Catena 218 (2022) 106552

Contents lists available at ScienceDirect

Catena

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

Diffuse reflectance mid infra-red spectroscopy combined with machine learning algorithms can differentiate spectral signatures in shallow and deeper soils for the prediction of pH and organic matter content

Felipe Bachion de Santana [a], 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, Beggar's Bush, Haddington Road, Dublin D04 K7X4, Ireland

ARTICLE INFO

Keywords:
Soil spectral library
Organic matter, pH
Mid-infrared
Cubist
Irish soils

ABSTRACT

Precision and sustainable agriculture requires information about soil pH and organic matter (OM) content at higher spatial and temporal scales than current agronomic sampling and analytical methods allow. This study examined the accuracy of spectral models using high throughput screening (HTS) in diffuse reflectance mode in mid Infra-red (MIR)/DRIFT combined with machine learning algorithms to predict soil pH($CaCl_2$) and %OM in shallow and deeper topsoils compared to laboratory methods. Models were developed from an archive of samples taken on a 4 km^2 grid from the northern half of Ireland (Terra Soil project), which includes 18,859 samples (9,396 shallow + 9,463 deeper). The application of Cubist models showed that for different depths there are minor different spectral group associations with pH and %OM values. These differences resulted in a loss of accuracy in the extrapolation of the topsoil model to predict values from deeper topsoils or vice versa. Therefore we recommend the use of samples from both depths to build a calibration model.

The proposed methodology was able to determine %OM and pH using a unique multivariate regression model for both depths, with RMSEP values of 1.12 and 0.89 %; RPIQ values of 42.34 and 38.48; R^2_{val} of 0.9989 and 0.9993 for %OM determinations in shallow and deeper topsoils, respectively. For pH determinations the RMSEP values obtained were 0.25 and 0.34; RPIQ values of 6.04 and 4.94; R^2_{val} 0.9385 and 0.8954. Both regression models are classified as excellent predictions models, yielding RPIQ values >4.05 for shallow and deeper topsoils. The results demonstrated the high potential of HTS-DRIFT combined with machine learning algorithms as a rapid, accurate, and cost-effective method to build large soil spectral libraries, displaying predicted results similar to two separate soil laboratory methods (pH and LOI).

1. Introduction

In the last decade the global agriculture sector was responsible for approximately 928 million jobs, representing 30.8 % of the jobs in the world (Agriculture, 2021). However, in recent years (2018) the number of agriculture sector jobs has fallen to about 866 million jobs, representing 26.5 % of the jobs in the world (Agriculture, 2021). Despite this drop, agricultural productivity has increased in recent years to meet the global demand for food. This increase can be associated with the high level of investment in the agriculture sector, resulting in new agricultural machinery, strategies, and technologies to implement precision agriculture and sustainable agricultural practices (de Santana et al., 2019).

Among the strategies adopted by precision agriculture and sustainable agricultural practices, we can highlight the acquisition of accurate information about the spatial distribution of soil properties, such as percentage of organic matter (%OM), soil organic matter content (SOM), pH, elemental concentrations, P, K, N, and soil particle size, among other soil attributes (Terra et al., 2015). Using this information, farmers can take decisions to improve productivity and sustainable soil management. For example, pH values are used to evaluate the liming requirement of acidic and neutral soils, for optimum soil health and fertility, and % OM values can be converted to SOM and % Carbon for estimates of baseline values.

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

https://doi.org/10.1016/j.catena.2022.106552
Received 8 November 2021; Received in revised form 26 May 2022; Accepted 26 July 2022
Available online 3 August 2022

However, conventional analytical methods for such analyses are expensive and time-consuming. As a consequence many farmers reduce the number of samples collected for analysis, resulting in unrepresentative soil maps and results that are not monitored over time (Allo et al., 2020). To address this problem, in the last decade several authors proposed the use of vibrational spectroscopy techniques, especially Near Infrared (NIR) and Mid-Infrared (MIR/FTIR) spectroscopy, combined with chemometrics/machine learning algorithms to quantify or semi-quantify these and other soil parameters using only a single spectrum. In order to avoid additional sample preparation, such as Potassium bromide (KBr) pellets, the soil spectra are commonly obtained in diffuse reflectance mode. This combination results in a rapid, cost-effective, accurate, and non-destructive analysis, contributing to the implementation of precise agriculture and sustainable agricultural practices, where fast, non-expensive and accurate methods to determine the soil composition are required (de Santana et al., 2019; Stenberg et al., 2010).

Due to all the highlighted advantages, in the last decade several soil spectral libraries (SSL) were created around the world but only a few of these SSLs are composed of a large number of soil samples (Baumann et al., 2021; Dangal et al., 2019; de Santana et al., 2019; Demattê et al., 2019; Orgiazzi et al., 2018; Viscarra Rossel et al., 2016; Sanderman et al., 2020; Terra et al., 2015). Most of these large SSLs use the vis-NIR or NIR spectrometers to collected the spectra, a fact that can be attributed to several factors such as price, a larger radiation spot, the minor influence of soil sample preparation, and the facility to couple bifurcated optical fibre probes in the instruments, enabling the customization of commercial spectrometers to collect the soil spectra (de Santana et al., 2019).

Among these large SSLs we can highlight the Land Use and Coverage Area Frame Survey (LUCAS) programme (Orgiazzi et al., 2018). This spectral library is composed of ~22,000 soils samples collected from 26 EU Member States, including Ireland. The spectral library consist in the vis-NIR spectra (400–2500 nm) and includes reference values of pH, SOM, cation exchange capacity (CEC), total nitrogen content, phosphorous content, among other soil attributes (Orgiazzi et al., 2018). In the LUCAS spectral library the spectra acquisition was carried out in soil samples air-dried and mechanically sieved to 2 mm.

Using this vis-NIR spectral library, Jiechao Yang et al. (2020) created a multivariate regression model based on the combination of convolutional neural networks (CNN) and recurrent neural networks (RNN) to determine the values of organic carbon (OC), N, CEC, and pH ($CaCl_2$). The authors used 75 % of the SSL to build the model and the remaining 25 % of samples was used to validate the regression model. The authors obtained good results in OC and pH predictions, with root mean square errors on prediction (RMSEP) values of 6.40 ($g.kg^{-1}$) and 0.35; R^2_{val} of 0.73 and 0.86 for OC and pH predictions, respectively. These results are consistent with other studies that uses large vis-NIR soil spectral libraries (Viscarra Rossel and Webster, 2012; Stevens et al., 2013).

We can also highlight the SSL created by Dangal et al. (2019) and subsequently updated by Sanderman et al. (2020). In the last study the authors used a large MIR-SSL to determine several parameters in United States soils, including pH_{Water} (35.297 samples) and OC (42.893 samples). Before the spectra acquisition in the diffuse reflectance infrared Fourier transform (DRIFT) spectrometer, the soil samples were air-dried, sieved to 2 mm and then ball-milled. The authors used 80 % of the samples to build regression models based on the Cubist algorithm and the remaining samples were used to validate the regression model. Using this methodology the authors obtained excellent accuracy, with RMSEP values of 0.36 and 0.69 %; R^2_{val} of 0.88 and >0.99 for pH and OC predictions, respectively.

The results obtained by both papers show that NIR and MIR (DRIFT) spectroscopy can be employed to build large spectral libraries to predict pH and OC values with good accuracy. Nevertheless, the results obtained using DRIFT spectroscopy were generally superior and more suitable for accurate soil analysis (Sabetizade et al., 2021). Compared to vis-NIR spectroscopy, DRIFT spectroscopy is more affected by the soil sample preparation, requiring ball-milling of the soil samples to <0.100 mm to reduce the specular reflectance and maximize the homogeneity of the sample (Soriano-Disla et al., 2014). It is important to emphasize that the ball-milling process is already used in loss-on-ignition (LOI) analysis to determine OM%, so employing DRIFT spectroscopy does not require an additional step in laboratories doing routine soil analysis (Metzger et al., 2021; Soriano-Disla et al., 2014).

Based on the above, this study employed DRIFT spectroscopy combined with machine learning regression models to determine pH and OM content in a large soil spectral library. This library was created as part of the Terra Soil project and comprises spectra of shallow and deeper topsoil samples collected in the northern half of Ireland as part of Geological Survey Ireland's Tellus programme - Terra Soil (gsi.ie).

The agri-food sector in Ireland makes a significant contribution to the economy, accounting for 7.1 % of total employment and providing 6.7 % of Gross National Income in 2019 (Department of Agriculture Food and Marine, 2020). In the coming years the central objective of Ireland's agri-food strategy (FoodVision 2030) is that Ireland becomes an international leader in sustainable food systems, while ensuring profitability, social sustainability and positive or neutral impact on the environment. Implementation of the common agricultural policy in Ireland post-2020 will also place a very strong emphasis on efficiency and the environment.

The proposed methodology will contribute to achieving the goals of Food Vision 2030, enabling generation of high-resolution maps of soil pH and %OM from agricultural and non-agricultural regions in Ireland. The methodology uses a rapid, cost-effective, robust, and accurate method that will help improve the efficiency of the Irish agri-food sector. Another innovation of this study is the use of a large SSL to assess differences of %OM and pH predictions in both depths using regression models derived from each depth. For example, regression models built using shallow topsoil samples were used to predict the %OM and pH in shallow and deeper topsoils (different depth). In addition, the accuracy of the calibration model built using samples from both depths was assessed.

