Sample selection for this step included an initial screening of the Tellus archive to exclude peat soils using a visual soil assessment from the field soil surveyors' reports. Samples described as having peat content were excluded, leaving 6,788 samples classed as MOM (non-peat) soil for analysis. From these MOM soil samples, a geospatially random selection was made in a GIS, yielding 984 samples (Fig. S1). These samples were used as reference samples for particle size analysis for building the spectral models.

2.4.2. Laboratory comparison of gravimetric methods; pipette and PARIO method

In the interest of reducing the time required for soil particle size analysis, the PARIO method was used instead of the Pipette method to determine the accuracy in the external samples. In order to ascertain if the two laboratory methods yield consistent results, fifteen samples that were analysed using the Pipette method were also analysed using the PARIO method. These samples were selected from the calibration set in order to cover low, mean and high levels of clay, sand and silt.

Since the results were carried out by two different laboratories using variations on the gravimetric method (pipette and sedimentation) it was necessary to compare the results obtained from the PARIO and Pipette methods. The Shapiro-Wilk test showed that individually, each soil particle size parameter (clay, sand and silt) followed a normal distribution, therefore the paired sample *t*-test was selected to assess statistical differences between both methods (Skoog et al., 2014). The hypothesis evaluated using the Paired Sample *t*-test were:

H_0: The mean difference between paired observations is zero.
H_1: The mean difference between paired observations is not zero.

2.4.3. External model validation

In order to check the accuracy of the unknown samples, a subset of

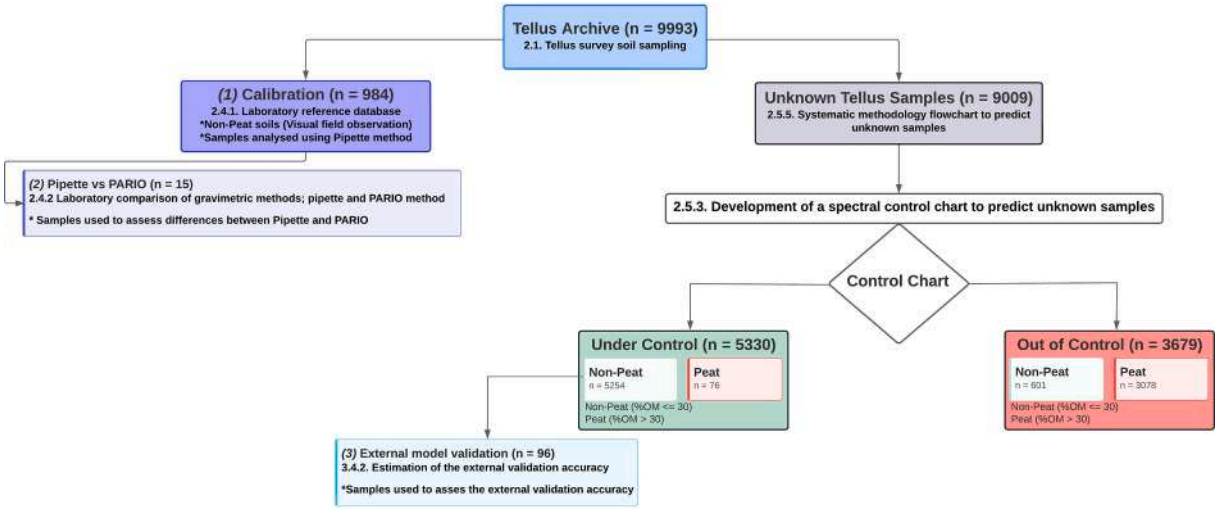


Fig. 1. Flow chart indicating how the subsets selected for traditional analysis were set.

96 MOM samples classified as under control by the control chart was selected. These samples were selected based on their predicted particle size values to represent the full range of sand, silt and clay values in the model. They were then subjected to laboratory analysis to determine particle size using the PARIO, a gravimetric method that is also compared with the pipette method in the study.

2.5. Data analysis

2.5.1. Centred Log-Ratio transformation

Log-ratio transformation can be used in compositional data to overcome the limitations of representing compositional data, such as soil particle size, as raw percentages or proportions. Each variable represents a proportion of the whole composition and by definition must sum to a constant, i.e. 100%. These relative measures are restricted to 0 – 100% numerical space, also known as the simplex. That is, if the sand content, for example, increases, the values of silt and clay must decrease to maintain the constant sum restriction. To deal with this Aitchison (1986) proposed the use of Log-ratio transformation between two or more fractions instead of the use of individual fraction, enabling the use of standard unconstrained multivariate statistics, like PLSR and SVM, among other chemometric algorithms (Jaconi et al., 2019).

Among the log-ratio transformations the additive log-ratio transformation (*alr*) and the centred log-ratio (*clr*) are the two most common transformations used in soil analysis (Mert et al., 2016; Ramirez-Lopez et al., 2019). The *clr* transformation uses the geometric mean of the sample components as the reference (denominator), while the *alr* transformation uses a single component, i.e. 100 – y, as the reference (Quinn et al., 2019). Since the soil particle size data can show large variations in their composition the *clr* transformation (Eq. (1)) was used in this study.

$$z_i = \left[ln\frac{y_{i,1}}{g(y_i)}, \cdots, ln\frac{y_{i,j}}{g(y_i)} \right] \tag{1}$$

Where y_i represents the *i*-th sample, $y_{i,1}$ represent the first parameter from the *i*-th sample, $y_{i,j}$ represents the j-th value from this sample and $g(y_i)$ represents the geometric mean of the *i*-th sample, that is: $\sqrt[j]{y_{i,1}\cdots y_{i,j}}$.

After obtaining the soil particle size predicted values the *clr* inverse function (Eq. (2)) is used to return the transformed data (z_i) into the original space (y_i). These values were normalised for their sum to be 100 and then used to build the predicted maps and calculate the accuracy parameters.

$$y_i = \frac{e^{z_{i,1}}}{\sum_{j=1}^{J} e^{z_{i,j}}}, \cdots, \frac{e^{z_{i,j}}}{\sum_{j=1}^{J} e^{z_{i,j}}} \tag{2}$$

Where *J* is the total number of columns.

Log-ratio transformations can occasionally give similar results to the models built without the transformations. However, in some cases, especially when dealing with values close to 0 or 100%, the transformations can improve the prediction results and are often used to overcome this. For this reason, this approach was used here and included in the proposed systematic approach for the soil particle size analysis.

2.5.2. Support vector regression and model accuracy

Similar to the support vector machine (SVM) the SVR uses a kernel transfer function, in this case radial basis function (Eq. (3)), to project the **X**-block in to a high-dimensional feature space. In Eq. (3) the x_i represents the spectral data of the *i*-th sample and γ defines the width of the Gaussian function (de Santana and Daly, 2022; Schölkopf and Smola, 2002).

$$K(x, x_i) = \exp\left(-\gamma||x - x_i|^2\right)\gamma > 0 \tag{3}$$

However, instead of calculating the hyperplane to separate binary classes like in the SVM for classification proposes, here the algorithm calculates a hyperplane w that fits the maximum number of points, minimising the distance between the hyperplane with the samples projected in the feature space (Eq. 4).

Minimise: $\frac{1}{2}\|w\| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$ (4).

Where (C) is the cost parameter that determines the influence of each sample, defining the model complexity, ξ_i and ξ_i^* are called slack variables and are used as a threshold distance between the hyperplane and the samples considered as support vectors (Schölkopf and Smola, 2002). The sample prediction is done using the hyperplane (w) as a slope that multiplies the result of the kernel function for the *i*-th sample x_i, as shown in Eq. (5).

$$y_i = w.K(x_i) + b \tag{5}$$

As shown in Eq. (6) during the hyperplane calculation the SVR also aims to minimize the prediction error ($y_i - \hat{y}_i$) with an accept error ε.

$$\text{Subject to}: \begin{Bmatrix} y_i - w.K(x_i) - b \le \varepsilon + \xi_i \\ w.K(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{Bmatrix} \tag{6}$$

The SVR hyperparameters ε, C and γ must be optimized to obtain

feasible results. However, instead of testing several combinations of these hyperparameters (grid search method), the Bayesian optimization algorithm was used in this work. This algorithm aims to minimize the objective function, by selecting hyperparameters combinations that show a higher probability of minimizing the objective function, which is directly related to the root mean square error (RMSE, Eq. (7)) (MathWorks, 2017).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2} \tag{7}$$

The accuracy of the regression models was evaluated using the RMSE in calibration (RMSEC) and validation (RMSEP), coefficient of determination (R^2, Eq. (8)), and ratio of performance to interquartile distance (RPIQ, Eq. (9)).

