European Journal of Soil Science

WILEY

RESEARCH ARTICLE OPEN ACCESS

Mid-Infrared Spectroscopy to Estimate Ratios of Soil Organic Carbon to Clay

Jonah M. Prout[1] | Timo S. Breure[2] | Keith D. Shepherd[3] | Gifty E. Acquah[4] | Franklin D. Hewlett[1] | Steve P. McGrath[1] | Guy J. D. Kirk[5] | Stephan M. Haefele[1]

[1]Soil Health and Management, Rothamsted Research, Harpenden, Hertfordshire, UK | [2]European Commission, Joint Research Centre (JRC), Ispra, Italy | [3]Innovative Solutions for Decision Agriculture (iSDA), Harpenden, Hertfordshire, UK | [4]Bruker AXS LLC, Madison, Wisconsin, USA | [5]Faculty of Engineering and Applied Sciences, Cranfield University, Cranfield, Bedford, UK

Correspondence: Jonah M. Prout (jonah.prout@rothamsted.ac.uk)

Received: 8 April 2026 | **Revised:** 28 April 2026 | **Accepted:** 8 May 2026

Keywords: estimation uncertainty | MIR spectroscopy | soil clay | soil organic carbon

ABSTRACT

The index based on the ratio of soil organic carbon to clay concentration (SOC/clay) has been confirmed as an effective tool for assessing the organic matter status of mineral soils. The index contains four classes: very good (SOC/clay ≥ 1/8), good (1/10–1/8), moderate (1/13–1/10) and degraded (≤ 1/13). Conventional analyses of SOC and clay concentrations are resource intensive, however. Here, we investigate the use of mid-infrared spectroscopy (MIRS) as an alternative for quantifying SOC/clay ratios and the index. We use data from the National Soil Inventory of England and Wales ($n = 1301$). We obtained RMSE values of 3.6 and 57 g kg^{-1} for estimations of SOC and clay concentrations by MIRS, respectively. We found that a simplified three-class index was satisfactory for predictions using MIRS. The accuracy of clay concentration estimations was a limiting factor, particularly for the combined middle class (approx. 60%–70% success). Over 80% of the samples in the other two classes were correctly predicted. The probability that a sample was in the estimated index class can be used to filter out estimates likely to be misclassified. The filtering removed at least twice the number of misclassified samples as the number of correctly classified that had below-threshold probability. We found that uncertainty estimates are more informative if calculated from the component estimates of SOC and clay rather than the SOC/clay ratio. An important source of uncertainty is the mismatch between MIRS estimates of clay concentration based on clay mineralogy and conventional estimates based on particle size fraction. The additional information on clay mineralogy given by MIRS means it provides a better characterisation of SOC–clay interactions affecting soil functioning.

1 | Introduction

Monitoring and assessment of soils is important for managing land and for achieving environmental goals. In previous work we have shown the value of an index of soil organic matter status based on the ratio of soil organic carbon concentration to clay concentration (SOC/clay) for soils across England and Wales (Prout et al. 2021, 2022). The SOC/clay index is a simple indicator and has three thresholds defining four classes (Table 1). The Soil Monitoring and Resilience Directive (COM 2025/2360) lists the SOC/clay ratio as an indicator of soil organic matter status. Consequently, recent literature discusses the applicability of the SOC/clay index as an indicator of soil status across a wider geographic range and varying soil types, stating that it is not generalisable across Europe (Poeplau and Don 2023; Rabot et al. 2024; Sauzet et al. 2024).

Despite its simplicity, the costs and speed of measuring SOC and clay concentrations using conventional "wet lab" methods may be prohibitive for large-scale applications. In this paper we

J. M. Prout and T. S. Breure contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *European Journal of Soil Science* published by John Wiley & Sons Ltd on behalf of British Society of Soil Science.