2. Material and methods

2.1. Soil sampling

During the period of 2011–2019 9,396 shallow topsoil (5–20 cm) and 9,463 deeper topsoil (35–50 cm) samples were collected from agricultural and non-agricultural lands in the Border, Midland, West and Periurban regions of Ireland as part of Geological Survey Ireland's Tellus programme (GSI, 2021). Regional samples were collected on a 4 km^2 grid, periurban samples on a 1 km^2 grid. Shallow and deeper topsoils were collected at each site (Figure S1).

The predominant soil types used in this study are classified as Peat, Grey Brown Podzolic, Gley, Brown Earth, Brown Podzolic, Complex, Podzol, Brown Earth L.B.S, among others, according to Gardiner and Radford 1980 and based on the U.S > Department of Agriculture (USDA) classification system (Gardiner and Radford, 1980; Survey, 1999). The topsoil composition largely reflects the composition of the underlying subsoils and the bedrock from which the subsoils are derived. Bedrock geology of the northern part of Ireland covered by the Tellus survey is very diverse and includes granite, gneiss, schist, sandstone and shale. The midlands are largely underlain by limestone. There are extensive areas of blanket bog in upland areas and of raised bog in the midlands.

2.2. Reference analysis and Log-ratio transformation

All soil reference analyses were performed on behalf of Geological Survey Ireland in analytical laboratories accredited to ISO/IEC 17025:2005. Prior to analysis, samples were prepared by drying at 30°, disaggregation, sieving to 2 mm and ball milling to a particle size nominally 95 % < 0.032 mm. The %OM was determined using LOI

analysis, for which 1 g of ball-milled soil was placed in a furnace at 450 °C for 4 h. The %OM was then calculated from the ratio of the weights of combusted (OM) and raw sample (OM + inorganic matter (IM)). The pH ($CaCl_2$) analyses were performed on oven-dried soil samples sieved to <2 mm. A $CaCl_2$ solution (0.01 M) was added to each sample and the pH was analysed using pH electrode. Full details of Tellus analytical methods can be found in Knights et al. (2020).

Since the OM was calculated by the ratio of combusted (OM) and raw soil sample (OM + IM), the %OM values are relative measures (compositions) and are restricted to the numerical space known as the simplex. The range of values in this simplex are restricted to 0–100 and are not independent of other measures expressed as compositions. In other words, if the OM or IM change the other value must change to maintain the proportion in the simplex. This characteristic is commonly referred to as the "closure" problem, which could result in incorrect regression models, since machine learning regression algorithms analyse each soil feature independently, not considering that OM + IM must add up to 100 % (Aitchison, 1982; Jaconi et al., 2019). A more recent review of dealing with compositional data is provided by Greenacre et al. (2022). To deal with this problem, the additive log-ratio (Eq. (1)) proposed by Aitchison (1982) was used as the **Y**-block to build the %OM models. The histogram containing the new distribution of %OM values after this transformation can be visualized as in Figure S2. Following this, the inverse function of additive log-ratio (Eq. (2)) was used in the predicted values to return the data in to the original space to build the predicted maps and calculate the accuracy parameters.

$$r = ln\left(\frac{\%OM}{100 - \%OM}\right) \quad (1)$$

$$\%OM = \frac{100 \times e^r}{e^r + 1} \quad (2)$$

2.3. Spectral acquisition

The soil sample spectra were obtained in the diffuse reflectance mode using the FT-IR Spectrometer INVENIO® (Bruker, Massachusetts - USA) with a mercury-cadmium-telluride (MCT) detector. This spectrometer was connected to a Bruker microplate reader extension for high-throughput screening (HTS-XT), which can analyse ~70 samples per hour operating with one analyst.

The ball-milled soil samples were filled into a sample cup with a 7 mm internal diameter and the surface was smoothed with a blade for further analysis in the HTS-XT system. The samples were scanned in the MIR range (4000–600 cm^{-1}) in diffuse reflectance mode (DRIFT), using 32 scans at a spectral resolution of 1.43 cm^{-1}. At the beginning of each batch of analysis of 23 samples, the background was obtained using a gold reference disk. Every day before commencement of the analysis, the instrument was checked using a reference compound (Acetaminophen – USP specification). This reference spectrum was monitored using an internal multivariate control chart based on PCA and D- and Q- statistics at significance level of probability of 0.05 (Laursen et al., 2011). The reference values of pH and %OM combined with the Mid-Infrared spectra were used to build the Tellus soil spectral library.

2.4. Spectral pre-processing

Spectral regions showing atmospheric CO_2 signals (2398–2281 cm^{-1}) were removed from the spectrum. Subsequently, classical pre-processing spectral algorithms for soil analysis using vibrational spectroscopy were evaluated. The pre-processing algorithms tested were multiplicative signal correction (MSC), extended multiplicative signal correction (EMSC), standard normal variate (SNV) and smoothing combined with first and second derivatives (Rinnan, 2014).

2.5. Selection of samples for calibration and validation sets

Shallow topsoil samples were organized in ascending order for %OM and pH values. Then, samples in even lines were selected for calibration and the remaining samples were selected to compose the validation set. This process generated four subsets for shallow topsoil samples: the $\%OM_{shallow}$ calibration and validation sets, and the $pH_{shallow}$ calibration and validation sets. The same strategy was applied for deeper topsoil samples, also generating four subsets, the $\%OM_{deeper}$ calibration and validation sets, and the pH_{deeper} calibration and validation sets. As shown in Fig. 1, when carrying out this strategy the calibration and validation sets will display the same histogram profile in the entire range evaluated.

We investigated the accuracy of regression models using these subsets. For this, samples from shallow topsoils (Library 1) were used to build regression models to determine the %OM and pH in shallow (Library 1) and deeper (Library 2) topsoils. The opposite strategy was also evaluated, i.e. deeper topsoils (Library 2) were used to build regression models to determine the %OM and pH in shallow (Library 1) and deeper (Library 2) topsoils. Moreover, we built a global model using samples from both depths (shallow calibration and deeper calibration – Library 3) and the accuracy of this model was evaluated using samples from both depths; in other words, shallow validation and deeper validation (Library 3). All the libraries used to determine %OM and pH in both depths are summarized in Table 1.

2.6. Data analysis

All calculations were performed in MATLAB R2020b environment (Mathworks) and R Studio, using an Intel Core i5-8250 CPU with four cores and eight threads with 8 GB memory. Support Vector Machine and PLS models were built using the "Statistics and Machine Learning Toolbox". Cubist and Random forest models were building using Cubist and Ranger package for R.

2.6.1. Partial least squares - PLS

PLS is considered a classical multivariate calibration method used in vibrational spectroscopy data. The PLS algorithm uses the correlation between the **X** block (infrared spectra) and **y** block (%OM or pH) to build the regression model. During this process the **X** block is projected in to lower dimensional space (latent variables - LV) to maximize the covariance between **X** and **y** blocks. During this process (calibration step), the number of LVs must be optimized to obtain feasible results in the validation set (Geladi and Kowalski, 1986).

In this study the optimum number of LVs was selected by analysing the root mean square error of cross-validation (RMSECV) using the venetian blind cross-validation method versus the number of LVs. The fewest number of LVs which could display the minimum RMSECV value was selected (Haaland and Thomas, 1988). Outliers were evaluated using the information from the **X** block. Samples with Hotelling T^2 and Q residuals that present simultaneously these values above the critical values, at a significance level of probability of 0.05, were considered outliers and excluded (Laursen et al., 2011).

2.6.2. Support vector machine

In the support vector machine (SVM) algorithm the **X** block is non-linearly mapped into a high-dimensional feature space using a specific transfer kernel function (Cortes and Vapnik, 1995). After, in this feature space a linear decision surface (hyperplane) is calculated to solve the regression or classification problem. Among the kernel functions used in the SVM models, the radial basis function (Eq. (3)) is the most used.

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

In this function, the kernel scale parameter (γ) controls the influence of a single sample in the calibration model, defining the width of the

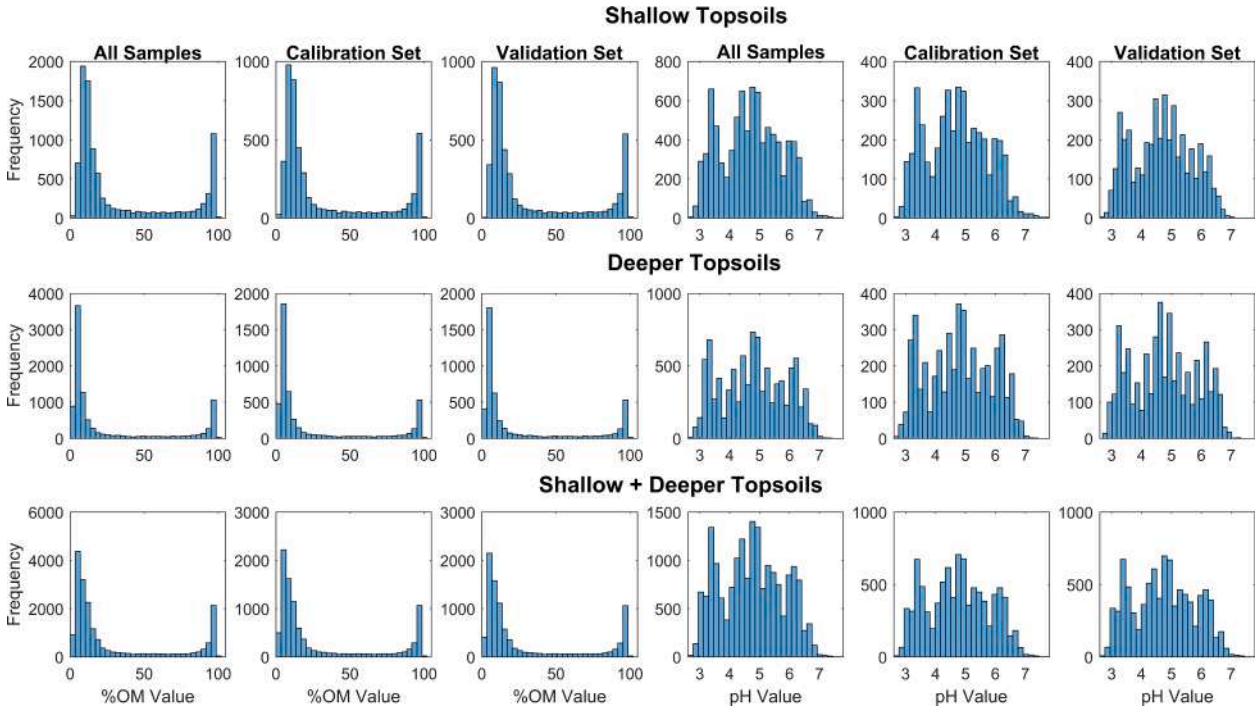

Fig. 1. Histogram of %OM and pH values for shallow and deeper topsoil in: all samples, calibration and validation sets.