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \overline{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{8}$$

$$\mathrm{RPIQ} = \frac{Q3 - Q1}{\mathrm{RMSE}} \tag{9}$$

In equations 7–8, $\widehat{y_i}$ is the predicted value, y_i is the reference value, $\overline{y_i}$ its average, and n is the total number of samples. Both R^2 and RMSE are the classical accuracy parameters used to assess the accuracy of regression models. The R^2 can be used to give an estimative about the relationship between the reference and the values predicted by the spectral models, while the RMSE is used to provide an estimative about the error in predictions using the spectral models (de Santana and Daly, 2022).

The RPIQ (Eq. (9)) is the accuracy parameter which summarises the quality of the regression model in a single value, being calculated by the ratio of the interquartile distance (Q3 – Q1) and the RMSE. Using the RPIQ values Ludwig et al. (2017) proposed thresholds to classify the quality of the regression models. Models with RPIQ > 4.05 can be classified as 'Excellent Model', models with RPIQ values between 3.38 and 4.04 are classified as 'Good Model', RPIQ values between 2.70 and 3.37 can be considered 'Approximate Quantitative Model' and regression models with RPIQ values between 2.02 and 2.69 should be used to distinguishing between high and low values, that is a 'Fair Model'.

2.5.3. Development of a spectral control chart to predict unknown samples

Principal Component Analysis (PCA) aims to reduce the dimensionality of the original data (**X**) by projecting the original data onto a new coordinate system, named principal components (PC). In the PCA space, each sample is represented by new coordinates denominated scores (**T**), and the linear combinations used to calculate **T** are denominated loadings (**P**). Geometrically the **P** are the cosines of the angle between the original variables and the PC axis, reflecting the importance of each variable in this new projection (Laursen et al., 2011). As shown in Eq. (10) the unmodelled part of the matrix is denominated error (**E**), which is calculated by the difference between **X** and $\widehat{\mathbf{X}}$, that is **X - T P**$^{\mathbf{T}}$.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{10}$$

In this work we used the PCA results (**T**, **P** and **E**) to build two control charts, the first one based on **T** which was used to calculate the Hotelling T^2 values, and the other one based on the **E** which was used to calculate the Q residual values. The Hotelling T^2 control chart measures the Mahalanobis distance from a new sample to the centre of calibration samples, to monitor systematic variations in the model. In other words, to check if the intensities of the peaks already presented in the calibration samples are present in the spectra of unknown samples with the same intensity (Laursen et al., 2011). If a spectral sample falls outside this chart this means there are one or more peaks with lower or higher intensity compared to the calibration samples. The Hotelling T^2 is calculated using Eq. (11) and the T^2 critical value can be calculated using Eq. (12).

$$T^2_{Hotelling} = \frac{\mathrm{t_i^T}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathrm{t_i}}{\mathrm{n}-1} \tag{11}$$

$$T^2_{critical} = \frac{K(n-1)}{n-1}F_{(K,n-K)\alpha} \tag{12}$$

Where t_i is the scores values for the i-th sample, **T** is the score matrix, n is the number of samples, K is the number of PCs and F is F-distribution value at a α significance level (in this study 0.01) with n by $n - K$ degrees of freedom.

The Q control chart monitors the residual vector (e_i) calculated by a new sample projected in the PCA model, which is built using the under control samples, in this case, the calibration samples. If a sample fails in this control chart, then there are one or more peaks in the spectra that were not presented in the calibration samples, enabling the monitoring of non-systematic variations in the model.

The Q_i residual is calculated using the Eq. (13), where the e_i represents the residual for each sample. The $Q_{critical}$ was calculated using Eq. (14), which considers that the Q residuals follow a normal distribution to approximate a weighted χ-square distribution:

$$Q_i = e_i e_i^T \tag{13}$$

$$Q_{critical} = \theta_1^{\,h_0}\sqrt{\left[1 - \theta_2 h_0\left(\frac{1-h_0}{\theta_1^2}\right) + \sqrt{\frac{z_\alpha 2\theta_2 h_0^2}{\theta_1}}\right]} \tag{14}$$

The covariance of **E**, represented here by $Cov(\boldsymbol{E})$, θ_1 is the sum of the values on the main diagonal of $Cov(\boldsymbol{E})$, defined as trace, θ_2 and θ_3 are the trace of $Cov(\boldsymbol{E}^{\circ 2})$ and trace of $Cov(\boldsymbol{E}^{\circ 3})$, respectively. Where $^{\circ 2}$ and $^{\circ 3}$ are the Hadamard product, also known as the element-wise product. The z_α is the standardized normal variable at a α significance level, in this case 0.01, and $h_0 = \frac{2\theta_1\theta_3}{3\theta_1^2}$. Spectra that fail in one of the control charts are considered unrepresentative and are not predicted by the regression models.

2.5.4. Development of a spectral control chart to check the instrument repeatability

The control charts based on PCA were also used to identify possible variations in the instrument over time. However, instead of using the soil samples in the control chart, Acetaminophen – USP specification was used as the reference compound. The acetaminophen was scanned in FTIR twice daily, in the morning before the first analysis and again in the evening after the last batch. Then after finishing the acquisition of the calibration samples, a PCA was built, and the repeatability of these spectra was monitored over time using Hotelling T^2 and Q residuals control charts at a significance level of probability of 0.01 (Laursen et al., 2011). The instrument was re-calibrated if the reference compound did not pass in both control charts.

2.5.5. Systematic methodology flowchart to predict unknown samples

The systematic methodology used to select samples for calibration, build the regression models, identify out of control samples, predict under control unknown samples (using), and subsample the under control unknown samples to be analysed using the reference method is summarised in Fig. 2.

The first step consists of identifying which samples represent the area to be analysed. This can be done using spectra information by selecting the most representative spectra, i.e. using the Kennard stone algorithm, or using information from previous samples collection, as we did in this work (***2.4.1. Laboratory reference database***). The next step consists of checking if the samples selected to build the calibration model present representative soil particle size values. In other words, only a few samples in a large range will not provide enough information for the spectra models to predict unknown samples in this same range, i.e. in Fig. 3a, there are only a few samples with clay content between 40 and

