Received: 1 November 2022 | Revised: 13 April 2023 | Accepted: 25 April 2023

DOI: 10.1111/ejss.13370

RESEARCH ARTICLE

European Journal of Soil Science

WILEY

Interpretable spectroscopic modelling of soil with machine learning

Alexandre M. J.-C. Wadoux

Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia

Correspondence
Alexandre M. J.-C. Wadoux, Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, New South Wales, Australia.
Email: alexandre.wadoux@sydney.edu.au

Abstract

Spectroscopic modelling of soil has advanced greatly with the development of large spectral libraries, computational resources and statistical modelling. The use of complex statistical and algorithmic tools from the field of machine learning has become popular for predicting properties from their visible, near- and mid-infrared spectra. Many users, however, find it difficult to trust the predictions made with machine learning. We lack interpretation and understanding of how the predictions were made, so that these models are often referred to as black boxes. In this study, I report on the development and application of a model-independent method for interpreting complex machine learning spectroscopic models. The method relies on Shapley values, a statistical approach originally developed in coalitional game theory. In a case study for predicting the total organic carbon from a large European mid-infrared spectroscopic database, I fitted a random forest machine learning model and showed how Shapley values can help us understand (i) the average contribution of individual wavenumbers, (ii) the contribution of the spectrum-specific wavenumbers, and (iii) the average contribution of groups of spectra taken together with similar characteristics. The results show that Shapley values revealed more insights than commonly used interpretation methods based on the variable importance. The most striking spectral regions identified as important contributors to the prediction corresponded to the molecular vibration of organic and inorganic compounds that are known to relate to organic carbon. Shapley values are a useful methodological development that will yield a better understanding and trust of complex machine learning and algorithmic tool in soil spectroscopy research.

KEYWORDS

chemometrics, deep learning, interpretable machine learning, mid-infrared, organic carbon, random forest, Shapley

https://doi.org/10.1111/ejss.13370

1 | INTRODUCTION

The reflectance spectra of soils in the visible, near and mid-infrared contain information on the interaction between chemical soil compounds and electromagnetic radiation. Analysis and modelling of high-dimensional spectral data is usually a challenging task so that information about the chemical composition or concentration of a soil property must be extracted with mathematical manipulation of the spectra and statistical modelling.

A common objective in soil spectroscopy studies is therefore to exploit the relationship between the complex spectral feature and laboratory-measured soil data and to produce predictive models of the property of interest. Recent examples of studies using this approach are Hutengs et al. (2019) for estimating the organic carbon content of bulk soil samples collected in the field with a handheld mid-infrared (MIR) spectrometer, or Cañasveras et al. (2010) to estimate aggregate stability in Mediterranean soil from visible and near-infrared spectra.

Conventionally, spectroscopic modelling is made with techniques based on spectra dimension reduction such as principal component or partial least squares regression (PLSR, Wold et al., 2001). High-dimensional spectra data can also be reduced into a smaller set of variables using, for example, wavelet multi-resolution analysis (Viscarra Rossel & Lark, 2009) or the Gaussian pyramid scale space (Behrens et al., 2022). In recent years, data-driven and non-linear algorithmic tools from the field of machine learning became popular for spectroscopic modelling. Examples of machine learning algorithms used in soil spectroscopy are random forest (de Santana et al., 2018), deep neural networks (Padarian et al., 2019), cubist (Minasny & McBratney, 2008), or support vector machine (Deiss et al., 2020). The reader is referred to Meza Ramirez et al. (2021) for an overview of machine learning in spectroscopy. Machine learning algorithms are usually more accurate than simple models and can find a pattern in the high-dimensional spectral data, but their structure is complex, lacks transparency, and is beyond human understanding. Information about their internal functioning cannot be readily obtained so these models are often referred to as black boxes (see, for example, the issues raised in McBride, 2022).

Interpreting a spectroscopic model can be made using model-specific variable importance indices that inform on the relative contribution of the wavelengths or wavenumbers to the prediction. In the popular PLSR model, one can interpret the loadings, the regression coefficients and the variable importance in projection. For example, Rienzi et al. (2014) used a variable importance metric to summarize the contribution of visible and near-infrared bands to the prediction of organic carbon with PLSR under varying moisture levels whereas Janik et al. (2007) interpreted the loadings of a PLSR model, fitted to understand the important spectral features that contributed to the prediction of soil-water properties. Alternatively, modelling based on simple regression trees such as cubist (Quinlan, 1992) generates a set of conditions and linear models that are readily interpretable. Interpretation, however, is more challenging for other algorithms including neural networks, ensemble of decision trees with random forest or support vector machines, which currently lack interpretability.

Highlights

- Complex machine learning and algorithmic tools are often used for spectroscopic modelling.
- Shapley values can help understanding how the prediction of properties from their spectra is made.
- Shapley values enable going beyond the average variable importance in prediction.
- The method can be applied to any complex machine learning and deep learning algorithm.

In the statistical and machine learning literature (Molnar, 2020) and in soil science (Wadoux & Molnar, 2022), several model-independent techniques were developed or applied to interpret complex models. These techniques are, for example, the variable importance using permutation (Molnar, 2020, section 8.5), the accumulated local effect (Apley & Zhu, 2020) or the Shapley values (Shapley, 1953) with their various forms of estimation (e.g., SHAP, Lundberg & Lee, 2017). The methods have a solid statistical foundation, can be applied to any fitted model and enable interpreting differentiable aspects of the model prediction. To date, only a few studies reported model-independent interpretation methods in soil spectroscopy. Chalaux Clergue et al. (2023), for example, conducted a permutation analysis on a cubist model fitted with mid-infrared data to reveal the main wavenumbers contributing to the estimation of a soil aggregate stability index in mainland France. In Haghi et al. (2021) and Zhong et al. (2021), the SHAP method is used to report on the wavenumbers contributing to the model prediction of various soil properties. In each of these studies, a model-independent technique is used to obtain a global importance metric of the wavenumber or wavelength in the prediction of a soil property. While useful, new developments in local methods of interpretation by means of Shapley values may yield insights that go beyond the variable importance in prediction. It is worthwhile to include these recent developments in interpretable machine learning for understanding how the predictions from complex spectroscopic models were made.

This paper aims to explore the use of Shapley values for the interpretation of complex spectroscopic models and to demonstrate its applicability with a large spectral dataset and machine learning. In a case study for predicting total organic carbon using mid-infrared soil spectroscopic data and a fitted random forest model, I show how Shapley values can yield insights beyond the variable importance in prediction through the average contribution of the individual wavenumbers, the contribution of the spectrum-specific wavenumbers, and the average contribution of groups of spectra with similar characteristics.

2 | MATERIALS AND METHODS

2.1 | Dataset

The soil samples from the freely available geochemical mapping of agricultural soils and grazing land of Europe (GEMAS) dataset (Reimann et al., 2014) are used hereafter. The GEMAS dataset comprises 4115 geo-referenced soil samples spanning 34 European countries, which represents a density of about 1 sample per 2500 km^2. About half of the samples were collected on agricultural soil (Ap-horizon, 0–20 cm, regularly ploughed fields) while the other half comes from land under permanent grass cover (grazing land soils, 0–10 cm). The soil samples are composite with support of 100 m^2 and were collected in 2008 following a standard field protocol described in Reimann et al. (2014). Mid-infrared spectra were measured on all air-dried and <2 mm sieved samples using a Perkin–Elmer Spectrum-One™ Fourier-transform infrared spectrometer (Perkin Elmer Inc., Mass. USA). The spectra had a spectral range of 7800–450 cm^{-1} and a spectral resolution of 2 cm^{-1}. Silicon carbide discs (Perkin–Elmer Life and Analytical Sciences Pty Ltd., Australia) were used for the background reference scan. Spectra were converted to pseudo-absorbance using $\log\left(\text{reflectance}^{-1}\right)$. More details on the scanning procedure are found in Soriano-Disla et al. (2013). A standard normal variate (SNV) transformation (Barnes et al., 1989) was applied to the spectra. For the modelling and interpretation, spectra were then resampled to a resolution of 8 cm^{-1}. Hereafter only the pre-processed spectra in the range smaller than 4000 cm^{-1} are used. The total soil organic carbon (TOC, in percent) is considered as our property of interest. The TOC content of the samples was obtained by dry combustion according to the ISO standard 10,694 (ISO 10694:1995, 1995). The TOC dataset had a minimum value of 0, a median of 2.1, a mean of 3.34, a standard deviation value of 4.9, a 1st and 3rd quartile value of 1.4 and 3.5, respectively, and a maximum value of 49%. The pre-processed spectra along with the TOC values in colour are shown in Figure 1.