Table 1
Libraries adopted to determine %OM and pH in shallow and deeper topsoils.

Name	Calibration Samples	Validation Samples
Library 1	Calibration Shallow topsoils	Validation Shallow topsoils + Validation Deeper topsoils
Library 2	Calibration Deeper topsoils	Validation Shallow topsoils + Validation Deeper topsoils
Library 3	Calibration Shallow topsoils + Calibration Deeper topsoils	Validation Shallow topsoils + Validation Deeper topsoils

Gaussian function (Filgueiras et al., 2014). The prediction of y_i values is done using Eq. (4), where the hyperplane is represented by (w), slope, and (*b*) is the offset of the regression line.

$$y_i = w.K(x_i) + b \tag{4}$$

The regression function is calculated by minimizing Eq. (5) subject to Eq. (6).

$$\text{Minimize}: \frac{1}{2}\|w\| + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{5}$$

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

Where (ε) represents the maximum accepted error during the calibration step and cost parameter (C) determines the trade-off between the model complexity and (ε), in other words the influence of each support vector. The slack variables ξ_i and ξ_i^* are used as a limit to determine the support vectors (Üstün et al., 2005). Comprehensive concepts regarding SVM can be found in Cortes and Vapnik (1995) and Üstün et al. (2005). During this whole process, the parameters ε, C and γ must be optimized. In this study Bayesian optimization was employed to reduce the time required to find this optimum combination (MathWorks, 2017).

2.6.3. Random forest

The Random forest (RF) algorithm (Breiman, 2001) uses an ensemble of decision trees combined with two randomization processes, random feature selection and bagging. The random feature selection consists in randomly selecting *mtry* variables that will be used at each node in each tree. For regression problems the default value for *mtry* is 33.33 % of the total number of variables; in other words, for spectra composed by 1500 variables, each tree will be built analysing 500 variables instead of analysing all 1500 variables. This process saves computational costs by reducing the dimensionality of the data (Breiman, 2001).

The bagging (bootstrap aggregation) was also introduced by Breiman and involves building each tree using a bootstrap replica of the training set with a replacement (Breiman, 1996). In this step approximately 66.67 % of the calibration samples are randomly selected to build each tree and the remaining calibration samples are used to evaluate the performance of that tree, similar to a cross-validation step. Since we have an ensemble of regression models, the prediction results of validation samples will be done by the average value predicted by the ensemble of trees (Breiman, 2001; de Santana et al., 2019). In this study the RF models were built using the default value of *mtry* and 300 trees.

2.6.4. Cubist

The Cubist algorithm was based on Quinlan's M5 algorithm, combining the decision trees to generate the nodes and linear multivariate regression models fitted to determine the y_i value at terminal nodes ("leafs") (Kuhn and Johnson, 2013). Employing this strategy the algorithm can separate a complex dataset into segments of linear data (decision tree). Multivariate regression models are then built in each segment.

In order to optimize the Cubist model we applied a boosting–like scheme called "committees" where iterative tree models are created in sequence. An initial Cubist model is built from which subsequent models attempt to reduce the errors of the previous Cubist model. For example: if the model predicted a sample with y_i value above the reference value, the next model will adjust downward this value (Kuhn and Johnson, 2013). It is important to highlight that, unlike traditional boosting, here the predictions are not made using a weight for each Cubist model. Like

RF, the prediction is the average value predicted by the Cubist committees (Kuhn et al., 2012).

Another implementation used in this study, to optimize the Cubist model, was the use of neighbours, which aims to perform additional corrections based on the nearest neighbours in the training set. In this approach a model is built followed by the prediction of a validation sample ($\widehat{y}$). The K most-similar neighbours are identified, as shown in Eq. (7), and a combination of the reference values of these K samples are used to calculate the final prediction (Quinlan, 1993):

$$\widehat{y} = \frac{1}{K}\sum_{l=1}^{k} wl[\widehat{y} + (tl - \widehat{tl})] \quad (7)$$

where tl is the reference value for a training set neighbour, $\widehat{tl}$ is the predicted value of that neighbour, and wl is a weight calculated using the distance of the training set neighbour to the validation sample (Kuhn and Johnson, 2013; Quinlan, 1993). In order to optimize the Cubist model, different combinations of numbers of committees (1, 10, 30 and 50) and neighbours (0, 1, 5 and 9) were tested using the grid search method with k-fold cross-validation (10 folds). The combination of committees and neighbours that present the lowest RMSECV were selected to build the optimized Cubist model.

2.6.5. Accuracy comparison and statistical tests

In order to compare the accuracy among the regression algorithms and libraries the RMSE and R^2 values were used. These parameters are the two metrics most used to assess the accuracy of regression models. Another parameter commonly used to evaluate the accuracy in spectroscopy models is the ratio of performance deviation (RPD). However, since the %OM and pH distributions did not follow a normal distribution, the ratio of performance to interquartile distance (RPIQ) was used instead. This accuracy parameter was calculated using the interquartile distance (Q3–Q1) divided by the RMSEP (Ludwig et al., 2017).

To assess statistical differences between accuracy parameters among the libraries used, the Wilcoxon signed-rank test and the randomization test suggested by van der Voet (1994) were used. Both tests were selected due to the non-normal distribution of the %OM and pH data.

The Wilcoxon signed-rank test was used to compare two paired populations, i.e. shallow topsoil samples predicted by library 3 and library 2. This is a nonparametric alternative test to the paired Student's t-tests; in other words, this test does require that the samples are normally distributed (Xia, 2020). The hypotheses evaluated were:

H0: $Y_{\text{Model A}} - Y_{\text{Model B}}$ comes from a distribution with zero median (accuracies are equal).

H0: $Y_{\text{Model A}} - Y_{\text{Model B}}$ does not comes from a distribution with zero median (accuracies are not equal).

The randomization test proposed by van der Voet (1994) was also used to compare the models accuracy. This test does not require any assumptions about normality or homoscedasticity of the data (distribution-free). More details about this test and its respective Matlab code can be found in the original paper (van der Voet, 1994). The hypotheses evaluated were:

H0: $\text{RMSEP}_{\text{Model A}} = \text{RMSEP}_{\text{Model B}}$ (accuracies are equal);

H1: $\text{RMSEP}_{\text{Model A}} \neq \text{RMSEP}_{\text{Model B}}$ (accuracies are not equal).

3. Results and discussion

3.1. Summary of reference measurements

The descriptive statistics of shallow and deeper topsoil samples collected from regional and periurban areas in Ireland are presented in Table 2. The pH ($CaCl_2$) values of shallow topsoil samples range between 2.60 and 7.78, with an interquartile range of 3.90–5.47 (50 % of samples), and mean and median of ~4.70. Deeper topsoil samples display a similar distribution to shallow samples (Fig. 1), with pH values between 2.60 and 7.40, an interquartile range of 3.90–5.70 and mean and median of ~4.80. Due to the greater proportion of organic material typically found in shallower soil and the acidification process that normally occur on the soil surface the interquartile range, mean and median pH values of deeper topsoils are slightly more alkaline than shallow topsoils.

The %OM values of shallow topsoil samples range between 0.63 and 99.90, with an interquartile range of 9.66–55.10, mean of 33.05 and median of 14.45. Compared to shallow topsoil samples, deeper soils have a considerably lower content of OM, with an interquartile range of 4.53–35.05, mean of 26.65 and median of 7.16. This is consistent with the observation that shallow topsoils typically contain more OM than deeper topsoils, as the proportion of vegetation and other organic matter decreases with depth in the soil profile. The large difference between mean and median values of %OM in both depths are attributed to the bimodal distribution of %OM values, which can be observed in the first column of Fig. 1. The bimodal distribution reflects the large variability in the region of peat soils, presenting mostly OM content >90 % in the sampled area.

3.2. MIR spectra

The original and pre-processed DRIFT spectra of shallow (Fig. 2 a–d) and deeper (Fig. 2 e–h) topsoils are shown in Fig. 2. The spectra were pre-processed using first derivative and smoothing (Savitzky-Golay) to reduce the baseline variation and enhance the spectral features (Rinnan, 2014). In order to facilitate the spectral interpretation a colour bar containing information about the reference parameters (%OM and pH) was used. For example, in Fig. 2a–b blue spectra have low contents of % OM and red spectra have high contents of %OM.

