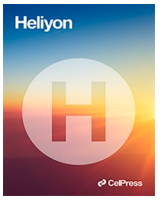
Heliyon 10 (2024) e30228

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Machine Learning and Feature Selection for soil spectroscopy. An evaluation of Random Forest wrappers to predict soil organic matter, clay, and carbonates

Francisco M. Canero [a,*], Victor Rodriguez-Galiano [a], David Aragones [b]

[a] Department of Physical Geography and Regional Geographic Analysis, Universidad de Sevilla, 41004, Seville, Spain
[b] Remote Sensing and Geographic Information Systems Lab (LAST-EBD), Doñana Biological Station, C.S.I.C., 41092, Seville, Spain

ARTICLE INFO

Keywords:
Random forest
Sequential flotant selection
Sequential flotant forward selection
Partial least squares regression
Wrapper methods
Sierra de las nieves

ABSTRACT

Soil spectroscopy estimates soil properties using the absorption features in soil spectra. However, modelling soil properties with soil spectroscopy is challenging due to the high dimensionality of spectral data. Feature Selection wrapper methods are promising approaches to reduce the dimensionality but are barely used in soil spectroscopy. The aim of this study is to evaluate the performance of two feature selection wrapper methods, Sequential Forward Selection (SFS) and Sequential Flotant Forward Selection (SFFS) built using the Random Forest (RF) algorithm, for dimensionality reduction of spectral data and predictive modelling of modelling soil organic matter (SOM), clay and carbonates. The reflectance of 100 soil samples, acquired from Sierra de las Nieves (Spain), was measured under laboratory conditions using ASD FieldSpec Pro JR. Four different datasets were obtained after applying two spectral preprocessing methods to raw spectra: raw spectra, Continuum Removal (CR), Multiplicative Scatter Correction (MSC), and a so-called "Global" dataset composed of raw, CR and MSC features. The performance of RF models built with feature selection methods was compared to that of Partial Least Squares Regression (PLSR) and RF (alone).

RF models built with SFS and SFFS outperformed PLSR and RF alone models: The best RF models with feature selection had a respective ratio of performance to interquartile distance of 1.93, 0.38 and 2.56. PLSR models had an accuracy of 1.41, 0.29 and 1.81 for SOM, carbonates, and clay, respectively. RF alone had a respective performance of 1.29, 0.29 and 1.81. The application of feature selection wrapper methods reduced the number of features to less than 1 % of the starting features. Features were selected across all spectra for SOM and clay, and around 900 nm, 1900 nm, and 2350 nm for carbonates. However, feature selection highlighted features around 1100 nm in SOM modelling, as well as other features around 2200 nm, which is considered a main absorption feature of clay. The application of feature selection with Random Forest was very important in improving modelling accuracy, reducing the redundant features and avoiding the curse of dimensionality or Hughes effect. Thus, this research showed an alternative to dimensionality reduction approaches that have been applied to date to model soil properties with spectroscopy and paves the way for further scientific investigation based on feature selection methods and machine learning.

* Corresponding author.
E-mail address: fcanero@us.es (F.M. Canero).

https://doi.org/10.1016/j.heliyon.2024.e30228
Received 5 October 2023; Received in revised form 19 April 2024; Accepted 22 April 2024
Available online 25 April 2024

1. Introduction

Spectroscopy deals with the absorption, emission and reflection of electromagnetic radiation by atoms and molecules [1]. Soil spectroscopy allows the estimation of spectrally active properties of soil [2,3] by studying the absorption features derived from their chemical composition and structure [4–6]. Some examples of spectrally active soil properties include minerals (clays, iron and carbonates), soil organic matter (SOM), moisture content and hygroscopic water [7]. Each soil property responds to different regions of the visible and near-infrared spectrum (vis-NIR-SWIR, i.e., 400–2500 nm). Clays and carbonates have their main absorption features around 2200 nm and 2350 nm, respectively [8,9], due to vibration overtones and combination modes of functional groups presents in clay and carbonates chemical structure. SOM and iron minerals could be primarily assessed with visible spectrum due to electron transition in atoms [10]. Moisture content leads to reduced reflectivity throughout the spectrum [11,12], with a greater impact on their relative absorption features, around 1400 and 1900 nm, and throughout the shortwave infrared spectrum [13].

A step of paramount importance in spectroscopic analysis prior to modelling is the application of spectral preprocessing (or pretreatment) methods [14]. These methods transform the spectral signal reducing the irrelevant information and improving model robustness [15]. Spectral preprocessing methods are also used to linearize the often non-linear relationship between spectral data and soil properties [16]. Furthermore, spectral preprocessing enhances absorption features and reduce the physical effects and noise in spectroscopy measurements [17]. Different spectral preprocessing methods are Continuum Removal [18], Multiplicative Scatter Correction [19], first derivative of spectra or Standard Normal Variate [20].

Predictive modelling of soil properties using spectroscopy has been performed using different linear modelling algorithms such as Multiple Linear Regression or Partial Least Squares Regression [21,22]. More recently, soil properties modelling with spectral data has benefited from the development of machine learning (ML) algorithms [23,24]. ML is a subfield of artificial intelligence inspired in biological learning [25]. ML groups together a series of algorithms for predictive modelling based on pattern recognition, such as Random Forest (RF) [25], Gradient Boosting (GB) [26] or Support Vector Machine (SVM) [27]. Different ML algorithms have been used in soil spectroscopy, including RF [28–30], SVM [31,32], GB [33], a comparison of different ML algorithms [34,35] and even the use of Deep Learning algorithms [36,37] such as Convolutional Neural Networks [38]. ML algorithms have a series of advantages compared to linear algorithms, including the ability to establish non-linear relationships among data and no assumption of normality for predictor features [39], which could have a positive impact, i.e. greater ML accuracy compared to linear algorithms in complex feature spaces [40].

Modelling soil properties using spectroscopy is challenging as spectral data are high-dimensional. Modern spectrometers have a spectral sampling near to 1 nm, resulting in more than two thousand features (i.e. wavelengths) in vis-NIR-SWIR range. However, as absorption features are located in specific parts of spectrum, most features are non-informative, which might explain the decrease in the accuracy of models if the overall spectrum is used [41]. This phenomenon is known as Hughes effect or "dimensionality curse" [42,43], which can be addressed through ML procedures. Two main approaches have been used to deal with Hughes effect in ML, feature extraction and feature selection [44]. Feature extraction reduces the number of features prior to modelling by creating new features from existing ones. Feature extraction is embedded in some modelling algorithms such as Principal Component Regression and PLSR, therefore its use has been common in spectroscopy studies. The main issue related to feature extraction is that non-informative features are not eliminated and thus are projected in the extracted features [45]. Feature selection is a process related to predictive modelling that selects a subset of original features with the aim of reducing the dimensionality of a dataset according to a specific criterion [46]. Selecting the optimal subset involves avoiding the Hughes effect, therefore improving modelling accuracy. Moreover, the optimal subset might be related with underlying physical processes that explain modelling results [47]. Feature Selection brings together three different methods: filters, embedded and wrapper [48]. Filters methods are independent from ML algorithms [49]. Embedded methods perform the selection during the training process, and are included into specific ML algorithms [50]. Wrapper methods combine a given predictive modelling method with a feature-search strategy, selecting the optimal subset based on a given criterion [50]. This latter method is considered of better performance than filter and embedded methods [51,52]. The use of different feature selection methods has been reported in soil spectroscopy: mutual information based filters [41,53], embedded feature selection methods built with PLSR, such as Variable Importance in Projection [54,55], feature selection wrapper methods built with PLSR, such as Competitive Adaptive Reweighted Sampling [34,56,57]; and genetic algorithms [58,59]. Moreover, recent studies have benchmarked different Features Selection methods [37,60], stating that the best Feature Selection methods depends on the dataset and modelling algorithm used.

This study provides a novel insight into the use of two feature selection wrapper methods, Sequential Forward Selection (SFS) and Sequential Flotant Forward Selection (SFFS) within the framework of soil modelling using vis-NIR-SWIR reflectance. Both methods are sequential forward search algorithms that select features one by one by adding features starting from an empty subset using a greedy procedure [61]. Forward search strategies are particularly computationally advantageous and robust against overfitting [62]. SFS and SFFS methods have not been used in soil vis-NIR-SWIR spectroscopy, although similar sequential Feature Selection methods have achieved good results [60]. Harefa and Zhou [63] did use four ML algorithms built with SFS using laser-induced breakdown spectroscopy to predict soil classes, with a better performance of models built with SFS. This study also examines the usefulness of combining different datasets (raw and preprocessed spectra) in a unique dataset. Raw and preprocessed spectra are usually modelled separately, as this may lead to a dramatic increase in the dataset's dimensionality. Therefore, the objective of his study is three-fold, where two of the objectives are modelling-related, and the latter one is related to dimensionality reduction. The first objective is assessing the performance of Sequential Forward Selection and Sequential Flotant Forward Selection methods built with Random Forest compared to RF alone (using an embedded Feature Selection method) and PLSR (with its own Feature Extraction method for

dimensionality reduction) for soil organic matter (SOM), clay and carbonates predictive modelling. The second objective was comparing the modelling performance of four different datasets, including a raw dataset, Continuum Removal dataset, Multiplicative Scatter Correction dataset, and a global dataset which combines raw, CR and MSC datasets. The third objective was evaluating the application of SFS and SFFS methods to reduce the high dimensionality of spectra data and identify key features in the modelling process.

2. Materials and methods

2.1. Study area

The area selected for this study is located in Sierra de las Nieves national park in Malaga, Spain (Fig. 1). The park has an area of approximately 229.79 km^2, with elevations ranging between 127 and 1917 m (summit of La Torrecilla). The main types of soils are leptosols, eutric cambisols, calcaric cambisols and chromic luvisols [64]. The landcover found in the study area is quite varied: there are forested areas, with species of the genus Abies, Pinus and Quercus [65]; scrublands; grasslands; bare soil and croplands. The lithological substrate is divided into two major sections: peridotite and carbonate rock [66]. These lithologies have a paramount importance in soil formation: cambisols are primarily associated with carbonate lithologies, and luvisols are associated with peridotites. The study area's climate is distinctly Mediterranean (Csa, under the Köppen-Geiger climate classification), with a dry summer period and abundant precipitation from September to May (mean annual rainfall of 954 mm) [67].

2.2. Data

2.2.1. Soil sampling

A total of 100 topsoil samples were taken between 12 and 20 October 2019. Spatial sampling was designed based on the Third Spanish National Forest Inventory (*Tercer Inventario Forestal Nacional de España*) [68], a systematic spatial sampling with a 1-km grid distance. Points were selected if they met two conditions: i) located at least 250 m away from roadways, ii) slope of less than 30 % based on the Digital Elevation Model of the National Geographic Institute of Spain (*Modelo Digital de Elevaciones del Instituto Geográfico Nacional de España*) at a resolution of 5 m. In the field, the exact location of the point was selected, depending on accessibility, in a location within a 100-m buffer of the point. Sampling intensity was also increased in areas with greater soil variability. Soil samples were taken at a depth of 0–10 cm.