2.2 | Spectroscopic modelling with random forest

The spectroscopic model was fitted with random forest (RF, Breiman, 2001a) to estimate the TOC from the MIR spectra. Random forest is a machine learning algorithm based on decision tree. A single tree is built by recursive partitioning of the data into non-overlapping regions using a splitting metric. The splitting metric is evaluated for different partitions. When the best cut point is determined the newly created partitions undergo the same procedure to grow the tree until a user-defined stopping criterion is met. The stopping criterion is usually the minimum of observations (node.size) in the last partition. Instead of a single tree, Breiman (1996) introduced the bagging procedure. In bagging, an ensemble of decision trees are fitted from a bootstrap sample of the training data. The predictions are aggregated by taking the average of all trees predictions. The RF algorithm builds on this and introduces an additional random perturbation during the splitting of the tree (Breiman, 2001a). In each partition, only a subset of predictors of size mtry from the original set of predictors is considered. The RF algorithm has therefore three user-defined parameters; the number of trees, node.size and mtry.

The RF predictions were assessed using validation statistics obtained through a random 10-fold cross-validation (CV) strategy. I created 10 folds of approximately equal sizes. Nine folds were used for fitting the RF algorithm and the remaining fold was used for validation. This procedure was repeated 10 times, each time leaving aside a different fold for validation. In each validation fold the predictions were computed. The validation statistics are calculated from the pairwise set of observations and predictions of all folds obtained from the CV. I report the mean error (ME), the root mean square error (RMSE), the Pearson's *r* correlation and the modelling efficiency coefficient (MEC, Janssen & Heuberger, 1995). The latter has an optimal value of 1 and can be negative if the model is a worse predictor than the mean of the observations taken as a predictor. The model used for interpretation in the next step is fitted on all observations.

2.3 | Interpretation with Shapley values

The RF model is interpreted with Shapley values (Shapley, 1953), a method initially developed within

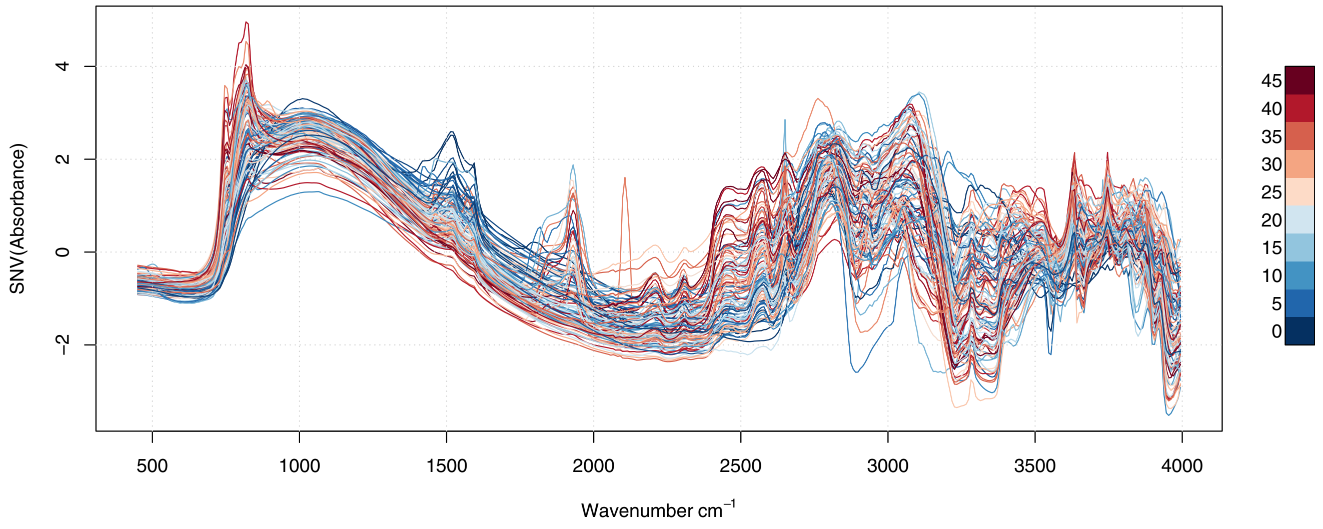


FIGURE 1 Example set of 100 pre-processed mid-infrared spectra coloured with the TOC values (in %). The 100 spectra were selected from the whole dataset of 4221 spectra using Kennard-Stone sampling (Kennard & Stone, 1969) on three principal component scores.

coalitional game theory. Consider a game where the prediction of the property of interest is the payout and each spectral band is a player. Spectral bands "collaborate" to obtain a gain (i.e. the prediction). Shapley values split the gain to the individual players according to their relative contribution to the outcome. The Shapley value of an individual spectral band is the average change in the prediction that the coalition receives when the spectral band is added to the game.

Consider the spectral matrix **X** of size $n \times b$ where is the number of spectra and b is the number of spectral bands. $\mathcal{S} \subseteq \{1,...,b\} \setminus \{j\}$ refers to any subset of size $1 < b$ of the spectral bands which excludes band j. The function $\widehat{f}(\mathbf{x}_{i,\mathcal{S}})$ is the RF prediction for sample i that describes the "payout" for the coalition of bands $\mathcal{S}$ contained in **X**. The Shapley value ϕ_j for band j and sample i is given by:

$$\phi_{i,j} = \sum_{\mathcal{S} \subseteq \{1,...,b\} \setminus \{j\}} \frac{|\mathcal{S}|!(b-|\mathcal{S}|-1)!}{b!} \left(\widehat{f}(\mathbf{x}_{i,\mathcal{S} \cup \{j\}}) - \widehat{f}(\mathbf{x}_{i,\mathcal{S}})\right), \tag{1}$$

where $\mathcal{S} \cup \{j\}$ is the subset $\mathcal{S}$ with the jth band added. Equation (1) has two components: the first is a weighted average, giving equal weight to each of the marginal contributions of all possible subsets of spectral bands whereas the second is the marginal contribution for a subset of bands. $\phi_{i,j}$ is the Shapley value that corresponds to the contribution of a spectral band to the prediction of the property of interest in a sample. Shapley values satisfy the following axioms (Molnar, 2020):

1. Efficiency: the sum of the Shapley values of all predictors equals the difference between the predicted value and the average of the property of interest from the calibration dataset.
2. Symmetry: the contribution of two predictors is the same if they contribute equally to the prediction.
3. Dummy: a predictor that does not contribute should have a Shapley value of 0.
4. Additivity: the Shapley value of a combined set of predictors is the sum of their individual contributions.

Obtaining an exact solution for Equation (1) requires estimating all possible sets of predictors with and without the jth band. This is computationally intractable if the number of predictors is large. A solution to approximate Shapley values from Equation (1) was proposed in Štrumbelj and Kononenko (2014) by means of Monte-Carlo (MC) sampling. The method relies on creating an ordered set of permutations of predictors, from which sampling is made. The Shapley value is then approximated by averaging the marginal contribution of the samples. This method relies on the user-defined number of Monte-Carlo samples, which should preferably be large to avoid the approximation error. Hereafter the method proposed in Štrumbelj and Kononenko (2014) is used to estimate the Shapley values.

A Shapley value is obtained for all wavenumbers of all spectra. The individual values are interpreted as the contribution of the wavenumber to the prediction of the target property, relative to the average of the target property in the calibration dataset. They are in the unit of the target properties and can be either negative or positive.