The Mid-Infrared (MIR) soil spectra display absorptions bands mainly at 3700, 2937, 2960, 2039, 2011, 1905, 1870, 1474, 1166, 1063, 821, 706 cm^{-1}, common to MIR soil spectra reported in previously studies (Baumann et al., 2021; Viscarra Rossel et al., 2008; Stenberg et al., 2010). The absorption bands at 3700–3600 cm^{-1} can be attributed to alumino-silicates, such as kaolinite, smectite and illite (Viscarra Rossel et al., 2008). The broad band around 3400 cm^{-1} can be attributed to the OH group of residual soil water in the structure of alumino-silicates (even after the oven-drying step) (Viscarra Rossel, et al., 2008). The absorption bands at 2930–2850 cm^{-1} and 1720 cm^{-1} are attributed to alkyl (—CH2) and carboxylic acid (—COOH), respectively, resulting in a high %OM content as shown in Fig. 2a–b and Fig. 2e–f. Proteins may be responsible for the absorption band at 1525 cm^{-1}, phenolics (C—OH) for that at 1285 cm^{-1}, and carbohydrates (C—O) for that at 1100 cm^{-1}. Again, in all these bands high %OM content is observed as shown in Fig. 2b and Fig. 2f. More details about each band can be found in the literature (Stenberg et al., 2010).

Fig. 2 a-b and Fig. 2 e-f display a high degree of similarity between

Table 2
Descriptive statistics of shallow and deeper topsoil samples.

	Parameter	No. Samples	Min.	Max.	Q1–Q3	IQR	Mean	Median	Std.
Shallow samples (5–20 cm)	%OM	9396	0.63	99.90	9.66–55.10	45.44	33.05	14.45	33.42
	pH	8790	2.60	7.78	3.90–5.47	1.57	4.71	4.70	1.00
Deeper samples (35–50 cm)	%OM	9463	0.30	99.70	4.53–35.05	30.52	26.65	7.16	34.61
	pH	9503	2.60	7.40	3.90–5.70	1.80	4.79	4.80	1.06

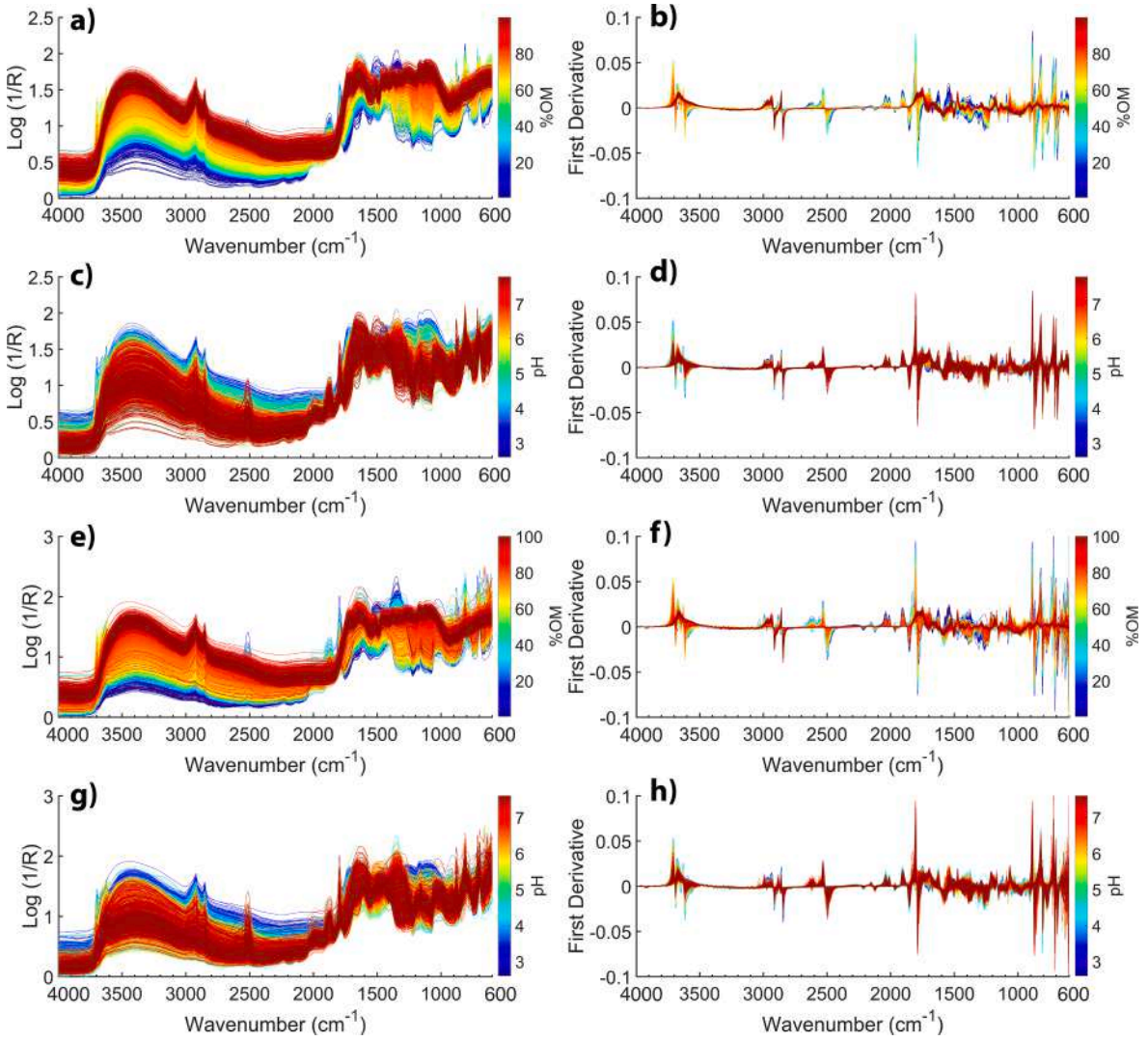

Fig. 2. Original and pre-processed DRIFT spectra of shallow (a–d) and deeper (e–h) topsoils samples.

the shallow and deeper topsoil spectra but there are a few differences related to %OM content at the 2500 cm^{-1}, 1200–1000 cm^{-1} and 780 cm^{-1} regions and at the end of the spectra at 750–600 cm^{-1}. The minor differences related to pH content in shallow and deeper topsoil spectra are shown in the derivative spectra, which enhance the spectral variations: Fig. 2d and Fig. 2h show few differences at 772, 739 cm^{-1} and at the tail end of the spectra at 750–600 cm^{-1}. The absorptions in this range can be attributed to the bending vibration of the B—O—B and Si—O—Si linkages of their segments, resulting in more intense bands in deeper topsoil samples due to the accumulation of B and sand in deeper depths (Karmakar, 2016; Krivoshein et al., 2020).

These minor differences could be an indication that there are different spectral signatures in shallow and deeper topsoil related to % OM and pH. These different spectral signatures could result in an inferior accuracy in the use of Library 1 to determine %OM or pH in deeper topsoil or in the use of Library 2 to determine %OM or pH in shallow topsoil.

3.3. Comparison of modelling approaches and variable importance

Pre-processing algorithms were compared with each regression method. The first derivative and smoothing (Savitzky-Golay) algorithm recorded the lowest RMSECV values for predicted %OM and pH values. The accuracy parameters obtained by SVM, RF, Cubist and PLS regression models for %OM and pH predictions of shallow and deeper topsoil are presented in Table 3.

In all cases the PLS regression method resulted in the lowest accuracy for %OM and pH predictions, in addition to excluding some validation samples as outliers, i.e. samples with Hotelling T^2 and Q residuals values above the limits calculated by the PLS model (Laursen et al., 2011). This fact is attributed to the large variability in composition of the Irish soil samples investigated, resulting in non-linear relationships between the soil parameters and the MIR spectra (Sabetizade et al., 2021).

Among the machine learning algorithms evaluated, the Cubist algorithm yielded the highest accuracy parameters in %OM predictions for both shallow and deeper topsoil using libraries 1, 2 and 3. The Cubist models were built using 30 committees and nine neighbours and this combination resulted in the lowest RMSECV errors using the grid search method. For pH determinations, the SVM and Cubist (30 committees and zero neighbours) regression models yielded similar RMSEP, RPIQ and R^2 values using the libraries 1, 2 and 3.

The models built using Library 1 predicted %OM and pH in shallow topsoils with similar accuracy to that obtained in the cross-validation, i.e. $RMSECV_{shallow} \sim RMSEP_{shallow}$. However, in deeper topsoil predictions the errors obtained were in general 50 % larger than those calculated in the cross-validation, i.e. $RMSEP_{deeper} \sim 1.5 \times RMSECV_{shallow}$.