Highlights

- We assess MIRS methodologies to estimate the SOC/clay index and its respective classes.
- MIRS would have time/cost benefits over conventional analyses for assessing SOC/clay index.
- We recommend a modified three-class SOC index for improved estimation by MIRS.
- Probability that an estimate belongs to an index class can suggest targeted conventional analyses.

explore the use of dry spectral methods for estimating SOC and clay concentrations on the basis that SOC and clay have distinct absorption features in the infrared-region (McCarty et al. 2002; Reeves 2010; Hermansen et al. 2016). Dry spectral methods, such as mid-infrared spectroscopy (MIRS), have the advantages that they require less sample preparation and labour than conventional methods, have high-throughput capabilities and can simultaneously determine several other soil properties (Nocita et al. 2015).

The use of MIRS to measure SOC concentration is well established and its accuracy has been shown to depend on the wet lab method (Viscarra Rossel et al. 2006, and references therein). The effectiveness of the SOC/clay index rests on how well clay concentration—to date, measured by particle size analysis—gauges SOC protection by its interaction with clay and the beneficial effects of SOC–clay interactions on soil structure. In fact these depend not only on clay concentration but also on its mineralogy and effective surface area. Further, particle size analysis, whether by the hydrometer method or laser diffraction, relies on the assumption that clay particles have spherical geometry and that the clay fraction is completely dispersed from organic matter (Thomas et al. 2021). MIRS gives a more specific measure of clay concentration through information on clay mineralogy (Madejová 2003; Tinti et al. 2015; Huang et al. 2025). We note this richer information could be exploited in improving the SOC/clay ratio index as a measure of soil organic matter status. However, the present study is focused on MIRS as a proxy for clay concentration measured by wet lab analysis. A challenge in using MIRS for wet-lab clay concentration arises from the basic mismatch between what is detected by MIRS and wet lab methods, and this introduces uncertainty. In principle, it would be possible to allow for mineralogy in the MIRS calibrations for wet-lab clay concentration given the necessary independent data on mineralogy. However, mineralogy is generally not measured in soil surveys and is not included in the National Soil Inventory of England and Wales (NSI) we used in this study. Despite these limitations, strong correlations between estimates of clay concentration by MIRS and by particle size analysis are reported (Terra et al. 2015; Dangal et al. 2019; Sanderman et al. 2020).

The accuracy and precision of MIRS for our purposes depend on creating a calibration model from MIR spectra against results of conventional analyses. Calibration can be done at a site-specific scale. But to save time and effort, spectral libraries developed for a range of soils can be used instead of or as supplements to local calibration data. Such spectral libraries have been developed globally for a range of soil properties, mostly using visible–near infrared spectroscopy (vis–NIRS) (Brown et al. 2006; Stevens et al. 2013; Viscarra Rossel et al. 2016) but also MIRS (Clairotte et al. 2016; Dangal et al. 2019; Grinand et al. 2012; Sanderman et al. 2020).

We emphasise the importance of quantifying uncertainties. If uncertainties are excessive it is necessary to supplement datasets with additional measurements from conventional lab analyses (Dangal et al. 2019). We use partial least squares regression (PLSR) and random forest (RF) models to estimate the values and uncertainties in SOC and clay concentrations and the SOC/clay ratio. Both methods can be used to estimate SOC and clay concentrations separately and hence SOC/clay index class. But RF can also be used to estimate the categorical index class directly.

The objectives of the work reported here were to test (1) if estimations of SOC/clay from MIRS calibrations are sufficiently accurate to classify soils using the index thresholds and (2) whether the associated uncertainty of the index classifications was informative for deciding which samples should be re-analysed by conventional laboratory methods.