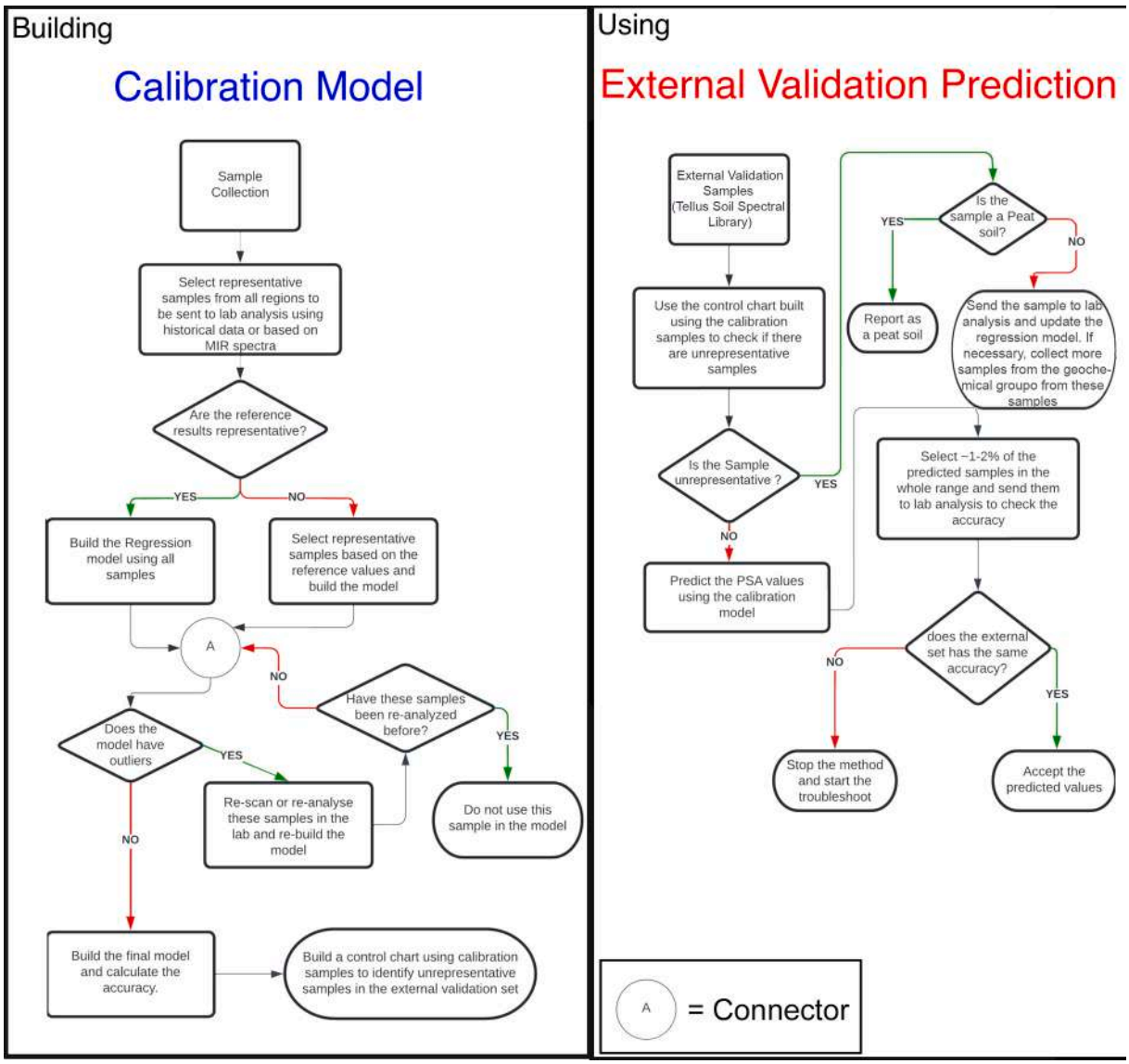


Fig. 2. Protocol used to create the spectroscopy models and predict external validation samples.

60%, so it is better to exclude this region from the model than include unrepresentative information in the calibration model, which will result in lower accuracy in the prediction of unknown samples in this range.

The next step is the outliers identification, samples classified as outliers should be re-analysed (spectra and reference values). Then, we can split these samples into calibration and internal validation to calculate the accuracy parameters from the regression models. After that, we merged these samples again and built a final calibration model. Using these calibration samples, a control chart should be built to identify unrepresentative samples during the prediction of unknown samples. This step will increase the confidence in the predicted results by identifying samples with spectral signatures outside the range contained in the calibration set. Samples considered unrepresentative/out of control must be analysed using the reference method, and the samples considered under control can be predicted using the spectroscopy methodology. Finally, to guarantee the accuracy of the prediction of unknown samples, we propose selecting 1–2% of the predicted samples and sending them to be analysed using the reference method. If these samples' accuracy is similar to that obtained in the internal validation set, we can accept the predicted values. If not, we need to start troubleshooting the methodology.

The proposed methodology was designed to monitor the instrument repeatability, identify samples with different spectral signatures and check the prediction accuracy. During the sample collection, the instrument stability was monitored using acetaminophen as a reference compound, and the whole spectra were monitored over time using two control charts based on the Hotelling T^2 and Q residuals values. The identification of samples with different spectral signatures (out of control) was also evaluated using Hotelling T^2 and Q residuals control charts using the soil calibration samples spectra were used to build the control chart. It is important to highlight that MIR spectroscopy will not fully replace the soil particle size reference method, because eventually non-representative samples will be identified by the control chart, and ~1–2% of the predicted samples should be analysed using the reference method to check the accuracy in the unknown samples.

2.5.6. *Software and packages used*

All calculations were performed in MATLAB version R2017a (Mathworks), using an Intel Core i5-8250 CPU with four cores and eight threads with 8 GB memory. Support Vector Machine models were built using the 'Statistics and Machine Learning Toolbox' version 11.1. The soil texture classification was performed using a Microsoft Excel Macro-enabled spreadsheet freely available to download at (de Santana, 2022).

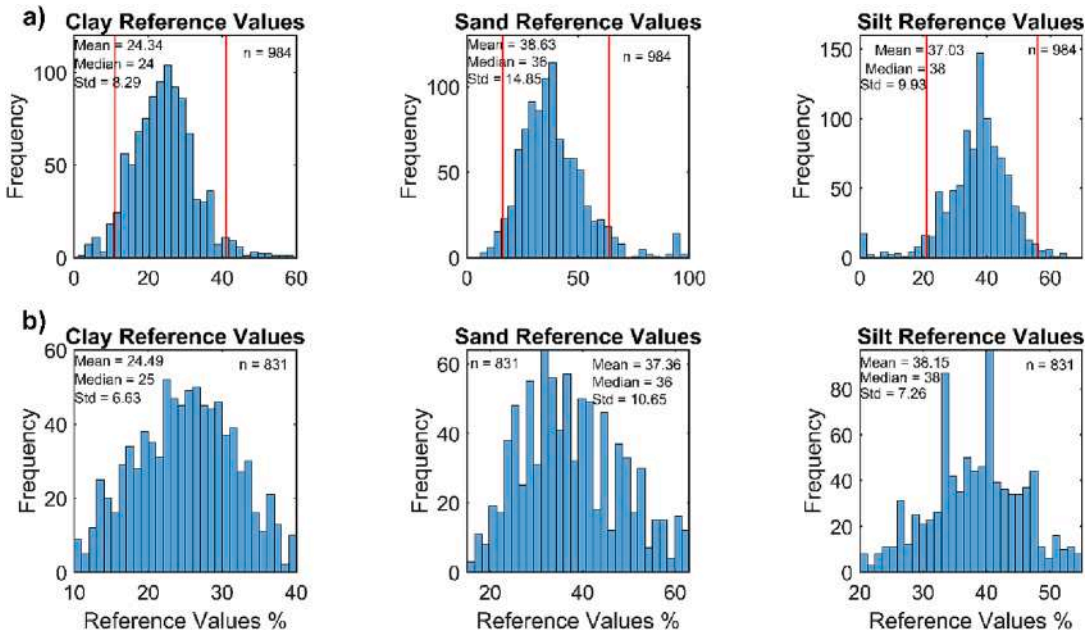


Fig. 3. Histogram of clay sand and silt reference values from (a) all 984 calibration samples and (b) after excluding unrepresentative and Peat samples.

3. Results

3.1. Laboratory reference data on soil particle size

The descriptive statistics for soil particle size determined on 984 soil samples selected from the wider Tellus survey are presented in Fig. 3a. Since there were not enough representative samples across the whole range of values (from 0 to 100%) for clay, sand, and silt values, a cut-off (red vertical lines in Fig. 3a) was used to define the range of values with sufficient representativeness to build the regression models. These cut-offs were selected based on the histogram analysis (Fig. 3), excluding regions that we have a large variation, with only a few samples. The proposed cut-offs are consistent with the texture classes in the Irish soil information system (ISIS), which mainly comprise fine loamy, coarse loamy and loamy soils (Creamer et al., 2016).

Although an initial screening of the Tellus survey samples was conducted to exclude samples described from field examination (visual inspection) as peat soils, 18 samples with %OM values > 30 (laboratory analysis) were still present in the calibration dataset. Since the model objective was only to predict particle size for MOM soils, peat soils (% OM > 30) were excluded from the calibration set. As peat soils consist mainly of OM at different stages of degradation, the unique structure of organic soils means that conventional soil texture and particle size methods designed for mineral soils do not adequately describe the texture of peat soils, which have specific nutrient managements (Schulte et al., 2018). This resulted in a final dataset composed of 831 samples (Fig. 3b), with mean and median values varying slightly from the original dataset.

3.2. MIR spectra

The first derivative and smoothing (Savitzky-Golay) spectra of three calibration samples with low, mean and high levels of clay, sand and silt are shown in Fig. 4. It is possible to visualise some relationship between each soil particle size and the chemical information in the spectra, i.e., the absorption bands at 3700 – 3600 cm^{-1} are mainly attributed to clay minerals, such as kaolinite, smectite, illite and other alumino-silicates, indicating higher clay content (Viscarra Rossel et al., 2008; Stenberg et al., 2010).

Clay minerals are mainly composed of silica, alumina, magnesia, and iron. Thus there are several spectral regions that reflect low clay content, i.e., at 2940–2900 cm^{-1} related to CH_3, CH_2, CH present in OM (Bruker Optics, 2009; Stenberg et al., 2010). Some clay minerals present intense peaks, such as the Kaolinite (1:1 phyllosilicate) at 1780 cm^{-1} and 930–920 cm^{-1} due to O-Fe of Fe oxides (goethite), both resulting in higher intensities according to the clay content. At 1620 cm^{-1} the absorptions are related to SiO_2 presenting in sand, resulting in low levels of clay and silt (Krivoshein et al., 2020; Mendes et al., 2022).