Using Library 2, we observed the same behaviour in %OM and pH

Table 3
Accuracy parameters obtained by SVM, RF, Cubist and PLS regression models in shallow and deeper topsoils for %OM and pH determinations.

		n_{cal}	RMSECV/ RPIQ	R^2_{cal}	n_{val} Shallow	RMSEP/RPIQ Shallow	R^2_{val} Shallow	n_{val} Deeper	RMSEP/RPIQ Deeper	R^2_{val} Deeper
Library 1 % OM	SVM	4763	1.22/37.22	0.9987	4633	1.35/35.10	0.9984	4604	1.84/25.76	0.9979
	RF	4763	1.51/30.07	0.9980	4633	1.44/32.91	0.9982	4604	2.02/23.46	0.9977
	Cubist	4763	0.90/50.46	0.9993	4633	1.08/43.88	0.9990	4604	1.55/30.57	0.9985
	PLS –19 LV	4763	1.80/25.23	0.9971	4449	1.87/25.34	0.9970	3783	2.13/22.25	0.9972
Library 2 % OM	SVM	4859	1.04/27.21	0.9991	4633	2.13/22.25	0.9964	4604	1.17/40.50	0.9989
	RF	4859	1.23/23.01	0.9987	4633	2.25/21.06	0.9967	4604	1.21/39.17	0.9988
	Cubist	4859	0.80/35.38	0.9995	4633	2.14/22.14	0.9968	4604	0.95/49.88	0.9993
	PLS – 19 LV	4859	1.76/16.08	0.9974	4516	2.85/16.63	0.9949	4435	1.61/29.43	0.9980
Library 3 % OM	SVM	9622	1.13/34.88	0.9989	4633	1.35/35.10	0.9984	4604	1.08/43.88	0.9991
	RF	9622	1.37/28.77	0.9984	4633	1.48/32.02	0.9981	4604	1.22/38.84	0.9988
	Cubist	9622	0.83/47.49	0.9994	4633	1.12/42.31	0.9989	4604	0.91/52.08	0.9993
	PLS – 19 LV	9622	1.97/20.01	0.9967	4542	2.05/23.12	0.9964	4350	1.74/27.24	0.9977
Library 1 pH	SVM	4471	0.21/7.57	0.9557	4319	0.23/6.52	0.9430	4632	0.41/3.66	0.8547
	RF	4471	0.28/5.68	0.9219	4319	0.27/5.56	0.9244	4632	0.43/3.49	0.8469
	Cubist	4471	0.21/7.57	0.9576	4319	0.23/6.52	0.9421	4632	0.41/3.66	0.8526
	PLS – 13 LV	4471	0.29/5.48	0.9140	4242	0.28/5.36	0.9128	4043	0.44/3.41	0.8234
Library 2 pH	SVM	4871	0.30/6.00	0.9183	4319	0.31/4.84	0.9060	4632	0.34/4.41	0.8981
	RF	4871	0.35/5.14	0.8901	4319	0.35/4.29	0.8817	4632	0.35/4.29	0.8905
	Cubist	4871	0.30/6.00	0.9228	4319	0.34/4.41	0.9012	4632	0.34/4.41	0.8986
	PLS – 16 LV	4871	0.37/4.86	0.8772	4238	0.34/4.41	0.8899	4521	0.38/3.95	0.8747
Library 3 pH	SVM	9342	0.28/6.07	0.9248	4319	0.24/6.25	0.9401	4632	0.35/4.29	0.8949
	RF	9342	0.33/5.15	0.8981	4319	0.28/5.36	0.9207	4632	0.36/4.17	0.8889
	Cubist	9342	0.27/6.30	0.9323	4319	0.25/6.00	0.9366	4632	0.34/4.41	0.8954
	PLS – 17 LV	9342	0.35/4.86	0.8860	4270	0.29/5.17	0.9126	4423	0.38/3.95	0.8739

predictions of deeper topsoil, i.e. $RMSECV_{deeper} \sim RMSEP_{deeper}$. For %OM predictions in another depth, $RMSEP_{shallow}$ values were almost twice as large as $RMSECV_{deeper}$. However, for pH predictions the errors obtained for both shallow and deeper topsoil were similar, i.e. $RMSECV_{shallow+deeper} \sim RMSEP_{deeper} \sim RMSEP_{shallow}$. These results shows that the extrapolation of the model to determine pH and %OM values from a different depth to those originally used in calibration, may result in a loss of accuracy, especially in %OM predictions.

The close values of RMSECV, $RMSEP_{deeper}$, $RMSEP_{shallow}$ obtained by the Cubist regression model built using samples from both depths (Library 3), indicates that the Cubist algorithm was able to extract information from both shallow and deeper topsoil samples simultaneously to predict %OM and pH. In order to check if there are statistical differences between the prediction results obtained by the Cubist models using Libraries 1, 2 and 3, two statistical tests were evaluated, the Wilcoxon signed-rank test and the randomization test.

Both tests were carried out to check accuracy differences in the following prediction models:

1 – Shallow topsoil predicted by the Cubist model using Library 3 vs shallow topsoil predicted by Cubist model using Library 2. In other words, we compared the RMSEP of 1.12 % vs 2.14 %, and RPIQ 42.31 vs 22.14 for %OM predictions; and the statistical differences between RMSEP of 0.25 vs 0.34, and RPIQ 6.00 vs 4.41 for pH predictions.

2 – Deeper topsoil predicted by the Cubist model using Library 3 vs deeper topsoil predicted by Cubist model using Library 1. That is, RMSEP of 0.91 % vs 1.55 %, and RPIQ 52.08 vs 30.57 for %OM; and RMSEP of 0.34 vs 0.41, and RPIQ 4.41 vs 3.66 for pH predictions.

The summary of the statistical test carried out to check accuracy differences in the regression models and their respective results are presented in Table 4. As noted all the p-values results obtained by Wilcoxon signed-rank and randomization tests were <0.001. Thus, with a significance level of probability of 0.05, we can conclude that the Cubist models built using Library 3 improved the accuracy in predicting the composition of samples from depths that are not represented in the calibration set. In other words, predicting the composition of deeper topsoil using shallow topsoil (Library 1) and predicting the composition of shallow topsoil using deeper topsoil (Library 2).

The ability to predict %OM and pH in samples from both depths using a unique model can minimize possible errors related to soil collection in the field, such as mislabelling samples depths, improving

Table 4
Wilcoxon signed-rank and the randomization test results.

Calibration model	Library Predicted Library 1 %OM RMSEP	RPIQ	p-value	pH RMSEP	RPIQ	p-value
			Wilcoxon* Randomization**			Wilcoxon* Randomization**
Cubist (library 3)	1.12	43.31	*<0.001	0.25	6.00	*<0.001
Cubist (library 2)	2.14	22.14	**<0.001	0.34	4.41	**<0.001
Calibration model	Library 2 %OM RMSEP	RPIQ	p-value	pH RMSEP	RPIQ	p-value
			Wilcoxon* Randomization**			Wilcoxon* Randomization**
Cubist (library 3)	0.91	52.08	*<0.001	0.34	4.41	*<0.001
Cubist (library 2)	1.55	30.57	**<0.001	0.41	3.66	**<0.001

the robustness of the proposed method, as well as eliminating the requirement to build models to determine only shallow or deeper topsoils.

This excellent performance of Cubist algorithm shows that the decision trees were able to separate the complex SSL containing samples from both depths in segments of homogeneous data, maximizing posteriorly the use of linear multivariate regression models in these segments. The scatter plots showing reference versus predicted values of % OM and pH in shallow and deeper topsoil in Library 3 using the Cubist regression algorithm are shown in Fig. 3. In order to facilitate the scatter plots analysis, a colour bar containing the recurrence of the predicted values for each reference value was inserted.

The results obtained in %OM predictions using DRIFT spectroscopy combined with the Cubistic algorithm are very close to those obtained by the laboratory LOI method, with RMSEP values of 1.12 and 0.91 %, RPIQ of 42.34 and 38.48; and R^2_{val} values of 0.9989 and 0.9993 in shallow and deeper topsoil predictions, respectively. Analyses of the recurrence of predicted values of %OM in Fig. 3 show that most of the values range between 0 and 20 % and 90–100 %. In addition, these samples were predicted with %OM values close to the reference ones.

The same methodology was able to determine pH values with acceptable accuracy, presenting RMSEP values of 0.25 and 0.34; RPIQ of 6.04 and 4.94; R^2_{val} of 0.9385 and 0.8954 in shallow and deeper topsoil predictions, respectively. Despite the dispersion in the predicted pH values in Fig. 3, the colour bar shows that most of the samples were predicted with pH values close to the reference ones (green, yellow and red). Only some samples, represented by dark blue, were predicted with pH values that were significantly different from the reference ones. In addition, both regression models were classified as excellent predictor models (RPIQ values >4.05) according to (Ludwig et al., 2017), yielding for shallow and deeper topsoil RPIQ values >4.90.

The reference and predicted maps of %OM and pH in shallow and deeper topsoil, using Library 3 combined with the Cubist algorithm are shown in Fig. 4. It is observed that %OM and pH maps display strong similarity between the reference and predicted values in both shallow and deeper topsoils, with a few variations across the map. These slight variations can be better visualized on the prediction error maps, where for %OM and pH most of the validation samples are displayed with prediction errors of ±1 % and ±0.25, respectively (yellow points), i.e. about ±1 × RMSEP for both models. This strong coherence between the maps shows the excellent accuracy obtained by the DRIFTS methodology in agriculture applications.

The distributions of %OM and pH patterns of soils shown in Fig. 4 reflect bedrock geology, climate, topography, the presence of living organisms and time. These soil-forming factors vary considerably across Ireland, hence the large number of soil types across the country, including the study area in this work. Ireland's climate is described as a temperate, west maritime with highest rainfall along the western seaboard and inland areas of high relief. The combination of bedrock geology and glaciation has influenced the spatial distribution of peatland (%OM ≥ 30) and organo-mineral (%OM < 30) soils in Ireland. These soils typically form in areas with impeded drainage and high rainfall. The predominant soils in the west and north-west of the study area are peat and organo-mineral; wet and acidic in nature characterised by high % OM and low pH values. The landscape and climate of the midland and eastern region of the study is different. These areas historically have lower rainfall with more hours of daily sunshine and 40 % of the midlands is underlain by Carboniferous rocks. Soil in these regions are derived from calcite-rich limestone that underlies much of the central and eastern parts of the study area. It is predominantly mineral soil, with a tendency towards higher soil pH and lower %OM (Gardiner and Radford, 1980; Hammond, 1981).