2 | Materials and Methods

2.1 | Soil Samples and Data

2.1.1 | National Soil Inventory of England and Wales Soil Spectral Library

The National Soil Inventory (NSI) was first sampled from 1978 to 1983 on a 5-km grid across England and Wales (Hallett et al. 2017; McGrath and Loveland 1992). The litter layer was removed and soils were then sampled to 15 cm depth. Clay concentration (mineral particles $< 2\,\mu m$) was measured using the sieve–pipette

TABLE 1 | The SOC/clay index of soil organic matter status (after Prout et al. 2021).

	x = SOC/clay ratio		
Index class	**Fraction**	**Decimal**	**Description**
Very Good	$x \geq 1/8$	$x \geq 0.125$	Good structure likely. Exceeding the likely capacity for SOC–mineral interaction; a proportion is particulate SOC
Good	$1/10 \leq x < 1/8$	$0.1 \leq x < 0.125$	Low risk of poor structure. Achieving the likely capacity for SOC–mineral interaction
Moderate	$1/13 \leq x < 1/10$	$0.0769 \leq x < 0.1$	Moderate risk of poor structure. Potential for SOC storage
Degraded	$x < 1/13$	$x < 0.0769$	High risk of poor structure. Higher potential for SOC storage

TABLE 2 | Summary of soil properties in the spectral library.

Type	*n*	Property[a,b]	Unit	Mean	Med.	Minimum	Maximum	25% quantile	75% quantile
Calibration	956	Clay	$g\,kg^{-1}$	260.58	246.50	4	879	163	338.25
		SOC	$g\,kg^{-1}$	41.63	35.73	3.23	118.13	22.16	55.46
		SOC/Clay	—	0.26	0.14	0.02	11.42	0.09	0.26
Validation	319	Clay	$g\,kg^{-1}$	268.64	253	10	704	188	338
		SOC	$g\,kg^{-1}$	38.55	34.29	8.13	114.66	24.77	47.71
		SOC/Clay	—	0.19	0.14	0.04	3.28	0.09	0.20

[a]Summary statistics for all properties were calculated before transforming as described in Spectroscopy—pre-processing of soil data and spectral data.
[b]Summary statistics for SOC/clay were calculated using individual SOC/clay data not from summary statistics of SOC divided by summary statistics of clay.

method on $<2\,mm$ (peroxide-treated) soil samples (Avery and Bascomb 1974). This method for particle size analysis does not consider the mineralogy of the particles in the size classes. SOC was originally measured using a modified Walkley–Black method (Kalembasa and Jenkinson 1973), but for consistency with modern laboratory methods we re-analysed the SOC of a subset of archived ground NSI soil samples, held at Rothamsted Research, by the Dumas method (LECO CN628 Combustion Analyser, LECO Corporation, St Joseph, Michigan, USA). For samples with pH >7, SOC was obtained by subtracting inorganic C (determined by Primacs, Skalar Analytical BV, Breda, Netherlands) from Dumas total C. The selection steps were as follows:

From soils with SOC $\leq 200\,g\,kg^{-1}$ (by Walkley–Black) 1500 samples were identified using a Kennard–Stone algorithm on spectra that had the intended pre-processing applied (see Spectroscopy). For some samples selected for re-analysis there was insufficient soil in the archive. Hence, the algorithm was run to identify 1800 samples (including the original 1500), providing a more directed way to select supplementary samples (preferentially by previously measured SOC, pH and major soil group). For the analyses presented here samples with SOC $>120\,g\,kg^{-1}$ or without soil texture values or both were excluded, resulting in 1301 samples. Therefore, we have clay concentration from the NSI data (held by Cranfield University; Hallett et al. 2017) and SOC concentration from our new analyses (linked by unique identifiers). These are referred to from now as the "conventional" data set (including SOC_{Con}, $clay_{Con}$ and $(SOC/clay)_{Con}$). Table 2 gives SOC, clay, SOC/clay summary statistics for this set.