In the fingerprint region (1500 – 600 cm^{-1}), we can highlight several bands, such as the absorptions at 1285–1200 cm^{-1} mainly attributed to OH and C-O, reflecting low contents of clay and silt (Bruker Optics, 2009). The absorptions at 910–900 cm^{-1} are associated with metal-OH bending presenting in the clay minerals, reflecting higher contents of clay (Schroeder, 2002).

Still in the fingerprint region, the band at 795 cm^{-1} can be associated with SiO_2 reflecting lower contents of clay and higher content of sand. At 700 cm^{-1} the SiO_2 peaks suggest a lower content of clay and higher contents of silt and sand. The peak at 655 cm^{-1} is also related to SiO_2, indicating a higher content of sand. These main chemical groups responsible for the soil absorptions related to clay, sand and silt contents are organized in Table 1.

3.3. Regression models

The spectral library for the laboratory reference data was divided into calibration and internal validation sets to build the model and evaluate its performance. To select samples for internal validation, the samples with reference values were reorganized by Tellus survey region and in ascending order according to clay content. Samples that were in line position multiples of 3 were selected to compose the internal validation set (n = 280) and the remaining samples were used to compose the calibration set (n = 551). The scatter plots showing the reference versus predicted values of clay, sand and silt are shown Fig. 5.

As shown in Fig. 5, the spectral models predicted the soil particle size with excellent accuracy, especially for sand and clay determinations, presenting $R^2_{val} \geq 0.84$, and ratio of performance to interquartile distance (RPIQ) in validation ≥ 3.71. The clay spectral model was classified as 'Excellent Model', with $RPIQ_{val} = 4.97$, similar calibration (red) and validation (blue) fitted lines, both close to the 1:1 fitted line (dotted black line). The SVR model to predict sand was classified as 'Good

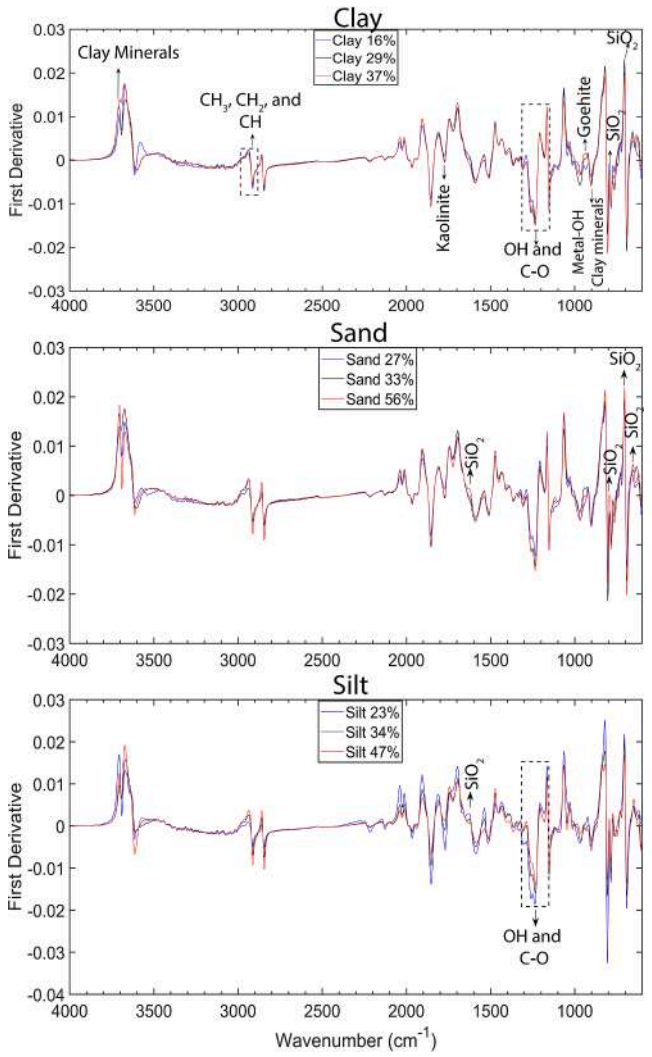


Fig. 4. First derivative and smoothing spectra of three soil samples with low, mean and high levels of clay, sand and silt.

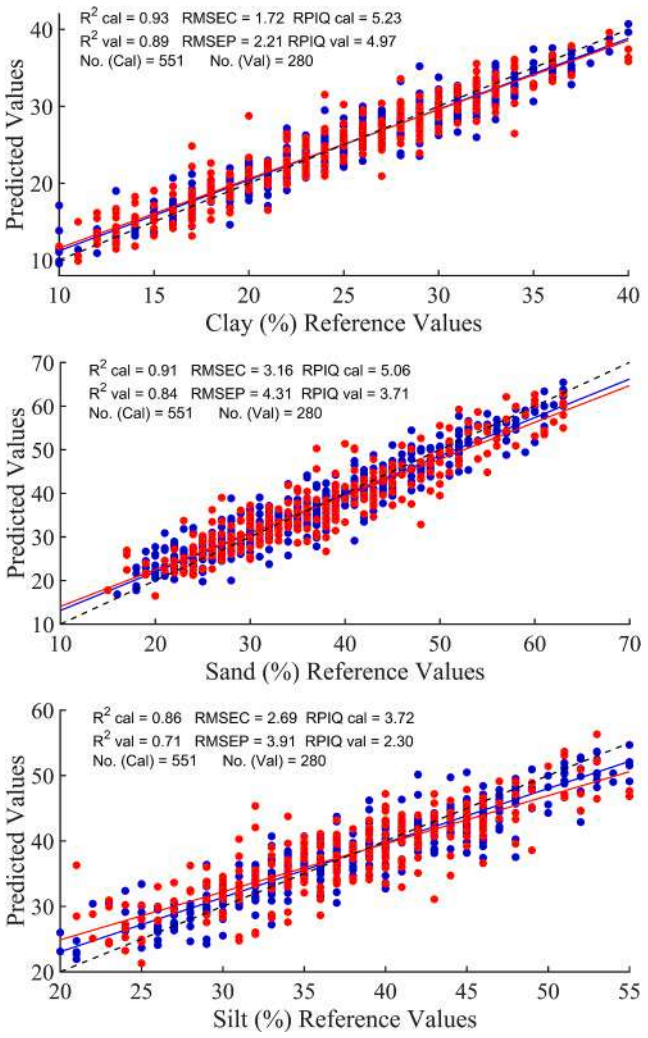


Fig. 5. Scatter plots showing reference versus predicted values of clay, sand and silt. Calibration samples are represented by blue and validation samples by red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Band assignments for fundamental MIR absorptions of soil constituents.

Band (cm^{-1})	Mainly chemical groups/ constituents	Reference
3700 – 3600	Kaolinite, smectite, illite and other alumino-silicates	(Viscarra Rossel et al., 2008)
2940 – 2900	CH_3, CH_2, CHf	(Bruker Optics, 2009; Stenberg et al., 2010)
1780	Kaolinite (1:1 phyllosilicate)	(Mendes et al., 2022)
1620	Both organic (C = O) and inorganic (SiO_2) constituent	(Krivoshein et al., 2020)
1285 – 1200	OH and C-O	(Bruker Optics, 2009)
930–920	O-Fe of Fe oxides (goethite)	(Mendes et al., 2022)
910–900	Metal-OH bending	(Schroeder, 2002)
795, 700, 655	SiO_2	(Krivoshein et al., 2020)

Model', with $RPIQ_{val} = 3.71$, just short of 'Excellent Model' classification. Similar to SVR model to predict clay, this model also presented similar calibration and validation fitted lines, with a slight slope compared to the 1:1 fitted line. The SVR model to predict silt was classified as 'Fair Model', with $RPIQ_{val} = 2.30$ and $R^2_{val} = 0.71$ (Ludwig et al., 2017).

The accuracy parameters and the reference versus predicted values of spectroscopy models built without the *clr* transformation are present in File S1. The TSL accuracy results were broadly similar using the *clr* transformation, indicating whether the proposed methodology is robust.