The four statistical axioms of Shapley values described previously enable to combine the values and understand various aspects of the fitted model. Hereafter I describe three example uses of Shapley values.

1. The average contribution of the wavenumbers to the prediction: this is similar to the average variable importance commonly reported in soil spectroscopy studies, but unlike the average variable importance it has a different interpretation. It is obtained by taking the wavenumber-specific average of the absolute Shapley values.
2. The contribution of the spectrum-specific wavenumbers: instead of taking the average of the absolute values I report the Shapley value of each individual wavenumber. The values can be negative or positive.
3. Contribution by groups of spectra with similar prediction characteristics. This is done by *k*-means clustering (Hartigan et al., 1979) of Shapley values. Grouping Shapley values enables us to find a pattern in a large number of Shapley values and identify spectra with similar characteristics in predicting the target soil property. The optimal number of clusters is evaluated using the sum of within-cluster squared errors as a criterion and the elbow method. I test several clusters between 2 and 20. The optimal number is indicated when the criterion has a minor further decrease when increasing the number of clusters.

2.4 | Practical implementation and computational aspects

All analyses were made in the R programming language (R Core Team, 2022). The pre-processing and dimensionality reduction of spectra were carried out with the prospectr (Stevens & Ramirez-Lopez, 2022) and resemble (Ramirez-Lopez et al., 2022) packages. Fitting of the RF algorithm was made with the ranger package (Wright & Ziegler, 2017) using the variance as splitting criterion. The number of trees was set to 100, whereas the two other parameters were set to their default value, that is, to a value of 5 for node.size and to the rounded down square root of the number of predictors for mtry. Estimation of the Shapley values was made with the package (Greenwell, 2020). A trial-and-error approach was used to verify that the MC sample size was sufficiently large to produce stable outcomes (after Nol et al., 2010, Figure 7). The Shapley values were estimated twice with the same MC sample size but with a different seed. The Shapley values of the two repetitions are represented in a scatter plot. The procedure is repeated for various MC sample sizes from 50 to 1000. I considered the number of MC samples to be sufficient when the results are close to the 1:1 line. To speed up processing, the procedure for estimating the Shapley values was parallelized. Calculations on a standard desktop computer with 8 cores took approximately 1 h for estimating Shapley values with an MC sample size of 50 and two days for an MC sample size of 1000.

3 | RESULTS

The RF model evaluated with a 10-fold CV shows no systematic over- or under-predictions (ME = 0.06%). It has an RMSE of 2.03%, a linear correlation coefficient $r = 0.91$ and a MEC of 0.83. The scatterplot of the measured and predicted TOC values are presented in Figure 2. Most points are scattered close to the 1:1 line, but few large TOC values are under-predicted (e.g. measured TOC is 48% and the predicted value is 19%). Overall, the visual inspection of Figure 2 and the validation statistics indicate that the fitted RF model is sufficiently accurate to predict the TOC and that it can serve as a basis for interpretation.

Figure 3 shows the Shapley values estimated with MC sample sizes of 50, 100, 250, 500 and 1000 plotted against the Shapley values of another repetition with sizes 50, 100, 250, 500 and 1000 but with a different seed. Even for small MC sample sizes, the points are generally always close to the 1:1 line. There is only a marginal improvement in increasing the number of MC samples

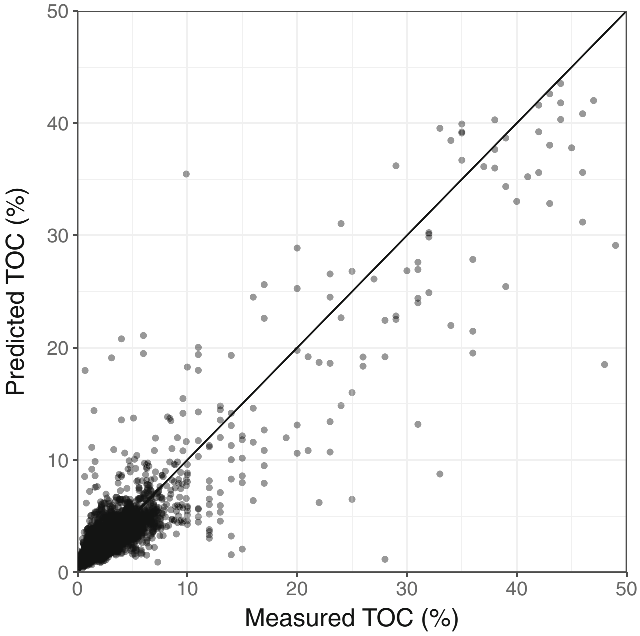


FIGURE 2 Scatterplot of measured versus predicted values of TOC (in %) by the fitted RF model. Predicted values are obtained by 10-fold CV.

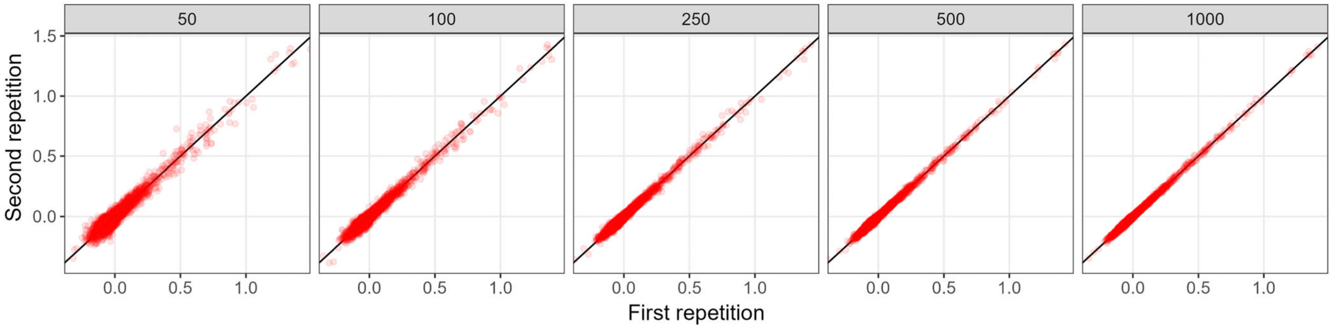


FIGURE 3 Scatter plot of Shapley values estimated with two repetitions with different seeds and for an MC sample size of 50, 100, 250, 500 and 1000.