Fig. 5 shows the residuals ($y_{reference} - \hat{y}_{predicted}$) obtained in the validation set by %OM and pH predictions in shallow and deeper topsoil. It displays a positive bias for soil samples with high %OM content (≥97.5 %), i.e. in this range (≥97.5 % & ≤100 %) the validation samples were predicted to have a %OM content slightly lower than the reference values. However, this is not problematic since most of these samples were predicted with errors ≤2 %. In addition, this range is composed of only 82 (1.77 %) and 119 (2.58 %) samples from shallow and deeper topsoil, respectively.

For pH predictions we observe a negative bias for samples from both depths with pH ≤ 3.0. Also, there is a positive bias for shallow topsoil with pH ≥ 6.1 and for deeper topsoil with pH ≥ 6.3. That is, there is some bias in the pH predictions for 467 (10.08 %) and 414 (9.58 %) samples from shallow and deeper topsoil, respectively. Despite this slight bias, as we can see in the predicted maps (Fig. 4), these differences are minimal and do not give rise to significant variations in the %OM and pH predicted maps.

During the construction of the Cubist regression models some variables were used more frequently to create the nodes or to fit the linear models. These variables are known as important variables and are calculated in percentages by their usage in the rule conditions of Cubistic model (Kuhn et al., 2012). The importance of each variable

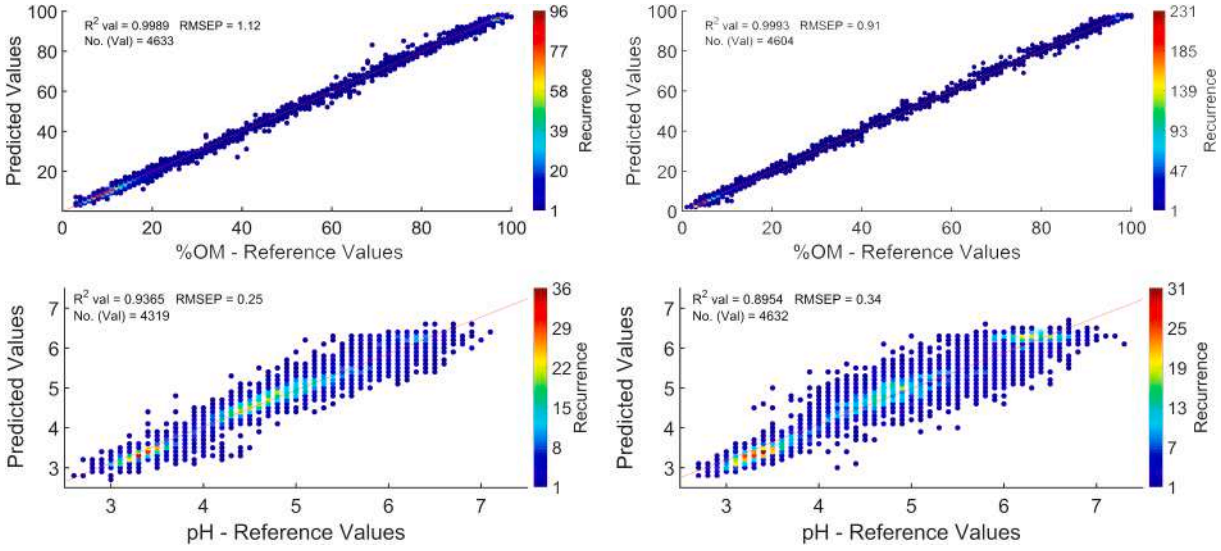

Fig. 3. Scatter plots showing reference versus predicted values of %OM (first row) and pH (second row) in shallow (first column) and deeper (second column) soils in Library 3 using the Cubist regression algorithm.

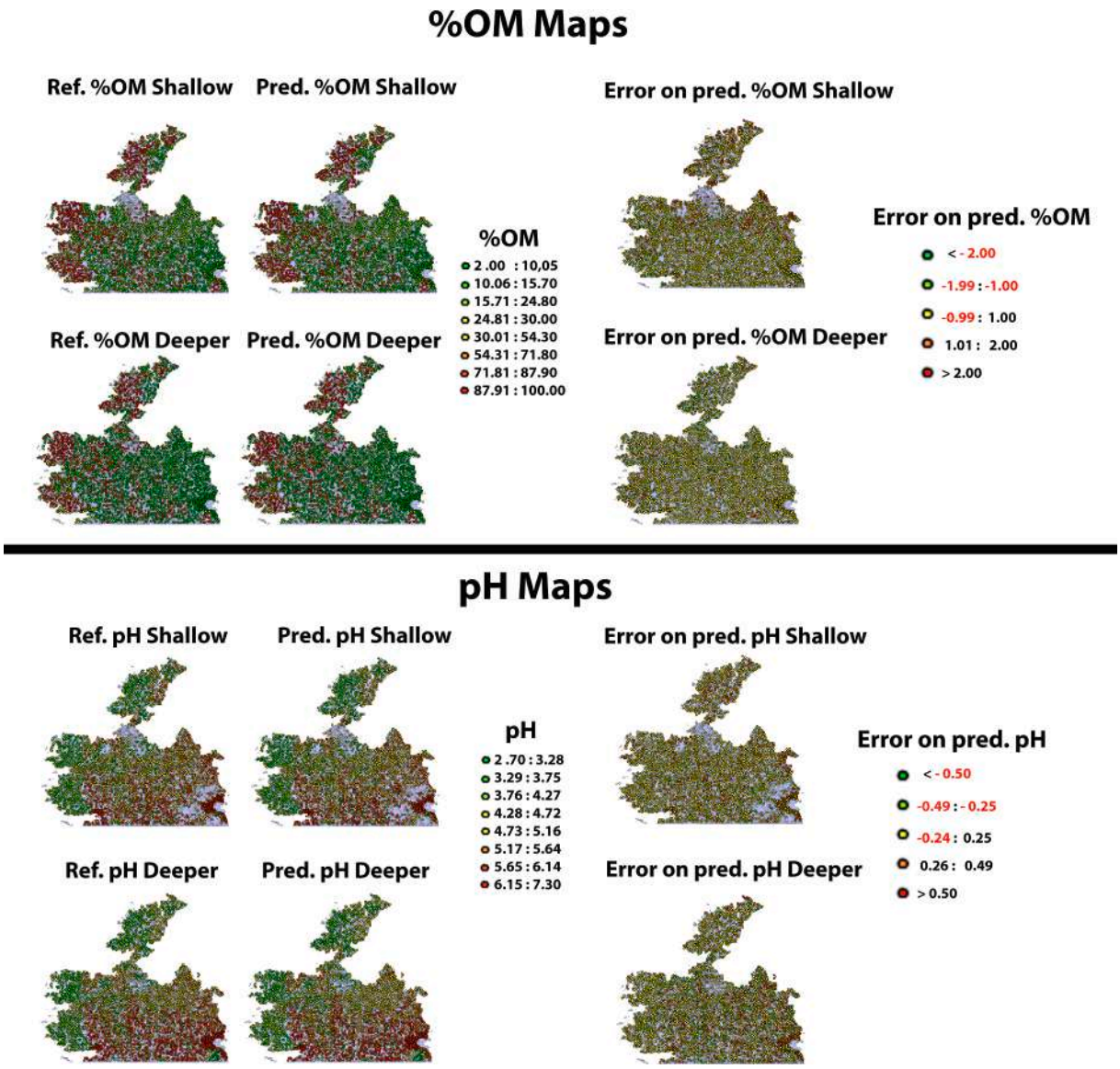

Fig. 4. Reference and predicted maps of %OM and pH and their respective prediction errors, in shallow and deeper topsoils using the DRIFT methodology.

calculated by the Cubist model using Libraries 1, 2 and 3 in %OM and pH determinations are illustrated in Fig. 6. The analysis of these variables indicates which spectral regions/vibrational groups were important to correlate the MIR spectra with the soil property evaluated, in this case % OM and pH.

In general, Libraries 1 and 2 resulted in a similar importance variable profile in %OM and pH predictions. However, for %OM predictions a close inspection of Fig. 6 reveals differences between the intensities of the most important variables, i.e. in the range 810–600 cm^{-1} the model built using Library 1 showed higher variable importance compared to Libraries 2 and 3. This can be attributed to higher content of SiO_2 accumulated in deeper topsoil with a higher mineral content. The high content of SiO_2 in some deeper topsoil samples can contribute spectral noise in this range, slightly changing the spectral signature of %OM content with the MIR spectra, as shown in Fig. 2b and Fig. 2f.

On the other hand, in the range 3000–2800 cm^{-1} the %OM model built using Library 2 presented higher variable importance than the model built using Library 1. The important bands at 2979 cm^{-1} and 2844 cm^{-1} obtained by Library 2 are also related to the higher content of SiO_2 in deeper topsoil, resulting in intense bands in this range, which are negatively correlated with %OM, as shown in Fig. 2f.

There are also some specific bands in Library 1 that were not selected by Library 2, such as the peaks at 1762, 666, and 610 cm^{-1}. The opposite behaviour was also observed, where the peaks at 2979, 1925, 1755 cm^{-1} were considered important by Library 2 but were not considered important by Library 1. For pH predictions the most important variables display a high degree of similarity, with few differences at 2883, 1794, 1218 and 729 cm^{-1}. These slight differences between the peaks and band intensities for the values of pH and %OM are responsible for the loss of accuracy in the use of Library 1 to determine %OM and pH in deeper topsoils or in the use of Library 2 to determine these values in shallow topsoils. The models built using Library 3 display an average intensity importance between Library 1 and Library 2, this is attributed to the use of samples from both libraries in Library 3. This behaviour is clearly evident in the variable importance of %OM models.