The soil particle size predicted results was used to determine the soil texture using the ADAS (https://adas.co.uk/) texture triangle (European Committee for Standardization, 2004). The proposed methodology correctly classified the soil texture of 231 out of 280 validation samples, representing an accuracy of 82.5%. It is important to emphasise here that most of the misclassified samples were classified into neighbouring texture classes, close to the correct class, e.g., 'Clay' instead of 'Clay Loam'. The list containing the soil texture classification for the whole validation set can be visualized in Table S1.

After calculating all the accuracy parameters, the calibration and validation samples were grouped to build the final calibration set and a new model containing all the 831 samples was created. This model was used to predict the soil particle size values of the 9009 unknown samples in the TSL.

3.4. Prediction of soil particle size for the Tellus spectral library

The spectra of the calibration samples and the unknown TSL samples are shown in Fig. 6a. Some of the unknown samples contained spectral intensities higher than those contained in the calibration samples. This

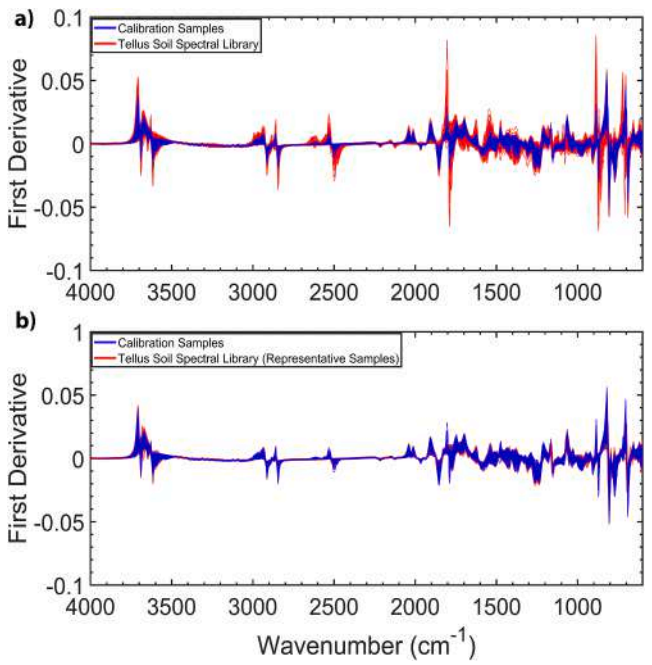


Fig. 6. First derivative and smoothing spectra of calibration and unknown samples from Tellus spectral library (a) and after excluding unrepresentative samples (b).

indicated that these samples could contain sand, silt or clay contents outside the concentration ranges or have different compositions, i.e. samples with higher OM content (peat soils), which are not present in the calibration set. Thus, these samples were not predicted by the spectral model and must be analysed using the classical method.

The identification of these unrepresentative samples was performed using two control charts based on the PCA, the Hotelling T^2 and Q residuals control charts at a significance level of probability of 0.01. Fig. 6b. shows the MIR spectra after removing the unrepresentative samples using both control charts.

From the 9009 samples contained in the TSL, 5330 samples (5254 non-peat soils + 76 peat soils) were considered representative and 3679 samples (601 non-peat soils + 3078 peat soils) were considered unrepresentative, and as consequence were not predicted by the model. These results show the excellent ability of the proposed methodology to mitigate the use of conventional analytical methods covering wide-scale heterogeneous areas. The control charts based on PCA correctly classified about 90% of non-peat soils, that is, No. of under control non-Peat soils / No. of non-Peat soils = [100*(5254/(601 + 5254))].

3.4.1. *Non-peat soils classified as out of control*

A geospatial analysis of the sample dataset indicated that the 601 non-peat soil samples classified as unrepresentative had some unique geological and landscape features. For example, coastal and alluvial soils with high sand content, soils in boundary regions between mineral and peat, especially in areas containing exposed bedrock or till. Some areas can vary significantly from the soil particle size values contained in calibration, such as urban soils and made-ground areas (Fig. 7a-b). In addition, several of these samples were collected in regions with lower sample densities (Fig. S1).

3.4.2. *Estimation of the external validation accuracy*

The accuracy of predicted values of % sand, silt and clay in the TSL unknown samples was evaluated by selecting 96 samples for laboratory analysis using the reference method to determine soil particle size. The samples were selected from the 5254 non-peat soil samples predicted by spectroscopy. These samples were chosen to cover the whole range of

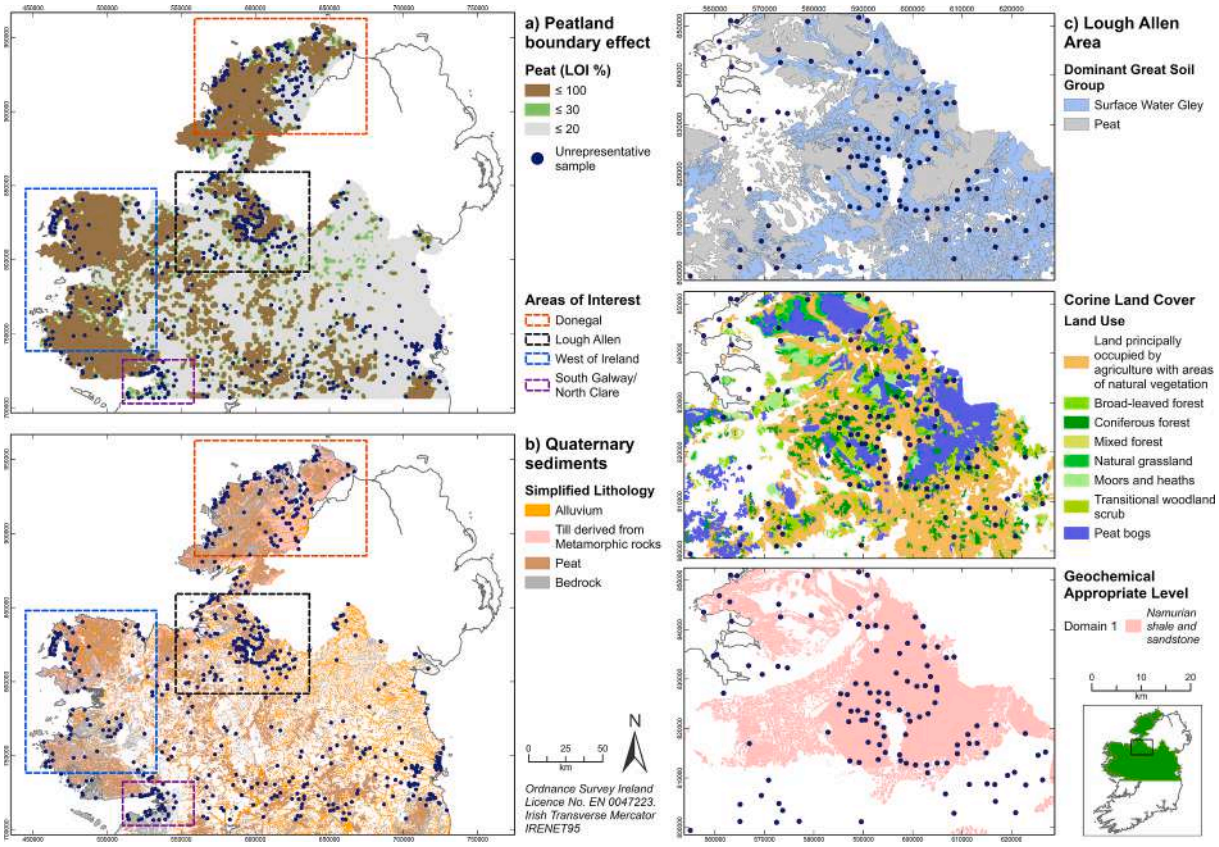


Fig. 7. Maps of non-Peat soils samples classified as unrepresentative by the control chart.

the predicted values. The PARIO method was used to determine soil particle size for these samples as this had replaced the pipette method for gravimetric analysis at Teagasc laboratories. As the reference calibration results were obtained in an external accredited laboratory using the gravimetric method with the pipette approach, both methods were checked for statistical comparison. The reference values obtained by 15 MOM samples analysed by both methods are shown in Table S2. The paired *t*-test resulted in p-values of 0.36, 0.75 and 0.55 for sand, silt, and clay, respectively, showing both methods present statically the same values. Consequently, the PARIO method was utilised to assess the accuracy of the external validation samples.

The soil texture class and the reference values of clay, sand and silt determined in the external validation (n = 96) by PARIO and the spectra models are presented in Table S3, while their respective accuracy parameters are given in Table 2. The RMSEPs calculated for the external validation set were in general 23% higher than those calculated by the internal validation set, while the $RPIQ_s$ values were, slightly higher for sand and silt; and 30.2% lower for clay determinations. The soil texture classification model was able to correctly classify 79.2% of the external validation set against 82.5% from the internal validation. These minor variations indicate that the proposed methodology was able to extract useful information from the calibration set and predicted the unknown samples with the same accuracy.