2.2 | Spectroscopy

2.2.1 | Sample Preparation and Spectral Data Collection

Soil samples, ground to $<100\,\mu m$ for wet chemistry analysis (McGrath and Loveland 1992), were obtained from the Rothamsted Research Sample Archive. Prior to plate-filling, sample bottles with lids removed were placed in an oven at 40°C for 8 h. Samples were placed in duplicates in 96-well plates. The first six wells of each plate as gold-reference cap, blank and two internal reference soils in duplicate. A Bruker Tensor II spectrometer (Bruker Scientific, Berlin, Germany) with an HTS-XT high-throughput accessory was used to acquire the spectra. The Bruker Tensor II has a humidity sensor to minimise the effect of water vapour on measurements; we kept the humidity to 10% as per the instrument guidelines. Diffuse Reflectance Infrared Fourier Transform (DRIFT) spectra were collected at $1.43\,cm^{-1}$ with a scan time of 32 s per sample. We note that DRIFT spectra were collected on undiluted samples (i.e., without the addition of Kbr), and thus our results should be interpreted in context of the spectral distortion due to band inversions for frequencies below about $1150\,cm^{-1}$ (Nguyen et al. 1991). Technical replicates were averaged, and the spectra were resampled over $4000–600\,cm^{-1}$ at $2\,cm^{-1}$ resolution by using the *prospectr* package in R (Stevens and Ramirez-Lopez 2022).

2.2.2 | Pre-Processing of Soil Data and Spectral Data

The SOC, clay, SOC/clay data were Box–Cox-transformed to better approximate a normal distribution. Lambda values for SOC, clay and SOC/clay were 0.1583, 0.5832 and −0.3284 respectively. Regions showing atmospheric CO_2 features ($2389–2268\,cm^{-1}$; Dangal et al. 2019) were removed as a precautionary measure. A first derivative Savitzky–Golay filter (window size = 21, polynomial = 3) was applied to the spectra. Two per cent of the samples (and corresponding spectra) were removed based on the residuals of 10-fold cross-validation to estimate SOC and clay (1% with the highest absolute residuals for each property). This attempted to account for potential error in data acquisition, in line with Sanderman et al. (2020).

2.2.3 | Calibration Regression

We used partial least squares regression (PLSR) and random forests (RF) to build spectral calibrations. A Kennard–Stone algorithm was applied to the spectral data to split the data into calibration (75%) and validation (25%) sets. The minimum number of principal components that explained 95% of the variance in the spectra was selected for Kennard–Stone splitting. For the PLSR, the number of components each model was run with was selected using the one-sigma approach (Hastie et al. 2013), and we considered a maximum of 15 components. This resulted in models with 14 components for SOC, six for clay and seven for SOC/clay. For the RF regression, we included the same principal components that were used for the Kennard–Stone sampling ($n = 22$) as predictors. Similarly to the PLSR, we predicted Box–Cox transformed SOC, clay and SOC/clay. We also performed a 10-fold cross-validation to select the best set of hyperparameters, using a grid search. We considered the following

hyperparameter ranges: number of trees (150–550), mtry (2–15), minimum node size (1–10) and maximum depth (5–40). We selected the hyperparameter values that showed the lowest root mean squared error (RMSE) on average across the 10 folds. We also tested RF as a classifier by predicting SOC/clay class probabilities. The same methodology as above was applied, except that we selected the highest classification accuracy on average across the 10 folds. For the SOC and clay RF regressions, the following hyperparameters were selected: number of trees (450), mtry (15), minimum node size (1) and maximum depth (40). For SOC/clay these were: number of trees (150), mtry (9), minimum node size (1) and maximum depth (20). For the direct 4 class classifications: number of trees (350), mtry (15), minimum node size (5) and maximum depth (40), and the direct three class classifications these were: number of trees (250), mtry (15), minimum node size (5) and maximum depth (10).

The estimated SOC, clay and SOC/clay values were back-transformed prior to statistical comparison with the back-transformed conventional data. We compared three different ways to estimate SOC/clay against the respective conventional values: (1) we estimated the decimal fraction from the MIR spectra ($(SOC/clay)_{Est}$), (2) we calculated the decimal fraction from the estimated SOC and clay ($SOC_{Est}/clay_{Est}$) and (3) we calculated the decimal fraction using estimated SOC and conventionally measured clay ($SOC_{Est}/clay_{Con}$). The latter method could represent a scenario in which soils have been revisited, and the clay concentration measured conventionally in the past is not expected to have changed. This also allowed us to see the effect that uncertainties associated with clay concentration had on the estimations of SOC/clay.