In %OM predictions, the most important variables are located at: 3698, 3625, 2979, 2926–2910, 2844, 2115, 1925, 1795–1691, 1560–1380, 1235, 1200, 1165, 1072, 906, 840, 810, 783 and 698 cm^{-1}. The variables at 3690–3620 cm^{-1} are associated with clay minerals, resulting in low value of %OM in Fig. 2b and 2f. Those at 2930–2840 cm^{-1} reflect —CH and SiO_2, while those at 2115 cm^{-1} can be attributed to alkyne. Variables at 2000–1900 cm^{-1} can be attributed to C—H

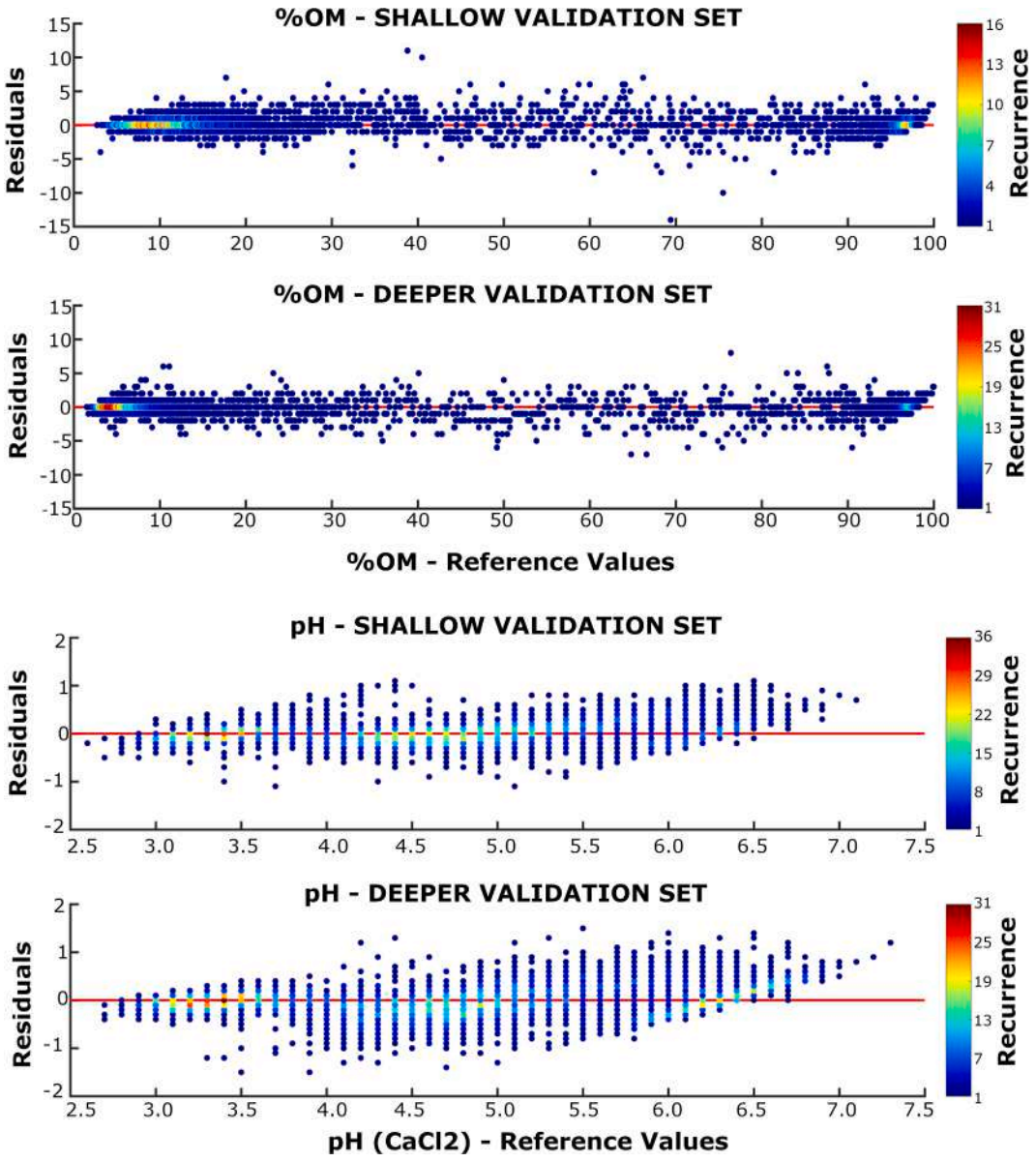

Fig. 5. Residuals obtained by Cubist regression models in %OM and pH determinations in shallow and deeper topsoils.

bending (aromatic), at 1800–1700 cm^{-1} to C═O, at 1535 cm^{-1} to protein amide, at 1480–1350 cm^{-1} to methyl —CH, at 1100–1000 to quartz (resulting in a low values of %OM in Fig. 2b and 2f), at 900–780 cm^{-1} to —CH and at 700–600 to iron oxides, B—O—B and Si—O—Si linkage (Baumann et al., 2021; Soriano-Disla et al., 2014; Stenberg et al., 2010; Krivoshein et al., 2020).

The most important variables in pH predictions are located at: 2926, 2862, 1794, 1714 (most intense), 1688, 1517–1510, 1479, 1389, 1218, 794 and 726 cm^{-1}. Most of these bands have already been discussed in %OM models (2926, 2862, 1479, 1389 and 794 cm^{-1}). The absorption band at 1517 cm^{-1} can be attributed to —CH and —NH, that at 1218 cm^{-1} may be attributed to C—O, that at 726 cm^{-1} may be attributed to benzene derivative and the most important bands at 1720–1680 cm^{-1} are related to C═O of carboxylic acid and esters, indicating high values of pH in this range Fig. 2d and 2 h (Ch'ng et al., 2011; Soriano-Disla et al., 2014; Stenberg et al., 2010).

The combination of the important variables (Fig. 6) with the spectra containing a colour bar related to pH and %OM content (Fig. 2), enables visualization of the spectral signatures related pH and %OM content in shallow and deeper topsoil. Also, the most important variables selected by the regression models are chemicals related to pH and %OM, attributing chemical, in addition to mathematical, meaning to the regression models. Finally, all of these results elucidate the differences in RMSEP and RPIQ values obtained by the use of samples from shallow topsoil to predict samples from deeper topsoil and vice-versa.

4. Conclusions

In this paper an accurate, rapid and cost-effective methodology to predict %OM and pH in shallow and deeper topsoil samples from agricultural and non-agricultural regions was developed. The spectral libraries collected in MIR range using HTS-DRIFT system were combined with laboratory data and used to assess differences in %OM and pH predictions in Irish shallow and deeper topsoil. Among the regression methods tested, the Cubist algorithm built using shallow and deeper topsoil achieved the best accuracy parameters in %OM and pH predictions in shallow and deeper topsoil using a unique multivariate regression model. The proposed methodology was able to determine the pH values with good accuracy, with RMSEP values of 0.25 and 0.34; RPIQ values of 6.04 and 4.94; R^2_{val} of 0.9385 and 0.8954 for shallow and

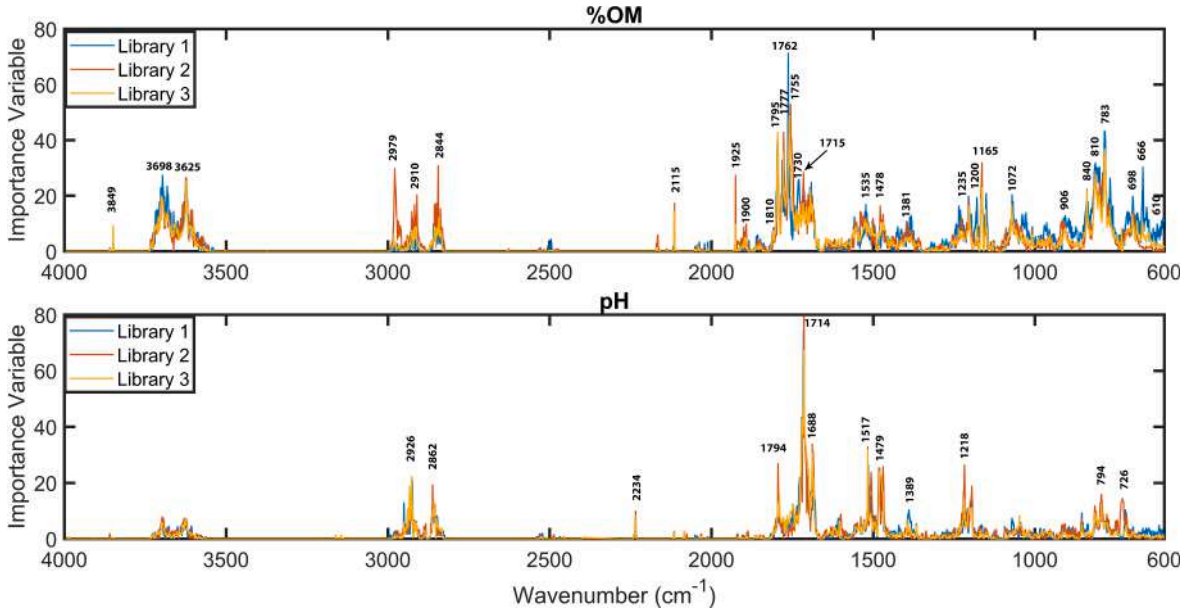

Fig. 6. Variable importance for %OM (first row) and pH (second row) determinations in libraries 1, 2 and 3 using the Cubist regression algorithm.

deeper topsoil, respectively. The %OM content was predicted with comparable accuracy to that of the laboratory LOI reference method, with RMSEP values of 1.12 and 0.91 %; RPIQ values of 42.34 and 38.48; R^2_{val} values of 0.9989 and 0.9993 in shallow and deeper topsoil, respectively. Both regression models are classified as excellent predictions models, yielding RPIQ values >4.05 for shallow and deeper topsoil.