3.5. Predicted maps

Since the external validation set presented accuracies ($RMSEP_{ext}$, $RPIQ_{ext}$ and R^2_{ext}) comparable to internal validation, the predicted values obtained by the spectra models were accepted and the soil particle size map for each attribute was built (Fig. 8). Using the predicted values of clay, sand and silt the soil texture map based on the ADAS system was also built (Fig. 8d).

4. Discussion

4.1. Spectral models accuracy

The accuracy results showed that the spectroscopy model could predict clay and sand with excellent and good accuracy, respectively. While for silt, the accuracy was intermediate, being classified as a fair model according to the threshold proposed by Ludwig et al. (2017). However, it is important to highlight that among the three soil particle sizes, the clay and sand contents are more frequently used to characterise soil drainage properties and as indicators of nutrient storage and release. For example, the ratio between the soil organic carbon and clay content can be used as an indicator of soil health, while sand and clay contents are used to calculate the soil drainage (Prout et al., 2021).

The proposed methodology can also be used without any log-ratio transformations. However, despite models built without the *clr* transformation (File S1) presented similar accuracy compared to spectral models built using the *clr* transformation, we recommend investigating whether the accuracy between the models built with and without the log-ratio transformations is the same, especially when dealing with values close to 0 and 100%.

The RMSEP obtained can be used to calculate the expected prediction error (confidence interval) during the prediction of external validation samples. For homoscedastic models (random residuals/prediction errors), it is expected that about 68.2% of the samples will be predicted with errors between ±1 × RMSEP and ~95.4% with ±2 × RMSEP. We can use these errors to evaluate whether or not the spectra models show enough accuracy for a specific application. For example, in clay predictions the spectra model should predict the clay content with ±2.21% of error for ~68.2% of the dataset and ±4.42% of error for ~95.4% of the dataset.

4.2. Prediction unknown samples

One of the challenges to implementing the spectroscopy methodology in standard soil laboratories is the ability to identify samples that present different spectral signatures compared to the calibration samples. This is especially problematic for predictions on large spectral libraries that cover wide spatial areas with heterogeneous soil cover, such as the TSL in this study representing 35,716 km^2 survey area. The presence of compounds not contained in the calibration samples could considerably affect one or more spectral ranges, resulting in lower accuracies. It is important to emphasise that predicting values outside the spectral signatures present in the calibration range is not recommended because the linear response is not guaranteed (ASTM E1655-17, 2017). Fig. 6a shows some of the TSL samples are outside the spectral variation present in the calibration and should not be predicted using the spectroscopy models.

Fig. 6b shows the use of both control charts enabled the identification of almost all the spectral variations not contained in the calibration samples. As a consequence, the under control samples show essentially the same spectral variations presented in the calibration samples.

4.3. Unrepresentative non-peat soil samples

Besides the samples we expect to have higher sand content (coastal and alluvial soils) and consequently be classified as unrepresentative by the control chart, other elements contributed to some soil samples being classified as unrepresentative. The organo-mineral soils (%OM between 20% and 30%), appears to be a limitation of the methodology. Of the samples classified as representative by the control chart, 7.73% were organo-mineral soils. While, in the unrepresentative samples, 29.62% were organo-mineral soils. Some of these unrepresentative organo-mineral are shown in Fig. 7a in Donegal and Lough Allen regions. This suggests that the calibration model does not contain enough representative organo-mineral and as consequence, several of these unknown soil sample were considered unrepresentative.

Spatially, organo-mineral soils tend to occur in peatland boundary areas, where there is a transition from peat to mineral soil and the mixing of the two either through natural or anthropogenic processes (Hammond, 1981; Schulte et al., 2018). As shown in Fig. 7b, some organo-mineral unrepresentative soils are located in areas that contain exposed bedrock or till alongside peat and along drainage pathways derived from such regions, i.e. Lough Allen (Leitrim, Roscommon, Cavan), West of Ireland and Donegal. This may indicate that the model is less able to handle organo-mineral soils with a 'peat-natural' interface rather than a 'peat-agriculture' interface. To minimise this limitation,

Table 2
Accuracy parameters calculated by calibration, validation and external validation sets.

Soil Attribute	Accuracy parameters									
	R^2_{cal}	RMSEC	R^2_{val}	RMSEP	$RPIQ_{val}$	Soil texture	R^2_{ext}	$RMSEP_{ext}$	$RPIQ_{ext}$	Soil texture
	$n_{cal} = 551$		$n_{val} = 280$				$n_{ext} = 96$			
Clay	0.93	1.72	0.89	2.21	4.97	231/280	0.92	2.73	3.47	76/96
Sand	0.91	3.16	0.84	4.31	3.71	(82.5%)	0.90	4.53	3.98	(79.2%)
Silt	0.86	2.69	0.71	3.91	2.30		0.79	4.51	2.64	

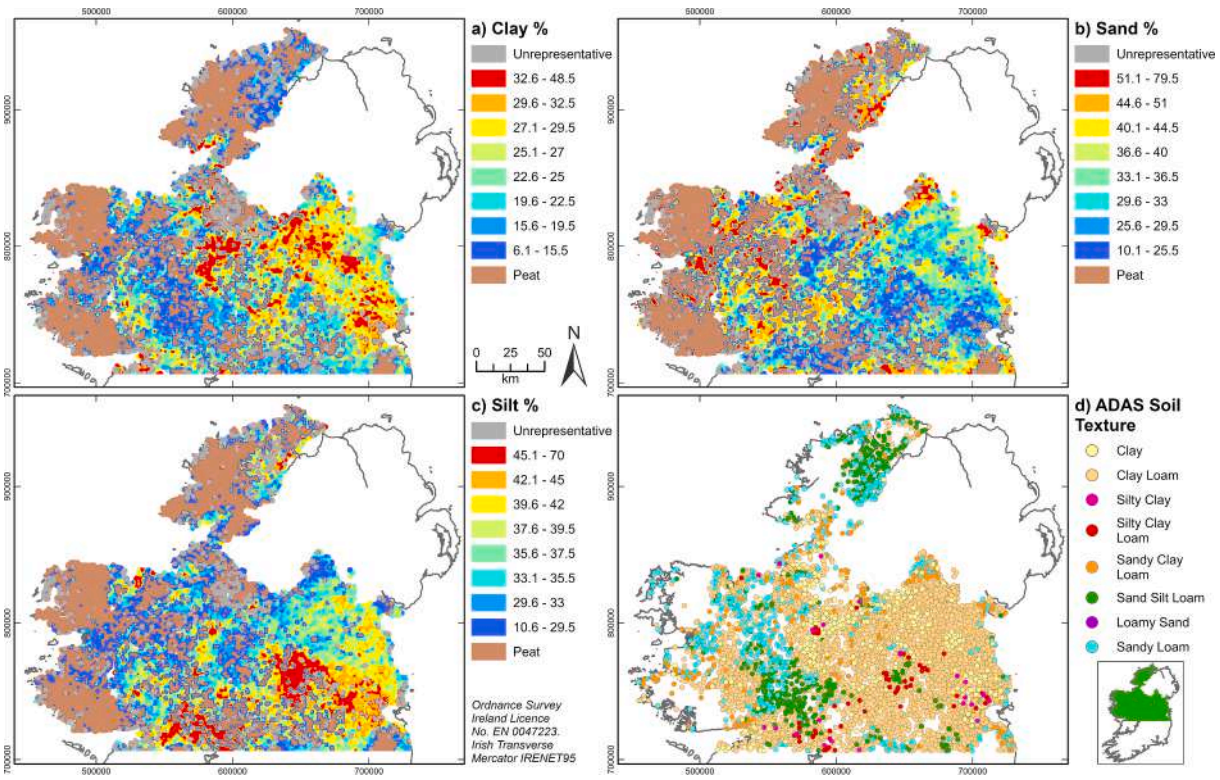


Fig. 8. Soil particle size and soil texture class maps predicted by the spectroscopy methodology.

organo-mineral samples from the 'peat-natural' interface can be added to the calibration model.