2.2.4 | Evaluation of Model Performance

We assessed the quantitative performance of the models using mean estimation metrics based on the root mean squared error (RMSE), *R*-squared (R^2), bias, Lin's concordance correlation coefficient (CCC) (Lin 1989); and the ratio of performance to interquartile distance (RPIQ) (Bellon-Maurel et al. 2010). We assessed the categorical index class (with respect to the thresholds in Table 1) from the proportion of observations falling in the corresponding index class range using chi-square tests and standardised residuals.

We also quantified the marginal uncertainties associated with each estimation. Given that the observations within the calibration set were selected using the Kennard–Stone sampling method, our calibration regression is controlled by design (Wehrens et al. 2000). Hence, we used the non-parametric bootstrapping by residuals method to estimate the uncertainty of PLSR estimations (Zhang 2014). The details are in the Supporting Information. The bootstrap was iterated 1000 times: the standardised residuals were sampled with replacement and added to the PLSR estimations. A PLSR was performed on this new response vector and the coefficients of this model were stored. The variance of the coefficients was considered as the estimated variance for each observation. For the random forest regression, we also performed bootstrapping by residuals and considered the variance in predictions as the estimated variance for each observation.

Estimations of SOC and clay based on MIR spectra each have associated estimation error variances. In computing the $SOC_{Est}/clay_{Est}$ and $SOC_{Est}/clay_{Con}$ indices, we propagated these errors into an estimation variance for the index as follows for each observation $j = 1, \ldots, n$:

$$\sigma_v = \frac{\widehat{SOC}}{\widehat{Clay}} \sqrt{\left(\frac{\sigma_{\widehat{SOC}}}{\widehat{SOC}}\right)^2 + \left(\frac{\sigma_{\widehat{Clay}}}{\widehat{Clay}}\right)^2} \quad (1)$$

The total error variances for each observation in the validation set were then computed by adding the bias variance:

$$\sigma_{tj} = \sigma_{vj} + \sigma_{y-\hat{y}} \quad \text{for} \quad j = 1, \ldots, n. \quad (2)$$

Lastly, we assessed the value of the bootstrapped estimation error variances from each method in the context of the SOC/clay index classes. That is, we computed the probability that an estimation belongs to its SOC/clay index class based on the cumulative distribution functions belonging to the estimated mean and variance:

$$\delta = \int_{-\infty}^{th_2} f\left(\hat{y}_j, \sqrt{\sigma_{vj}}\right) - \int_{-\infty}^{th_1} f\left(\hat{y}_j, \sqrt{\sigma_{vj}}\right) \quad (3)$$

where th_1 takes the minimum SOC/clay threshold value and th_2 the upper SOC/clay threshold for a SOC/clay index class (as per Table 1). For example, in the case of the Degraded SOC/clay index class, $th_1 = 0$ and $th_2 = 1/13$. Results for the estimated class of samples can then be assessed against how certain we want to be about a sample, its specific probability threshold. Filtering based on the uncertainty of an estimate allows the analyst to decide whether IR spectroscopy provides a satisfactory result. In applied settings, the estimate uncertainty is the only measure from which to determine whether it is necessary that the sample should be reanalysed by conventional laboratory methods (Dangal et al. 2019). We considered probability thresholds ranging from 0.6 to 0.99 probability (where 1 equals 100%) and evaluated the effect of filtering based on this threshold. That is, observations in the validation set were removed when they were uncertain (i.e., below the set probability threshold). We computed the total matches and mismatches in the validation set based on the conventionally determined index class. We then evaluated what percentage of the matches and mismatches was filtered by applying the set probability threshold. Lastly, we evaluated the ratio of mismatches/matches in the filtered data. This ratio indicates how informative estimation uncertainties are in filtering out index estimates that do not match the conventional index class.