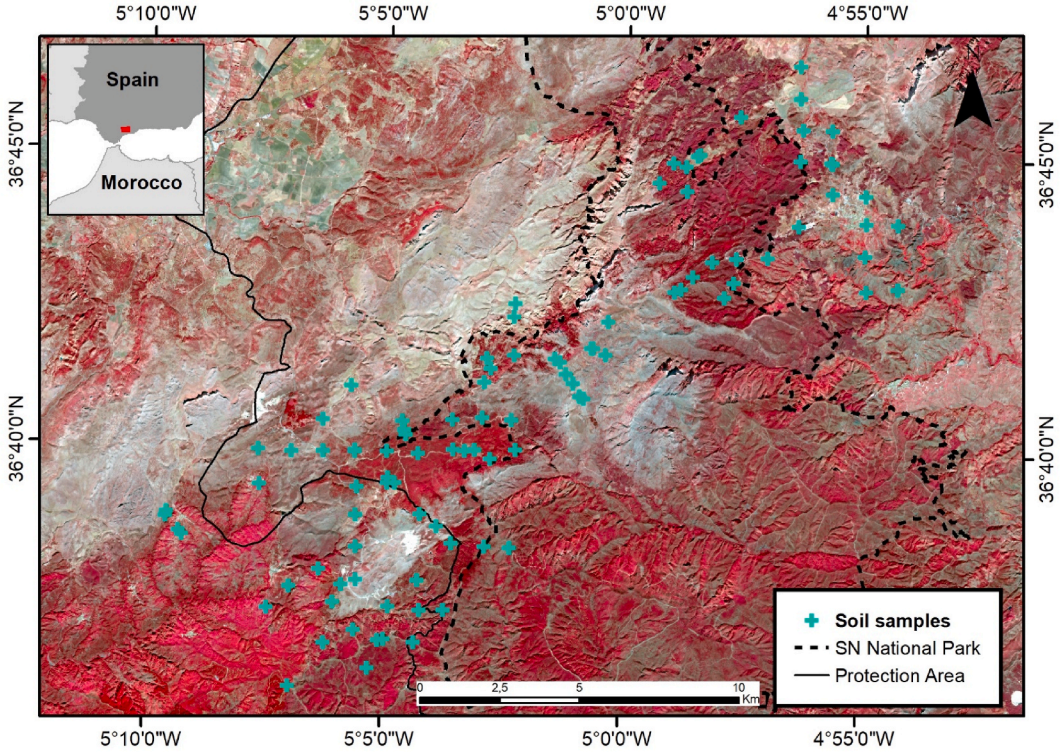


Fig. 1. Map showing the location of Sierra de las Nieves national park. CRS: WGS84. False colour composition of a Sentinel-2 image from October 2019 (near infrared, red and green bands). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Three soil properties were selected as target features: SOM, carbonates content and clay content (measured in percentage). The SOM percentage was determined using loss-on-ignition calcination method, measured as the percentage of weight loss of soil before and after burning away its organic matter (as per UNE-EN 13039:2012, "Soil improvers and growing media - Determination of organic matter content and ash"). The percentage of carbonates was extracted using the Bernard calcimeter method [69,70], which consisted of quantifying the percentage of CO_2 released when the sample was treated with hydrochloric acid. The percentage of clay was taken alongside the other two textural fractions, silt, and sand. Textural fractions were determined using the hydrometer method, also termed densimeter or Bouyoucos method [71]. A total of 100 measurements were taken for SOM and carbonates, and 99 measurements were taken for clay due to an error that occurred when measuring the textural fractions in one sample.

2.2.2. *Spectroscopic measurements*

Spectroscopic measurements were carried out using an ASD FieldSpec Pro JR spectrometer (Analytical Spectral Devices Inc., Boulder, CO, USA). This spectrometer can detect electromagnetic energy across three spectral ranges: the first is in the visible and near infrared range (VNIR, 350–1000 nm), and the latter two are in the shortwave infrared range (SWIR1, 1000–1800 nm) and (SWIR2, 1800–2500 nm). The sensor has a spectral resolution of 3 nm (@ 0.7 μm) and 30 nm (@ 1.4 μm, 2.1 μm), resampled to 1 nm. 250 g of the soil samples were placed into 10-cm Petri dishes. The light source was directed at an angle of incidence of 75° from the horizontal plane at a distance of 60 cm from the soil sample. Radiance was converted to reflectance using a Spectralon™ white reflectance panel with a reflectivity close to 100 %, and reflectance was recalibrated after every ten soil samples. Ten reflectivity samples were taken for each soil sample and the average was calculated using ViewSpecPro software. The ends of each spectrum (350–399 nm and 2451–2500 nm) were removed due to the noise generated by the spectroradiometer [72]. A total of 2051 predictor features (i.e. wavelength measurements) per spectrum were obtained for each sample. Spectroscopic measurements were taken before the samples were dried out and sieved so that spectral data would resemble data measured under natural conditions via remote sensing or field spectroscopy.

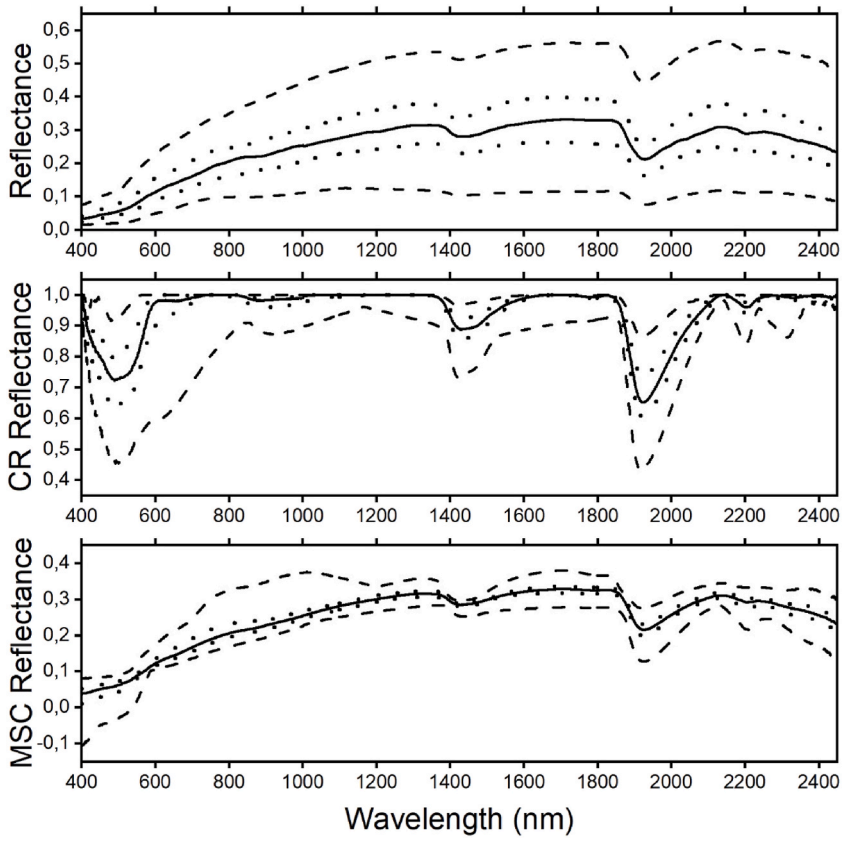


Fig. 2. a Median (solid line), minimum and maximum (dashed lines) and p25 and p75 values (dotted lines) by wavelength for raw (top), continuum removal-processed (centre) and multiplicative scatter correction-processed (bottom) spectra.
Fig. 2b. Spectra of the minimum, 25th percentile, average, 75th percentile and maximum sample value for SOM (above), carbonates (center) and clay (below). The intensity of the colour refers to position: the whitest spectra is the minimum, while the blackest one is the maximum. Solid line is the average, dotted line refers to 25th and 75th percentiles, and dashed lines refers to minimum and maximum.

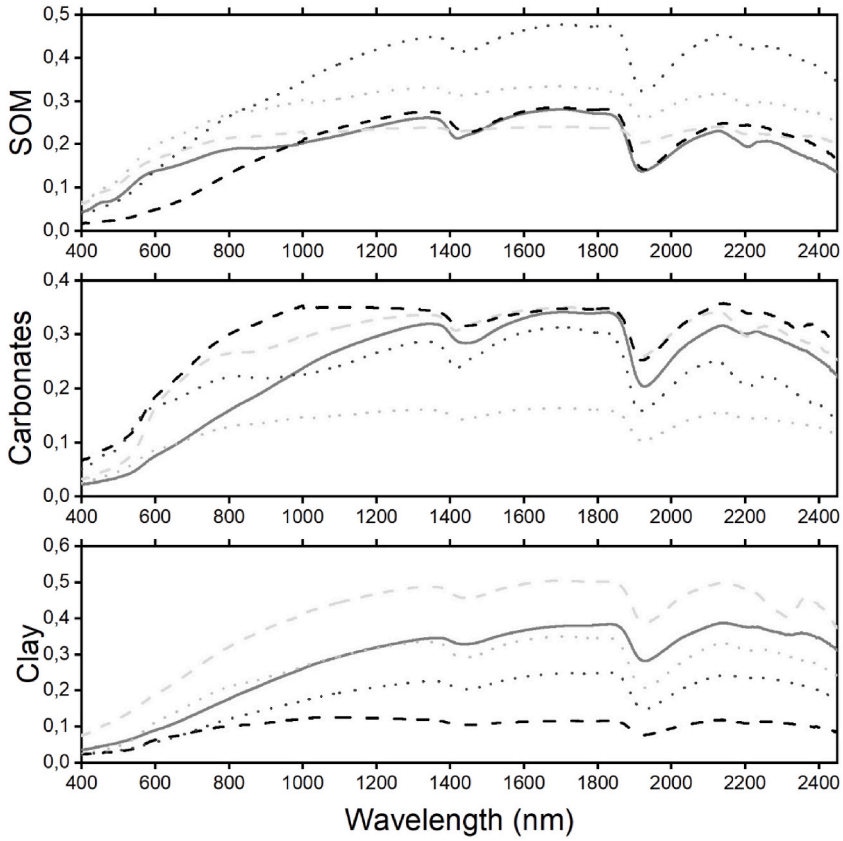


Fig. 2. (*continued*).

2.3. Methods

2.3.1. Spectral preprocessing

Two spectral preprocessing methods were used: Continuum Removal (CR) and Multiplicative Scatter Correction (MSC). CR normalises spectra in order to have a common baseline to compare individual absorption features [18] and it also enhances absorption features independent of observation scales and conditions [73]. MSC is a process that corrects the noise generated by physical conditions when measuring reflectance, reducing the internal variability of predictor features [14]. CR and MSC were applied using R [74] with the prospectr 0.2 and pls 2.7–3 packages [75]. Fig. 2a shows the median, maximum, minimum, and 25th and 75th percentile values by wavelength for three datasets: raw spectra, CR spectra and MSC spectra. Some negative values were found around 400–600 nm for MSC spectra, derived from the reference spectrum used in this study, the average spectra. Fig. 2b represents the spectra of the median, minimum, maximum 25th and 75th percentiles samples for each soil property. A bivariate correlations analysis was performed between each soil property and each wavelength (i.e. predictor feature), differentiating between datasets. A fourth dataset, "Global", was generated by combining the three individual datasets (raw, CR and MSC) containing 6153 predictor features (2051 features per dataset).

