Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 279 (2022) 121441

Contents lists available at ScienceDirect

Spectrochimica Acta Part A:
Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

# 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

Felipe Bachion de Santana, Karen Daly *

*Environment, Soils and Land Use Department, Teagasc, Johnstown Castle Research Centre, Wexford Y35 TC97, Ireland*

ARTICLE INFO

*Keywords:*
Irish soil information system
Soil particle size
Soil analysis
Machine learning
Support Vector Machine
PLS

ABSTRACT

This study evaluated the influence on predicted physical and chemical parameters of soil particle sizes commonly used in the infrared spectra acquisition, < 0.100 mm (ball-milled) and < 2 mm for MIR and NIR ranges, respectively. The influences were evaluated through the accuracy (RMSEP and RPIQ) results and the chemical information extracted by multivariate classification and regression models. For this a national population of soils containing 888 samples from 225 modal soil profiles, each with the reference values of sand, silt, clay, $pH_{(CaCl2)}$, $pH_{(Water)}$, total carbon, organic carbon (OC), cation exchange capacity, nitrogen, aluminium and bulk density, was used. Spectra were collected in MIR and NIR ranges using samples with both particle sizes. For each soil attribute, 29 random calibration and validation sets were generated and SVM, PLS and Cubist regression models were built. This same strategy was used to classify the soil samples according to their respective horizons (1 or 2-7) using SVM, PLS-DA and random forest algorithms. Results obtained by the randomised calibration and validation set did not present positive or negative bias on the RMSEP and RPIQ values based on soil particle sizes. In general, random variations of the RMSEP and RPIQ values were observed for all soil attributes. In addition, ball-milled and < 2 mm spectral models did not present large differences in both accuracy parameters simultaneously. The median Matthews correlation coefficient values calculated by the classification models showed minor variations of 2.61% and 0.65% for samples from both particle sizes in MIR and NIR ranges, respectively. The 'Variable Importance in Projection' or VIP scores, calculated by PLS and PLS-DA models, did not show any large variation in the chemical information extracted from MIR and NIR spectra for models built using samples from both particle sizes. The results from this study show that scanning ball-milled or < 2 mm sieved soil samples will result in spectra models in MIR and NIR ranges with the same accuracy and same chemical information. This suggests there is a big potential to eliminate the ball-milling sample step in soil laboratories that use MIR and NIR vibrational spectroscopy techniques to predict soil attributes, thereby reducing the time and costs associated with soil analysis.

## 1. Introduction

The United Nations (UN) estimate that the world's population will reach 9.7 billion by 2050 [1], feeding this population will require an increase in global agricultural production by around 60–70% [2]. To achieve this goal, it will be necessary to improve agricultural efficiency, expand production to non-agricultural areas and monitor the soil security in agricultural and non-agricultural areas [3]. To monitor soil security, analytical determination of soil attributes, such as: soil texture, pH, total carbon (TC), organic carbon (OC), phosphorus, nitrogen, cation exchange capacity (CEC), among other parameters are required. However, conventional analytical methods are based on traditional wet analysis, being considered time-consuming and expensive, for example the pH, OC, N, P and K determinations could take up to 2 weeks, while the determination of soil texture could take up to 4–8 weeks depending on the structure of the laboratory. Delays and bottlenecks in routine soil laboratories are common when high volumes of samples require different assays for multiple parameters, which increases the costs associated with monitoring soil security over large areas. In addition to these limitations, wet chemical methods generate chemical wastes, which require disposal adding to the cost of analysis. To minimize these problems and promote the move toward 'Green Chemistry', several soil

* Corresponding author.
*E-mail address:* Karen.Daly@teagasc.ie (K. Daly).

https://doi.org/10.1016/j.saa.2022.121441
Received 7 October 2021; Received in revised form 9 May 2022; Accepted 26 May 2022
Available online 30 May 2022

**Table 1**
Soil particle sizes used to collected NIR and MIR spectra in reflectance mode.

| Reference | Range | Soil particle size | Samples | | Soil parameter | Range | RMSEP | $R^2_{val}$ |
|---|---|---|---|---|---|---|---|---|
| | | | Cal | Val | | | | |
| Deiss, 2020 [4] | MIR | < 2 mm | 300 | 100 | OC | 0.10 – 9.10 % | 0.16 | 0.93 |
| Baumann, 2021 [5] | MIR | Ball-milled | 4295 | – | TC | 0.1–58.3 % | 0.84* | 0.97* |
| Hutengs, 2018 [6] | MIR | Ball-milled | 40 | – | OC | 0.62 – 2.70 % | 0.16* | 0.85* |
| Knox, 2015 [7] | MIR | Ball-milled | 696 | 296 | TC | 1.33 – 523 (g/kg) | 0.24 | 0.95 |
| Dangal, 2019 [8] | MIR | Ball-milled | 33,991 | 8,498 | OC | 0 – 60% | 0.69 | >0.99 |
| Wang, 2018 [9] | MIR | Ball-milled | 212 | 100 | OC | 0.97 – 39.93 (g/kg) | 0.27 | 0.94 |
| Barthès, 2016 [10] | MIR | < 2 mm | 97 | – | TC | 2 – 121 (g/kg) | 5.7* | 0.98* |
| Barthès, 2016 [10] | MIR | Ball-milled | 97 | – | TC | 2 – 121 (g/kg) | 9.8* | 0.93* |
| Terra, 2015 [11] | MIR | Ball-milled | 881 | 378 | OC | 1.2 – 40.6 (g/kg) | 0.13 | 0.77 |
| Terra, 2015 [11] | NIR | < 2 mm | 881 | 378 | OC | 1.2 – 40.6 (g/kg) | 0.16 | 0.65 |
| de Santana, 2019 [12] | NIR | < 2 mm | 27,581 | 13,791 | SOM | 0 – 50 (g/dm$^3$) | 5.03 | 0.72 |
| Yang, 2020 [13] | NIR | < 2 mm | 13,272 | 4,000 | OC | 0 – 166 (g/kg) | 6.40 | 0.73 |
| Shi, 2014 [14] | NIR | < 2 mm | 1581 | – | SOM | 1.7 – 75 (g/kg) | 3.66 | 0.90 |

*Cross validation

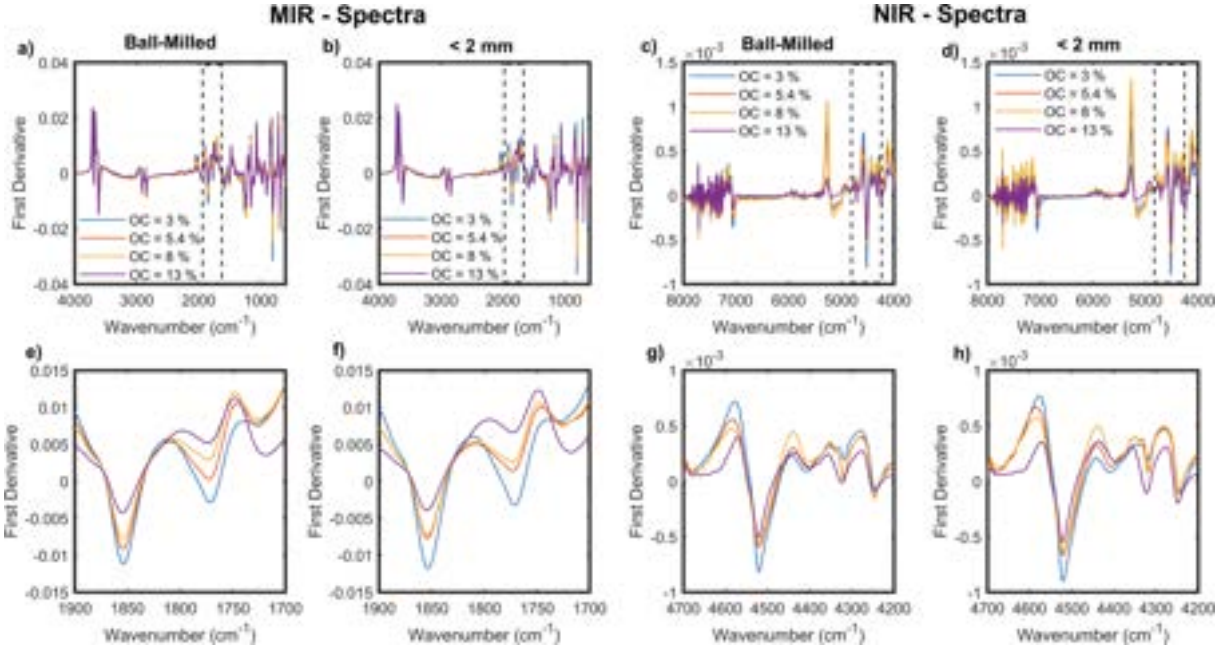


**Fig. 1.** First derivative spectra of four soil samples obtained in two different particle size and their respective amplified regions in MIR and NIR ranges.

**Table 2**
Descriptive statistics of soil samples.

| Parameter | No. Samples | Min. | Max. | Q1 – Q3 | IQR | Mean | Median | Std. |
|---|---|---|---|---|---|---|---|---|
| Clay % (w/w) | 860 | 2.00 | 64.00 | 11 – 27 | 16 | 19.36 | 18 | 10.59 |
| Sand % (w/w) | 860 | 2.00 | 95.00 | 36 – 60 | 24 | 48.41 | 48 | 18.28 |
| Silt % (w/w) | 860 | 1.00 | 77.00 | 26–39 | 13 | 32.27 | 33 | 11.01 |
| pH $CaCl_2$ | 887 | 2.87 | 8.54 | 4.73 – 5.90 | 1.16 | 5.43 | 5.31 | 1.00 |
| pH Water | 888 | 3.71 | 9.15 | 5.46–6.56 | 1.10 | 6.11 | 6.01 | 1.00 |
| OC % (w/w) | 888 | 0.04 | 42.47 | 0.28 – 2.81 | 5.53 | 2.62 | 0.81 | 5.22 |
| CEC (cmol kg$^{-1}$) | 878 | 0.11 | 52.56 | 3.67 – 12.62 | 9.95 | 9.37 | 7.44 | 8.10 |
| Nitrogen % | 888 | 0.00 | 2.62 | 0.05 – 0.32 | 0.27 | 0.24 | 0.12 | 0.32 |
| Aluminium (g kg$^{-1}$) | 602 | 0.02 | 13.16 | 1.21–2.87 | 1.66 | 2.31 | 1.86 | 1.80 |
| TC % (w/w) | 888 | 0.06 | 42.47 | 0.55 – 4.31 | 3.76 | 3.47 | 1.64 | 5.40 |
| Bulk Density (g/cm$^3$) | 438 | 0.20 | 1.95 | 0.91 – 1.42 | 0.51 | 1.14 | 1.12 | 0.34 |

*Q1 – first quartile, Q3 - third quality, IQR - interquartile range

scientists proposed the use of near (NIR) or mid (MIR) infrared in tandem with chemometrics/machine learning algorithms as rapid and low-cost analytical methodology to quantify or semi-quantify, several soil attributes from a single spectrum. In order to speed up and facilitate spectral acquisition, spectra are obtained in reflectance mode using dried sample, sieved to < 2 mm or ball-milled/finely-ground to < 0.100 mm. Table 1 shows a literature review containing the soil particle size used to collect spectra in reflectance mode using MIR/DRIFT (diffuse reflectance infrared Fourier transform spectroscopy) and NIR ranges to determine TC or OC or SOM, which are common soil attributes used to