2.2.5 | Software

For data handling and visualisations, we used the *tidyverse* (Wickham et al. 2019) and *ggplot2* (Wickham 2016) packages in R. For the analyses we used the following R packages: *pls* (Liland et al. 2022), *stats* (R Core Team 2021), *epiR* (Stevenson et al. 2023), *resemble* (Ramirez-Lopez et al. 2022) and *prospectr* (Stevens and Ramirez-Lopez 2022), and *ranger* (Wright and Ziegler 2017).

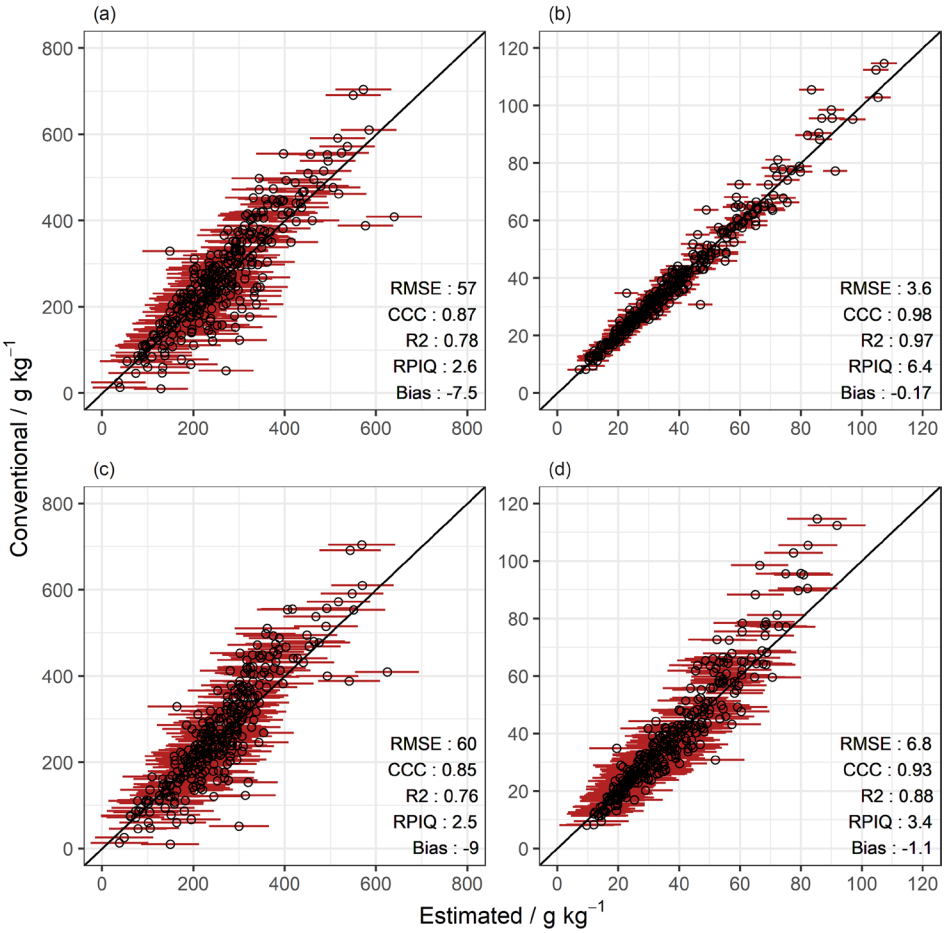


FIGURE 1 | Estimations of clay (a) and SOC (b) using partial least squares regression (PLSR) on a validation holdout set ($n = 319$) and (c) and (d) correspond to clay and SOC predictions by random forest (RF). Horizontal lines ($^-$) are the total estimation error variances (Equation 2).