2.3.2. Modelling algorithms

PLSR is a two-step parametric regression algorithm combining a Feature Extraction method that first extracts latent features from the predictor, and then applies a multivariate linear regression using the latent features [75]. The feature extraction of PLSR performs an iteration using different projections of the predictor dataset to extract the scores and aims to optimise covariance between the extracted scores in feature extraction and the target feature. Targeting the covariance between predictor features and the target feature to improve predictive accuracy differentiates PLSR feature extraction from Principal Component Analysis [76]. PLSR is used for regression problems in disciplines where only a small number of observations are available with a higher dimensionality [77]. PLSR may be more effective than multiple linear regression and other parametric algorithms when a greater number of features are available compared to the number of observations. These datasets may also present multicollinearity [78] or a linear combination could exist between two features, which would make it impossible to use multiple linear regression and other linear algorithms for modelling. The PLSR algorithm was applied in R using the pls 2.7–3 package [79] with the following parameters: *plskernel* method, a maximum of 10 latent features and a 10-fold cross validation. The one-sigma algorithm was used to select the optimum number of latent features [79]. Variance Importance in Projection (VIP) scores [77] were used to assess feature importance in PLSR modelling.

Random Forest (RF) is a machine learning algorithm based on a decision tree ensemble [25,40]. RF is based on a bagging process (or bootstrapping aggregation): each individual tree of the ensemble is grown with different training data subsets that use a random selection of features and observations from the original dataset [42]. The terminal nodes of each individual tree in the RF have an associated simple regression model that applies to that node only. RF computes the output value by averaging the resulting value for all trees (as observed in Fig. 3). The subset composed of unused samples is called "out-of-bag" (OOB), which can be used by individual trees for evaluation purposes. By averaging the individual error of each tree, RF can compute an unbiased and internal estimation of the generalisation error [80].

The advantages of RF with regards to individual decision trees are largely a result of the bagging process [81], that is, its ability to handle complex data structures [82] and its relatively simple hyperparameter tuning process, only requiring two hyperparameters to be tuned: the number of trees (*ntree*) and the number of selected random features per tree (*mtry*). The strategy for hyperparameter optimisation in RF was carried out in two steps due to the application of feature selection. The hyperparameter *ntree* was optimised in a range of 100–2000 trees, at an interval of 100 trees. The second step consisted of applying the feature selection algorithms using the *ntree* number selected in the previous step for each soil property and dataset combination.

2.3.3. Feature selection

Two feature selection wrapper methods were selected, Sequential Forward Selection (SFS) and Sequential Flotant Forward Selection (SFFS). Feature selection wrapper algorithms select a relevant feature subset, evaluating prediction within a modelling algorithm [48]. The process can be outlined in three steps: i) establish an evaluation metric to serve as a feature selection criterion (i.e. Root Mean Square Error (RMSE)), ii) select a search algorithm to choose the order in which feature subsets are evaluated, and iii) train the model. There are several different search algorithms, such as an exhaustive search, genetic algorithms, random search, or sequential search, with the latter being the method used in this study. Sequential searching is defined by the iterative nature of the algorithm [83] and can be run in several ways. Sequential Forward Selection method starts with a set without features, and progressively adds features until an improvement in the accuracy of the models is no longer observed or until a specific number of features is reached. SFFS method works similarly to SFS, but when a subset is defined, a sequential backward selection is applied until the best subset of features is obtained, and SFS will begin again in the event this does not occur. A sequential search was chosen over other search algorithms because there was a better trade-off between performance and computational cost [50], selecting local optimum instead of global optimum. The alpha parameter (search detection threshold) was set at 0.001. RMSE was used as optimisation criteria, and the combination with the lowest RMSE was selected. Both steps were run in R using the mlr 2.17.1 package [84].

2.3.4. Model evaluation

Four performance measures were used: R-squared (R^2), RMSE, ratio of performance to deviation (RPD) and ratio of performance to interquartile distance (RPIQ) and the following equations were used for the metrics:

$$R^2 = 1 - \frac{RSS}{TSS}$$

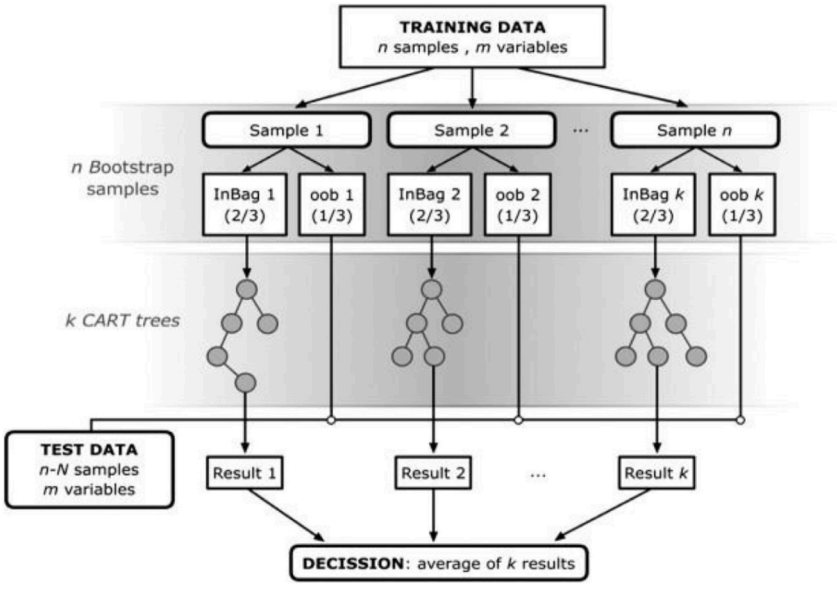


Fig. 3. RF algorithm diagram. CART=Classification And Regression Trees. Rodriguez-Galiano, Chica-Olmo [80].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - o_i)^2}{n}}$$

$$RPD = \frac{sd}{RMSE}$$

$$RPIQ = \frac{Q_3 - Q_1}{RMSE}$$

where: RSS = Residual sum of squares, TSS = Total sum of squares, n = number of observations, p_i = predicted ith value, o_i = observed value for the ith observation, sd = standard deviation and Q_3 and Q_1 correspond to third and first quartiles. It is worth mentioning that the possible values for R^2 range from negative infinity (worst result) to 1 (best result).

An internal model validation was used due to the limited number of samples (n = 100). PLSR used 10-fold cross validation. In Random Forest, RMSE was calculated by averaging the measures of error for each tree using out-of-bag data. The optimal hyper-parameters and feature subset were identified as those with a lower RMSE. RPIQ was also considered to evaluate the performance of models, as RPD is correlated to R^2 [85].

3. Results

3.1. Soil analysis and correlations with spectra

The boxplots of modelled soil properties are shown in Fig. 4. SOM showed a skewed distribution, with a mean value of 12.19 %, a range of 3.9–42.18 % and a standard deviation of 7.37, with a median value of 10.14 % and values of the first and third quantile of 7.87 % and 13.96 %. Carbonate samples showed a highly skewed distribution, due to the high number of samples with no carbonate content. Mean value of carbonates was 3.99 %, median of 0.25 %, a range of 0–69.6 %, a standard deviation of 9.96 % and an interquartile range of 2.325 %. Clay samples showed a normal distribution, with an average value of 28.55 % and a median of 26 %, a range of 2–64 % and a standard deviation of 13.77 %. First and third quantile of clay were 18 and 37 %. Pearson's correlation coefficients were −0.03 for SOM and carbonates, 0.39 for SOM and clay, and 0.27 for carbonates and clay.

Fig. 5 shows bivariate correlations between each soil property and predictor features (i.e. wavelength) by dataset. Correlation analysis showed higher correlations for SOM and carbonates in their main absorption features (visible region and 2350 nm), while clay had a low correlation with its main absorption feature (around 2200 nm). Moreover, most important correlations between each soil property and spectral data had negative values. The highest correlation for SOM was found around the red visible region (600–700 nm) for all three datasets, with negative values. SOM had its higher correlation in the visible region with the CR dataset than with the MSC and raw datasets, with correlation values greater than 0.7 (being a negative correlation). SOM also showed a high and negative correlation around 2150 and 2250 nm with the CR dataset. Carbonates only had a high and negative correlation around 2350 nm (main absorption feature of carbonates) with the CR dataset. Also, relevant correlations were found around 1200–1400 nm and 1500–1800 nm with the CR dataset, and around 800–1200 nm and 1400–1900 nm with the MSC dataset. Clay had a high and negative correlations with the red visible region of the raw dataset, and around 1400 and 1900 nm with the CR dataset. A high and positive correlation between MSC and clay was found around 1600–1800 nm with the CR dataset.

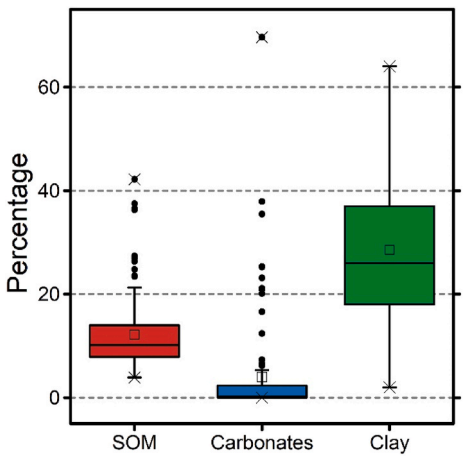


Fig. 4. Boxplot of modelled soil properties.

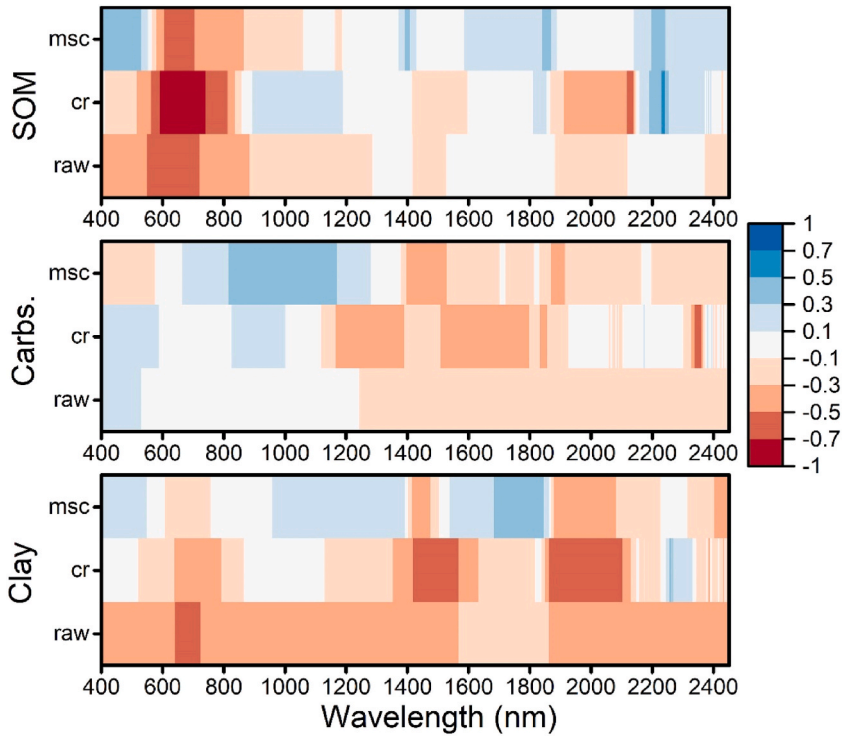


Fig. 5. Pearson's correlation coefficients between each wavelength and soil property by dataset. RAW = raw spectra, CR = continuum removal, MSC = multiplicative scatter correction.