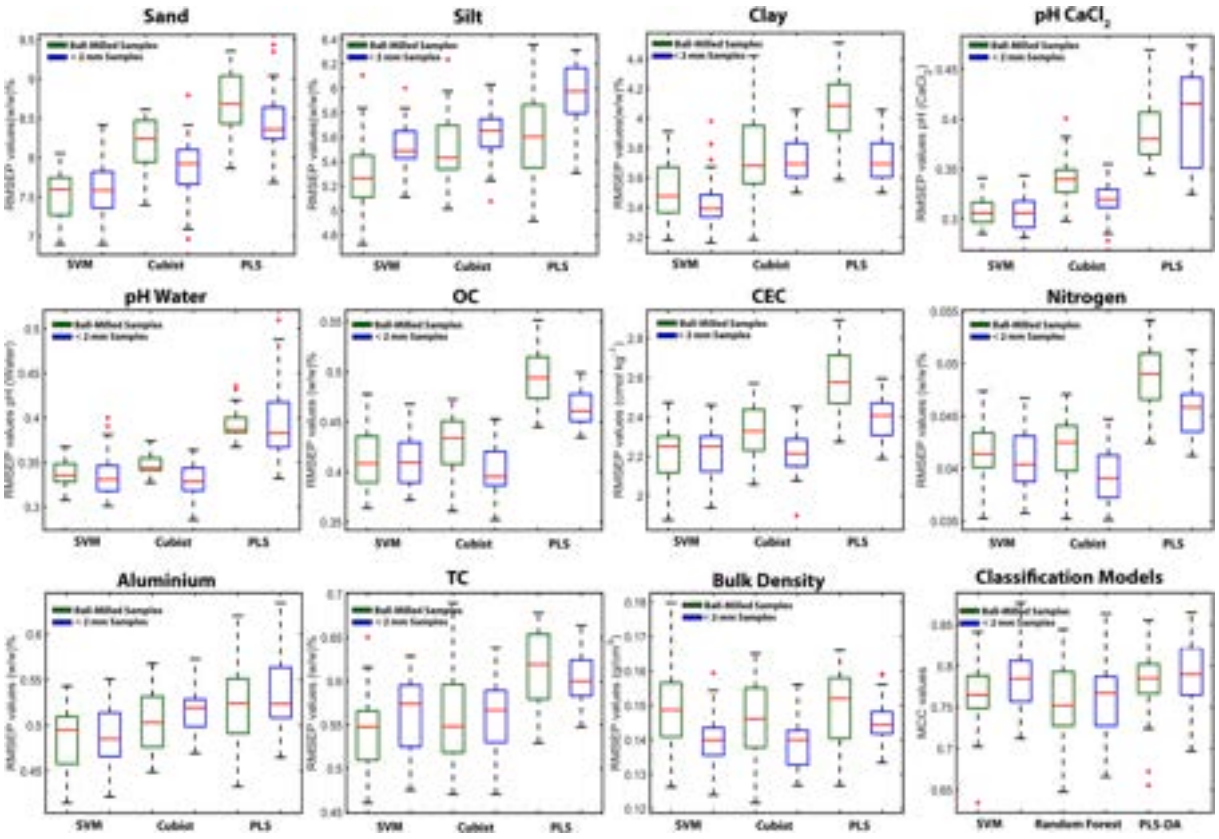


**Fig. 2.** Boxplots of RMSEP values obtained for SVM, Cubist and PLS regression models for ball-milled and < 2 mm samples collected in MIR range; Boxplot of MCC values obtained by SVM, RF, and PLS-DA classification models using ball-milled and < 2 mm samples. Median is represented by the red line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

evaluate carbon stocks. This search was conducted using the Web of Knowledge platform with the follow terms for MIR range: (DRIFT) or (MIR) or (mid-infrared) and (soil) and (carbon). For NIR range the follow terms were used: (NIR) or (vis-NIR) or (visible and NIR) or (NIRS) or (near infrared) and (soil) and (carbon).

As is apparent in Table 1, there is a preference in NIR studies for scanning soil samples with particle size of < 2 mm, while for diffuse reflectance MIR spectroscopy, also known as mid-DRIFT or DRIFT spectroscopy, the samples tend to be ball-milled/grinding soil [15]. This could be attributed to the smaller radiation spot of MIR spectrometers (~2–4 mm) compared to NIR spectrometers (~5–10 mm), which increases the probability of specular reflectance effects in soil samples with particle size of < 2 mm. In addition, ball-milling samples increases the homogeneity of soil samples, which should increase the precision/reproducibility, thus mitigating the impact of the smaller radiation spot of MIR spectrometers. Despite this consensus, there are no studies in literature that perform a systematic comparison on the accuracy of spectral models built using ball-milled and < 2 mm soil samples in NIR or MIR ranges. To the best of our knowledge, there are only studies that investigated the variations of the ball-milling step on the MIR spectra [16,17] or compared the accuracy of partial least squares (PLS) regression models using the MIR spectra collected before and after the ball-milling step [18]. Thus, a systematic comparison using different chemometrics regression and classification methods using the spectra collected in MIR and NIR ranges; and selecting different samples to compose the calibration and validation sets is required. This information is extremely important for researchers and routine laboratories, since ball-milling the soil samples requires specific machinery and additional preparation time, increasing the time and costs associated with soil analysis using infrared spectroscopy techniques.

Ball-milling is completed as an additional preparation step following drying at 30°, disaggregation and sieving to 2 mm. The ball-mill step typically takes 1 – 5 min, depending on the configuration of available equipment (e.g. number of ball mills available, number of milling vessels on each ball mill) and the number of operators deployed. For example, for soil surveys such as Tellus (www.tellus.ie) over 20,000 samples were ball-milled in preparation for analysis and spectral acquisition. The ball-milling step in this survey to a particle size nominally 95 % <0.032 mm accounts for>25 % of the total sample preparation time for pre-dried Tellus soil samples. Moreover, the planetary ball mills and agate vessels used for Tellus samples are relatively expensive equipment that can add significantly to sample preparation costs. Therefore a systematic study involving the improvement/gain in accuracy achieved by ball-milling the soil samples in NIR and MIR ranges is essential for projects with large sample archives requiring high throughput analysis, in order to assess the potential reduction in costs (machinery) and time associated with eliminating the ball-milling.

To that end, the objective of this study was to evaluate the influence of sample particle size on the accuracy of soil parameters predicted in both NIR and MIR ranges. To meet this objective, a national population of soils was used and prediction models for multiple parameters (sand, silt, clay, $pH_{(CaCl2)}$ , $pH_{(Water)}$, TC, OC, CEC, nitrogen, aluminium and bulk density) were developed from MIR and NIR spectra. Soil spectra were collected in both ranges and from samples representing two particle sizes, namely; < 2 mm and < 0.100 mm (ball-milled), resulting in four soil spectral libraries. Furthermore, the potential for ball milling to improve the accuracy of classifying samples by soil horizons was also investigated.

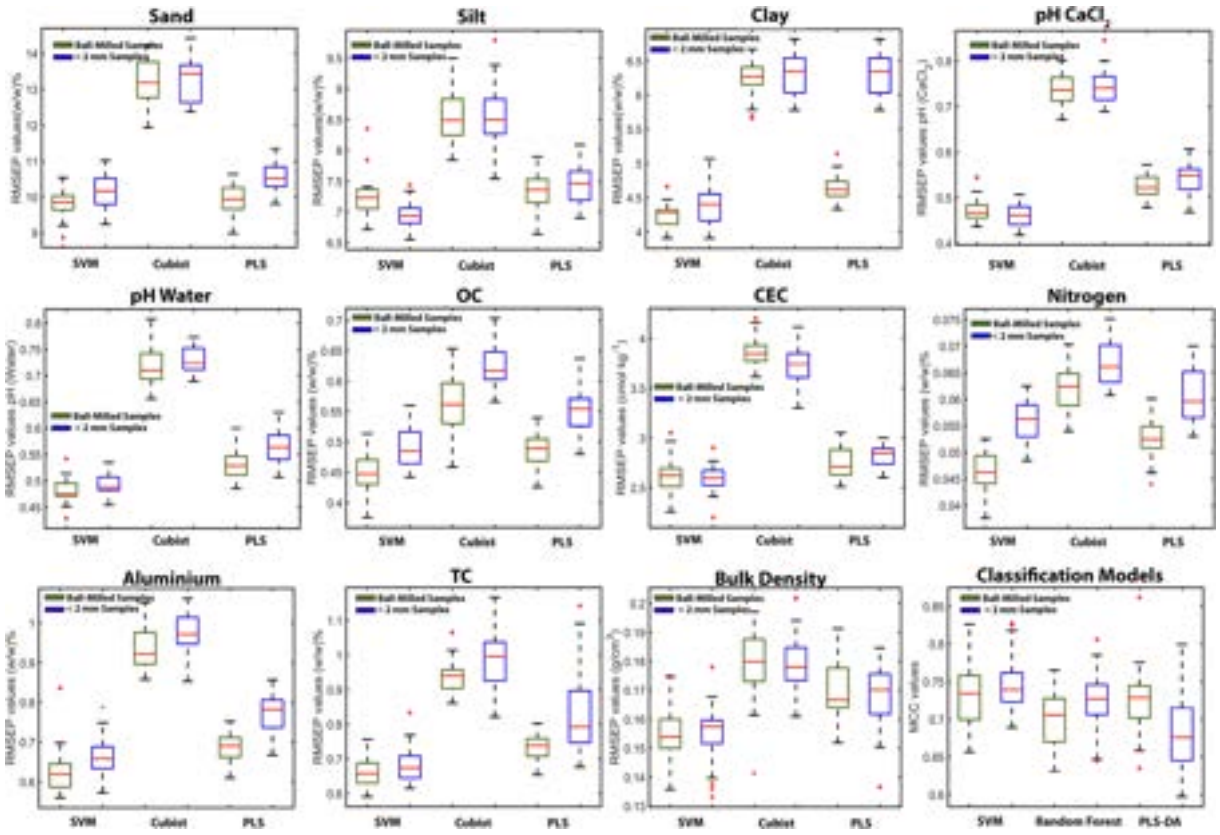


**Fig. 3.** Boxplots of RMSEP values obtained for SVM, Cubist and PLS regression models for ball-milled and < 2 mm samples collected in NIR range; Boxplot of MCC values obtained by SVM, RF, and PLS-DA classification models using ball-milled and < 2 mm samples. Median is represented by the red line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

## 2. Material and methods

### 2.1. Soil archive

A total of 888 soil samples were collected from 225 modal profiles from 14 counties of Ireland: Carlow, Clare, Kildare, Laois, Leitrim, Limerick, Meath, Offaly, Tipperary South, Waterford, Westmeath, Wexford, West Cork, West Mayo and West Donegal at a scale of 1:250,000, which is economically feasible for a soil map of Europe according to European Soil Bureau Network (ESBN) Technical Working Group [19]. These samples belong to an existing archive of the Soil Information System (SIS) Ireland [19] located at Johnstown Castle, Wexford. Soil samples were taken from each soil horizon to a depth of 1 m, resulting in 226, 220, 209, 150, 65, 12 and 1 sample for the first, second, third, fourth, fifth, sixth and seventh horizon respectively. Figure S1 shows the sample locations.

According to Gardiner and Radford's (1980) [20] classification system, the first horizon soil samples used in this study represented 11 Great Groups, mostly from the classes Brown Earth and Surface Water Gley. The predominant soil types, with at least five samples, were classified as: Anthropic Brown Earths, Gleyic Brown Earths, Humic Groundwater Gleys, Humic surface-water Gleys, Stagnic Brown Earths, Stagnic Luvisols, Typical Brown Earths (mostly), Typical Brown Podzolics, Typical calcareous Brown Earths, Typical Lithosols, Typical Luvisols, and Typical Surface-water Gleys. A detailed description about the Irish SIS can be found in the Environmental Protection Agency Report no. 204 [19].

### 2.2. Reference analysis

All soil analyses were performed at the Agriculture and Food Development Authority (Teagasc) laboratory as part of the database for Soil Information System for Ireland [19]. Soil samples were oven-dried at 30 ∘C and mechanically sieved to < 2 mm prior to analysis. Soil texture (sand, clay and silt) was determined using the sieve-pipette method, employing hydrogen peroxide to remove all organic matter and sodium hexametaphosphate as dispersing agent [21]. The $pH_{CaCl2}$ and $pH_{water}$ analyses were conducted using a pH electrode following the addition of $CaCl_2$ solution (0.01 M) or deionized water to the soil samples [22].

Aluminium ($Al^{+3}$) was determined by inductively coupled plasma optical emission spectrometry (IPC-OES) following extraction with sodium dithionite and sodium citrate [21]. CEC values ($Ca^{2+}$, $Mg^{2+}$, $K^{+}$ and $Na^{+}$) were determined by IPC-OES using $BaCl_2$ solution as extraction agent [21]. Soil OC, TC and nitrogen were determined using a LECO TrueSpec CN elemental analyser following the method described at ISO 10694: 1995 – Determination of organic and total carbon after dry combustion [21]. Bulk density was determined using the An Foras Talúntais method, based on ISO 11272:1998 – Determination of dry bulk density, with corrections on the stone mass and volume corrections [21].