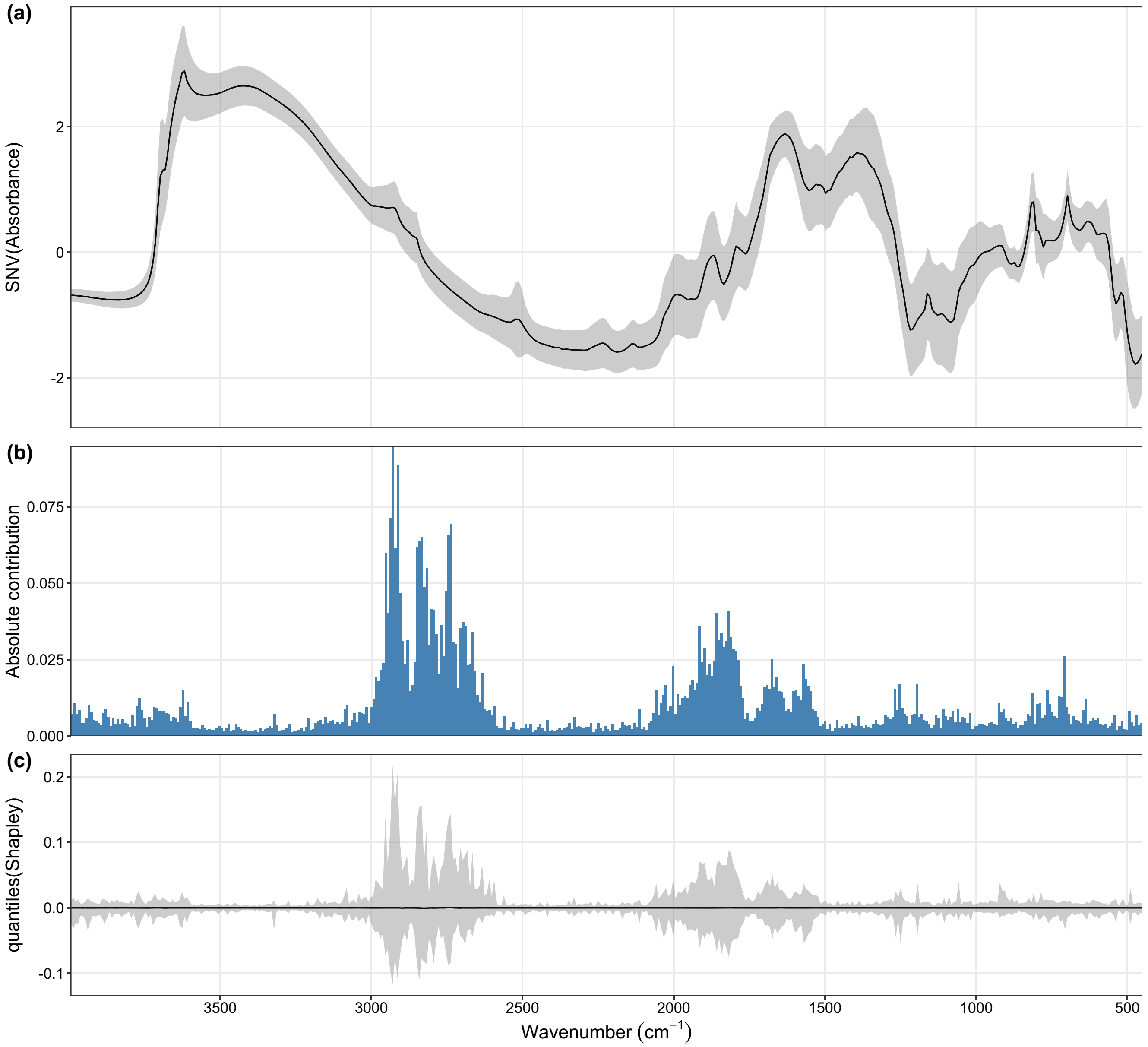


FIGURE 4 Mean and standard deviation of the pre-processed spectra (a) along with the average of the absolute contribution of the spectral regions to the TOC predictions (b) and the 0.05th and 0.95th percentiles of the individual Shapley values obtained for each wavenumber (c). Note that the plot in (b) is the band average of the absolute values shown in plot (c).

from 250 to 1000. I considered a sample size of 1000 to be sufficient to obtain a stable outcome because the results are very close to the 1:1 line. Hereafter, all the interpretation results are presented with Shapley values estimated with an MC sample size of 1000.

Figure 4a shows the mean (black line) and standard deviation (grey shaded area) of the pre-processed spectra, whereas Figure 4b,c show the magnitude of the spectral band contribution to the TOC prediction. Note that Figure 4b is the average of the absolute Shapley values reported in Figure 4c. In absolute values (Figure 4b) the two regions around 1730 and 2930 cm^{-1} are the most important contributors to the prediction of TOC, with values up to 0.08% and 0.04%, respectively. Small regions around 1620, 2020, 3600 cm^{-1} have also a relatively small contribution to the prediction. Figure 4c further shows the 0.05th and 0.95th percentiles of the individual Shapley values obtained for each wavenumber. Contributions are mostly positive (i.e. when the wavenumber value is higher, the contribution to the TOC prediction is also higher), but the two main spectral regions contributing to the TOC prediction also have large negative contributions up to 0.1%.

I further discriminate three groups of spectra having high TOC content (i.e. TOC $\geq$ 30%), low TOC content but high clay content (i.e. TOC $\leq$ 5% and clay $\geq$ 35%) and low TOC content but high sand content (i.e. TOC $\leq$ 5% and sand $\geq$ 85%). These values were chosen to have similar numbers of spectra in each of the three groups. Figure 5 shows the pre-processed average spectra for the three groups (a) and their relative wavenumber contribution to the TOC prediction for each group (b). The relative contribution is taken as the average of the absolute Shapley values by group. Figure 5 shows how the model adjusts the prediction of TOC by group of spectra. The group with high TOC content has several spectral regions contributing to the prediction, whereas the group with low TOC and high clay content uses few spectral regions (i.e. mostly at around 1800 and 2930 cm^{-1}). The group with low TOC and high sand content has a similar pattern of contribution that the group with low TOC and high clay content, but the relative contribution of the regions are smaller.

The Shapley values of all wavenumbers against the ranked measured TOC values (Figure 6) shows that for TOC values higher than 3%, the region at around 1800 always contributes positively to the prediction. Below this limit, it has a negative contribution (i.e. the absence of a contribution in this region results in TOC values close to the average value of the calibration dataset). For large TOC values, more regions contribute to the prediction, the most important of which are the regions at 1800 and 2930–3000 cm^{-1}. For large TOC values (i.e. TOC 30%), additional positive contributions are made from a region

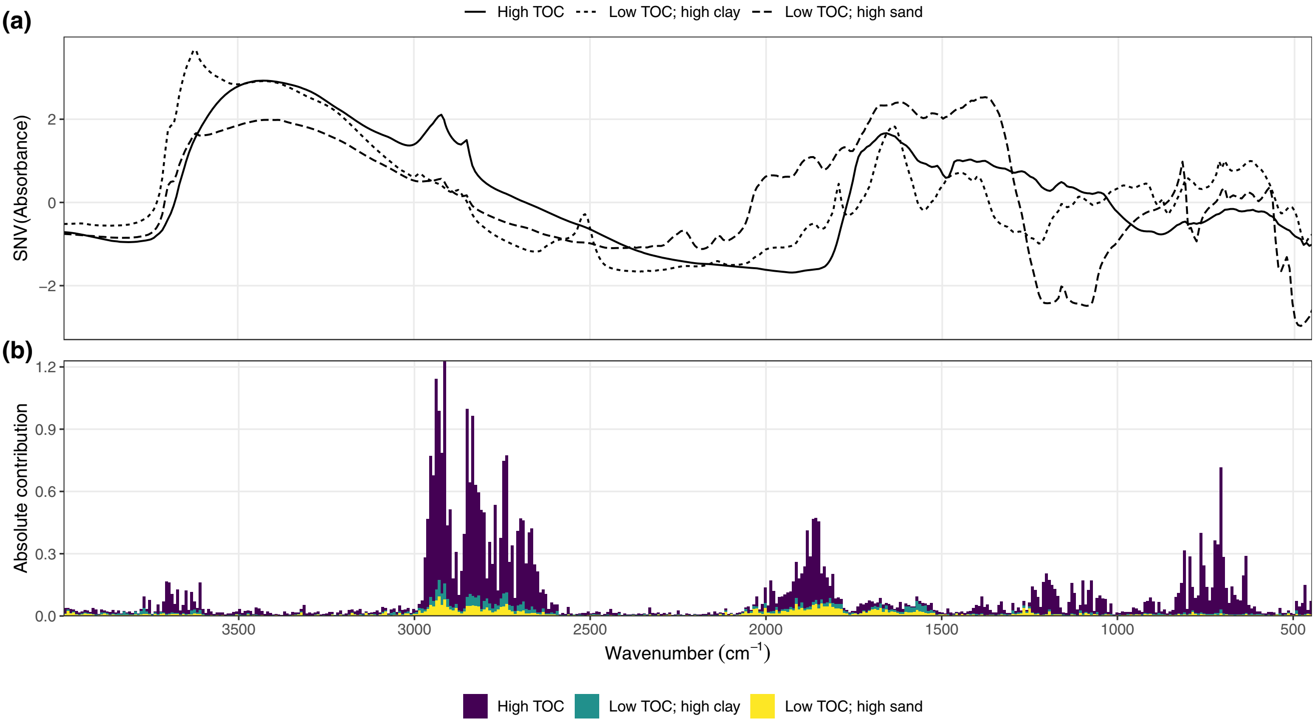


FIGURE 5 Average pre-processed spectra for the three groups (a) and average absolute contribution of the wavenumbers to the TOC prediction for each group (b). The contribution is expressed in the group average of the absolute Shapley values for each wavenumber.

at around 1050 cm^{-1}, a narrow region around 650 cm^{-1} and a very narrow region around 3620 cm^{-1}. The pattern of contributions is investigated further in the Discussion.