3.2. Accuracy assessment of the models

The results of PLSR models are shown in Table 1. SOM models outperformed clay and carbonates models according to R^2 and RDP, independently of the dataset. However, clay models were more accurate according to RPIQ. SOM-Global was the most accurate model for SOM ($R^2 = 0.65$, RPIQ = 1.41), but with similar performance of the other datasets. PLSR models for carbonates showed low values for R^2 (0.12–0.32) and RPIQ (0.25–0.29). The most accurate carbonates model was Carbonates-CR ($R^2 = 0.32$, RPIQ = 0.29), but with minimal differences with respect to other predictions. Clay models had similar accuracy, with the R^2 and RPIQ values for Clay-PLSR models ranging between 0.33–0.43 and 1.69–1.83, respectively. The Global clay model was the most accurate model ($R^2 = 0.43$, RPIQ = 1.83).

Table 2 summarises modelling results with RF alone, RF with Sequential Forward Selection (RF-SFS) and RF with Sequential Flotant Forward Selection (RF-SFFS). The RF models with feature selection were more accurate than the RF alone models for all three soil properties. SOM models had a higher R^2 than clay and carbonates, but clay models showed higher RPIQ. SOM models had a R^2 range from 0.19 to 0.71. SOM models with RF and feature selection outperformed models with RF alone. RF-SFS models showed a RPIQ of between 1.26 and 1.93. being the Global dataset the model with the best performance ($R^2 = 0.7$). The accuracy of RF-SFFS models

Table 1
Performance results for PLSR models. LF = latent features, RAW = raw dataset, CR = continuum removal, MSC = multiplicative scatter correction.

Soil property	Spectral preprocessing	No. of LFs	R^2	RMSE	RPD	RPIQ
SOM	RAW	7	0.53	4.97	1.48	1.22
	CR	5	0.60	4.64	1.59	1.31
	MSC	7	0.51	5.12	1.44	1.19
	Global	8	0.65	4.3	1.71	1.41
Carbonates	RAW	4	0.16	9.07	1.09	0.26
	CR	8	0.32	8.14	1.22	0.29
	MSC	4	0.27	8.45	1.18	0.28
	Global	5	0.12	9.39	1.06	0.25
Clay	RAW	3	0.33	11.23	1.23	1.69
	CR	5	0.41	10.57	1.3	1.80
	MSC	5	0.39	10.74	1.29	1.77
	Global	4	0.43	10.37	1.33	1.83

Table 2
Performance results for Random Forest (RF), RF with Sequential Forward Selection (RF-SFS) and RF with Sequential Flotant Forward Selection (RF-SFFS). ntree = number of trees, raw = raw spectra, CR = continuum removal, MSC = multiplicative scatter correction.

		ntree	RF				RF-SFS				RF-SFFS			
			R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ
SOM	raw	1300	0.28	5.42	1.36	1.12	0.35	4.82	1.52	1.26	0.15	5.14	1.43	1.18
	CR	300	0.43	5.35	1.38	1.14	0.68	3.37	2.19	1.81	0.71	3.4	2.17	1.79
	MSC	100	0.37	4.7	1.57	1.29	0.54	4.09	1.8	1.49	0.44	4.15	1.77	1.47
	Global	400	0.19	4.96	1.48	1.23	0.7	3.15	2.33	1.93	0.34	3.47	2.12	1.75
Carbonates	raw	100	−23.4	11.97	0.83	0.19	−61.3	9.49	1.05	0.24	−6.18	9.48	1.05	0.25
	CR	200	−0.07	8.08	1.23	0.29	0.05	6.07	1.64	0.38	0.22	6.27	1.59	0.37
	MSC	400	−3.49	9.93	1	0.23	−6.1	7.12	1.39	0.33	−7.53	6.57	1.47	0.35
	Global	500	−2.2	8.81	1.12	0.26	0.48	6.42	1.55	0.36	0.13	6.34	1.57	0.37
Clay	raw	600	0.09	12.26	1.13	1.55	0.31	10.84	1.27	1.75	0.04	10.86	1.27	1.75
	CR	1300	0.12	11.21	1.23	1.69	0.37	9.08	1.52	2.09	0.45	9.73	1.41	1.95
	MSC	1400	0.21	10.8	1.28	1.76	0.34	9.67	1.42	1.96	0.55	8.79	1.57	2.16
	Global	800	0.34	10.49	1.31	1.81	0.66	7.86	1.76	2.42	0.66	7.41	1.86	2.56

varied between a R^2 of 0.15–0.71 and a RPIQ of 1.18–1.79. In contrast to RF-SFS, the best model was achieved with CR dataset (R^2: 0.71, RPIQ: 1.75). RF models for carbonates had a negative R^2 value, with a RPIQ range from 0.19 to 0.29. RF-SFS models had RPIQ values ranging from 0.24 to 0.38, and RF-SFFS models ranged between a RPIQ of 0.25–0.37. Carbonates-CR models with RF-SFS and RF-SFFS were the most accurate (RPIQ = 0.38 and 0.36, respectively), although the model built with Global dataset had a better R^2 value (0.48). Clay models with feature selection outperformed RF alone models; RF-SFS models had a RPIQ between 1.75 and 2.42 and RF-SFFS between 1.75 and 2.56, while RF models had a RPIQ between 1.55 and 1.81. Clay models built with Global dataset were more accurate than the other models (RPIQ of 2.42 and 2.56 for RF-SFS and RF-SFFS models respectively).

RF with feature selection outperformed not only RF but also PLSR models. However, PLSR models had a similar or even slightly better performance than RF models. The best R^2 value was found in CR-RF-SFFS for SOM (0.71), Global-RF-SFS for carbonates, 0.48, and both Global models for clay, 0.66. Raw models had a similar accuracy for all the modelling algorithms, while the accuracy of models by preprocessed spectra (CR, MSC and Global) was increased using RF built with SFS and SFFS in comparison with PLSR and RF alone. For all the modelling algorithms it was found a higher R^2 for SOM, but a higher RPIQ for clay models.

Fig. 6 shows scatterplots of observed and predicted values for the best PLSR and RF models by soil property. RF plots were less scattered (closer to the 1:1 line) due to their superior accuracy compared to PLSR plots. SOM plots showed less scatter in RF in the 0–20 % SOM range, which is where most observations were found. The best Carbonates-PLSR model (with the CR dataset) predicted negative values in some samples that were out of range (0–100 %), while the RF model (CR-RF-SFS) was able to set that constraint. The clay RF model showed more samples closer to the 1:1 line (better prediction), and all models underestimated high values and overestimated low values.

3.3. Feature evaluation in PLSR results

Figs. 7 and 8 show the results of the feature extraction of PLSR, plotting the VIP scores using the most accurate PLSR models for the three soil properties: the Global dataset for SOM and clay, and the CR dataset for carbonates.

Fig. 7 showed the importance of CR dataset over raw and MSC datasets in global models using PLSR, given that the features of that dataset were the most important in SOM and clay models. The SOM model was built using the first eight latent features, comprising 98.7 % of variance (Table 3). The highest VIP scores were found in the visible region of CR dataset, especially around 680 nm (red region), with values over 3. This region was also highlighted in raw dataset, with the highest values within the raw features, whereas no region of MSC dataset was found to have a VIP score higher than 1. The clay global model had 4 extracted features achieving a 92.11 % of variance explained, being more parsimonious than SOM and carbonates models. The higher VIP scores were found around 2000 nm region in CR dataset, with values higher than 3, and in the visible region, with values over 2. The raw dataset had a higher average VIP value than MSC data, but no region could be highlighted as important. The carbonates model explained the 98.32 % of variance using the first 8 latent features. The highest VIP scores for carbonates (see Fig. 8) were found around the 2350 nm absorption features, related to carbonates content. Other relevant spectral regions according to VIP scores were the visible region around the 450 nm and the 2000 nm region.

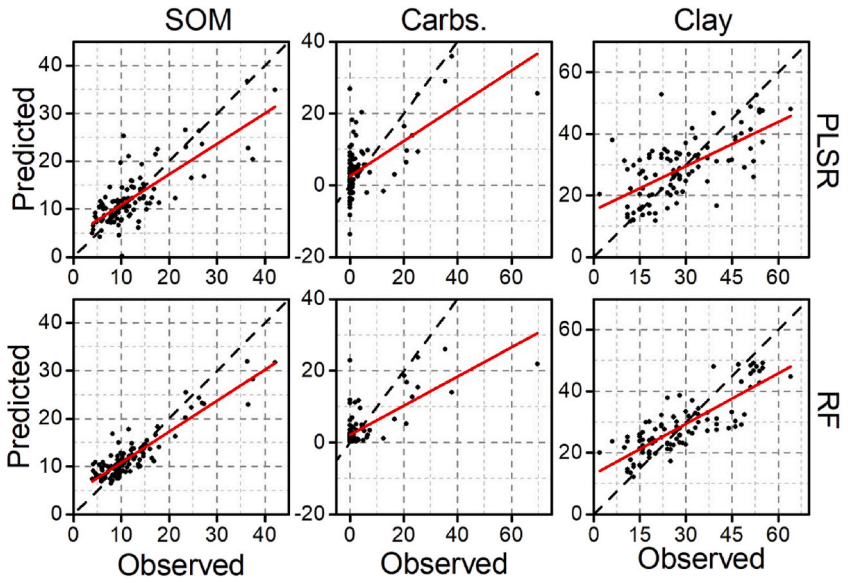


Fig. 6. Scatterplots of observed and predicted values for the best soil property models by modelling algorithm. Top, PLSR models: SOM-Global (left), Carbonates-CR (centre) and Clay-Global (right); Bottom, RF models: SOM-Global-RF-SFS (left), Carbonates-CR-RF-SFS (centre) and Clay-Global-RF-SFFS (right). Red line = line of best fit. Black dashed line = 1:1 plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

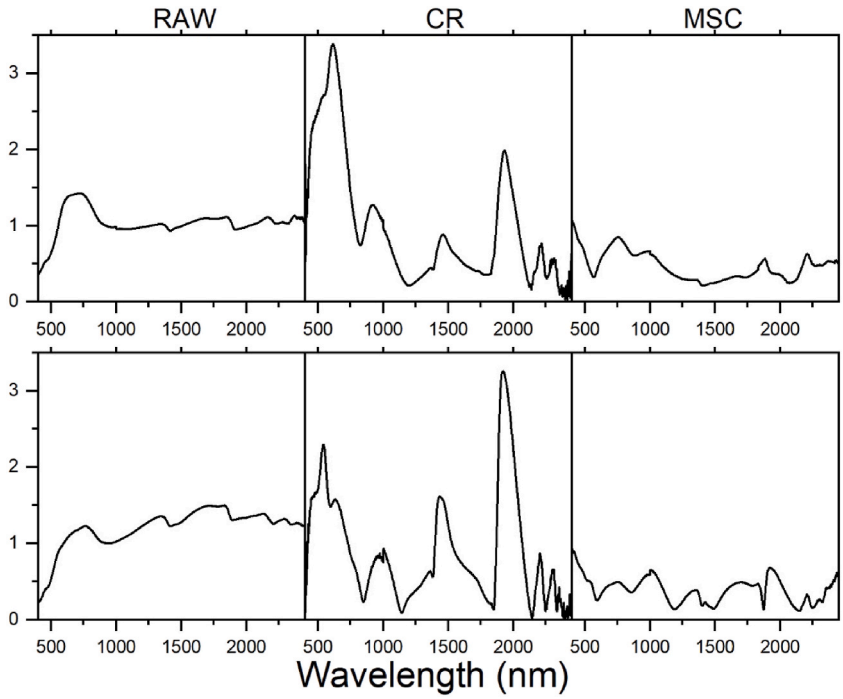


Fig. 7. VIP scores of the global models for SOM (top) and clay (bottom).