### 2.3. MIR and NIR spectral acquisition

Soils were scanned in both MIR and NIR ranges in the diffuse reflectance mode using the FT-IR Spectrometer INVENIO S® (Bruker, Massachusetts - USA) with mercury cadmium telluride (MCT) detector for the diffuse reflection measurements in MIR and NIR ranges. This spectrometer uses two different sources, a globar (u-shaped silicon carbide - SiC) in the MIR range and a tungsten halogen lamp in the NIR range. Switching between the radiation sources is done automatically by the instrument through a moving mirror. The radiation beam strikes the soil samples at an angle of 90° and has a diameter of approximately 2.5

**Table 3**
Median of regression model results obtained by SVM using ball-milled and < 2 mm soil samples in MIR and NIR ranges.

| Parameter | Particle size | MIR | | | | | | NIR | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RMSEP | $R^2_{val}$ | No. Cal/Val. Samp. | RMSEP Gai* | RPIQ | RPIQ Class | RMSEP | $R^2_{val}$ | No. Cal/Val. Samp. | RMSEP Gain* | RPIQ | RPIQ Class |
| Clay | Ball-milled | 3.47 | 0.87 | 538/268 | −2.31% | 4.46 | EM | 4.28 | 0.83 | 532/261 | 2.80% | 3.97 | GM |
| | < 2 mm | 3.39 | 0.88 | 524/261 | | 4.72 | EM | 4.40 | 0.80 | 536/263 | | 3.63 | GM |
| Sand | Ball-milled | 7.6 | 0.82 | 536/267 | −0.26% | 3.42 | GM | 9.85 | 0.69 | 539/271 | 3.15% | 2.33 | FM |
| | < 2 mm | 7.58 | 0.80 | 535/259 | | 2.90 | AQM | 10.16 | 0.68 | 547/266 | | 2.41 | FM |
| Silt | Ball-milled | 5.26 | 0.65 | 505/245 | 4.38% | 2.28 | FM | 7.23 | 0.38 | 518/245 | −4.15% | 1.66 | NRM |
| | < 2 mm | 5.49 | 0.63 | 519/253 | | 2.19 | FM | 6.93 | 0.41 | 514/259 | | 1.84 | NRM |
| pH $CaCl_2$ | Ball-milled | 0.31 | 0.88 | 546/274 | 0.00% | 3.56 | GM | 0.47 | 0.72 | 550/271 | −2.13% | 2.25 | FM |
| | < 2 mm | 0.31 | 0.89 | 536/265 | | 3.63 | GM | 0.46 | 0.73 | 552/276 | | 2.03 | FM |
| pH Water | Ball-milled | 0.33 | 0.87 | 550/270 | 0.00% | 3.24 | AQM | 0.48 | 0.77 | 554/268 | 2.08% | 2.40 | FM |
| | < 2 mm | 0.33 | 0.86 | 535/263 | | 2.63 | FM | 0.49 | 0.72 | 544/278 | | 1.95 | NR |
| OC | Ball-milled | 0.41 | 0.92 | 518/254 | 0.00% | 4.49 | EM | 0.45 | 0.93 | 517/256 | 6.67% | 4.45 | EM |
| | < 2 mm | 0.41 | 0.93 | 516/252 | | 4.34 | EM | 0.48 | 0.90 | 518/259 | | 3.89 | GM |
| CEC | Ball-milled | 2.25 | 0.85 | 523/266 | 0.00% | 3.69 | GM | 2.62 | 0.79 | 535/262 | −0.76% | 2.97 | AQM |
| | < 2 mm | 2.25 | 0.82 | 528/249 | | 3.68 | GM | 2.60 | 0.78 | 535/268 | | 3.17 | AQM |
| Nitrogen | Ball-milled | 0.04 | 0.94 | 515/253 | 0.00% | 5.08 | EM | 0.05 | 0.92 | 517/256 | 20.00% | 4.20 | EM |
| | < 2 mm | 0.04 | 0.93 | 510/246 | | 5.05 | EM | 0.06 | 0.88 | 528/259 | | 4.21 | EM |
| Aluminium | Ball-milled | 0.49 | 0.82 | 353/169 | −2.04% | 2.93 | AQM | 0.62 | 0.72 | 359/173 | 6.45% | 2.34 | FM |
| | < 2 mm | 0.48 | 0.83 | 349/179 | | 2.88 | AQM | 0.66 | 0.71 | 366/183 | | 2.32 | FM |
| TC | Ball-milled | 0.41 | 0.92 | 518/254 | 0.00% | 6.36 | EM | 0.66 | 0.92 | 531/256 | 3.03% | 4.29 | EM |
| | < 2 mm | 0.41 | 0.93 | 516/252 | | 5.04 | EM | 0.68 | 0.92 | 529/258 | | 5.46 | EM |
| Bulk Density | Ball-milled | 0.15 | 0.78 | 265/127 | −6.67% | 3.33 | AQM | 0.15 | 0.74 | 269/131 | 6.67% | 2.84 | AQM |
| | < 2 mm | 0.14 | 0.79 | 266/134 | | 3.72 | GM | 0.16 | 0.72 | 270/135 | | 3.10 | AQM |

*MSEP gain in ball-milling the samples = -100 × ($RMSEP_{ball\text{-}milled}$ − $RMSEP_{<2\ mm}$) /$RMSEP_{ball\text{-}milled}$
EM – Excellent Model, GM – Good Model, AQM – Approximate Quantitative Model, FM – Fair Model, NRM – Non-Reliable model.

mm.

The spectrometer was coupled with an autosampler, allowing analysis of ~ 60 samples per hour in both ranges. The autosampler is composed of a microplate reader extension for high-throughput screening (HTS-XT) from Bruker (Bruker, Massachusetts - USA). As shown in Fig. S2 the soil samples were filled into micro-plates of 7 mm internal diameter and the surface of each was smoothed with a blade before the spectrum acquisition. As shown in Fig. S2 the ball-milled spectra can be easily smoothed, however the < 2 mm soils could not present a smoothed surface like the ball-milled samples. The spectra were collected in MIR (4,000–600 $cm^{-1}$) and NIR (10,000–4,000 $cm^{-1}$) ranges with spectral resolution of 1.43 $cm^{-1}$, employing 32 scans to maximize the signal-to-noise ratio. At the beginning of each analysis the background spectrum was collected using a gold reference disk.

Soil spectra were collected once in both ranges, without replicates. In order to evaluate the spectrometer repeatability the instrument was checked every day using a reference compound (Acetaminophen – USP specification), both in the morning prior to the first analysis and again in the evening. The MIR and NIR Acetaminophen spectra were monitored using a multivariate control chart based on the principal component analysis (PCA) using D- and Q- statistics at significance level of probability of 0.05 [23]. This step enables monitoring of the spectrometer by checking the repeatability of the acetaminophen spectra over time, increasing the confidence and accuracy of the spectral models.

### 2.4. Spectral pre-processing

The reflectance NIR and MIR spectra were converted to pseudo absorbance log (1/Reflectance), after both spectra were pre-processed using different pre-processing algorithms: smoothing, first and second derivative Savitzky–Golay smoothing, standard normal variate (SNV) and multiplicative signal correction (MSC) [24,25]. For each pre-processing algorithm tested a PLS or partial least squares for discriminant analysis (PLS-DA) model was built to identify which pre-processing provided the most accurate results.

### 2.5. Selection of samples for calibration and validation sets

Four soil spectral libraries were created from (1) ball-milled samples in MIR range, (2) < 2 mm samples in MIR range, (3) ball-milled samples in NIR range and (4) < 2 mm samples in NIR range. These were used to compare the accuracy of ball-milled and < 2 mm soil samples on the prediction of sand, silt, clay, $pH_{(CaCl2)}$, $pH_{(Water)}$, TC, OC, CEC, nitrogen, aluminium and bulk density values. These soil spectral libraries were also used to classify the soil samples by horizon, i.e. horizon 1 or horizons 2–7.

In order to increase the variability in the results and obtain representative accuracy values for each spectral library, calibration and validation samples were randomly selected 29 times for each soil attribute. In other words, for regression purposes 319 (11 soil attributes × 29 random selections) calibration and validation sets were generated for each spectral library, while for classification purposes 29 training and test sets were generated for each spectral library.

### 2.6. Regression models

In order to investigate the influence of ball-milled and < 2 mm soil samples on model accuracy, a range of chemometrics and machine

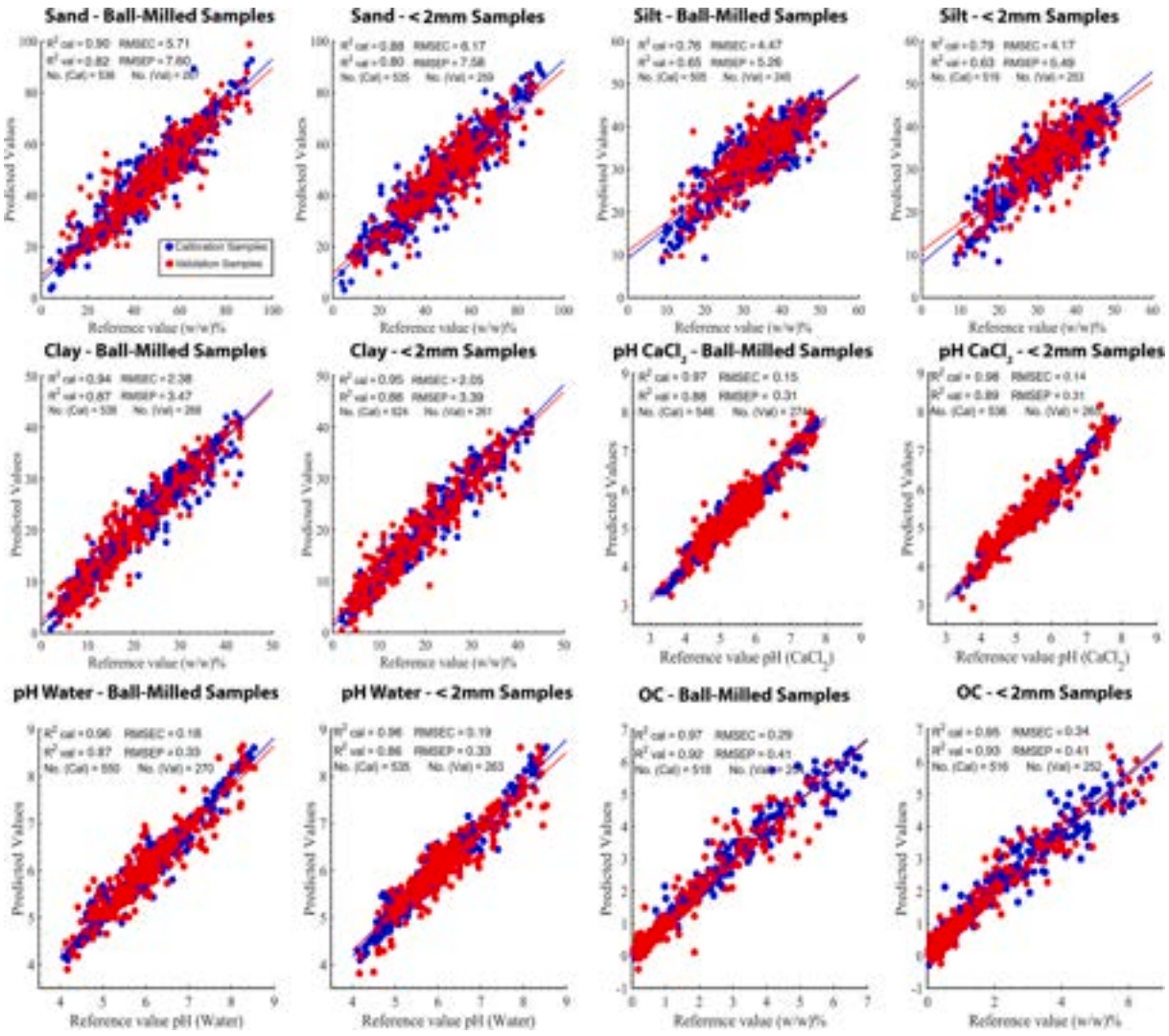


**Fig. 4.** Plot of reference versus predicted values from MIR spectra by SVM regression models for: sand, silt, clay, pH $CaCl_2$, pH Water, and OC.

learning methods were evaluated. For each calibration set generated, a SVM, Cubist and PLS regression model was built and the validation set was predicted. These algorithms were selected due to their wide use in the literature to build regression models from MIR and NIR soil spectral libraries [4,8,12]. To standardize calibration and validation sets for all regression models for each calibration set generated, outlier samples were identified by the Q residues and Hotelling $T^2$ values calculated by the PCA. Samples with Q and $T^2$ values above the critical limits calculated at significance level of probability of 0.01 were removed as outliers, then the regression models were built [26].