The land-use type, specifically 'natural' vs 'anthropogenic' was seen to influence whether a sample was classified as a representative or not. For example, 14.58% of representative samples were found to have Corine (Agency, 2018) classifications of either natural (alluvial, marsh, dunes etc.), forested, or agricultural land with significant natural vegetation, and 83.90% with either other agricultural or man-made classifications (Fig. 7c). While in the unrepresentative group, 49.42% of the samples are classified as 'natural'. The majority of the calibration samples (Fig. S1) were taken from managed agricultural land, so as a consequence samples from more 'natural' settings are proportionally over-represented in the unrepresentative sample group.

As we can see in Fig. 7, four areas display significant concentrations of unrepresentative samples. As well as organic soils and land use, other geogenic factors may also contribute to this apparent clustering, such as:

1. Lough Allen area.

There are some organo-mineral soils located on 'Land principally occupied by agriculture with significant areas of vegetation' (Fig. 7a). Fig. 7c presents additional Corine classifications in the area, comprising peat bogs, forests, moor and heathland, and transitional woodland-shrub, which correspond with previous observations on natural vs anthropogenic land use. As observed in Fig. 7c, unrepresentative samples also align well with the Surface Water Gleys mapped in the area, and could be influenced by associated drainage patterns. Furthermore, the samples are situated in Domain 1 of the Soil Recovery Facility (SRF) classifications, which is geochemically distinct and underlain by Namurian metasediments (shales and sandstones) (EPA, 2020). This Domain represents a relatively small section of the survey area (Fig. S1) used to produce the regression model, so it is possible that inclusion of further samples from this Domain, plus rock type, could minimize the number of non-representative samples.

2. South Galway/North Clare.

Significant bedrock outcrop associated with karst landscapes appear to correlate with a number of samples that have been classed as unrepresentative. Extensive bedrock outcrop of this type is localized to this part of the survey area, and sample density is somewhat sparse in patches, which could result in variance of composition compared to calibration samples. Additionally, some samples located close to the coastline could be affected by coastal processes and windblown material, even where the main land type and use is not classed as coastal.

3. West of Ireland.

Many organo-mineral unrepresentative samples located in counties Galway and Mayo, with upland areas dominated by bedrock exposure alongside peatlands (Fig. 7b), coast or influenced by coastal processes.

4. Donegal.

Unrepresentative samples are predominantly concentrated in a zone underlain by till derived from metamorphic rocks, including schists, pelites, quartzites etc. As a result, these may contain higher proportions of sand or silt, and, combined with exposed bedrock, organic material and alluvium in steep drainage channels in parts of the area, could explain why so many samples are classed as unrepresentative.

Generally, soils underlain by metasediments, granites, quartzites, conglomerates or shales, and tills derived from these, appeared to have a higher number of unrepresentative samples than soils underlain by carbonates. For example, to the north of Dublin, which is underlain by some conglomeratic units, and to the south which is underlain by granites (GSI, 2022). This could be a factor of the data used to generate the model, since the majority is from agricultural land through the Irish Midlands, and so might become less evident if the calibration set was expanded to include samples from other locations outside of the current survey area. Through the Irish Midlands, aside from peat influence, local

variations in the calcareous till may be responsible for samples being classed as unrepresentative (GSI, 2022).

Understanding the geological factors underpinning the unrepresentative samples is essential. By including this analysis, we can minimise the limitation of the methodology by collecting/adding more samples to the calibration from specific regions or soil classifications, not originally included. For example, in our case, the geological factors showed collecting more organo-mineral soils, samples with natural vegetation, coastal soils, samples located in Domain 1, and soils derived from metamorphic rocks could minimise the number of unrepresentative samples in the future, which must be sent to the laboratory for classical analysis.

4.4. Cost-effectiveness of the spectral model

The accuracy results (Table 2) and the large number of samples predicted using the proposed methodology shows the high potential for this systematic approach to be integrated with digital soil mapping to produce soil information at regional and national scales in a cost-effective and fast way. The cost-effectiveness of the model was evaluated by assuming an approximate cost of €35.00 per sample (excluding taxes) for soil particle size. Analysis of 5254 samples using the proposed methodology, resulted in a saving of €182,890 (€35.00 * 5254 samples). In addition, considering no investment in the laboratory infrastructure, the classical laboratory reference method would require four years to complete the analysis of 5254 samples. The proposed methodology predicted the soil particle size values from 5254 samples in <6 months and correctly identified 97.59% of the peat soils (3078 out of 3154 samples) using the proposed spectral control chart.

5. Conclusion

This study proposed a systematic approach to improve confidence in the spectral predictions of a problematic soil attribute (particle size) over a large regional scale with a heterogeneous landscape. The proposed methodology moves away from a global model and confines the prediction to mineral and organo-mineral soils within a specified range. The results revealed that the multivariate control charts based on PCA proposed here could identify unrepresentative samples, which must not be predicted by the spectra models. This is an important conclusion given the challenge of including the full extent of soil variation in spectra models. Therefore methods to identify unrepresentative samples based on their spectra are necessary to identify unknown samples which cannot be correctly predicted by MIR spectroscopy. Furthermore, the subsample strategy used to check the accuracy of the unknown samples, classified by the control chart as under control, increased the confidence in the predicted results. This provided an opportunity to compare this subset's accuracy with the internal validation set. Significant differences in accuracy must be investigated by troubleshooting, i.e. checking the instrument repeatability, soil sample preparation, etc.

Although this study only evaluated the proposed methodology in sand, silt and clay determinations, the same systematic approach could be used for other soil parameters. Improving the confidence of the spectra model predictions will enable high-resolution mapping of soil attributes in a cost-effective, fast and environmentally friendly way. Furthermore, the systematic approach described in this study can be used as standard operating procedures for implementing spectroscopy methods in routine laboratories receiving unknown soil samples from large spatial areas with high heterogeneity. This approach can advance digital soil maps and also precision agricultural techniques aligned with Agriculture 4.0 and soil health monitoring requirements under the forthcoming Soil Health Law in Europe.

Declaration of Competing Interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

This work has been jointly supported by Geological Survey Ireland and Teagasc (Terra Soil Collaborative Agreement 2018). The authors are grateful to Ms. Courtney Doyle, Patricia Berry and Dr. Anna Fenelon for technical support in spectroscopy; and Dr. Luis Lopes-Sangil for technical support in soil particle size analysis at Teagasc Johnstown Castle laboratories. Geological Survey Ireland authors publish with the permission of the Director of Geological Survey Ireland.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2023.116491.

References

Agency, E.E., 2018. Copernicus Land Monitoring Service Datasets. URL https://land.copernicus.eu/pan-european/corine-land-cover/clc2018.
Aitchison, J. (John), 1986. The statistical analysis of compositional data. London.
ASTM E1655-17, 2017. ASTM E1655-17 Standard Practices for Infrared Multivariate Quantitative Analysis. ASTM Int. 05, 30. https://doi.org/10.1520/E1655-05R12.2.
Bachion de Santana, F., Daly, K., 2022. A comparative study of MIR and NIR spectral models using ball-milled and sieved soil for the prediction of a range soil physical and chemical parameters. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 279, 121441.
Brodský, L., Vašát, R., Klement, A., Zádorová, T., Jakšík, O., 2013. Uncertainty propagation in VNIR reflectance spectroscopy soil organic carbon mapping. Geoderma 199, 54–63. https://doi.org/10.1016/j.geoderma.2012.11.006.
Bruker Optics, 2009. Analytical Life Science Process Guide for Infrared Spectroscopy think forward Spectroscopy. URL https://www.ccmr.cornell.edu/wp-content/uploads/sites/2/2015/11/GuideforInfraredspectroscopy.pdf.
European Committee for Standardization, 2004. Geotechnical investigation and testing - Laboratory testing of soil - Part 4: Determination of particle size distribution.
Creamer, R.E., Simo, I., O'Sullivan, L., Reidy, B., Schulte, R.P.O., Fealy, R.M., 2016. Irish Soil Information System: Soil Property Maps.
de Santana, F.B., de Souza, A.M., Poppi, R.J., 2018. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 191, 454–462. https://doi.org/10.1016/j.saa.2017.10.052.
de Santana, F.B., de Souza, A.M., Poppi, R.J., 2019. Green methodology for soil organic matter analysis using a national near infrared spectral library in tandem with learning machine. Sci. Total Environ. 658, 895–900. https://doi.org/10.1016/j.scitotenv.2018.12.263.
de Santana, F.B., Otani, S.K., de Souza, A.M., Poppi, R.J., 2021. Comparison of PLS and SVM models for soil organic matter and particle size using vis-NIR spectral libraries. Geoderma Reg. 27, e00436.
de Santana, F.B., 2022. Felipebachion – Github. URL https://github.com/felipebachion/PSA (accessed 7 November, 2022).
Dunne, K.S., Holden, N.M., O'Rourke, S.M., Fenelon, A., Daly, K., 2020. Prediction of phosphorus sorption indices and isotherm parameters in agricultural soils using mid-infrared spectroscopy. Geoderma 358, 113981. https://doi.org/10.1016/j.geoderma.2019.113981.
Durner, W., Iden, S.C., von Unold, G., 2017. The integral suspension pressure method (ISP) for precise particle-size analysis by gravitational sedimentation. Water Resour. Res. 53, 33–48. https://doi.org/10.1002/2016WR019830.
Durner, W., Iden, S.C., 2021. The improved integral suspension pressure method (ISP+) for precise particle size analysis of soil and sedimentary materials. Soil Tillage Res. 213, 105086 https://doi.org/10.1016/j.still.2021.105086.
EPA, 2020. Guidance on waste acceptance criteria at authorised soil recovery facilities 1–20. URL https://www.epa.ie/publications/compliance–enforcement/waste/Guidance-on-Waste-Acceptance-Criteria-at-Authorised-Soil-Recovery-Facilities.pdf (accessed 7 November, 2022).
Gardiner, M.J., Radford, T., 1980. Soil associations of Ireland and their land use potential: Explanatory bulletin to soil map of Ireland, National Soil Survey of Ireland. Mount Salus Press Limited, Dublin, Ireland. URL https://www.teagasc.ie/media/website/environment/soil/General.pdf (accessed 7 November, 2022).
GSI, 2021. GSI, Geological Survey Ireland (2021).Terra Soil. URL https://www.gsi.ie/en-ie/programmes-and-projects/tellus/activities/tellus-product-development/smart-agriculture/Pages/Terra-Soil.aspx (accessed 7 November, 2022).