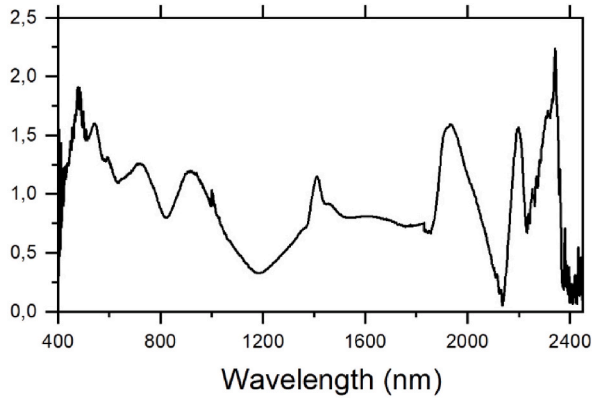


Fig. 8. VIP scores of the CR model for carbonates.

Table 3
Cumulative variance explained by each latent feature for the best PLSR model by soil property. Values in bold = variance explained by the predictor dataset used for modelling (number of latent features selected). LF = Latent feature.

	1 LF	2 LF	3 LF	4 LF	5 LF	6 LF	7 LF	8 LF
SOM-Global (%)	40.24	80.33	85.00	87.97	91.76	97.38	98.29	**98.7**
Carbonates-CR (%)	34.74	68.08	82.57	94.35	95.13	96.70	97.74	**98.32**
Clay-Global (%)	64.22	78.95	88.71	**92.11**	95.36	97.22	98.41	98.58

3.4. Feature selection

The application of Sequential Forward Selection (SFS) and Sequential Flotant Forward Selection (SFFS) methods reduced the number of features in modelling by 99.41–99.86 % for SOM, 99.56–99.9 % for carbonates and 99.6–99.93 % for clay, compared to the number of starting features. Selected features for RF with feature selection models with raw, CR and MSC datasets are shown in Fig. 9. Features were selected in different spectral regions by SOM models, i.e. in the visible region, around 1100 nm; and in the SWIR region,

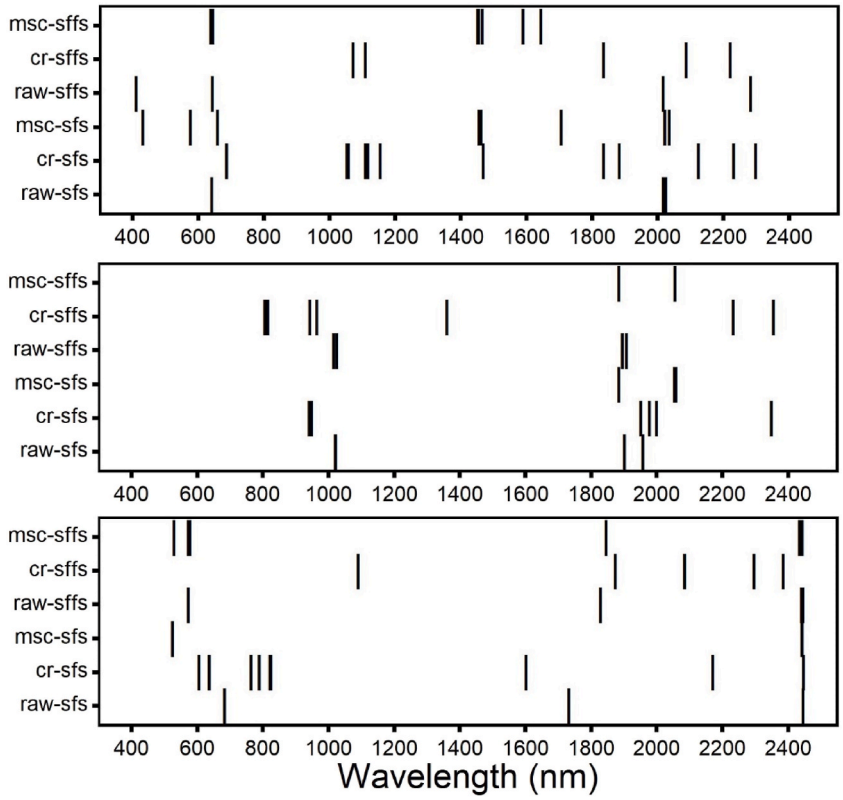


Fig. 9. Selected features in individual models by soil property (soil organic matter: top, Carbonates: centre, Clay: bottom), by dataset and feature selection method combination (Y axis). SFS = Sequential Forward Selection, SFFS = Sequential Flotant Forward Selection, CR = Continuum Removal, MSC = Multiplicative Scatter Correction.

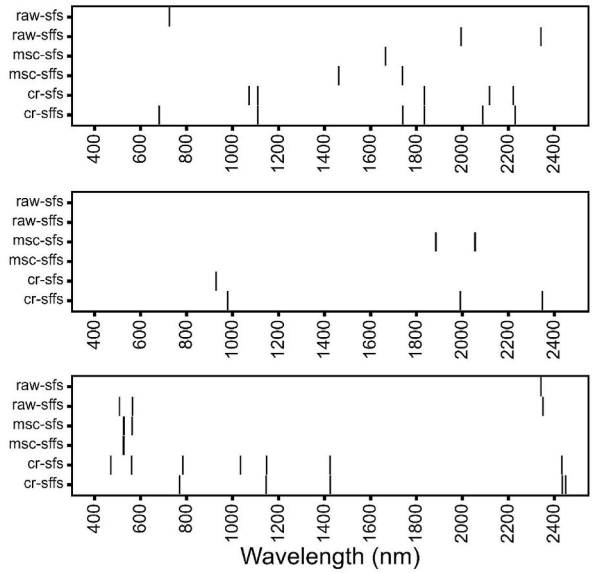


Fig. 10. Selected features in global models by soil property (soil organic matter: top, Carbonates: centre, Clay: bottom), by dataset and feature selection method combination (Y axis). SFS = Sequential Forward Selection, SFFS = Sequential Flotant Forward Selection, CR = Continuum Removal, MSC = Multiplicative Scatter Correction.

with hardly any differences between SFS and SFFS for the datasets. The raw and MSC models selected more features in the visible region than the CR-RF-SFS model (one feature) and CR-RF-SFFS model (zero features). However, only the latter models selected features around 1100 nm. The features selected in carbonates models were similar, concentrated around 800–1000 nm and 1900 nm. Features were selected between 800 and 1000 nm for raw models (around 1000 nm) and CR models (around 800 nm with SFFS and 950 nm with both algorithms) and around 1900–2050 nm for all models except CR-RF-SFFS. Models with CR dataset were the only models where features were selected in other regions: 1361 and 2233 nm by CR-SFFS, and 2350 nm by both models for the main absorption feature of carbonates. Clay models selected features mainly around 500–800 nm and 1600–2450 nm, showing slight differences between the feature selection algorithms. CR dataset models selected features in the NIR region: CR-RF-SFS selected some features around 800 nm, and CR-RF-SFFS selected one feature near 1100 nm. All models selected features in SWIR, but only the CR-RF-SFS model selected a feature at 2170 nm near the absorption feature of clay around 2200 nm. All models selected features around 2450 nm, which could be a key feature.

Fig. 10 shows the features selected by RF with feature selection models for the three soil properties in Global models. There were few differences in selected features by feature selection method and by soil property. The models selected 7 features for SOM, 5 for carbonates and 12 for clay using SFS method. Models with SFFS method selected 10 features for SOM, 4 for carbonates and 11 for clay out of the 6153 total starting features in the Global models. The SOM models selected features in the red visible region, around 1100 nm and 1650–2350 nm. Selected features primarily coincided with features from the CR dataset for both algorithms. Features selected in clay models were similar for both SFS and SFFS, with selected features in the visible and NIR regions, and 2350–2450 nm. For the visible region, features were selected from the CR and MSC datasets with RF-SFS, and from the raw and MSC datasets with RF-SFFS. All features were selected from the CR dataset in the NIR and SWIR regions, except 2342 and 2431 nm (RF-SFS with the raw dataset) and 2350 nm (RF-SFFS with the raw dataset). For carbonates models, features were selected around NIR (900–1000 nm), 1900 nm and carbonates-related absorption feature (2350 nm) by Global-RF-SFFS. All features were selected from the CR dataset, except 1885, 2054 and 2056 nm, which were selected by RF-SFS from the MSC dataset. The Carbonates-Global-SFS model did not select a feature in carbonates-related absorption feature.

There were differences between the features that were selected in individual and Global models for each soil property. SOM selected more features in individual models, coinciding with the selected regions and with similar accuracy observed between models using the CR and Global datasets. Clay-Global models were more accurate than individual models with less selected features. NIR features were selected in the Global model (CR dataset), and only Clay-CR-SFFS selected a feature around that spectral region. Carbonates RF-FS models selected features in a similar region, but the Global model selected less features than individual models.

4. Discussion

4.1. Accuracy assessment

4.1.1. Feature selection models achieved an improved modelling performance

The results of this study suggest that the accuracy of the modelling of SOM, carbonates, and clay with soil spectroscopy could be improved by combining modelling algorithms and feature selection wrapper methods: RF with feature selection models performed better than RF alone and PLSR models. This may be due to the fact that feature selection methods select the subset of key features in datasets with high dimensionality, thereby increasing accuracy by avoiding the Hughes effect [43,86]. Moreover, results suggested that SFS and SFFS not only improved the accuracy of modelling, but also its parsimony, that is, improve the accuracy of models while making them simpler. According to Xu, Hong [87], this aspect has been little studied in the literature. However, the emerging of new complex approaches such as Feature Selection will contribute to further understanding of models and underlying phenomena. Feature selection methods for soil properties modelling using spectroscopy are relatively unexplored in the literature. Some works, such as Hong, Chen [88], Vohland, Ludwig [89] and Vohland, Ludwig [90] reported an improved accuracy when feature selection was applied. Hong, Chen [88] modelled soil organic carbon (SOC) from croplands in Iowa, USA using RF alone and RF with two feature selection methods, i.e. continuous wavelet transform (RF-CWT) and competitive adaptive reweighted sampling (RF-CARS). They observed improved performance in the RF-CWT model (RMSE = 0.151) over the RF model (RMSE = 0.183). Vohland, Ludwig [89] used PLSR and PLSR with CARS to model SOC and other soil properties of croplands in Germany using lab-based spectroscopy (vis-NIR and MIR), achieving an accuracy of RPD = 1.58 with PLSR, while the PLSR-CARS model had a RPD of 1.98 — an improvement of 25.31 %. In this same study area, Vohland, Ludwig [90] used PLSR with CARS, a generic algorithm and "iteratively retaining informative variables" to model SOC and other soil properties. This was done by comparing HyMap and laboratory-based spectroscopy (resampled to HyMap spectral resolution), which improved the accuracy for SOC prediction with laboratory-based soil spectroscopy (a RPD increase from 2.36 to 3.08). The previous studies showed an improved performance of machine learning algorithms with feature selection, regardless of the feature selection method used. Therefore, it is advised to use any feature selection method when modelling with ML in high-dimensionality spaces, because the multicollinearity and Hughes effect would lead to a worsen performance.