The PLS algorithm can be considered the standard multivariate regression algorithm used in vibrational spectroscopy data. Furthermore, PLS is commonly used to determine quantitative physical or chemical soil parameters [12,15,26]. The PLS algorithm performs the decomposition of the spectral data matrix, **X**, into the product of two matrices, scores and loadings. During this process the matrix **X** is correlated with a feature of interest, **y** vector (PLS – 1), in this case a soil attribute. This last process generates the other two vectors, the weighting (**w**) and regression coefficients (**b**) [27]. The number of latent variables (LV) must be optimized to obtain feasible results by extracting useful information from the calibration step. For this, a *k*-fold cross-validation (10 folds) method was used, and the LV that gave the minimum value of root mean square error of cross-validation (RMSECV) was selected [28]. Comprehensive concepts regarding PLS and its implementation in spectroscopy applications can be found at [27].

The SVM regression algorithm mapped the matrix **X** into a non-linear high-dimensional feature space using a kernel function, the radial basis function most used for soil spectral libraries applications [29]. Using this kernel function the hyper-parameters Epsilon, Cost function and Sigma must be optimized. In this study Bayesian optimization was employed in each calibration set to reduce the computer costs required to find this optimum combination [30].

The Bayesian algorithm aims to minimize the objective function, which is directly related to the RMSECV, by testing different combinations of the hyper-parameters. However, instead of testing all the combinations like the grid search method, this algorithm selects the hyper-parameter combination which provides most potential improvement (higher probability) on the objective function [30,31]. Details about the SVM for regression proposes can be found at [32].

The Cubist regression algorithm is an extension of Quinlan's M5 algorithm and can be briefly summarized as a combination of decision trees to generate the nodes, grouping the samples according to their similarity, separating the complex dataset into groups of lower variability to then build linear multivariate regression models in each segment [33]. In order to optimize the accuracy of the results, a boosting–like scheme called "committees" was employed while the corrections in predictions were based on nearest neighbours of calibration samples. To find the optimum combination of both hyper-parameters

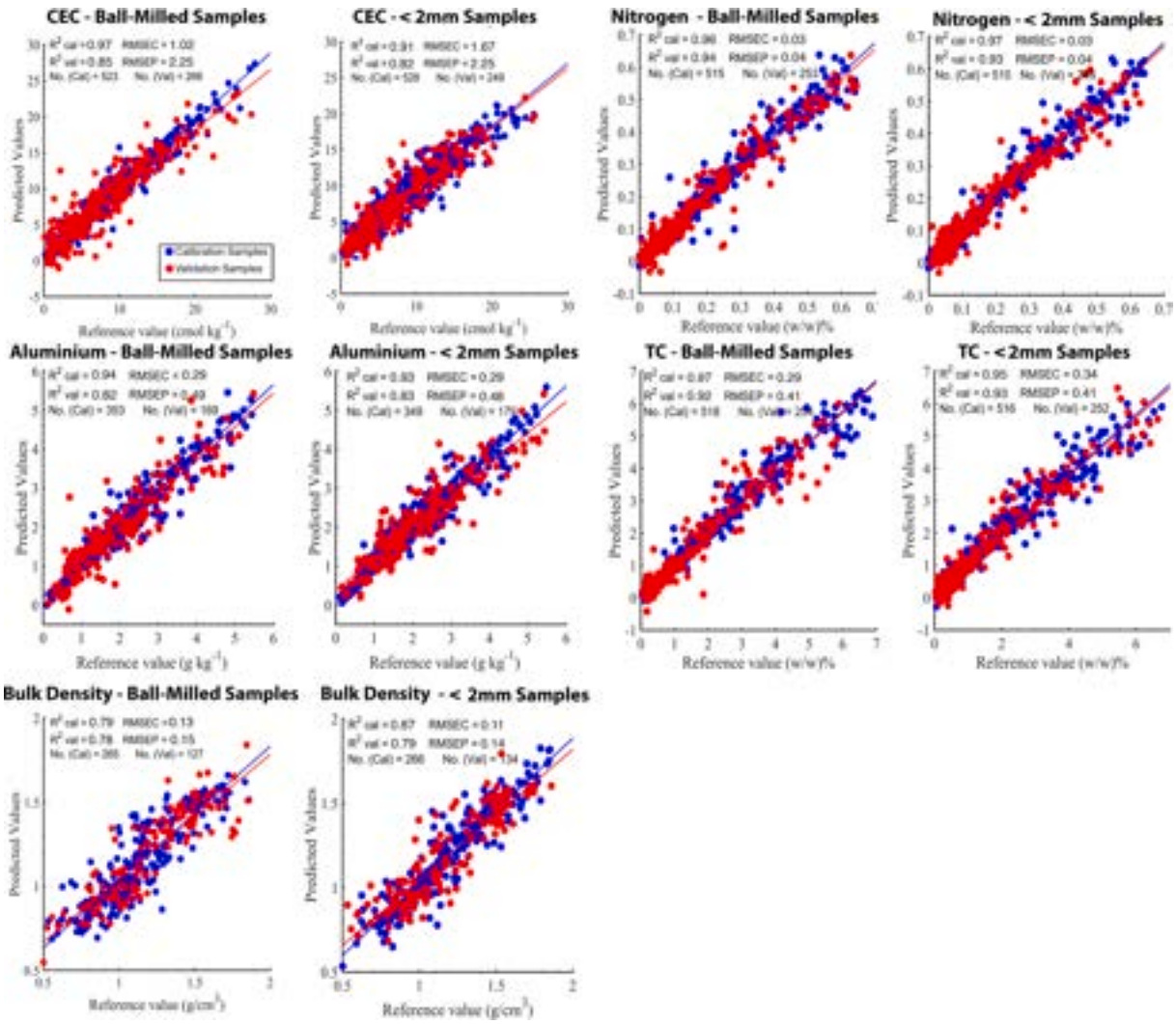


**Fig. 5.** Plot of reference versus predicted values from MIR spectra by SVM regression models for: CEC, nitrogen, aluminium, TC, and bulk density.

the grid search method with *k*-fold cross-validation (10 folds) was used for each soil parameter. Comprehensive details about the Cubist algorithm, committees and corrections based on nearest neighbours can be found at [33].

### 2.7. Classification models

The same approach used for regression models was used to build the classification models. For each training and test set randomly generated, 29 for each soil spectral library, SVM, Random Forest (RF) and PLS-DA classification models were built and the test set was predicted.

The PLS-DA classification algorithm is widely used in vibrational spectroscopy data and is present in several statistical software packages, like the PLS algorithm. The procedure used to build the PLS-DA classification model is the same as that used by the PLS but, instead of using a **y** vector containing a feature of interest, a discrete variable to describe the class of each sample was used. Typically, a value of 1 is assigned to the class of interest and a value of 0 or −1 is assigned to another class. Despite using a discrete variable, the **y** vector is modelled as a continuous variable, resulting in predicted values not exactly −1 or 1, requiring the calculation of a threshold value to separate the classes. In this study Bayes theorem was used to calculate the optimum threshold between the classes using *k*-fold cross-validation (10 folds). Samples predicted with **y** values above the calculated threshold are predicted as class 1, the remainder are predicted as class −1 [34].

Similar to the SVM regression algorithm, the SVM classification algorithm mapped the matrix **X** into a non-linear high-dimensional space using a kernel function, in this case the radial basis function. In this high-dimensional space a hyperplane is calculated to separate samples from both classes. The optimization of Cost function and Sigma hyperparameters were performed using Bayesian optimization for each training set [30]. Details about the SVM for classification proposes can be found at [35].

Random Forest is an ensemble method based on decision trees combined with two randomization processes, the bagging and random feature selection [36,37]. The bagging randomization process consists of building each individual tree (classifier) using a bootstrap replica of the training set with a replacement. Two thirds of bootstrap samples, in bag samples, are selected to build each tree and the remaining samples, out-of-bag samples, are used to estimate the model performance of this individual tree, similar to a cross-validation process [30,36]. Instead of using all variables to build each tree, the random feature selection randomly select *m* variables, also known as *mtry*, to be used at each node to build each tree. The default value of *mtry* for classification is the square root of the total number of variables/features [36]. In this study the ensemble of trees was composed of 300 trees, the *mtry* default value was used and samples predictions were made by the majority classification vote of the ensemble of trees.

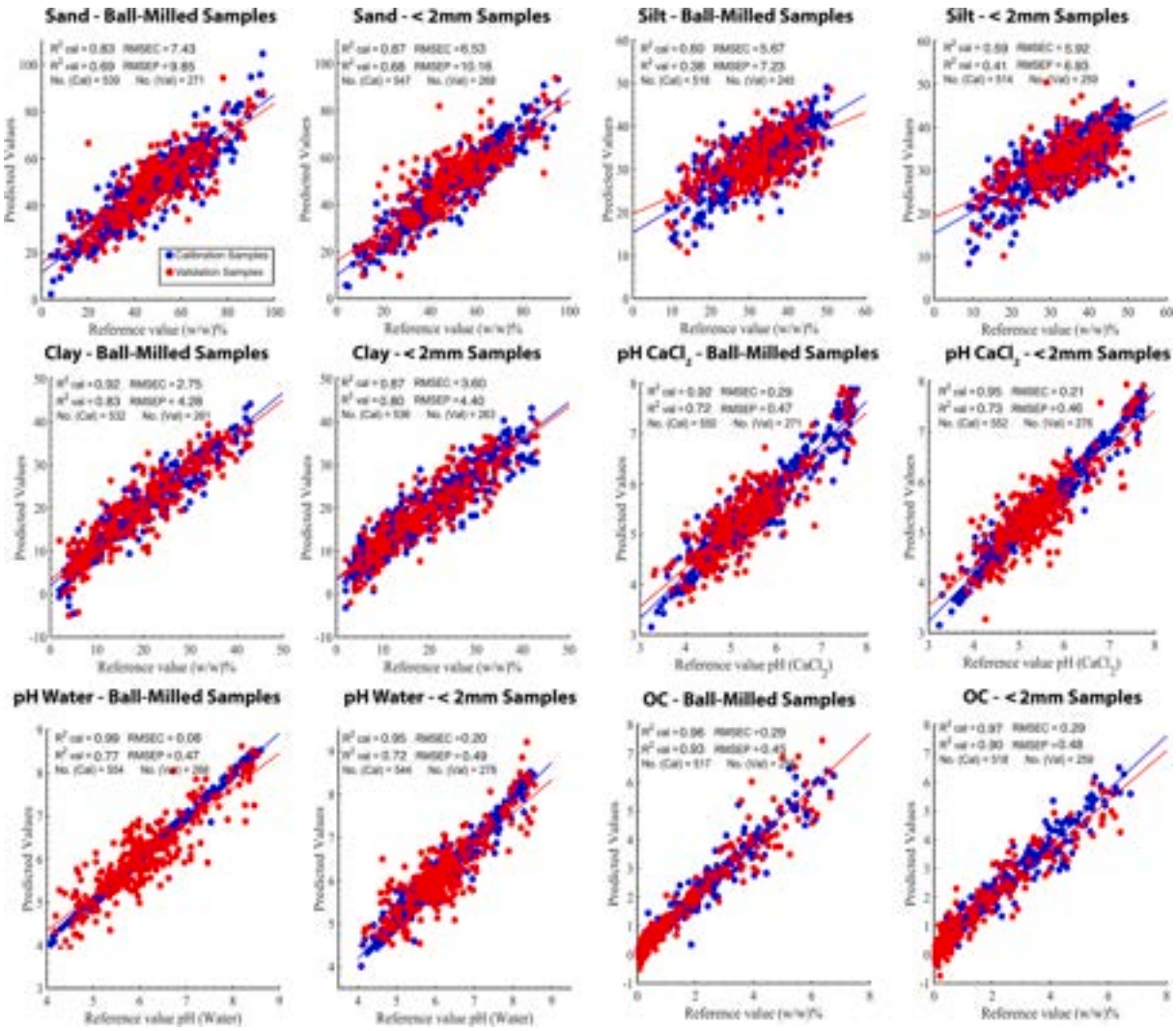


**Fig. 6.** Plot of reference versus predicted values from NIR spectra by SVM regression models for: sand, silt, clay, pH $CaCl_2$, pH Water, and OC.