The important variable analysis of Cubist models combined with the first derivative of the spectra ensures that the important variables are chemically related to the %OM and pH in soil. This confirms that the Cubist algorithm was able to extract the appropriate chemical information from the SSL to build the regression models. This study contributes towards the understanding of the analysis of soil from different depths using Mid-Infrared spectroscopy, demonstrating that for different depths there are minor different spectral group associations with the values of pH and %OM, which resulted in an inferior accuracy in the extrapolation of the model built using shallow topsoil to predict samples from deeper topsoil or vice versa. The proposed methodology shows high potential to be used as a routine rapid alternative or complementary method in soil laboratories to determine %OM and pH. The speed of analysis can lead to the development of accurate, high-resolution maps of soil composition in agricultural and non-agricultural regions, enabling ongoing cost-effective monitoring. Finally, the proposed methodology can contribute to achieving the goals of the Irish 2030 agri-food strategy goals.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Karen Daly reports financial support was provided by Irish Agriculture and Food Development Authority. Karen Daly reports was provided by Geological Survey of Ireland.

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 at Teagasc Johnstown Castle laboratories. Geological Survey Ireland authors publish with the permission of the Director of Geological Survey Ireland.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.catena.2022.106552.

References

Agriculture, G., 2021. Industrial Agriculture and Small-scale Farming. Available from: https://www.globalagriculture.org/report-topics/industrial-agriculture-and-small-scale-farming.html#:~:text=Approximately 3.4 billion people – or, derive their livelihoods from agriculture (accessed 02 November 2021).

Aitchison, J., 1982. The Statistical Analysis of Compositional Data. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-009-4109-0.

Allo, M., Todoroff, P., Jameux, M., Stern, M., Paulin, L., Albrecht, A., 2020. Prediction of tropical volcanic soil organic carbon stocks by visible-near- and mid-infrared spectroscopy. Catena 189, 104452. https://doi.org/10.1016/j.catena.2020.104452.

Baumann, P., Helfenstein, A., Gubler, A., Keller, A., Meuli, R.G., Wächter, D., Lee, J., Viscarra Rossel, R.A., Six, J., 2021. Developing the Swiss soil spectral library for local estimation and monitoring. Soil Discuss. 1–32 https://doi.org/10.5194/soil-2020-105.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Ch'ng, H.-Y., Ahmed, O.H., Ab. Majid, N.M., 2011. Qualitative assessment of soil carbon in a rehabilitated forest using Fourier transform infrared spectroscopy. Sci. World J. 11, 532–545. https://doi.org/10.1100/tsw.2011.54.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1023/A:1022627411411.

Dangal, S.R.S., Sanderman, J., Wills, S., Ramirez-Lopez, L., 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Syst. 3, 1–23. https://doi.org/10.3390/soilsystems3010011.

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.

Demattê, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M.V., Dalmolin, R.S.D., de Araújo, M.d. S.B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., de Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barretto Sampaio, E.V., Menezes, R.S.C., de Souza, J.J.L.L., Abrahão, W.A.P., Coelho, R.M., Grego, C.R., Lani, J.L., Fernandes, A. R., Gonçalves, D.A.M., Silva, S.H.G., de Menezes, M.D., Curi, N., Couto, E.G., dos Anjos, L.H.C., Ceddia, M.B., Pinheiro, É.F.M., Grunwald, S., Vasques, G.M., Marques Júnior, J., da Silva, A.J., Barreto, M.C.d.V., Nóbrega, G.N., da Silva, M.Z., de Souza, S.F., Valladares, G.S., Viana, J.H.M., da Silva Terra, F., Horák-Terra, I., Fiorio, P.R., da Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M.F., de Souza Junior, V.S., Brefin, M.D.L.M.S., Ruivo, M.D.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., de Sousa Mendes, W., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., e Souza, A.B., Quesada, C.A., do Couto, H.T.Z., 2019. The Brazilian

Soil Spectral Library (BSSL): a general view, application and challenges. Geoderma 354, 113793. https://doi.org/10.1016/j.geoderma.2019.05.043.
Department of Agriculture Food and Marine, 2020. Fact Sheet on Irish Agriculture - September 2020. Available from: https://www.gov.ie/en/publication/3ec3a-fact-sheet-on-irish-agriculture-september-2020/ (accessed 02 November 2021).
Filgueiras, P.R., Alves, J.C.L., Poppi, R.J., 2014. Quantification of animal fat biodiesel in soybean biodiesel and B20 diesel blends using near infrared spectroscopy and synergy interval support vector regression. Talanta 119, 582–589. https://doi.org/10.1016/j.talanta.2013.11.056.
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.
Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.
Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., Quinn, T., 2022. Aitchison's Compositional Data Analysis 40 Years On: A Reappraisal. Available from: http://arxiv.org/abs/2201.05197 (accessed 24 May 2022).
GSI, Geological Survey Ireland (2021). Terra Soil. Available from: https://www.gsi.ie/en-ie/programmes-and-projects/tellus/activities/tellus-product-development/smart-agriculture/Pages/Terra-Soil.aspx (accessed 02 November 2021).
Haaland, D.M., Thomas, E.V., 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60 (11), 1193–1202. https://doi.org/10.1021/ac00162a020.
Hammond, R.F., 1981. The Peatlands of Ireland. Soil Survey Bulletin No. 35. An Foras Taluntais, Dublin.
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 (1), 127–139. https://doi.org/10.1111/ejss.12761.
Karmakar, B., 2016. Fundamentals of glass and glass nanocomposites. In: Glass Nanocomposites. Elsevier, pp. 3–53. https://doi.org/10.1016/B978-0-323-39309-6.00001-8.
Knights, K.V., Szpak, M., Mather, J., Collins, L., 2020. Tellus geochemical survey: shallow topsoil data from the border and west of Ireland.
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.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling with Applications in R. Available from: http://appliedpredictivemodeling.com/s/Applied_Predictive_Modeling_in_R.pdf (accessed 02 November 2021).
Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2012. Cubist Models for Regression. R Packag. *Packag.* Available from: https://mran.microsoft.com/snapshot/2016-09-04/web/packages/Cubist/vignettes/cubist.pdf (accessed 02 November 2021).
Laursen, K., Rasmussen, M.A., Bro, R., 2011. Comprehensive control charting applied to chromatography. Chemom. Intell. Lab. Syst. 107 (1), 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.
MathWorks, 2017. Statistics and Machine Learning Toolbox™ User's Guide R 2017 a. The MathWorks Inc, Apple Hill Drive.
Metzger, K., Zhang, C., Daly, K., 2021. From benchtop to handheld MIR for soil analysis: predicting lime requirement and organic matter in agricultural soils. Biosyst. Eng. 204, 257–269. https://doi.org/10.1016/j.biosystemseng.2021.01.025.
Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69 (1), 140–153. https://doi.org/10.1111/ejss.12499.
Quinlan, J.R., 1993. Combining instance-based and model-based learning. In: Machine Learning Proceedings 1993. Elsevier, pp. 236–243. https://doi.org/10.1016/B978-1-55860-307-3.50037-X.
Rinnan, Å., 2014. Pre-processing in vibrational spectroscopy – when, why and how. Anal. Methods 6 (18), 7124–7129. https://doi.org/10.1039/C3AY42270D.
Sabetizade, M., Gorji, M., Roudier, P., Zolfaghari, A.A., Keshavarzi, A., 2021. Combination of MIR spectroscopy and environmental covariates to predict soil organic carbon in a semi-arid region. Catena 196, 104844. https://doi.org/10.1016/j.catena.2020.104844.
Sanderman, J., Savage, K., Dangal, S.R.S., 2020. Mid-infrared spectroscopy for prediction of soil health indicators in the United States. Soil Sci. Soc. Am. J. 84 (1), 251–261. https://doi.org/10.1002/saj2.20009.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M. J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49 (2), 139–186. https://doi.org/10.1080/05704928.2013.811081.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, 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.
Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., Chen, H.YH., 2013. Prediction of soil organic carbon at the European scale by visible and near InfraRed reflectance spectroscopy. PLoS One 8 (6), e66409. https://doi.org/10.1371/journal.pone.0066409.
Survey, S., 1999. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys (second ed.). US Agriculture Handbook, U.S. Government Printing Office, Washington, DC. https://doi.org/10.1007/BF01574372.
Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: comparing vis-NIR and mid-IR reflectance data. Geoderma 255–256, 81–93. https://doi.org/10.1016/j.geoderma.2015.04.017.
Üstün, B., Melssen, W.J., Oudenhuijzen, M., Buydens, L.M.C., 2005. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. Anal. Chim. Acta 544 (1-2), 292–305. https://doi.org/10.1016/j.aca.2004.12.024.
van der Voet, H., 1994. Comparing the predictive accuracy of models using a simple randomization test. Chemom. Intell. Lab. Syst. 25 (2), 313–323. https://doi.org/10.1016/0169-7439(94)85050-X.
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.
Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63, 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Sci. Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012.
Xia, Y., 2020. Correlation and association analyses in microbiome study integrating multiomics in health and disease. Prog. Mol. Biol. Transl. Sci. 171, 309–491. https://doi.org/10.1016/BS.PMBTS.2020.04.003.
Yang, J., Wang, X., Wang, R., Wang, H., 2020. Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using vis–NIR spectroscopy. Geoderma 380, 114616. https://doi.org/10.1016/j.geoderma.2020.114616.