The better performance of RF with feature selection compared to PLSR could be due to the selection being replaced with a feature extraction prior to linear regression modelling being performed, as features with more noise could be introduced during feature extraction [45]. However, using feature extraction in highly dimensional spaces, even when linear algorithms are subsequently used, may yield results that are similar to or even more accurate than those obtained using non-linear algorithms based on ML without dimensionality reduction. This was the case in this study when comparing results from PLSR and RF alone (PLSR: RPIQ = 1.41, 0.29 and 1.83 for SOM, carbonates and clay, respectively; RF: 1.29, 0.29 and 1.81 for SOM, carbonates and clay, respectively). Moura-Bueno, Dalmolin [91] reported similar results when modelling SOC with soil spectroscopy using PLSR (R^2 = 0.74, RMSE% =

0.56) and RF alone (R^2 = 0.72, RMSE% = 0.56). Knox, Grunwald [92] compared PLSR and RF alone to model soil carbon fractions using the vis-NIR and mid-infrared regions, with a RF performance that was slightly better than the PLSR model. Castaldi, Hueni [93] evaluated Sentinel 2 and airborne imaging spectroscopy to estimate SOC in croplands using PLSR and RF. They observed that PLSR had an accuracy that was similar to or better than RF for all study areas except one.

Different results were obtained for the soil properties according to the performance measure considered. Predictive modelling for SOM yielded better results than for clay or carbonates using R^2 and RPD, being these two performance measures correlated. Using RPIQ instead, clay models yielded better performance than SOM and carbonates. The reported difference might depend on how performance is measured rather than the predictive ability of the models themselves. RPD uses standard deviation to normalize RMSE to make results comparable between different target features (i.e., soil properties). However, standard deviation does not describe of the spread of population when the population is skewed. Bellon-Maurel, Fernandez-Ahumada [94] proposed RPIQ to measure predictive capacity in relation with the data spread, regardless of the distribution, by using interquartile range instead of standard deviation. Therefore, as clay showed a gaussian distribution and SOM and carbonates did not, clay models are more accurate than SOM and carbonates models when considering RPIQ.

Difference in performance results by soil properties did not appear to be linked to soil properties themselves [95], with disparate results being reported in the bibliography. Sierra de las Nieves has heterogeneous landcover and geology. It also has a complex forest ecosystem with a wide range of SOM with values between 3.9 and 42 % (Table 1), and a high value could mask absorption features in soil spectra [96]. Furthermore, the amount of humidity detected in the samples during experimental measurements could have a negative effect on soil properties estimation [97], particularly in the case of clay and carbonates. Volkan Bilgili, van Es [98] worked with various soil properties applying multivariate adaptive regression splines and PLSR in a semiarid region of Turkey, with better results for clay (RPD = 3.08) than for SOM (RPD = 1.94) and carbonates (RPD = 1.93). Ostovari, Ghorbani-Dashtaki [95] used PLSR to model SOM and carbonates in a semiarid region of Iran, with similar validation results: RMSE = 0.30 and RPD = 1.6 for SOM, and RMSE = 5.24 and RPD = 1.6 for carbonates. Overall, these results suggest that difference in results by soil property depended on several geographical, measurement and methodological factors.

4.1.2. Evaluation of preprocessing methods in performance

The combination of spectral preprocessing methods in the Global dataset only led to accuracy improvement in clay modelling. SOM reached similar accuracy with the Global dataset and the CR dataset, and carbonates models with the CR dataset outperformed those with Global dataset. Therefore, different preprocessing methods should be tested [99], as no consensus could been reached on which spectral preprocessing method is the most appropriate [100]. Volkan Bilgili, van Es [98] found similar modelling results using raw

Table 4
Summary of studies that apply feature selection methods in soil spectroscopy, the number of starting features and the number of selected features. Only soil properties that were similar to those in this study were selected.

Source	Feature selection method	Modelling algorithm	Soil property	Number of starting features	Number of selected features	% selected features
Hong, Chen [88]	CWT	RF	SOC	296	21	7.09
	CARS	RF	SOC		28	9.46
Vohland, Ludwig [90]	CARS	PLSR	SOC	125	15	12
	GA	PLSR	SOC		19	15.2
	IRIV	PLSR	SOC		11	8.8
Vohland, Ludwig [89]	CARS	PLSR	SOC	411	10–47	2.43–11.47
Shi, Chen [102]	SPA		SOC	681 (Yixing)	62 (Yixing)	9.10
				511 (Honghu)	58 (Honghu)	11.35
	GA	PLSR	SOC		145 (Yixing)	21.29
					96 (Honghu)	18.79
Raj, Chakraborty [53]	AMI		TC (SOC)	2150	221	10.28
	AMI		Clay		62	2.88
Gomez and Coulouma [55] *	VIP + beta coefficients	(PLSR)	Clay	2051	336	16.36
	VIP + beta coefficients		Carbs		174	8.48
Adeline, Gomez [103] *	VIP + beta coefficients	(PLSR)	Clay	1961	254	12.95
	VIP + beta coefficients		Carbs		103	5.25
Viscarra Rossel and Behrens [24]	VIP	(PLSR)	SOC	876	29	3.31
	MARS				14	1.6
	VIP	(PLSR)	Clay		31	3.54
	MARS				13	1.48
Wang, Qiao [104]	SPA		SOC	Not reported	13	
	SPA		Clay		10	

CWT: Continuous Wavelet Transform. CARS: Competitive Adaptive Reweighted Sampling. GA: Genetic Algorithm. IRIV: Iteratively Retains Informative Variables. SPA: Successive Projection Algorithm. AMI: mutual information based adjacency. VIP: Variance Importance Projection. MARS: multivariate adaptive regression splines. RF: Random Forest. PLSR: Partial Least Squares Regression. SOC: Soil Organic Carbon. TC: Total Carbon. Carbs: Carbonates. Studies with * only used feature selection for interpretation purposes and not for modelling..

spectra, first derivative of spectra, and the combination for clay, SOM, carbonates and other soil properties in a semi-arid area of Turkey. Tiecher, Moura-Bueno [100] combined raw spectra with various spectral preprocessing methods to model sediment sources (first derivative with Savitzky-Golay, second derivative with Savitzky-Golay, Standard Normal Variate, MSC and normalisation), and two combinations of the latter two methods: MSC with first derivative and Savitzsky-Golay and normalisation with first derivative with Savitzky-Golay. The best result was obtained using first derivative with Savitzky-Golay, followed by a combined dataset (normalisation with first derivative with Savitzky-Golay).

4.2. Dimensionality reduction and analysis of selected features

Feature selection wrapper methods achieved a noticeable reduction in dimensionality, with the total number of selected features representing less than 1 % of the starting features for all cases. The nature and number of selected features depended on the feature selection method that was used, i.e. Sequential Forward Selection (SFS) or Sequential Flotant Forward Selection (SFFS), with both methods beginning with an empty set and gradually adding key features [48]. Feature selection wrapper methods selected a lower number of features because the predictive features (i.e. spectral data) may display multicollinearity [101]. The number of selected features in our study, in relative terms, was less than other studies that applied feature selection methods to predict soil properties using spectroscopy (see Table 4). In those studies, the number of selected features ranged between 1.48 and 21.29 %, with no feature selection method appearing to be better than others in terms of reducing dimensionality.

The subset of features selected for SOM included features from various spectral regions. SOM comprises several chemical compounds in its composition, including carbohydrates, lignin, or cellulose [105]. Those compounds had their own absorption features related to their chemical structure (e.g., NH, CH, or CO), explaining the selection of features along all spectral regions [106]. To illustrate this, the most accurate model (Global-RF-SFS) selected features at 1109, 1741 and 2118 nm, which may have been selected due to the presence of aromatic compounds (absorption feature at 1109 nm), alkyl compounds (1754 nm) and polysaccharide compounds (2137 nm) [24]. In general, feature selection methods highlighted features around 1100 nm instead of in the visible region in most accurate models. Conversely, SOM had a higher correlation with the visible region than with the 1100 nm region (see Fig. 5), which may be suggestive of a non-linear association. These results are not in line with other studies that applied feature selection methods to predict SOM with soil spectroscopy [53,89,90,102,104], where the visible region was most significant for SOM prediction. However, none of those studies applied a feature selection wrapper method build with RF or another type of non-linear ML algorithm. Wang, Qiao [104] used a feature selection filter method, Successive Projection Algorithm, to identify key features for SOM and selected the following wavelengths: 410, 450, 550, 625, 780, 850, 1410, 1670, 1730, 1860, 1910, 1960 and 2250 nm. Raj, Chakraborty [53] used different feature selection methods for modelling total carbon in Romania based on airborne imaging spectroscopy using PLSR and SVM. The best feature selection method was mutual information based adjacency, which selected features primarily in the visible and SWIR regions, choosing a subset representing 10.28 % of the starting features. Vohland, Ludwig [89] applied PLSR with Competitive Adaptive Reweighted Sampling, selecting between 10 and 17 out of 411 starting features (2.43–11.47 %) primarily associated with the visible region around 1995 and 2200 nm. Vohland, Ludwig [90] used PLSR with Competitive Adaptive Reweighted Sampling, genetic algorithms and "iteratively retains informative variables" with the three feature selection methods selecting between 8.8 and 15.2 % of the starting features in the visible and SWIR regions. Shi, Chen [102] used PLSR with a Successive Projection Algorithm and a genetic algorithm to model SOC in mixed croplands in China. The Successive Projection Algorithm selected features primarily in the visible region and between 2100 and 2400 nm. However, the minimum number of features selected was 58 (11.35 % of the total), which is much higher than in this study.