### 2.8. Accuracy comparison

#### 2.8.1. Comparison between regression models

The RMSEP value is the metric most used to assess the accuracy of multivariate regression models compared to the reference method, and is calculated using (Eq. (1)), where $n$ is the number of samples in the validation set, $\widehat{y}_i$ is the predicted value and $y_i$ is the reference value. This value is used to estimate the errors on the predictions using the proposed method, i.e. for homoscedastic models we expected that ~ 68.2% of the samples should be predicted with errors ± 1 × RMSEP, 95.4% with ± 2 × RMSEP and 99.7% with ± 3 × RMSEP.

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

Another accuracy parameter commonly used in spectroscopy models for soil analysis with wide variation in composition is the ratio of performance to interquartile distance (RPIQ), which is calculated as the interquartile distance (Q3 – Q1) divided by the RMSEP [38]. Using the RPIQ values, regression models can be classified using the thresholds proposed by Ludwig et al. (2017) [38]. In other words: "Excellent Model – EM" with RPIQ > 4.05, "Good Model - GM" with RPIQ ≥ 3.38 & ≤ 4.05, "Approximate Quantitative Model - AQM" with RPIQ ≥ 2.70 & < 3.38, "Fair Model - FM" (model can distinguishing between high and low values) with RPIQ ≥ 2.02 & < 2.70; and "Non-Reliable Model – NRM" for RPIQ < 2.02 [38]. As we can see in the thresholds proposed by Ludwig et al. (2017), regression models show an improvement in accuracy when the differences between RPIQ values are higher than 0.68, that is 3.38–2.70 = 0.68 RPIQ [38]. In this paper, both accuracy parameters were used to check for differences between ball-milled and < 2 mm spectra models. Only models with RMSEP gain (Eq. (2)) higher than 5% and RPIQ differences higher than 0.68, simultaneously, were considered different.

$$\mathrm{RMSEP\,gain} = -100 \times \frac{\mathrm{RMSEP_{ball-milled}} - \mathrm{RMSEP_{2mm}}}{\mathrm{RMSEP_{ball-milled}}} \tag{2}$$

#### 2.8.2. Comparison between classification models

The Matthews correlation coefficient (MCC) was used instead of the classical accuracy (Eq. (3)) to evaluate the quality of binary classification models due to the presence of unbalanced classes, 226 samples from horizon 1 (Class 1) and 657 samples from the other horizons (2 – 7) (Class −1). The MCC is a geometric mean, calculated using (Eq. (4)), where the TP (true positive) are the samples from the horizon 1 classified as samples from horizon 1 and TN (true negative) are the samples from the other horizons (2–7) classified as samples from the other horizons. False positive, FP are samples from the other horizons (2 – 7) classified as samples from horizon 1 and FN (false negative) are the samples from horizon 1 classified as samples from other horizons (2 – 7).

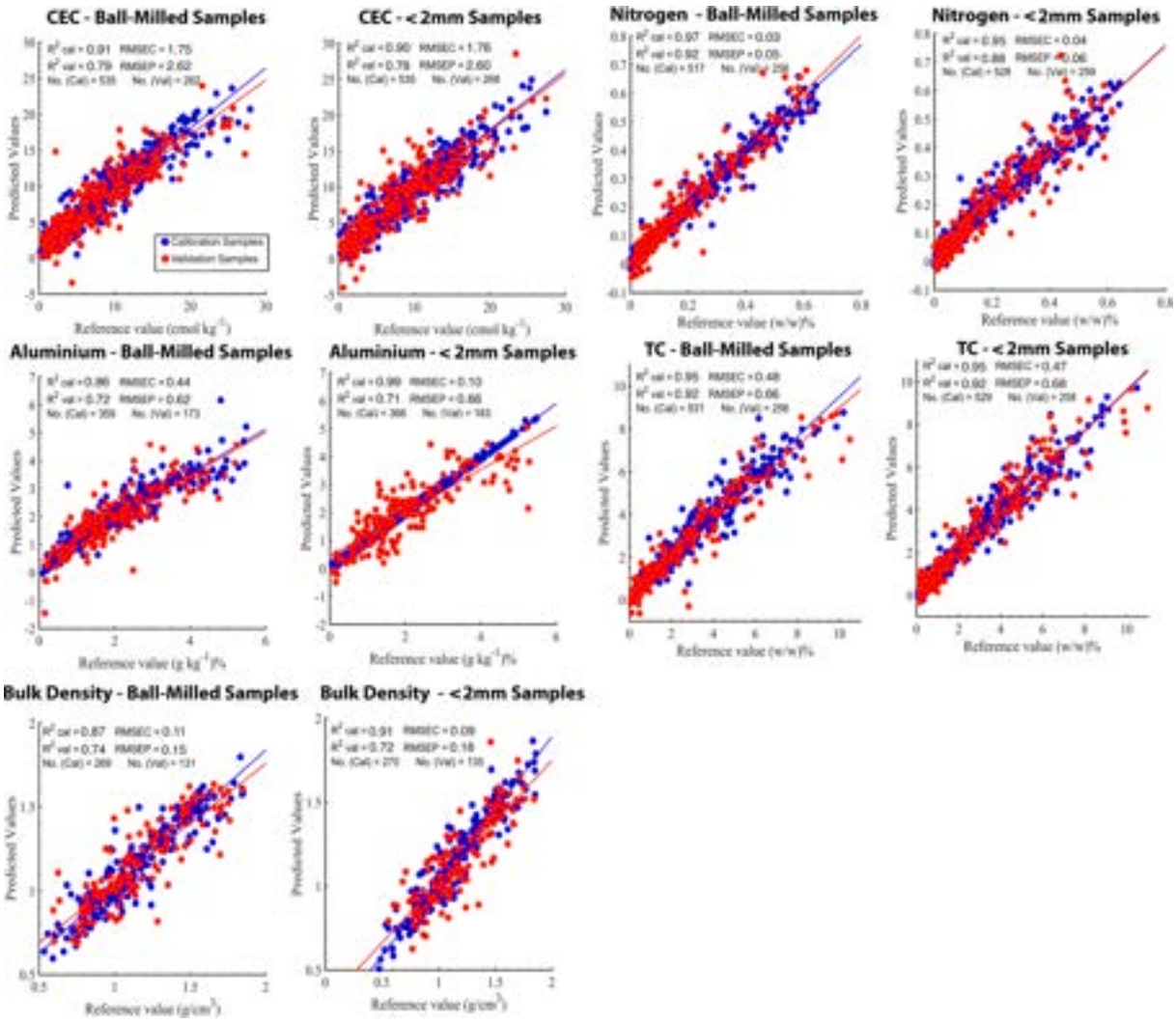


**Fig. 7.** Plot of reference versus predicted values from NIR spectra by SVM regression models for: CEC, nitrogen, aluminium, TC, and bulk density.

**Table 4**
Median of classification model results obtained by SVM using ball-milled and < 2 mm soil samples in MIR and NIR ranges.

| Parameter | Particle size | MIR | | | NIR | | |
|---|---|---|---|---|---|---|---|
| | | MCC | No. Cal/Val. Samp. | Gain* | MCC | No. Cal/Val. Samp. | Gain* |
| Horizon Classification | Ball-milled | 0.76 | 578/285 | −2.63% | 0.73 | 577/285 | −1.37% |
| | < 2 mm | 0.78 | 573/283 | | 0.74 | 578/285 | |

*MCC gain in ball-milling the samples = (MCC $_{\text{ball-milled}}$ − MCC $_{<2\text{ mm}}$) / MCC $_{\text{ball-milled}}$

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

The MCC results can vary between − 1 and + 1, where the value of + 1 represents a completely accurate classification, 0 a random prediction, and − 1 a completely inaccurate classification [34].

### 2.9. Data analysis

All calculations were performed in RStudio version 4.0.5 and MATLAB version R2017a (Mathworks), using an Intel Core i5-8250 CPU with four cores and eight threads with 8 GB memory. Cubist, Random Forest, PLS and PLS-DA models were built using Cubist version 0.3.0, Ranger version 0.13.1 and PLS version 2.8–0 packages for R, respectively. Support Vector Machine models were built using the "Statistics and Machine Learning Toolbox" version 11.1.

## 3. Results and discussion

### 3.1. MIR and NIR spectra

In order to visualise spectral differences between ball-milled and < 2 mm samples, four soil samples with different OC content were selected from first horizon profiles and illustrated in Fig. 1. This figure shows the MIR and NIR spectra of these samples after using the first derivative Savitzky–Golay smoothing algorithm, with a window size of 11 points and a second order polynomial. This pre-processing was selected to enhance the spectral features and reduce the baseline variation caused

**Table 5**
Prediction classes obtained by SVM classification models for ball-milled and < 2 mm samples in MIR and NIR ranges.

| MIR | | | | |
|---|---|---|---|---|
| | Ball-milled Samples | | <2 mm Samples | |
| | Actual Class | | Actual Class | |
| Predicted class | Horizon 1 | Horizons (2–7) | Horizon 1 | Horizons (2–7) |
| Horizon 1 | 67 (TP) | 21 (FN) | 64 (TP) | 14 (FN) |
| Horizons (2–7) | 7 (FN) | 190 (TN) | 10 (FN) | 195 (TN) |
| NIR | | | | |
| | Ball-milled Samples | | <2 mm Samples | |
| | Actual Class | | Actual Class | |
| Predicted class | Horizon 1 | Horizons (2–7) | Horizon 1 | Horizons (2–7) |
| Horizon 1 | 59 (TP) | 14 (FN) | 61 (TP) | 16 (FN) |
| Horizons (2–7) | 15 (FN) | 197 (TN) | 13 (FN) | 195 (TN) |

by the differences in radiation penetration and light paths due to the differences in particle sizes. The range from 10,000 to 8,000 $cm^{-1}$ (1,000 – 1,200 nm) was trimmed from the NIR spectra due to the presence of high instrumental noise.

The ball-milled and < 2 mm soil spectra show absorption bands common to MIR and NIR soil spectra reported in previous studies [5,11,12,37–39]. In the MIR range the absorption bands identified at 3700 – 3600 $cm^{-1}$ were attributed to kaolinite, smectite, illite and other alumino-silicates and those at 3400 $cm^{-1}$ were attributed to –OH residual soil water and residual water at alumina-silicates structures. The absorptions at 2930 – 2850 $cm^{-1}$ may be attributed to alkyl groups, at 2000 – 1820 $cm^{-1}$ range the absorptions are related to quartz, and at 1750 – 1680 $cm^{-1}$ to C = O of carboxylic acid and esters [18]. Proteins, phenolics (C–OH) and carbohydrates (C–O) are responsible for absorptions bands at 1680 and 1525; 1285 and 1100 $cm^{-1}$, respectively [15,39–41]. The band at 1100 $cm^{-1}$ can also be attributed to $SiO_2$, related to quartz [39].

The absorption bands identified in the NIR range at 7800 – 7020 $cm^{-1}$ (1282 – 1425 nm) were attributed to the first overtone of structural –OH stretching mode. This band and the band at 5263 $cm^{-1}$ (1900 nm) are also related to a combination of vibrations of –OH water bound in the interlayer lattices as hydrated cations and water adsorbed on particle surfaces [38].

The absorption band at 4950 – 4900 $cm^{-1}$ (2020 – 2041 nm) can be attributed to amides and those at 4705 $cm^{-1}$ (2125 nm) can be related to the C–H stretch/C–O stretch combination and N-H deformation/N-H stretch combination. The –OH stretch and metal–OH bend are responsible for the absorptions bands at 4545 – 4000 $cm^{-1}$ (2200 – 2500 nm), while the band at 4425 $cm^{-1}$ (2260 nm) is due to Al–OH (gibbsite). That at 4545 $cm^{-1}$ (2200 nm) is related to kaolinite, smectite and illite and that at 4277 $cm^{-1}$ (2338 nm) is attributed to calcite ($CaCO_3$) [11,40,41].