Figure 7 shows how the spectrum-specific contribution to the TOC prediction varies at wavenumber, for three spectral regions of importance. In the three bottom plots, each dot is an individual spectrum value at the wavenumber. The grey shaded areas in the upper plot of Figure 7 are the 0.05, 0.25, 0.45, 0.55, 0.75 and 0.95th percentiles of the pre-processed spectra. In the region (a) between 2882 and 2946 cm^{-1}, an increasing value at wavenumber relates to an increasing Shapley value, that is, a higher predicted value of TOC. The two other regions, (b) between 1802 and 1874 cm^{-1} and (c) between 682 and

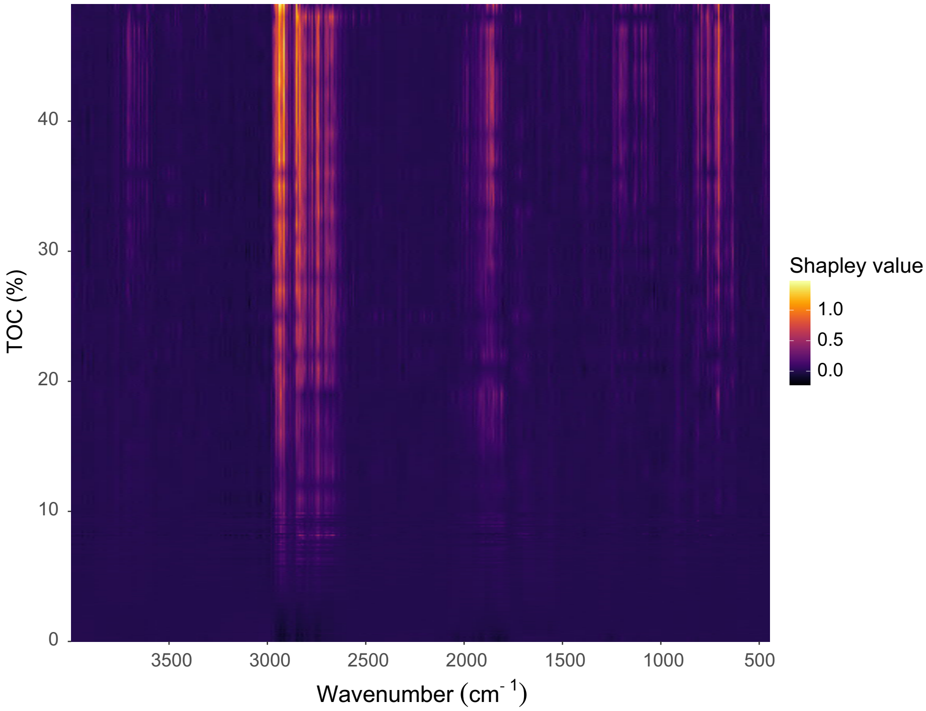


FIGURE 6 Shapley values of the spectra ranked by measured TOC values (in %). The colour indicates the contribution, in the unit of the TOC.

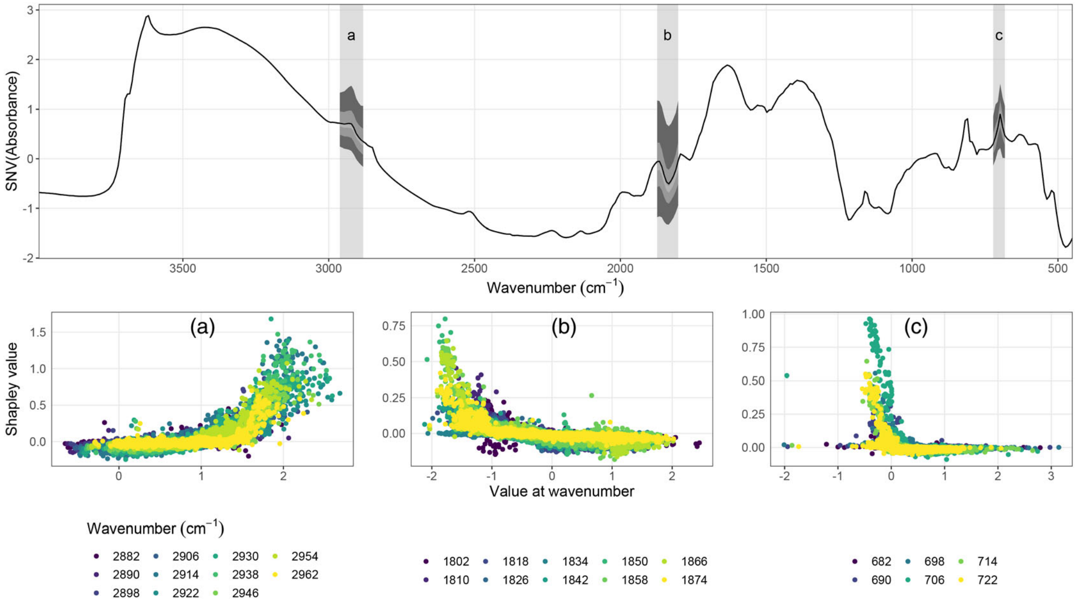


FIGURE 7 Wavenumber contribution to the TOC prediction, for three spectral regions of interest at (a) 2882–2946 cm^{-1}, (b) 1802–1974 cm^{-1} and (c) 682–722 cm. The grey shaded areas in the upper plot are 0.05, 0.25, 0.45, 0.55, 0.75 and 0.95th percentiles of the pre-processed spectra.