GSI, 2022. Geological Survey Ireland Spatial Resources: Public Data Viewer Series. URL https://dcenr.maps.arcgis.com/apps/MapSeries/index.html?appid=a30af518e87a4c0ab2fbde2aaac3c228 (accessed 7 November, 2022).

Hammond, R.F., 1981. The Peatlands of Ireland. accessed 7 November, 2022 Soil Surv. Bull. No. 35, 1–63. https://www.teagasc.ie/media/website/environment/soil/Peatlands-of-Ireland.pdf.

Hatten, J., Liles, G., 2019. A 'healthy' balance – The role of physical and chemical properties in maintaining forest soil function in a changing world 36, 373–396. https://doi.org/10.1016/B978-0-444-63998-1.00015-X.

Jaconi, A., Poeplau, C., Ramirez-Lopez, L., Van Wesemael, B., Don, A., 2019. Log-ratio transformation is the key to determining soil organic carbon fractions with near-infrared spectroscopy. Eur. J. Soil Sci. 70, 127–139. https://doi.org/10.1111/ejss.12761.

Krivoshein, P.K., Volkov, D.S., Rogova, O.B., Proskurnin, M.A., 2020. FTIR photoacoustic spectroscopy for identification and assessment of soil components: Chernozems and their size fractions. Photoacoustics 18, 100162. https://doi.org/10.1016/j.pacs.2020.100162.

Laursen, K., Rasmussen, M.A., Bro, R., 2011. Comprehensive control charting applied to chromatography. Chemom. Intell. Lab. Syst. 107, 215–225. https://doi.org/10.1016/j.chemolab.2011.04.002.

Ludwig, B., Vormstein, S., Niebuhr, J., Heinze, S., Marschner, B., Vohland, M., 2017. Estimation accuracies of near infrared spectroscopy for general soil properties and enzyme activities for two forest sites along three transects. Geoderma 288, 37–46. https://doi.org/10.1016/j.geoderma.2016.10.022.

Massey, P., O'Connor, C., Sills, P., Fenelon, A., Maloney-Finn, L., Stone, D., Reidy, B., Creamer, R., 2013. Irish Soil Information System : Laboratory Standard Operating Procedures Final Technical Report 7.

MathWorks, 2017. Statistics and Machine Learning Toolbox ™ User ' s Guide R 2017 a. The MathWorks Inc, Apple Hill Drive (2017).

Mendes, W.d.S., Demattê, J.A.M., Rosin, N.A., Terra, F.d.S., Poppiel, R.R., Urbina-Salazar, D.F., Boechat, C.L., Silva, E.B., Curi, N., Silva, S.H.G., José dos Santos, U., Souza Valladares, G., 2022. The Brazilian soil Mid-infrared Spectral Library: The Power of the Fundamental Range. Geoderma 415, 115776.

Mert, M.C., Filzmoser, P., Hron, K., 2016. Error Propagation in Isometric Log-ratio Coordinates for Compositional Data: Theoretical and Practical Considerations. Math. Geosci. 48, 941–961. https://doi.org/10.1007/s11004-016-9646-x.

Metzger, K., Zhang, C., Ward, M., Daly, K., 2020. Mid-infrared spectroscopy as an alternative to laboratory extraction for the determination of lime requirement in tillage soils. Geoderma 364, 114171. https://doi.org/10.1016/j.geoderma.2020.114171.

O'Rourke, S.M., Minasny, B., Holden, N.M., McBratney, A.B., 2016. Synergistic Use of Vis-NIR, MIR, and XRF Spectroscopy for the Determination of Soil Geochemistry. Soil Sci. Soc. Am. J. 80, 888–899. https://doi.org/10.2136/sssaj2015.10.0361.

Polakowski, C., Ryżak, M., Sochan, A., Beczek, M., Mazur, R., Bieganowski, A., 2021. Particle size distribution of various soil materials measured by laser diffraction—the problem of reproducibility. Minerals 11 (5), 465.

Poppiel, R.R., Francine, A., Dematt, M., 2022. Geoderma Bridging the gap between soil spectroscopy and traditional laboratory: Insights for routine implementation 425. https://doi.org/10.1016/j.geoderma.2022.116029.

Prout, J.M., Shepherd, K.D., McGrath, S.P., Kirk, G.J.D., Haefele, S.M., 2021. What is a good level of soil organic matter? An index based on organic carbon to clay ratio. Eur. J. Soil Sci. 72, 2493–2503. https://doi.org/10.1111/ejss.13012.

Quinn, T.P., Erb, I., Gloor, G., Notredame, C., Richardson, M.F., Crowley, T.M., 2019. A field guide for the compositional analysis of any-omics data. Gigascience 8, 1–14. https://doi.org/10.1093/gigascience/giz107.

Ramirez-Lopez, L., Wadoux, A.M.J.C., Franceschini, M.H.D., Terra, F.S., Marques, K.P.P., Sayão, V.M., Demattê, J.A.M., 2019. Robust soil mapping at the farm scale with vis–NIR spectroscopy. Eur. J. Soil Sci. 70, 378–393. https://doi.org/10.1111/ejss.12752.

Rinnan, Å., van den Berg, F., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends Anal. Chem. 28, 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047.

Schölkopf, B., Smola, A.J., 2002. Support Vector Machines and Kernel Algorithms. Handb. Brain Theory Neural Networks 1119–1125.

Schroeder, P., 2002. Infrared spectroscopy in clay science. Teach. clay Sci. 11, 181–206.

Schulte, R., O'Sullivan, L., Creamer, R., 2018. The Soils of Ireland, World Soils Book Series. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-71189-8.

Skoog, D.A., West, D.M., Holler, F.J., Crouch, S.R., 2014. Fundamentals of Analytical Chemistry, 10th ed. Cengage.

Stenberg, B., Viscara Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and Near Infrared Spectroscopy in Soil Science. Adv. Agron. 107, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.

Valle, S.S., Kienzle, J., 2020. AGRICULTURE 4.0 Start Agricultural robotics and automated equipment for sustainable crop production. Integr. Crop Manag. 24, 1–40.

Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B., 2008. Using a legacy soil sample to develop a mid-IR spectral library. Aust. J. Soil Res. 46, 1–16. https://doi.org/10.1071/SR07099.

Zhang, J., Li, G., Ding, S., Tian, H., Ren, X., Liu, M., Zheng, Y., 2022. Distribution Characteristics of Soil Particles and Their Relationships With Soil Organic Carbon Components in the Alluvial/Sedimentary Zone in the Lower Reaches of the Yellow River. Front. Environ. Sci. 10, 1–13. https://doi.org/10.3389/fenvs.2022.849565.