The MIR spectra of four soil samples before (<2 mm - Fig. 1b) and after (Fig. 1a) the ball-milling step are similar, with correlation coefficients (r) of 0.9924, 0.9838, 0.9842 and 0.9752 for the same sample analysed in both particle sizes. These samples showed good agreement between the spectra and OC content in the full and amplified region

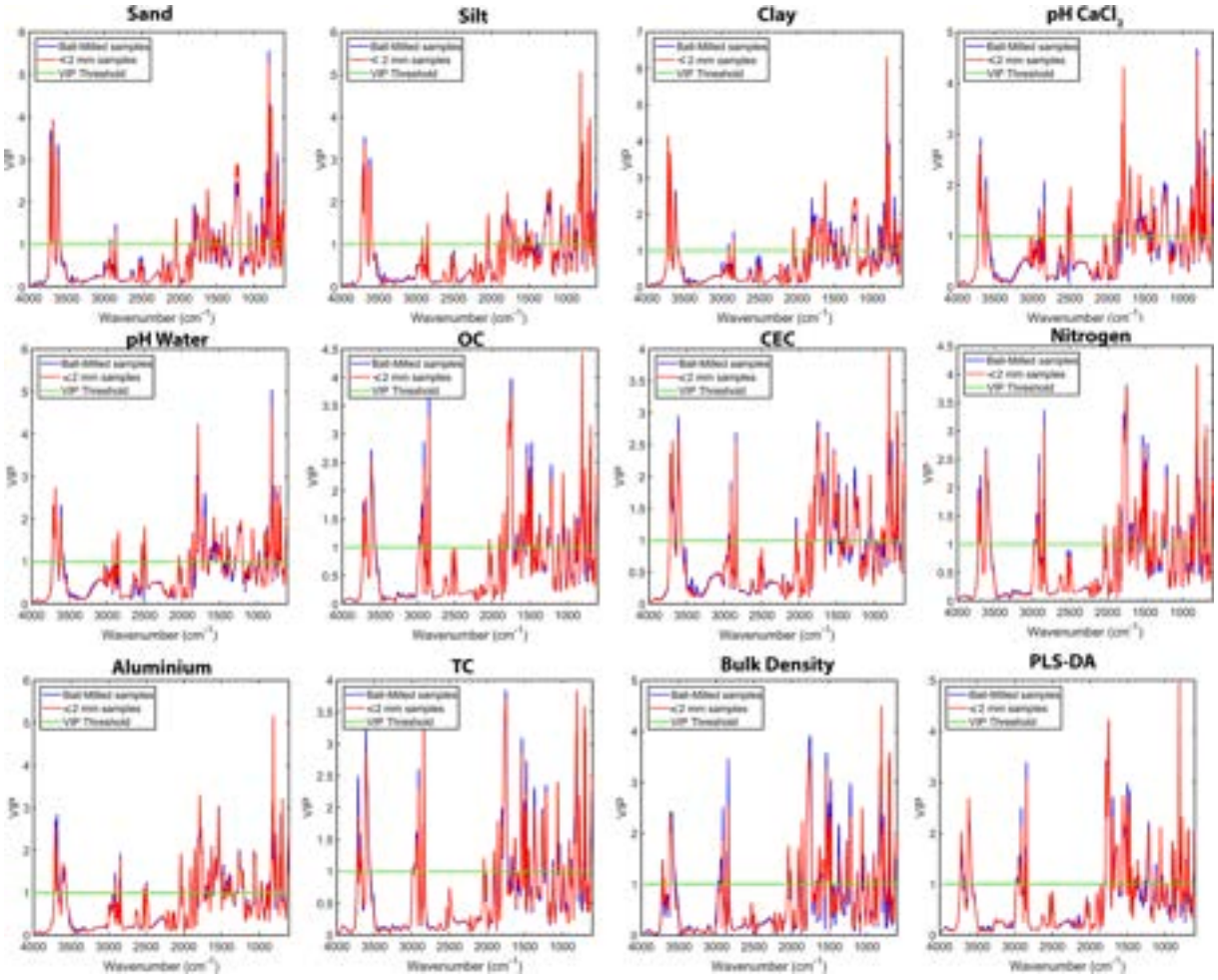


**Fig. 8.** Variable importance in projection (VIP) for PLS regression models using ball-milled and < 2 mm samples collected in MIR range; VIP values for PLS-DA classification models using ball-milled and < 2 mm samples.

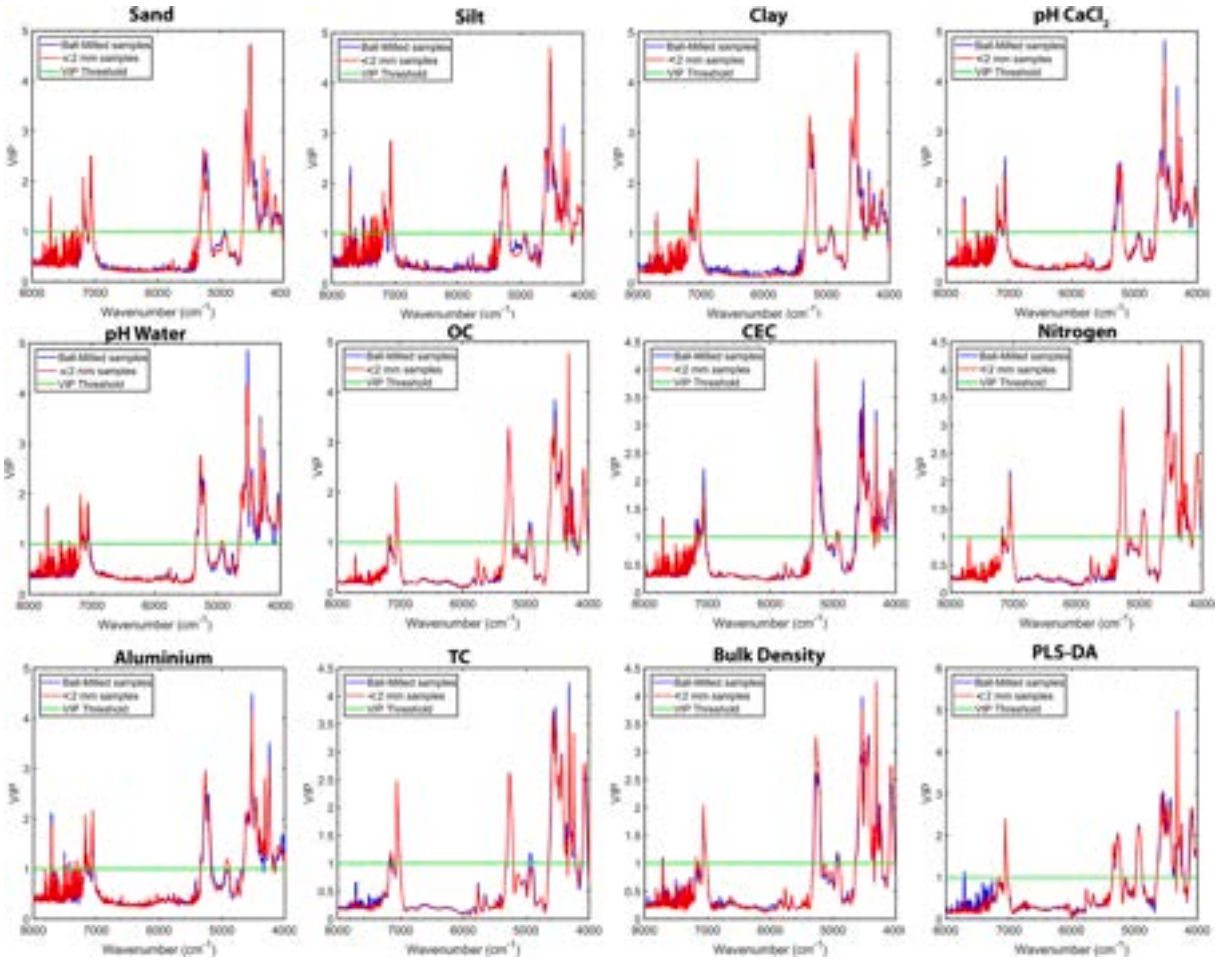


**Fig. 9.** Variable importance in projection (VIP) for PLS regression models using ball-milled and < 2 mm samples collected in NIR range; VIP values for PLS-DA classification models using ball-milled and < 2 mm samples.

**Table 6**
Most important variables selected by VIP scores and their respective soil components responsible for the absorptions.

| MIR | | | NIR | | | |
|---|---|---|---|---|---|---|
| Wavenumber ($cm^{-1}$) | Soil Component | Ref. | Wavenumber ($cm^{-1}$) | Wavelength (nm) | Soil Component | Ref. |
| 3730–3600 | Clay Minerals (kaolinite, smectite and illite) | [15] | 7800–7020 | 1282–1425 | Clay Minerals (–OH) and –OH water | |
| 2930 – 2850 | Organic Matter (alkyl, –CH2), $SiO_2$ | [15,43] | 5670–5210 | | | |
| 2540 | Carbonates (C = O) | [5] | 4950–4900 | 2020 – 2041 | Organic Matter (amide) | |
| 2044 | Organic Matter (N = C = S) | [44] | 4545–4000 | 2200–2500 | –OH and Clay Minerals (Metal-OH) | [15] |
| 1850–1800 | Organic Matter (C = O) | [43] | 4545 | 2200 | Clay Minerals (kaolinite, smectite and illite) | [15] |
| 1780 | $SiO_2$ | [5,43] | 4425 | 2260 | Clay Minerals (Al–OH, gibbsite) | |
| 1750 | Organic Matter (carboxylic acid, –COOH) | [15] | 4277 | 2338 | Clay Minerals (Calcite, $CaCO_3$) | [11,15] |
| 1670 | Organic Matter (Protein amide) | [15] | 4113 | 2431 | Clay Minerals (Al–OH) and –OH | [45] |
| 1627 | Water associated | [15] | 4090 | 2445 | Clay Minerals (illite) | [15] |
| 1525 | Organic Matter (protein amide) | [15] | | | | |
| 1480–1350 | Organic Matter (–CH) | | | | | |
| 1285 | Organic Matter (phenolics, C–OH) | [15] | | | | |
| 1100–1000 | $SiO_2$ and Organic Matter (Carbohydrates) | [15,43] | | | | |
| 930–910 | $SiO_2$ | [43] | | | | |
| 815 | Clay Minerals (Al-OH) and Organic Matter (cellulose) | [43] | | | | |
| 705 | Water, Organic Matter (C-H) and iron Oxides | [15,43] | | | | |

(Fig. 1e and Fig. 1f). The high intensities of the bands at 1850 $cm^{-1}$ and 1770 $cm^{-1}$ (Fig. 1e and Fig. 1f) are attributed to samples with low OC content. Samples from both particle sizes also yielded similar results in the NIR range (Fig. 1c and Fig. 1d), with r values of 0.9948, 0.9922, 0.9969 and 0.9700 for the same four ball-milled and < 2 mm samples. These samples also display similar agreement between the spectra and OC content, i.e. samples with low OC content have high-intensity bands at 4575 $cm^{-1}$ (2186 nm) and 4525 $cm^{-1}$ (2210 nm) (Fig. 1e and Fig. 1f).

Despite displaying excellent agreement between the spectra and OC content, as shown at 1860 and 1770 $cm^{-1}$ (MIR range) and at 4575 and

4525 $cm^{-1}$ (NIR range), the proportional relationship between the spectral intensities and OC content were not exactly the same, especially in samples with low %OC (5.4 % and 8.0 %). This observation indicated that the relationships between the spectral intensity and the soil attributes, such as OC, may not be the same for both particle sizes. This suggests the need to build regression or classification models for each soil particle size to correlate the spectral intensities with the soil attributes, which are likely to be different for the ball-milled and < 2 mm samples.

### 3.2. Regression models

#### 3.2.1. Descriptive statistics

The descriptive statistics of all 888 soil samples collected from 225 modal profiles representing 14 counties of Ireland are summarized in Table 2. The large ranges for the soil attributes analysed reflect the wide variation in type, horizon and depth of soil samples from around Ireland.

The histogram profile of soil attributes are shown in (Fig. S3). It is noted there were some samples not representative in the whole range, so instead of building the regression models using unrepresentative samples (few samples for a large range of soil attributes) we built the regression models using representative samples only, these samples are indicated inside the red rectangle.

#### 3.2.2. Comparison of model performance

All pre-processing algorithms were tested using the PLS regression algorithm. In this case the pre-processing with lowest RMSECV for most soil attributes analysed was the first derivative and smoothing (Savitzky-Golay) using a window size of 11 and a second order polynomial [25]. The boxplots containing the MCC values calculated by SVM, RF, and PLS-DA classification models built using 29 randomly training and test sets are shown in Fig. 2 (MIR) and Fig. 3 (NIR).