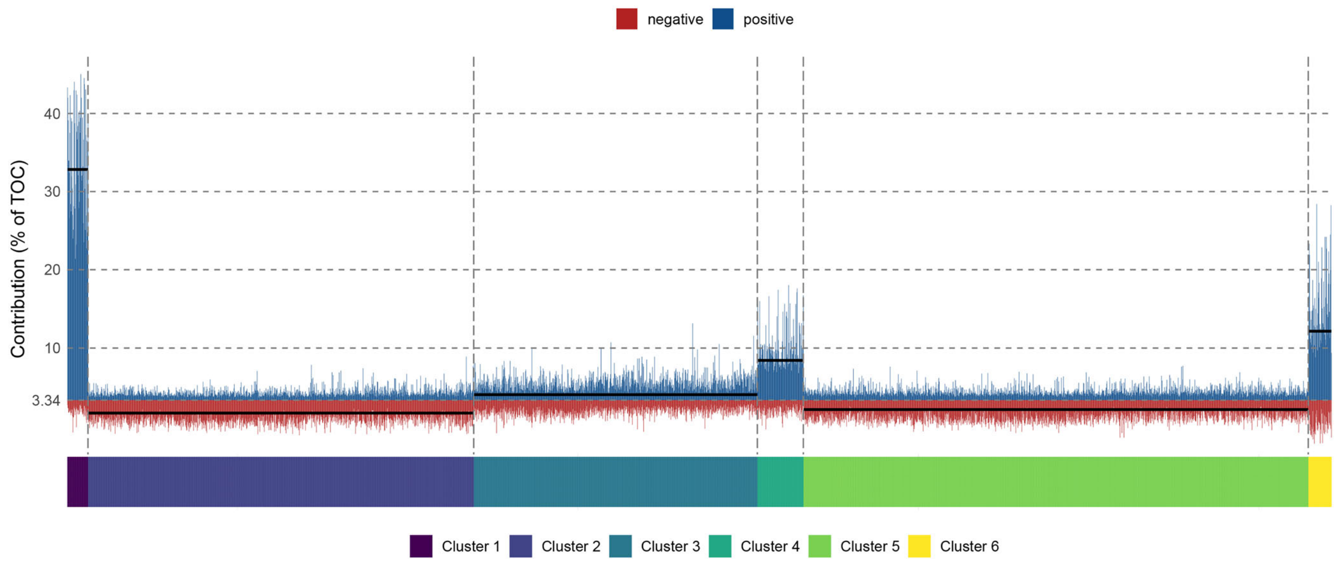
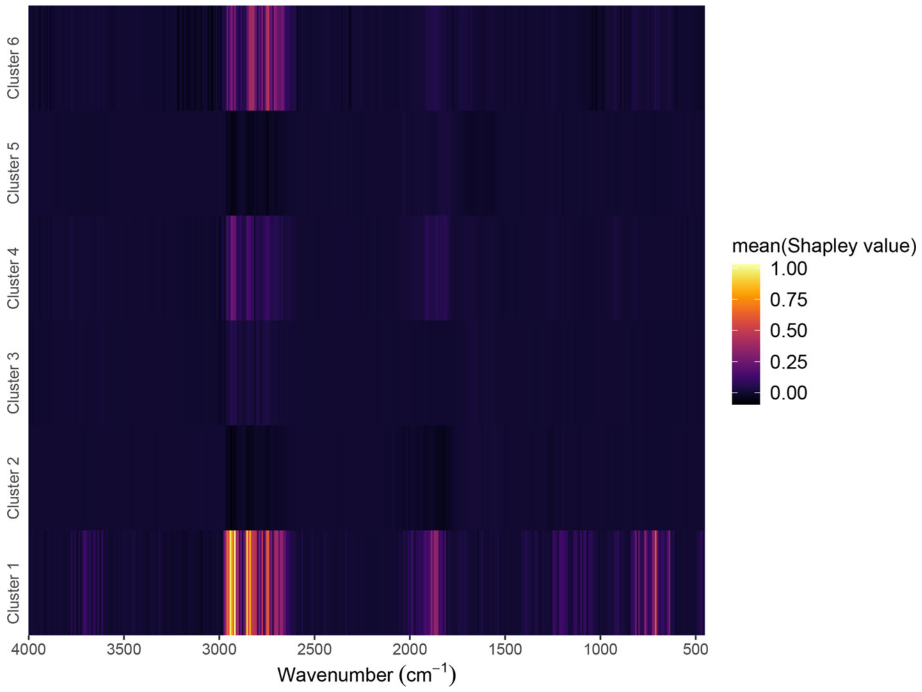


FIGURE 8 Cluster-specific contribution to the TOC prediction. The horizontal black line indicates the average predicted value of TOC by cluster.

FIGURE 9 Average of the Shapley values by wavenumber and for the six clusters.



722 cm^{-1} have a similar pattern but opposite to that observed in the region (a): for decreasing value of the SNV absorbance at a wavenumber, there is an decreasing contribution to the TOC prediction. For example, in the region (b), at wavenumber 1858 cm^{-1} the spectra that have an absorbance value higher than -1 have a contribution to the TOC prediction close to 0 but for values smaller than -1 of absorbance the contribution to the TOC prediction of the spectra is positive up to about 0.75%.

The elbow method and visual inspection of the criterion curve indicated that 6 classes were adequate to cluster the Shapley values. Figure 8 shows the cluster-specific contribution to the TOC prediction. Recall that the mean of the measured TOC is 3.34%, and that the sum of the positive and negative contributions results in the predicted value. The horizontal black line indicates the averaged predicted value of TOC by cluster. Figure 8 shows a clear pattern by cluster. For clusters 1, 3, 4, and 6 the positive contributions outweigh the negative contributions. An opposite pattern is found in cluster 2 where the negative contributions are more important than the positive ones, resulting in predictions of TOC that are

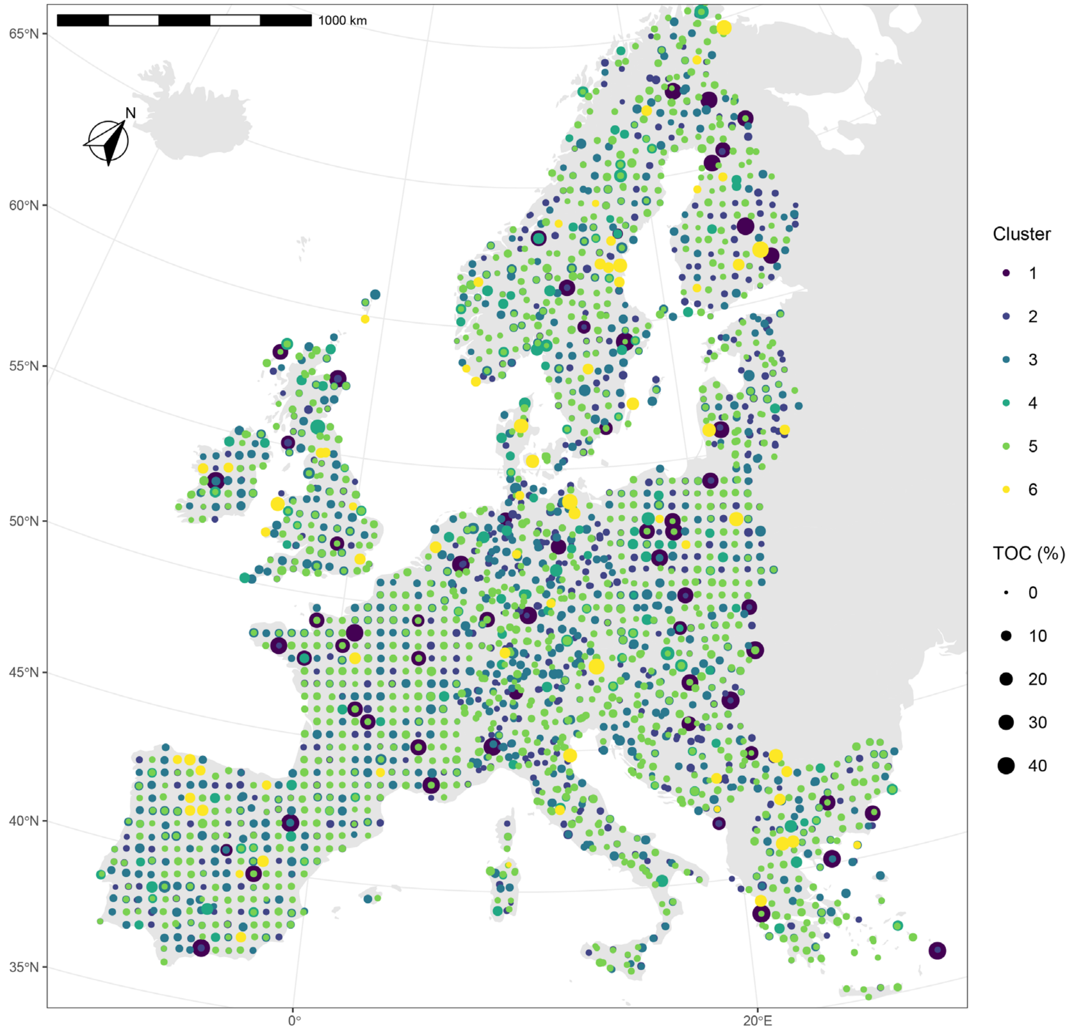


FIGURE 10 Spatial pattern of the clusters of Shapley values showing where spectra have similar prediction characteristics. The colour represents the cluster number and the dot size is the predicted TOC content (in %).

smaller than the average value. In cluster 5, the predicted values of TOC are close to the measured averaged values of TOC: the negative and positive contributions nearly cancel each other out.

The average of Shapley values by wavenumber and for the 6 clusters is shown in Figure 9. Cluster 2 has negative contributions from both 1800 and 2950 cm^{-1} while it is the opposite for clusters 1, 3, 4 and 6. For clusters 1 and 6, corresponding to the clusters predicting the highest TOC values (see also Figure 8), the bands at around 2950 cm^{-1} are important positive contributors to the TOC prediction. The only major difference in importance between clusters 5 and 6 is the spectral region at around 1900 cm^{-1} which contributes positively in cluster 1 but is not present in cluster 6.