Carbonates were shown to be more dependent on their main absorption feature (around 2350 nm) than other soil properties [107]. The only strong correlation between carbonates and spectra was found in their main absorption feature. Moreover, features around 2350 nm were only selected by models with improved accuracy. This absorption feature is related to the overtone of CO_3 bound around 2336 nm [24], related to the $CaCO_3$ structure of carbonates. Thus, the importance of 2350 nm in predicting carbonates has been reported in some studies [55,103]. Gomez and Coulouma [55] used Variable Importance in Projection and PLSR beta coefficients to identify key features in predicting carbonates, identifying 174 key features (8.48 % of the starting features). Most key features were identified around the absorption feature of carbonates. Adeline, Gomez [103] combined Variable Importance in Projection and beta coefficients to evaluate the relevance of features in modelling carbonates and other soil properties. That study identified 103 features (5.25 %), underpinning the importance of the absorption feature of carbonates. Both studies also identified the visible region as being important. In our study, features were also selected around 900 and 1950 nm outside the visible region. The selection of features around 900 and 1950 nm may be related to spurious relationships or due to a CO-related absorption feature at 1998 nm [108].

The features selected for clay corresponded to several spectral regions, given that clay prediction with soil spectroscopy has been shown to not only be dependent of its main absorption feature, but also affected by the general shape of the spectrum and correlations with other soil properties [73]. That is why it may be possible to model clay based on the absorption features of iron minerals or SOM in the visible region. The influence that clay has on SOM storage could explain its bivariate correlation (0.39) [109,110]. Other studies, such as Adeline, Gomez [103], stated that key features located in the visible region were the result of iron minerals being present, with correlations of 0.53 with clay. Global models selected features around 1100, 1400, 2350 and 2450 nm in the NIR and SWIR regions. These features can be associated with different minerals within clay, such as kaolinite (associated absorption feature at 1415 nm, related to OH bond) and illite, with two related absorption features at 2340 and 2450 nm [24], associated with the OH bonds in the illite chemical structure. Nevertheless, the 2450 nm absorption feature is considered to be poorly defined and might be derived from instrumental noise. Only a single model selected one feature around 2200 nm (CR-RF-SFS at 2170 nm), which is considered to be the main absorption feature of clay [8]. This absorption feature is associated with the presence of an overtone of AlOH as well as the

presence of various minerals in clay, such as smectite and kaolinite. Other studies have selected key features in both the main absorption feature of clay and other related absorption features [24,104,111]. Coblinski, Giasson [111] used Cubist, a decision-tree-based algorithm with an embedded feature selection method, to estimate textural fractions (clay, silt, and sand) in southern Brazil. The wavelengths 1415, 2200 and 2480 nm were significant in the modelling process, with 2480 nm having greater relative importance. Wang, Qiao [104] used Successive Projection Algorithm to select key features for clay prediction, which selected 10 features, and 3 of the features (1410, 2250 and 2400 nm) could be associated with both the main absorption feature of clay and with the absorption features selected in our study. Viscarra Rossel and Behrens [24] used two feature selection methods (Variable Importance in Projection and multivariate adaptive regression splines), selecting 31 and 13 key features, respectively. Both algorithms selected features near 1395 nm (absorption feature of kaolinite), 2172, 2212 and 2228 nm (close to the main absorption feature) and 2432 nm (absorption feature of illite).

Perhaps the most controversial point of feature selection may be the fact that various plausible models made of different features can be obtained with a similar level of performance. This is known in the literature as the Rashomon effect or the multiplicity of good models [25] and was especially noticeable in our case because spectroscopy data is characterised by its multicollinearity, particularly when raw spectra of less than 1 nm are used. Our results have yielded various feature combinations that enable predictive modelling of each soil property. Furthermore, small disturbances in training data, or during wrapper configuration, e.g., sequential searching strategies or the prediction algorithm, yield different feature subsets. However, this should be viewed merely as several different possibilities instead of as a disadvantage per se. This phenomenon is not exclusive to feature selection models with ML; it also occurs with other commonly used techniques in soil spectroscopy studies, such as PLSR, or even in multivariate linear regression.

5. Conclusions

This study evaluated the performance of two feature selection wrapper methods, Random Forest with Sequential Forward Selection (SFS) and with Sequential Flotant Forward Selection (SFFS) for reducing dimensionality and modelling three soil properties with soil spectroscopy (SOM, clay, and carbonates). RF models with feature selection outperformed PLSR and RF alone in predicting all soil properties, thus avoiding the Hughes effect. The best RF model built with any feature selection method had a RPIQ values of 1.93 for SOM (RF–SFS–Global), 0.38 for carbonates (RF–SFS–CR) and 2.56 for clay (RF–SFFS–Global). PLSR-based models had a RPIQ of 1.41, 0.29 and 1.81, and RF alone models had a RPIQ of 1.29, 0.29 and 1.81 for SOM, carbonates, and clay, respectively. Feature selection wrapper methods were reported to be capable of handling the high dimensionality and multicollinearity of spectral data and selected less than 1 % of the original starting features. The features that were selected varied depending on the soil properties. The features selected for SOM were found throughout the spectrum, but feature selection highlighted the importance of features around 1100 nm. The main absorption feature of carbonates (2350 nm) was found to be crucial in applying feature selection to predict carbonates and was selected in models with improved accuracy. The features selected for clay were found throughout the spectrum due to several factors, such as the absorption features associated with clay (1415, 2200 and 2450 nm) as well as other absorption features that correspond to other soil properties and which clay also correlates with. However, the main absorption feature of clay (2200 nm) was only selected by one model with feature selection. This study provides new insights into soil properties modelling with spectroscopy in combination with different spectral preprocessing and feature selection methods. Combining different dataset (raw, Continuum Removal and Multiplicative Scatter Correction spectra) in a Global dataset resulted in improved accuracy for clay modelling, but not for SOM and carbonates modelling. Feature selection wrapper methods have shown a promising potential for the modelling of soil properties with spectroscopy. This opens an interesting field of research in comparing different feature selection wrapper methods (exhaustive, random search, genetic algorithm) and machine learning algorithms (Support vector machines, artificial neural networks, gradient boosting trees or convolutional neural networks) using spectroscopic data. These algorithms could be effective in reducing highly dimensional data from laboratory, airborne and spaceborne spectroscopy.

Data availability statement

The data used in this study can be found in 10.5281/zenodo.10222011.

CRediT authorship contribution statement

Francisco M. Canero: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Victor Rodriguez-Galiano:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **David Aragones:** Writing – review & editing, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The first author is a FPU grant holder funded by the Spanish Ministerio de Universidades (Reference FPU18/04274). The authors would like to express their gratitude for the financial support provided under projects 2914/2022, funded by the Ministerio para la Transición Ecológica y el Reto Demográfico, and AGRARIA (MIA.2021.M01.004), funded by the Ministerio de Asuntos Económicos y Transformación Digital (Subdirección General de Inteligencia Artificial y Tecnologías Habilitadoras Digitales). We acknowledge Mr. Aragones and ICTS-RBD for providing logistic support, which falls under CSIC-PTI TELEDETECT activities.

References

[1] J.M. Hollas, Modern Spectroscopy, John Wiley & Sons, 2004.
[2] F. Riedel, et al., Prediction of soil parameters using the spectral range between 350 and 15,000 nm: a case study based on the Permanent Soil Monitoring Program in Saxony, Germany, Geoderma 315 (2018) 188–198.
[3] X.Q. Xia, et al., Reflectance spectroscopy study of Cd contamination in the sediments of the changjiang river, China, Environ. Sci. Technol. 41 (10) (2007) 3449–3454.
[4] C.-W. Chang, D.A. Laird, Near-infrared reflectance spectroscopic analysis of soil C and N, Soil Sci. 167 (2) (2002) 110–116.
[5] G.W. McCarty, et al., Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement, Soil Sci. Soc. Am. J. 66 (2) (2002) 640–646.
[6] R.A. Viscarra Rossel, et al., Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties, Geoderma 131 (1–2) (2006) 59–75.
[7] C. Gomez, P. Lagacherie, Mapping of Primary Soil Properties Using Optical Visible and Near Infrared (Vis-NIR) Remote Sensing (2016) 1–35.
[8] S. Chabrillat, et al., Use of hyperspectral images in the identification and mapping of expansive clay soils and the role of spatial resolution, Rem. Sens. Environ. 82 (2–3) (2002) 431–445.
[9] P. Lagacherie, et al., Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements, Rem. Sens. Environ. 112 (3) (2008) 825–835.
[10] S. Chabrillat, et al., Imaging spectroscopy for soil mapping and monitoring, Surv. Geophys. 40 (3) (2019) 361–399.
[11] S. Diek, et al., Minimizing soil moisture variations in multi-temporal airborne imaging spectrometer data for digital soil mapping, Geoderma 337 (2019) 607–621.
[12] B. Somers, et al., Modelling moisture-induced soil reflectance changes in cultivated sandy soils: a case study in citrus orchards, Eur. J. Soil Sci. 61 (6) (2010) 1091–1105.
[13] D.B. Lobell, G.P. Asner, Moisture effects on soil reflectance, Soil Sci. Soc. Am. J. 66 (3) (2002) 722–727.
[14] Å. Rinnan, F.v.d. Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC, Trends Anal. Chem. 28 (10) (2009) 1201–1222.
[15] A. Gobrecht, J.-M. Roger, V. Bellon-Maurel, Major Issues of Diffuse Reflectance NIR Spectroscopy in the Specific Context of Soil Carbon Content Estimation, 2014, pp. 145–175.
[16] S. Xu, et al., Integrating hyperspectral imaging with machine learning techniques for the high-resolution mapping of soil nitrogen fractions in soil profiles, Sci. Total Environ. 754 (2021) 142135.
[17] A. Gholizadeh, et al., Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features, Soil Water Res. 10 (4) (2016) 218–227.
[18] R. Clark, T. Roush, Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications, J. Geophys. Res. 89 (B7) (1984) 6329–6340.
[19] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Appl. Spectrosc. 42 (7) (1988) 1273–1284.
[20] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (5) (1989) 772–777.
[21] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2) (2001) 109–130.
[22] X. Yu, et al., Evaluation of MLSR and PLSR for estimating soil element contents using visible/near-infrared spectroscopy in apple orchards on the Jiaodong peninsula, Catena 137 (2016) 340–349.
[23] A.M. Mouazen, et al., Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy, Geoderma 158 (1–2) (2010) 23–31.
[24] R.A. Viscarra Rossel, T. Behrens, Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma 158 (1–2) (2010) 46–54.
[25] L. Breiman, Statistical modeling: the two cultures, Stat. Sci. 16 (3) (2001) 199–231.
[26] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[27] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[28] R. Vašát, et al., Combining reflectance spectroscopy and the digital elevation model for soil oxidizable carbon estimation, Geoderma 303 (2017) 133–142.
[29] K. Tan, et al., Random forest–based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data, Environ. Monit. Assess. 191 (7) (2019).
[30] K. Tan, et al., Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest, J. Hazard Mater. 382 (2020) 120987.
[31] A. Morellos, et al., Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy, Biosyst. Eng. 152 (2016) 104–116.
[32] G. Naibo, et al., Near-infrared spectroscopy to estimate the chemical element concentration in soils and sediments in a rural catchment, Catena (2022) 213.
[33] L. Liu, M. Ji, M. Buchroithner, Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra, Rem. Sens. 9 (12) (2017) 1299.
[34] K. Tan, et al., Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning, J. Hazard Mater. 401 (2021) 123288.
[35] F.B. de Santana, et al., 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, Catena (2022) 218.
[36] D. Ou, et al., Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction, Geoderma 385 (2021) 114875.
[37] Y. Wang, et al., A comparison of multiple deep learning methods for predicting soil organic carbon in Southern Xinjiang, China, Comput. Electron. Agric. (2023) 212.
[38] Y. Hong, et al., Data mining of urban soil spectral library for estimating organic carbon, Geoderma 426 (2022).
[39] W.M. Brown, et al., Artificial neural networks: a new method for mineral prospectivity mapping, Aust. J. Earth Sci. 47 (4) (2000) 757–770.
[40] V.F. Rodriguez-Galiano, et al., An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS J. Photogrammetry Remote Sens. 67 (2012) 93–104.
[41] Y. Zhang, et al., Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm, Geoderma 333 (2019) 23–34.