The boxplots containing the root mean square error on prediction (RMSEP) values obtained by the SVM, Cubist and PLS regression models built using 29 randomly calibration and validation sets are also shown in Fig. 2 (MIR) and Fig. 3 (NIR).

In the MIR range, the PLS regression algorithm yielded the worst median RMSEP results for all 11 soil attributes using soil samples with both particle sizes. Furthermore, the variation on the RMSEP results obtained by PLS models were, in general, higher than those obtained by SVM and Cubist models. This can be mainly attributed to two reasons: first, outliers samples negatively affect PLS more than Cubist and SVM algorithm. The use of a set of trees present in Cubist and the use of slack variables present in SVM were able to minimize the influence of outliers samples in the regression model. The second reason can be attributed to the lower capability of the PLS algorithm in generalizing non-linear relationships between the soil spectra and their respective attributes, resulting in higher variations/dispersion on the RMSEP values related the samples selected to build the regression models.

For most soil attributes, the SVM regression algorithm recorded the lowest median RMSEP, in addition to minor variations on the RMSEP values for samples with both particle sizes. In the soil horizon classification, SVM and PLS-DA algorithms yielded similar median MCC values, presenting both superior performance compared to RF algorithm.

In the NIR range, the Cubist algorithm yielded the highest RMSEP values for all soil attributes predicted using samples from both particle sizes, while the SVM regression algorithm gave the lowest median RMSEP, presenting minor variations on the RMSEP values. In the soil horizon classification, the SVM algorithm gave higher MCC values (best performance) than the RF and PLS-DA models built with soil samples of both particle sizes.

#### 3.2.3. SVM predictions using ball-milled and < 2 mm samples in MIR and NIR

Since the SVM algorithm yielded the best median RMSEP and MCC values in both spectral ranges and particle sizes, the estimation of the influence of the ball-milling step in regression and classification models accuracy was done by analysing the results of median SVM models. Table 3 shows the accuracy results of the median SVM regression models and Figs. 4 – 7 show the scatter plots of reference versus predicted values for each model.

Despite the calibration and validation sets being randomly generated, samples from both sets presented a similar distribution along the adjusted line in the median regression models. The slopes of the calibration (blue) and validation (red) adjusted regression lines were equivalent for both particle sizes, especially in the MIR range (Fig. 4 and Fig. 5). The median RMSEP and $R^2_{val}$ values indicated that ball-milling the soil samples before the spectra acquisition in the MIR range does not introduce any positive or negative biases on the RMSEP/accuracy. This is demonstrated by the observed random RMSEP variations of ± 5%. The exception is for the bulk density predictions, where ball-milling the soil samples increased the RMSEP by 6.67%. However, among all the soil attributes evaluated, the bulk density data were based on the lowest number of samples (438). Moreover, the RMSEP values for ball-milled and < 2 mm samples only differ by 0.01 ($g/cm^3$), that is 0.15 vs 0.14.

The quality of the MIR spectra models using only the RPIQ classes differentiate for sand, $pH_{water}$ and bulk density predictions using ball-milled and < 2 mm soils. However, in all these cases, the RPIQ values were close to the threshold, i.e. for sand predictions using ball-milled samples the RPIQ value was 3.42, being classified as GM, which is very close to AQM (≥2.70 & < 3.38). The same behaviour was observed for $pH_{water}$ and bulk density, with RPIQ values of 2.63 (close to AQM class) and 3.33 (close to GM class, RPIQ ≥ 3.38 & ≤ 4.05), respectively. In addition, for all these spectra models the differences between the RPIQ values were inferior than the threshold that separated the quality classes (0.68 RPIQ) [38]. The only soil attribute that shows RPIQ differences > 0.68 was the TC. However, as shown in Table 3, both models were classified in the same quality class (EM), with similar RMSEP values and $R^2_{val}$ values. Since both accuracy parameters did not present higher variations simultaneously, we can conclude that there are no accuracy differences between MIR spectra models built using ball-milled and < 2 mm soils for the soil attributes evaluated.

There are also random positive and negative variations of 5% on the RMSEP gain for 7 of 11 soil attributes analysed in the NIR range. For the other four parameters, OC, nitrogen, aluminium and bulk density, ball-milling the samples showed an improvement on the RMSEP values of 6.67%, 20.00%, 6.45% and 6.67%, respectively. Among these soil attributes, nitrogen, aluminium and bulk density spectra models were classified in the same quality class; and the OC spectra models presented differences in the RPIQ values < 0.68. For TC predictions, NIR spectra models showed RPIQ differences > 0.68, however both models were classified as EM with RMSEP gain < 5%; and similar $R^2_{val}$ values. Thus we can conclude that NIR spectra models presented the same accuracy for ball-milled and < 2 mm soil samples.

The MCC values and the predicted class calculated by SVM classification models using samples from both particle sizes are shown in Table 4 and Table 5. It is clear that there was no considerable improvement in the MCC results by ball-milling the soil samples as this process only reduced the MCC by 2.63% and 1.37% in MIR and NIR ranges, respectively. The predicted class results in Table 5 indicated that the difference in TP, TN, FP and FN values were minimal. Moreover, where the classification model presented a high TP value the corresponding FP was also high. The model built using ball-milled samples collected in the MIR range present a TP value of 67 compared to 64 obtained from < 2 mm samples in MIR. However this model yielded a FP value of 21 for ball-milled samples compared to 14 obtained for < 2 mm samples. This same behaviour was also observed for the models built in the NIR range.

The results obtained are good indicators that no general positive or negative bias on the RMSEP/MCC results was generated by ball-milling the soil samples before the spectra acquisition in MIR or NIR ranges. However, it was also necessary to evaluate if the chemical information

extracted from MIR and NIR spectra were the same for ball-milled and < 2 mm samples. Since the models yielded similar RMSEP, RPIQ and MCC values the chemical information (spectral bands) should be the same for both particle sizes. In this study, this evaluation was performed by analysing the importance of each variable in regression and classification models. However, there is no default method to identify these variables using the SVM algorithm, which was the algorithm that yielded the best performance results. To overcome this, the Variable Importance in Projection (VIP) was calculated by the PLS and PLS-DA models, which are traditional algorithms used for regression and classification proposes. The VIP scores (Eq. (5)) summarize the accumulated importance of each variable *j* according to the weights from each latent variable [42].

$$VIP_j = \sqrt{p \sum_{a=1}^{A} \left[ SS_a \left( \frac{W_{aj}}{\|W_a\|} \right)^2 \right] \Big/ \sum_{a=1}^{A} SS_a} \quad (5)$$

where $SS_a$ is the sum of squares explained by the *a*th latent variable and $\left(\frac{W_{aj}}{\|W_a\|}\right)^2$ represents the importance of the *j*th variable for the *a*th latent variable [42].

The VIP scores calculated by the PLS models for each soil attribute and by PLS-DA to classify the soil samples in 1st or 2 − 7th horizons are shown in Fig. 8 (MIR) and Fig. 9 (NIR).

All wavenumbers with VIP scores higher than 1 (recommended threshold) were considered important to build the regression or classification models [42].

The most important wavenumbers selected by VIP scores are summarized in Table 6. Most of these variables were already discussed in section 3.1. It is noted that for each soil attribute the chemical variables are associated with the parameter evaluated, i.e. for pH predictions using MIR spectra models the band at 1750 $cm^{-1}$ is one of the most important for the regression model, which is associated to carboxylic acid (–COOH) absorptions and chemically related to organic matter and pH.

Determinations of bulk density, a structural soil property, from MIR/NIR spectra was only possible due to the correlation of bulk density with soil attributes that shows a signal in the MIR/NIR spectra. In this case the bulk density values are negatively correlated to nitrogen, TC, OC and aluminium contents (Fig. S4). A complete description of the organic and inorganic compounds that are responsible for absorptions in MIR and NIR range can be found at [5,11,15,43].

The VIP scores calculated by PLS regression models using MIR spectra showed the most important variables were the same regardless of particle size. This same behaviour was also observed by PLS-DA models to classify the soil samples into horizon 1 or not (horizons 2–7).

In the NIR range, the most important variables selected by PLS models were similar for both particle sizes, with the exception of a few variables at 4950–4900 $cm^{-1}$ (2020 – 2041 nm), corresponding to amide, considered important for ball-milled samples and not important for < 2 mm samples in % TC predictions. The most important variables for all the four soil attributes (OC, nitrogen, aluminium and bulk density) that yielded a RMSEP gain of>5% also presented few differences. The variables at 4950–4900 $cm^{-1}$ (2020 – 2041 nm), 5135 $cm^{-1}$ (1947 nm), were only considered important for aluminium and nitrogen spectra models built using < 2 mm samples. For OC and bulk density spectra models the band at 7180 $cm^{-1}$ (1393 nm) was only considered important for < 2 mm samples. Although these variables were not considered important for ball-milled samples, it is important to highlight that the VIP scores of these variables were very close to 1, indicating the chemical information extracted by samples from both particle sizes are the same. This small discrepancy in the VIP scores can be attributed to minor difference in the intensities in the ball-milled and < 2 mm spectra collected in the NIR range. A small correction on the VIP threshold value to 0.90 would result in the same variables being classified as most important for all regression models.

VIP scores were also very similar for both particle sizes in the classification models, except for the band at 7701 $cm^{-1}$ (~1300 nm) which was considered important for ball-milled samples but not for < 2 mm samples. This band is located in a region with considerable noise in the NIR region (7950–7220 $cm^{-1}$) and since this band was considered important to classify the soil samples in their respective horizons (1 or 2 – 7) we can infer the ball-milling step was able to minimize this noise, enabling extraction of slightly more information to classify the soil samples in their respective horizons.

The results obtained challenge the broad consensus that soil samples should be ball-milled before the spectra acquisition in the MIR range due to the small radiation spot, largely because advances in detector sensitivity in modern spectrometers have overridden the need for ball-milling. For example, using MCT instead of deuterated-triglycine sulfate (DTGS); coupled with machine learning regression methods with high generalization power, such as SVM or Cubist instead PLS has improved results. The systematic comparison carried out in this study indicates that this combination was able to reduce/address this problem for all the 11 soil attributes evaluated in this paper. Compared to DTGS, the MCT detector shows higher sensitivity for systems where low rates of IR radiation reach the detector. In other words, the low IR radiation resulting from the interaction with soils with big particle sizes can be better quantified using this detector, improving its precision/representativeness [46]. In addition, since the accuracy results were the same for a spectrometer with a small spot size (~2.5 mm), which is the worst case scenario, the accuracy of results should be the same for spectrometers with bigger spot sizes (~5 – 10 mm).

Finally, it is important to highlight that despite showing the same accuracy and the same chemical information extracted by spectra models, it is necessary to build spectra models for each soil particle size. For example, ball-milled spectra cannot be used for predicting parameters from a spectra model built using < 2 mm samples or vice-versa. As shown in Fig. 1, the soil samples showed the same tendency on the spectra with regards to OC content but the spectral intensities and the band proportions are not the same for samples from both particle sizes, resulting in high errors on predictions.

## 4. Conclusions

This study carried out a systematic investigation about the influence of soil particle sizes on the accuracy of regression and classification models built using MIR and NIR spectra. The infrared spectra of both soil particle sizes, < 0.100 mm (ball-milled) and < 2 mm samples, were similar in both MIR and NIR ranges, giving excellent agreement between the spectra and a random soil attribute, in this case OC content, with correlation coefficient values above 0.9700 for the same soil samples analysed in both particle sizes and ranges. The median results obtained by SVM models on the 29 randomizations performed for each soil attribute did not indicate a bias on the RMSEP and RPIQ values related to soil particle sizes. The spectra models showed, in general, small random variations for both accuracy parameters. In addition, for both IR ranges no spectra model showed higher differences on the RMSEP and RPIQ simultaneously. In the classification models, both particle sizes obtained very similar MCC results, with minor variations of 2.61% (MIR) and 0.65% (NIR) to classify the soils into horizon 1 or horizons 2 – 7.

The VIP scores obtained by all PLS models showed that the most important variables were the same for spectra collected from both soil particle sizes in the MIR range, with minimal variations for aluminium, nitrogen, OC and bulk density in the NIR range. This was attributed to the minor difference in intensities in the ball-milled and < 2 mm spectra collected in the NIR range, shifting a few VIP scores to values close to 0.90 – 0.95.