The spatial pattern of the clusters is shown in Figure 10 with the same colour palette as Figure 8. Cluster 2 which corresponds to small TOC values covers most of Europe. It is difficult to distinguish a pattern for a specific cluster, but some general observations can be made. For example, there are areas where clusters 2 and 3 appear more frequently, such as in the Eastern part of Sweden and Slovakia, and in the Northern part of Germany. Clusters 1 and 6 clearly stand out because they correspond to large predicted TOC, although the TOC prediction in the two clusters is made using slightly different bands (see also Figure 9).

4 | DISCUSSION

The interpretation results rely on the empirical relationships captured by the machine learning model. In our case, the RF model fitted on the MIR spectra yielded an accurate prediction of TOC. It is no surprise because distinct and intense fundamental molecular vibrations of organic and inorganic compounds occur in the MIR range of the electromagnetic spectrum (Viscarra Rossel et al., 2006). While it is difficult to compare the RF prediction results to previous studies, the validation statistics are in line with published works. For example, the MEC value is in the range of R2 values between 0.84 and 0.97 for carbon and organic matter prediction with MIR reported in Soriano-Disla et al. (2014). In a study at the European scale to predict soil organic carbon content of the topsoil with vis-near infrared spectroscopy, Nocita et al. (2014) found R2 values of 0.79, 0.81 and 0.79 for prediction made on samples from cropland, grassland and woodland, respectively. In the prediction of soil properties with spectroscopy, the quality of predictions may vary greatly depending on factors such as the sample preparation before spectrum acquisition (e.g., fine grinding or sieving, see Wijewardane et al., 2021), the complexity of the spectral data, the accuracy and mismatch between the laboratory methods of the reference data, or the prediction method itself. Overall, the validation statistics and the comparison with existing works indicate that the RF model fitted in this study is sufficiently accurate to serve as a basis for the interpretation with Shapley values.

The most striking spectral regions identified as important correspond to the molecular vibration of organic and inorganic compounds abundantly reported in the literature. I obtained strong positive contributions from the spectral regions corresponding to C=O stretching of carbonyl C and aliphatic CH vibration at around 1720 and 2930 cm^{-1} (Tinti et al., 2015), respectively, and the regions corresponding to the OH stretching of clay mineral at around 3622 cm for large TOC values (i.e., TOC > 30%) (Viscarra Rossel & Behrens, 2010). The positive contribution at 1050 cm^{-1} is assigned to the C-O stretching of carbohydrates. Both have positive contributions for large TOC values only. The fingerprint region <1500 cm^{-1} is more difficult to interpret and it is unclear why this region contributes to the TOC prediction. However, previous studies have found a similar pattern of importance in the region around 750–900 cm^{-1} for the prediction of TOC. Haghi et al. (2021, Figure 10), for example, reported a positive contribution of this region for predicting TOC with a cubist model. In this study, this could be related to either the presence of carbonates that are spectrally active in this region (Tatzber et al., 2007) or to high-TOC content accompanying quartz-rich sediments in some parts of Europe (Xu et al., 2019), but this requires further investigations.

One of the valuable findings of this study was to highlight that Shapley values revealed more insights than commonly used interpretation techniques reporting the average variable importance based on, for example, the cubist conditions (e.g. Butler et al., 2018) or the importance obtained by permutation (e.g. Chalaux Clergue et al., 2023). Shapley values yielded information on the either positive or negative contribution of a spectral band to the prediction, for all spectral bands. This enables interpretation by groups of spectra with similar characteristics (e.g. Figure 5) or by geographical or spectral regions (e.g. Figure 10). In practice this means that one can understand how the model adjusts the prediction of the property of interest; not all spectral regions are used equally for the prediction, and the importance of the regions varies between spectra (Figure 6) and spatially (Figure 10). This has several potential implications, such as for determining the domain of prediction (Wadoux et al., 2021a) on spectral regions instead of on the whole spectra.

It should be stressed that the Shapley values belong to the family of permutation-based interpretation methods (e.g. partial dependence plots, permutation) that are sensitive to cross-correlated predictor variables (Molnar et al., 2022). While MIR spectra are strongly cross-correlated, it did not appear to be a problem in this study. Alternatively, one may consider a prior step to de-correlate the spectra (e.g. using principal component analysis) or interpretation methods that are not sensitive to dependence between predictor variables (e.g. conditional feature importance). The analysis of groups of spectra, as is done in Figure 5, also appears to be an effective solution to address this issue. However, more research is needed to identify means of dealing with correlated predictor variables, some of which are already in development in the machine learning literature (e.g. Mase et al., 2019).

In the literature, the choice of modelling strategy is usually based on the dilemma between prediction accuracy and interpretability. Complex models, such as artificial neural networks, are usually more accurate than simple models (e.g. stepwise linear regression) but their internal functioning is beyond human understanding. This has led to several criticisms of using machine learning and algorithmic tools in spectroscopic modelling. For example, McBride (2022) showed a high degree of scepticism towards machine learning, arguing that using black-box models lead to a severe risk of making an accurate prediction based on chance relationships found in the data. Several works, therefore, make a substantial variable selection, reducing the spectra containing several hundreds of wavenumbers to only a few, which combination is nearly as informative as the whole spectra (e.g. Wang et al., 2022). I contend, however, like Breiman (2001b) in the statistical literature and Wadoux et al. (2021b, section 5) in soil science, that a model need not be simple to provide reliable information. A complex model is in fact often more accurate than a simple model and carries supposedly a better representation of the system under study. Posing the question of modelling as a trade-off between accuracy and interpretability is wrong; accuracy is the objective and interpretation is only a means to obtain the information. This study is a first step to shed some light on black-boxes models and to move towards a better understanding of complex spectroscopic tools which are widely developed and applied in soil research.

5 | CONCLUSION

The Shapley values that I described and tested for spectroscopic modelling represent a new method for improving our understanding of complex models based on machine learning and deep learning algorithms. I showed in a case study for predicting the total organic carbon of soil samples from a large mid-infrared spectral library that Shapley values highlighted the most important spectral regions contributing to the prediction. From the results and discussion, I draw the following conclusions:

- Shapley values revealed more insight than commonly used variable importance metrics reported in machine learning spectroscopic studies.
- In a case study, I showed how Shapley values are used to understand the average contribution of individual spectral bands to the prediction and the positive or negative contribution of spectrum-specific wavenumbers to the prediction.
- It is possible to aggregate Shapley values by groups of spectra with similar characteristics. In a case study, I grouped Shapley values by spectra which correspond to soil samples with similar total organic carbon, clay and sand contents, and by groups with similar prediction characteristics using *k*-means clustering.
- The most striking spectral features identified for the prediction of the total organic carbon at 1720 and 2930 cm^{-1} corresponded to the molecular vibration of organic soil compounds reported in the literature.

The results of this study add to the growing body of literature that emphasizes the importance of going beyond prediction in modelling with machine and deep learning applied to soil research. The statistical literature has developed several methods that are not dependent on any model type and which can readily be applied to soil modelling research. I recommend systematically using an interpretation method such as Shapley values so that the high accuracy of machine learning and deep learning can be linked to discernible features in the spectra.

AUTHOR CONTRIBUTIONS

Alexandre M. J.-C. Wadoux: Conceptualization; investigation; methodology; validation; formal analysis; visualization; writing – review and editing; writing – original draft.

ACKNOWLEDGEMENTS

The author thanks the GEMAS Project Team for collecting, preparing and analysing the GEMAS samples and sharing the mid-infrared soil data. Open access publishing facilitated by The University of Sydney, as part of the Wiley - The University of Sydney agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in GSI GEMAS European Geochemical Data at https://data.gov.ie/dataset/gsi-gemas-european-geochemical-data. These data were derived from the following resources available in the public domain: - http://www.gtk.fi/publ/foregsatlas, http://gemas.geolba.ac.at/.