[42] V.F. Rodriguez-Galiano, et al., Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture, Rem. Sens. Environ. 121 (2012) 93–107.
[43] R. Bellman, in: Dynamic Programming, 2 ed., 2003. Mineola, NY.
[44] C. Sammut, G.I. Webb, Encyclopedia of Machine Learning, Springer, 2010.
[45] Z. Xiaobo, et al., Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (1–2) (2010) 14–32.
[46] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1–2) (1997) 245–271.
[47] M. Cocchi, A. Biancolillo, F. Marini, Chemometric Methods for Classification and Feature Selection, vol. 82, 2018, pp. 265–299.
[48] V.F. Rodriguez-Galiano, et al., Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods, Sci. Total Environ. 624 (2018) 661–672.
[49] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[50] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[51] D. Effrosynidis, A. Arampatzis, An evaluation of feature selection methods for environmental data, Ecol. Inf. 61 (2021).
[52] A. Cardenas-Martinez, et al., Predictive modelling benchmark of nitrate vulnerable zones at a regional scale based on machine learning and remote sensing, J. Hydrol. 603 (2021) 127092.
[53] A. Raj, et al., Soil mapping via diffuse reflectance spectroscopy based on variable indicators: an ordered predictor selection approach, Geoderma 314 (2018) 146–159.
[54] F. Castaldi, et al., Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon, Rem. Sens. Environ. 179 (2016) 54–65.
[55] C. Gomez, G. Coulouma, Importance of the spatial extent for using soil properties estimated by laboratory VNIR/SWIR spectroscopy: examples of the clay and calcium carbonate content, Geoderma 330 (2018) 244–253.
[56] J. Wang, et al., Spectral variable selection for estimation of soil organic carbon content using mid-infrared spectroscopy, Eur. J. Soil Sci. 73 (4) (2022).
[57] X. Shi, et al., Improving soil organic matter estimation accuracy by combining optimal spectral preprocessing and feature selection methods based on pXRF and vis-NIR data fusion, Geoderma 430 (2023).
[58] Y. Hong, et al., Combining fractional order derivative and spectral variable selection for organic matter estimation of homogeneous soil samples by VIS–NIR spectroscopy, Rem. Sens. 10 (3) (2018) 479.
[59] N.L. Tsakiridis, et al., A genetic algorithm-based stacking algorithm for predicting soil organic matter from vis–NIR spectral data, Eur. J. Soil Sci. 70 (3) (2019) 578–590.
[60] X. Zhang, et al., Towards optimal variable selection methods for soil property prediction using a regional soil vis-NIR spectral library, Rem. Sens. 15 (2) (2023).
[61] Y. Wang, J. Wang, H. Che, Two-timescale neurodynamic approaches to supervised feature selection based on alternative problem formulations, Neural Network. 142 (2021) 180–191.
[62] Y. Liu, F. Tang, Z. Zeng, Feature selection based on dependency margin, IEEE Trans. Cybern. 45 (6) (2015) 1209–1221.
[63] E. Harefa, W. Zhou, Performing sequential forward selection and variational autoencoder techniques in soil classification based on laser-induced breakdown spectroscopy, Anal. Methods 13 (41) (2021) 4926–4933.
[64] Junta de Andalucía, Mapa de Suelos de Andalucía a escala 1:40.000, Atlas de Andalucía, 2005.
[65] J.A. Luque-Espinar, et al., Karst and Vegetation: Biodiversity and Geobotany in the Sierra de las Nieves Karst Aquifer, 2020, pp. 11–22. Málaga, Spain.
[66] S. Mazzoli, et al., The evolution of the footwall to the Ronda subcontinental mantle peridotites: insights from the Nieves Unit (western Betic Cordillera), J. Geol. Soc. 170 (3) (2013) 385–402.
[67] A. Fernandez-Cancio, et al., Climate classification of Abies pinsapo boiss. Forests in southern Spain, Investigación Agraria: Sistemas y Recursos Forestales 16 (3) (2007) 222.
[68] J. Villanueva, Tercer Inventario Forestal Nacional (1997–2007), Ministerio de Medio Ambiente, Madrid, 2005.
[69] J. Porta Casanellas, Determinación de carbonatos totales en suelos mediante calcimetría de Bernard, in: C.O.d.I.A.d. Cataluña (Ed.), Técnicas y Experimentos en Edafología, 1986, pp. 69–76.
[70] F. Lamas, et al., Selection of the most appropriate method to determine the carbonate content for engineering purposes with particular regard to marls, Eng. Geol. 81 (1) (2005) 32–41.
[71] G.W. Gee, J. Bauder, A. Klute, Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods, Soil Science Society of America, American Society of Agronomy, 1986.
[72] X.-X. Qiao, et al., Hyperspectral estimation of soil organic matter based on different spectral preprocessing techniques, Spectrosc. Lett. 50 (3) (2017) 156–163.
[73] C. Gomez, P. Lagacherie, G. Coulouma, Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements, Geoderma 148 (2) (2008) 141–148.
[74] R Core Team, R: A Language and Environment for Statistical Computing, 2013. Vienna, Austria.
[75] B. Mevik, R. Wehrens, The pls package: principal component and partial least squares regression in R, J. Stat. Software 18 (2) (2007).
[76] S. Mahesh, et al., Comparison of partial least squares regression (PLSR) and principal components regression (PCR) methods for protein and hardness predictions using the near-infrared (NIR) hyperspectral images of bulk samples of Canadian wheat, Food Bioprocess Technol. 8 (1) (2014) 31–40.
[77] T. Mehmood, et al., A review of variable selection methods in Partial Least Squares Regression, Chemometr. Intell. Lab. Syst. 118 (2012) 62–69.
[78] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS Regression), WIREs Computational Statistics 2 (1) (2010) 97–106.
[79] B.-H. Mevik, et al., Package 'pls', 2020.
[80] V.F. Rodriguez-Galiano, M. Chica-Olmo, M. Chica-Rivas, Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain, Int. J. Geogr. Inf. Sci. 28 (7) (2014) 1336–1354.
[81] E. Briscoe, J. Feldman, Conceptual complexity and the bias/variance tradeoff, Cognition 118 (1) (2011) 2–16.
[82] J. Rogan, et al., Land-Cover change monitoring with classification trees using landsat TM and ancillary data, Photogramm. Eng. Rem. Sens. 69 (7) (2003) 793–804.
[83] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.
[84] B. Bischl, et al., Mlr: machine learning in R, J. Mach. Learn. Res. 17 (1) (2016) 5938–5942.
[85] A. McBratney, B. Minasny, Why you don't need to use RPD, Pedometron 33 (2013).
[86] A. Bommert, et al., Benchmark for filter methods for feature selection in high-dimensional classification data, Comput. Stat. Data Anal. 143 (2020) 106839.
[87] L. Xu, et al., Estimation of organic carbon in anthropogenic soil by VIS-NIR spectroscopy: effect of variable selection, Rem. Sens. 12 (20) (2020).
[88] Y. Hong, et al., Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: feature selection coupled with random forest, Soil Tillage Res. 199 (2020) 104589.
[89] M. Vohland, et al., Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection, Geoderma 223–225 (2014) 88–96.
[90] M. Vohland, et al., Quantification of soil properties with hyperspectral data: selecting spectral variables with different methods to improve accuracies and analyze prediction mechanisms, Rem. Sens. 9 (11) (2017) 1103.
[91] J.M. Moura-Bueno, et al., Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions, Geoderma 337 (2019) 565–581.
[92] N.M. Knox, et al., Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy, Geoderma 239–240 (2015) 229–239.
[93] F. Castaldi, et al., Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands, ISPRS J. Photogrammetry Remote Sens. 147 (2) (2019) 267–282.
[94] V. Bellon-Maurel, et al., Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy, TrAC, Trends Anal. Chem. 29 (9) (2010) 1073–1081.

[95] Y. Ostovari, et al., Towards prediction of soil erodibility, SOM and CaCO 3 using laboratory Vis-NIR spectra: a case study in a semi-arid region of Iran, Geoderma 314 (2018) 102–112.
[96] E.R. Stoner, M.F. Baumgardner, Characteristic variations in reflectance of surface soils, Soil Sci. Soc. Am. J. 45 (6) (1981) 1161–1165.
[97] M. Nocita, et al., Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy, Geoderma 199 (2013) 37–42.
[98] A. Volkan Bilgili, et al., Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey, J. Arid Environ. 74 (2) (2010) 229–238.
[99] M.S. Askari, S.M. O'Rourke, N.M. Holden, Evaluation of soil quality for agricultural production using visible–near-infrared spectroscopy, Geoderma 243–244 (2015) 80–91.
[100] T. Tiecher, et al., Improving the quantification of sediment source contributions using different mathematical models and spectral preprocessing techniques for individual or combined spectra of ultraviolet–visible, near- and middle-infrared spectroscopy, Geoderma 384 (2021) 114815.
[101] H. Pasternak, Y. Edan, Z. Schmilovitch, Overcoming multicollinearity by deducting errors from the dependent variable, J. Quant. Spectrosc. Radiat. Transf. 69 (6) (2001) 761–768.
[102] T. Shi, et al., Soil organic carbon content estimation with laboratory-based visible–near-infrared reflectance spectroscopy: feature selection, Appl. Spectrosc. 68 (8) (2014) 831–837.
[103] K.R.M. Adeline, et al., Predictive ability of soil properties to spectral degradation from laboratory Vis-NIR spectroscopy data, Geoderma 288 (2017) 143–153.
[104] C. Wang, et al., Hyperspectral estimation of soil organic matter and clay content in loess plateau of China, Agron. J. 113 (3) (2021) 2506–2523.
[105] A. Samuel Obeng, et al., Soil organic matter carbon chemistry signatures, hydrophobicity and humification index following land use change in temperate peat soils, Heliyon 9 (9) (2023) e19347.
[106] B. Stenberg, et al., Visible and Near Infrared Spectroscopy in Soil Science 107 (2010) 163–215.
[107] F. Gogé, et al., Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma 213 (2014) 1–9.
[108] L. Zhong, et al., Soil properties: their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks, Geoderma 402 (2021) 115366.
[109] A. Gholizadeh, et al., Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging, Rem. Sens. Environ. 218 (2018) 89–103.
[110] F.J. Matus, Fine silt and clay content is the main factor defining maximal C and N accumulations in soils: a meta-analysis, Sci. Rep. 11 (1) (2021).
[111] J.A. Coblinski, et al., Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths, Catena 189 (2020).