The results obtained shows that collecting MIR and NIR spectral libraries using spectrometers with high sensitive detectors, like MCT, and

combined with regression and classification models with higher generalization power, will not result in accuracy differences or loss of chemical information extracted by ball-milled or < 2 mm sieved soils. This study will facilitate the removal of the ball-milling step from laboratories that use vibrational spectroscopy techniques to predict certain soil attributes, reducing the time and costs associated with sample preparation.

*CRediT authorship contribution statement*

**Felipe Bachion de Santana:** Conceptualization, Methodology, Writing – original draft. **Karen Daly:** Supervision, Conceptualization, Methodology, Writing – review & editing.

## 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.

## Acknowledgements

The authors are grateful to Ms. Courtney Doyle, Patricia Berry and Dr. Anna Fenelon for technical support in spectroscopy at Teagasc Johnstown Castle laboratories. The authors are also grateful to Geological Survey of Ireland; Dr. Eric C. Grunsky, Dr. Mairéad M. Fitzsimons and Dr. Vincent Gallagher for their contributions and discussions. The authors wish to acknowledge funding for this work by the Irish Government Department of Agriculture Food and the Marine Research Stimulus Fund under grant ID 2019R463.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2022.121441.

## References

[1] U. Nations, World Population Prospects 2019, (2019). https://population.un.org/wpp/Download/Standard/CSV/ (accessed 30 September 2021).
[2] G. Silva, Feeding the world in 2050 and beyond – Part 1: Productivity challenges, Michigan State Univ, Ext, 2018 https://www.canr.msu.edu/news/feeding-the-world-in-2050-and-beyond-part-1 (accessed 30 September 2021.
[3] A. McBratney, D.J. Field, A. Koch, The dimensions of soil security, Geoderma 213 (2014) 203–213, https://doi.org/10.1016/j.geoderma.2013.08.013.
[4] L. Deiss, A.J. Margenot, S.W. Culman, M.S. Demyan, Optimizing acquisition parameters in diffuse reflectance infrared Fourier transform spectroscopy of soils, Soil Sci. Soc. Am. J. 84 (2020) 930–948, https://doi.org/10.1002/saj2.20028.
[5] P. Baumann, A. Helfenstein, A. Gubler, A. Keller, R.G. Meuli, D. Wächter, J. Lee, R. V. Rossel, J. Six, Developing the Swiss soil spectral library for local estimation and monitoring, Soil Discuss. (2021) 1–32, https://doi.org/10.5194/soil-2020-105.
[6] C. Hutengs, B. Ludwig, A. Jung, A. Eisele, M. Vohland, Comparison of portable and bench-top spectrometers for mid-infrared diffuse reflectance measurements of soils, Sensors (Switzerland). 18 (2018) 1–17, https://doi.org/10.3390/s18040993.
[7] N.M. Knox, S. Grunwald, M.L. McDowell, G.L. Bruland, D.B. Myers, W.G. Harris, Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy, Geoderma 239–240 (2015) 229–239, https://doi.org/10.1016/j.geoderma.2014.10.019.
[8] S.R.S. Dangal, J. Sanderman, S. Wills, L. Ramirez-Lopez, Accurate and precise prediction of soil properties from a large mid-infrared spectral library, Soil Syst. 3 (2019) 1–23, https://doi.org/10.3390/soilsystems3010011.
[9] X. Wang, J. Sanderman, K. Yoo, Climate-dependent topographic effects on pyrogenic soil carbon in southeastern Australia, Geoderma 322 (2018) 121–130, https://doi.org/10.1016/j.geoderma.2018.02.025.
[10] B.G. Barthès, E. Kouakoua, P. Moulin, K. Hmaidi, T. Gallali, M. Clairotte, M. Bernoux, E. Bourdon, J. Toucet, T. Chevallier, Studying the physical protection of soil carbon with quantitative infrared spectroscopy, J. Near Infrared Spectrosc. 24 (2016) 199–214, https://doi.org/10.1255/jnirs.1232.
[11] F.S. Terra, J.A.M. Demattê, R.A.V. Rossel, Spectral libraries for quantitative analyses of tropical Brazilian soils: comparing vis-NIR and mid-IR reflectance data, Geoderma 255–256 (2015) 81–93, https://doi.org/10.1016/j.geoderma.2015.04.017.
[12] F.B. de Santana, A.M. de Souza, R.J. Poppi, Green methodology for soil organic matter analysis using a national near infrared spectral library in tandem with learning machine, Sci. Total Environ. 658 (2019) 895–900, https://doi.org/10.1016/j.scitotenv.2018.12.263.
[13] J. Yang, X. Wang, R. Wang, H. Wang, Combination of Convolutional Neural Networks and Recurrent Neural Networks for predicting soil properties using Vis–NIR spectroscopy, Geoderma 380 (2020), 114616, https://doi.org/10.1016/j.geoderma.2020.114616.
[14] Z. Shi, Q.L. Wang, J. Peng, W.J. Ji, H.J. Liu, X. Li, R.A.V. Rossel, Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations, Sci. China Earth Sci. 57 (2014) 1671–1680, https://doi.org/10.1007/s11430-013-4808-x.
[15] J.M. Soriano-Disla, L.J. Janik, R.A.V. Rossel, L.M. MacDonald, M.J. McLaughlin, The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties, Appl. Spectrosc. Rev. 49 (2014) 139–186, https://doi.org/10.1080/05704928.2013.811081.
[16] J.A. Baldock, B. Hawke, J. Sanderman, L.M. MacDonald, Predicting contents of carbon and its component fractions in Australian soils from diffuse reflectance mid-infrared spectra, Soil Res. 51 (2013) 577–595, https://doi.org/10.1071/SR13077.
[17] B. Stumpe, L. Weihermüller, B. Marschner, Sample preparation and selection for qualitative and quantitative analyses of soil organic carbon with mid-infrared reflectance spectroscopy, Eur. J. Soil Sci. 62 (2011) 849–862, https://doi.org/10.1111/j.1365-2389.2011.01401.x.
[18] L.J. Janik, J.M. Soriano-Disla, S.T. Forrester, M.J. Mclaughlin, Effects of soil composition and preparation on the prediction of particle size distribution using mid-infrared spectroscopy and partial least-squares regression, Soil Res. 54 (2016) 889–904, https://doi.org/10.1071/SR16011.
[19] R.E. Creamer, I. Simo, L.O. Sullivan, B. Reidy, R.P.O. Schulte, R.M. Fealy, accessed 30 September 2021, Irish Soil Information System: Soil Property Maps (2016), http://erc.epa.ie/safer/reports.
[20] M.J. Gardiner, T. Radford, Soil associations of Ireland and their land use potential: Explanatory bulletin to soil map of Ireland, Mount Salus Press Limited, Dublin, Ireland, 1980 https://www.teagasc.ie/media/website/environment/soil/General.pdf (accessed 30 September 2021.
[21] P. Massey, C. O'Connor, P. Sills, A. Fenelon, L. Maloney-Finn, D. Stone, B. Reidy, R. Creamer, Irish Soil Information System : Laboratory Standard Operating Procedures Final Technical Report 7 (2013). https://eparesearch.epa.ie/safer/iso19115/display?isoID=3062.
[22] E.O. Mclean, Soil pH and Lime Requirement, in: Methods Soil Anal. Part 2 Chem. Microbiol. Prop., Academic Press, Inc, New York, NY, 1983: pp. 199–224. https://doi.org/10.2134/agronmonogr9.2.2ed.c12.
[23] K. Laursen, M.A. Rasmussen, R. Bro, Comprehensive control charting applied to chromatography, Chemom. Intell. Lab. Syst. 107 (2011) 215–225, https://doi.org/10.1016/j.chemolab.2011.04.002.
[24] Å. Rinnan, Pre-processing in vibrational spectroscopy – when, why and how, Anal. Methods. 6 (2014) 7124–7129, https://doi.org/10.1039/C3AY42270D.
[25] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639, https://doi.org/10.1021/ac60214a047.
[26] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831, https://doi.org/10.1039/c3ay41907j.
[27] ASTM E1655-17, ASTM E1655-17 Standard Practices for Infrared Multivariate Quantitative Analysis, ASTM Int. 05 (2017) 30. https://doi.org/10.1520/E1655-05R12.2.
[28] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202, https://doi.org/10.1021/ac00162a020.
[29] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1023/A:1022627411411.
[30] MathWorks, Statistics and Machine Learning Toolbox ™ User ' s Guide R 2017 a, (2017). The MathWorks, Inc., Apple Hill Drive.
[31] E. Brochu, V.M. Cora, N. de Freitas, A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning, (2010). http://arxiv.org/abs/1012.2599.
[32] B. Üstün, W.J. Melssen, M. Oudenhuijzen, L.M.C. Buydens, Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization, Anal. Chim. Acta. 544 (2005) 292–305, https://doi.org/10.1016/j.aca.2004.12.024.
[33] M. Kuhn, S. Weston, C. Keefer, N. Coulter, Cubist Models For Regression, R Packag. (2012). https://mran.microsoft.com/snapshot/2016-09-04/web/packages/Cubist/vignettes/cubist.pdf (accessed 30 September 2021)vignettes/cubist.pdf (accessed 30 September 2021).
[34] F.B. de Santana, S.J. Mazivila, L.C. Gontijo, W.B. Neto, R.J. Poppi, Rapid Discrimination Between Authentic and Adulterated Andiroba Oil Using FTIR-HATR Spectroscopy and Random Forest, Food Anal. Methods. 11 (2018) 1927–1935, https://doi.org/10.1007/s12161-017-1142-5.
[35] A.B. Techniques, Nonlinear Modeling, Springer US, Boston, MA, 1998. 10.1007/978-1-4615-5703-6.
[36] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[37] J. Liu, Z. Dong, J. Xia, H. Wang, T. Meng, R. Zhang, J. Han, N. Wang, J. Xie, Estimation of soil organic matter content based on CARS algorithm coupled with random forest, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 258 (2021), 119823, https://doi.org/10.1016/j.saa.2021.119823.
[38] B. Ludwig, S. Vormstein, J. Niebuhr, S. Heinze, B. Marschner, M. Vohland, Estimation accuracies of near infrared spectroscopy for general soil properties and

enzyme activities for two forest sites along three transects, Geoderma 288 (2017) 37–46, https://doi.org/10.1016/j.geoderma.2016.10.022.
[39] F.B. de Santana, A.M. de Souza, R.J. Poppi, Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 191 (2018) 454–462, https://doi.org/10.1016/j.saa.2017.10.052.
[40] B. Stenberg, R.A.V. Rossel, A.M. Mouazen, J. Wetterlind, Visible and Near Infrared Spectroscopy in Soil Science, Adv. Agron. 107 (2010) 163–215, https://doi.org/10.1016/S0065-2113(10)07005-7.
[41] A. Fakhry, O. Osman, H. Ezzat, M. Ibrahim, Spectroscopic analyses of soil samples outside Nile Delta of Egypt, Spectrochim, Acta Part A Mol. Biomol. Spectrosc. 168 (2016) 244–252, https://doi.org/10.1016/j.saa.2016.05.026.
[42] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69, https://doi.org/10.1016/j.chemolab.2012.07.010.
[43] P.K. Krivoshein, D.S. Volkov, O.B. Rogova, M.A. Proskurnin, FTIR photoacoustic spectroscopy for identification and assessment of soil components: chernozems and their size fractions, Photoacoustics. 18 (2020), 100162, https://doi.org/10.1016/j.pacs.2020.100162.
[44] I.R. Merck, accessed 9 February, Spectrum Table & Chart. (2022), https://www.sigmaaldrich.com/IE/en/technical-documents/technical-article/analytical-chemistry/photometry-and-reflectometry/ir-spectrum-table.
[45] A. Lazaar, A.M. Mouazen, K. EL Hammouti, M. Fullen, B. Pradhan, M.S. Memon, K. Andich, A. Monir, The application of proximal visible and near-infrared spectroscopy to estimate soil organic matter on the Triffa Plain of Morocco, Int. Soil Water Conserv. Res. 8 (2020) 195–204, https://doi.org/10.1016/j.iswcr.2020.04.005.
[46] J.C. Kenneth D. Kempfert, Eric Y. Jiang, Sherwin Oas, Detectors for Fourier Transform Spectroscopy, Thermo Nicolet Appl. Note. (2001). https://mmrc.caltech.edu/FTIR/Nicolet/Nicolet%20Tech%20Notes/DetectorsforFTIR1204.pdf (accessed 4 May, 2022).