REFERENCES

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *82*, 1059–1086.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, *43*, 772–777.

Behrens, T., Viscarra Rossel, R. A., Ramirez-Lopez, L., & Baumann, P. (2022). Soil spectroscopy with the gaussian pyramid scale space. *Geoderma*, *426*, 116095.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*, 199–231.

Butler, B. M., O'Rourke, S. M., & Hillier, S. (2018). Using rule-based regression models to predict and interpret soil properties from X-ray powder diffraction data. *Geoderma*, *329*, 43–53.

Cañasveras, J. C., Barrón, V., Del Campillo, M. C., Torrent, J., & Gómez, J. A. (2010). Estimation of aggregate stability indices in Mediterranean soils by diffuse reflectance spectroscopy. *Geoderma*, *158*, 78–84.

Chalaux Clergue, T., Saby, N. P. A., Wadoux, A. M. J.-C., Barthès, B. G., & Lacoste, M. (2023). Estimating soil aggregate stability with infrared spectroscopy and pedotransfer functions. *Soil Security*, *8*, 223–235.

de Santana, F. B., de Souza, A. M., & Poppi, R. J. (2018). Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *191*, 454–462.

Deiss, L., Margenot, A. J., Culman, S. W., & Demyan, M. S. (2020). Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma*, *365*, 114227.

Greenwell, B. (2020). Package “fastshap”. R package version 0.0.7 (Accessed Febraury 10, 2023). https://CRAN.R-project.org/package=fastshap

Haghi, R. K., Pérez-Fernández, E., & Robertson, A. H. J. (2021). Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. *Geoderma*, *396*, 115071.

Hartigan, J. A., Wong, M. A., et al. (1979). A k-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., & Vohland, M. (2019). In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. *Geoderma*, *355*, 113900.

ISO 10694:1995. (1995). Soil quality — Determination of organic and total carbon after dry combustion (elementary analysis). Standard: International Organization for Standardization Geneva, CH.

Janik, L. J., Merry, R. H., Forrester, S. T., Lanyon, D. M., & Rawson, A. (2007). Rapid prediction of soil water retention using mid infrared spectroscopy. *Soil Science Society of America Journal*, *71*, 507–514.

Janssen, P. H. M., & Heuberger, P. S. C. (1995). Calibration of process-oriented models. *Ecological Modelling*, *83*, 55–66.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*, 137–148.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In v. L. Ulrike, G. Isabelle, B. Samy, W. Hanna, & F. Rob (Eds.), *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). Curran Associates Inc..

Mase, M., Owen, A. B., & Seiler, B. (2019). Explaining black box decisions by Shapley cohort refinement. arXiv:1911.00467.

McBride, M. B. (2022). Estimating soil chemical properties by diffuse reflectance spectroscopy: Promise versus reality. *European Journal of Soil Science*, *73*, e13192.

Meza Ramirez, C. A., Greenop, M., Ashton, L., & Rehman, I. U. (2021). Applications of machine learning in spectroscopy. *Applied Spectroscopy Reviews*, *56*, 733–763.

Minasny, B., & McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *94*, 72–79.

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Lulu Press.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI – Beyond explainable artificial intelligence. Lecture notes in artificial intelligence* (pp. 55–84). Springer.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., & Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry*, *68*, 337–347.

Nol, L., Heuvelink, G. B. M., Veldkamp, A., de Vries, W., & Kros, J. (2010). Uncertainty propagation analysis of an n2o emission model at the plot and landscape scale. *Geoderma*, *159*, 9–23.

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional*, *16*, e00198.

Quinlan, J. R. (1992). Learning with continuous classes. In A. Adams & L. Sterling (Eds.), *5th Australian joint conference on artificial intelligence* (pp. 343–348). World Scientific.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna https://www.R-project.org/

Ramirez-Lopez, L., Stevens, A., Viscarra Rossel, R., Lobsez, C., Wadoux, A., & Breure, T. (2022). Resemble: Regression and Similarity Evaluation for Memory-Based Learning in Spectral Chemometrics. R package version 2.2.1. (Accessed September 23, 2022). https://CRAN.R-project.org/package=resemble

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., & O'Connor, P. (Eds.). (2014). *Chemistry of Europe's agricultural soils, part a*. Schweizerbart Science Publishers.

Rienzi, E. A., Mijatovic, B., Mueller, T. G., Matocha, C. J., Sikora, F. J., & Castrignanò, A. (2014). Prediction of soil organic carbon under varying moisture levels using reflectance spectroscopy. *Soil Science Society of America Journal*, *78*, 958–967.

Shapley, L. S. (1953). A value for n-person games. In K. H. William & T. A. William (Eds.), *Contributions to the theory of games* chapter 17 (pp. 31–40). Princeton University Press Princeton volume 28 of *Annals of Mathematics Studies*.

Soriano-Disla, J. M., Janik, L., McLaughlin, M. J., Forrester, S., Kirby, J., Reimann, C., & The EuroGeoSurveys GEMAS Project. (2013). The use of diffuse reflectance mid-infrared spectroscopy for the prediction of the concentration of chemical elements estimated by X-ray fluorescence in agricultural and grazing European soils. *Applied Geochemistry*, *29*, 135–143.

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. *Applied Spectroscopy Reviews*, *49*, 139–186.

Stevens, A., & Ramirez-Lopez, L. (2022). An introduction to the prospectr package. R package version 0.2.6. (Accessed September 23, 2022). https://CRAN.R-project.org/package=prospectr

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*, 647–665.

Tatzber, M., Stemmer, M., Spiegel, H., Katzlberger, C., Haberhauer, G., & Gerzabek, M. H. (2007). An alternative method to measure carbonate in soils by FT-IR spectroscopy. *Environmental Chemistry Letters*, *5*, 9–12.

Tinti, A., Tugnoli, V., Bonora, S., & Francioso, O. (2015). Recent applications of vibrational mid-infrared (IR) spectroscopy for studying soil components: A review. *Journal of Central European Agriculture*, *16*, 1–22.

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*, 46–54.

Viscarra Rossel, R. A., & Lark, R. M. (2009). Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. *European Journal of Soil Science*, *60*, 453–464.

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, *131*, 59–75.

Wadoux, A. M. J.-C., Malone, B., Minasny, B., Fajardo, M., & McBratney, A. B. (2021). Exploratory soil spectral analysis. In *Soil spectral inference with R* (pp. 81–113). Springer.

Wadoux, A. M. J.-C., & Molnar, C. (2022). Beyond prediction: Methods for interpreting complex models of soil variation. *Geoderma*, *422*, 115953.

Wadoux, A. M. J.-C., Román-Dobarco, M., & McBratney, A. B. (2021). Perspectives on data-driven soil research. *European Journal of Soil Science*, *72*, 1675–1689.

Wang, J., Liu, T., Zhang, J., Yuan, H., & Acquah, G. E. (2022). Spectral variable selection for estimation of soil organic carbon content using mid-infrared spectroscopy. *European Journal of Soil Science*, *73*, e13267.

Wijewardane, N. K., Ge, Y., Sanderman, J., & Ferguson, R. (2021). Fine grinding is needed to maintain the high accuracy of mid-infrared diffuse reflectance spectroscopy for soil property estimation. *Soil Science Society of America Journal*, *85*, 263–272.

Wold, S., Sjöström, M., & Eriksson, L. (2001). Pls-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 109–130.

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*, 1–17.

Xu, H., Demetriades, A., Reimann, C., Jiménez, J. J., Filser, J., Zhang, C., & Team GEMAS Project. (2019). Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-Central Europe using hot spot analysis based on GEMAS project data. *Science of the Total Environment*, *678*, 94–104.

Zhong, L., Guo, X., Xu, Z., & Ding, M. (2021). Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. *Geoderma*, *402*, 115366.

How to cite this article: Wadoux, A.M.J-C. (2023). Interpretable spectroscopic modelling of soil with machine learning. *European Journal of Soil Science*, *74*(3), e13370. https://doi.org/10.1111/ejss.13370