3 | Results

3.1 | SOC and Clay Estimation

SOC was estimated well by PLSR (RMSE = 3.6 g kg^{-1} and CCC = 0.98) (Figure 1b) and less so by RF (RMSE = 6.8 g kg^{-1} and CCC = 0.93) (Figure 1d). Clay had larger total error variances compared to SOC for both PLSR and RF but still had reasonable estimation metrics (RMSE = 57 and 60 g kg^{-1} and CCC = 0.87 and 0.85 for PLSR and RF, respectively) (Figure 1a,c). The PLSR loadings for the SOC and clay models are included in Figure S1.

3.2 | SOC/Clay Estimation

Mean estimation metrics for the direct estimation of SOC/clay value ($(SOC/clay)_{Est}$) and calculation using separate estimated SOC and clay values ($SOC_{Est}/clay_{Est}$) were almost the same for PLSR (Figure 2a,b). For RF, there were slightly better results for ($SOC_{Est}/clay_{Est}$) than for $(SOC/clay)_{Est}$ but the differences were small (Figure 2c,d). Considering the index thresholds, the SOC/clay ranges of the Good and Moderate classes are 0.025 and 0.023. So, the RMSE values for SOC/clay were approximately eight times greater than these ranges. Using estimated SOC with the conventionally measured clay value improved estimation metrics with PLSR such that the RMSE of $SOC_{Est}/clay_{Con}$ was less than the range of Good or Moderate class. With RF, however, the RMSE remained approximately twice these ranges.

The index class with values furthest from 0 (Very Good) has an open-ended range; there is thus a wider margin for estimation error to correctly classify a sample in this class. We calculated the performance statistics for soils which had conventionally measured SOC/clay below this upper threshold (SOC/clay < 0.125) to assess the effect of removing these samples (Table 3). The RMSE values using either model and any of the methods reduced to less than 0.023 and the CCC values increased (except for $SOC_{Est}/clay_{Con}$ for which CCC was already close to or equal to 1).

3.3 | Index Class Determination

For both PLSR and RF models, the agreement between the estimated index class ($Index_{Est}$) of an observation and its conventional index class ($Index_{Con}$) was highest for the Very Good class, followed by Degraded, Moderate and Good classes (Table 4). This was the trend for all three SOC/clay determination methods, except $SOC_{Est}/clay_{Est}$ with RF. The poorer agreement for Good and Moderate classes was likely due to their smaller ranges. A combined Good–Moderate class was created which gave agreement results more in line with the Very Good and Degraded

classes. For $(SOC/clay)_{Est}$, $SOC_{Est}/clay_{Est}$ and $SOC_{Est}/clay_{Con}$, respectively, the percentage of observations correctly estimated for this combined class were: 68%, 70% and 88% with PLSR and 66%, 66% and 75% with RF, respectively. As expected, using conventionally measured clay concentration gave the highest agreement. RF also allows for a regression directly against the SOC/clay index classes (RF *Index* in Table 4). This showed improved performance for the highest and lowest class but poor performance for Good or Moderate. Regression against the 3-class index gave correct estimations of: 90% for the Very Good, 67% for the Good–Moderate class and 67% for the Degraded class.

The standardised residuals from chi-squared analyses, showed highest positive values (greatest agreement) for the corresponding estimated and conventional classes (Figure 3). There were three exceptions to this between the Good and Moderate classes (Figure 3c) and the Moderate and Degraded classes (Figure 3b,d). The sign and magnitude of the standardised residuals suggested that misclassifications were more likely with adjacent classes. The highest contrasts were achieved with the $SOC_{Est}/clay_{Con}$ method. There looked to be a closer relationship between the Good and Moderate classes (e.g., the above noted exception in the $SOC_{Est}/clay_{Est}$ method) than other class combinations and combining them into a single class (Good–Moderate) gave a clearer separation (Figure S2). Similar trends resulted from the direct estimation of categorical index class by RF (Figure S3). Based on these findings, and the improved estimation percentages, we used the combined Good–Moderate class for our subsequent analyses. Additionally, based on the mean prediction metrics and chi-squared analysis, it was clear that RF underperformed compared to PLSR. For a more

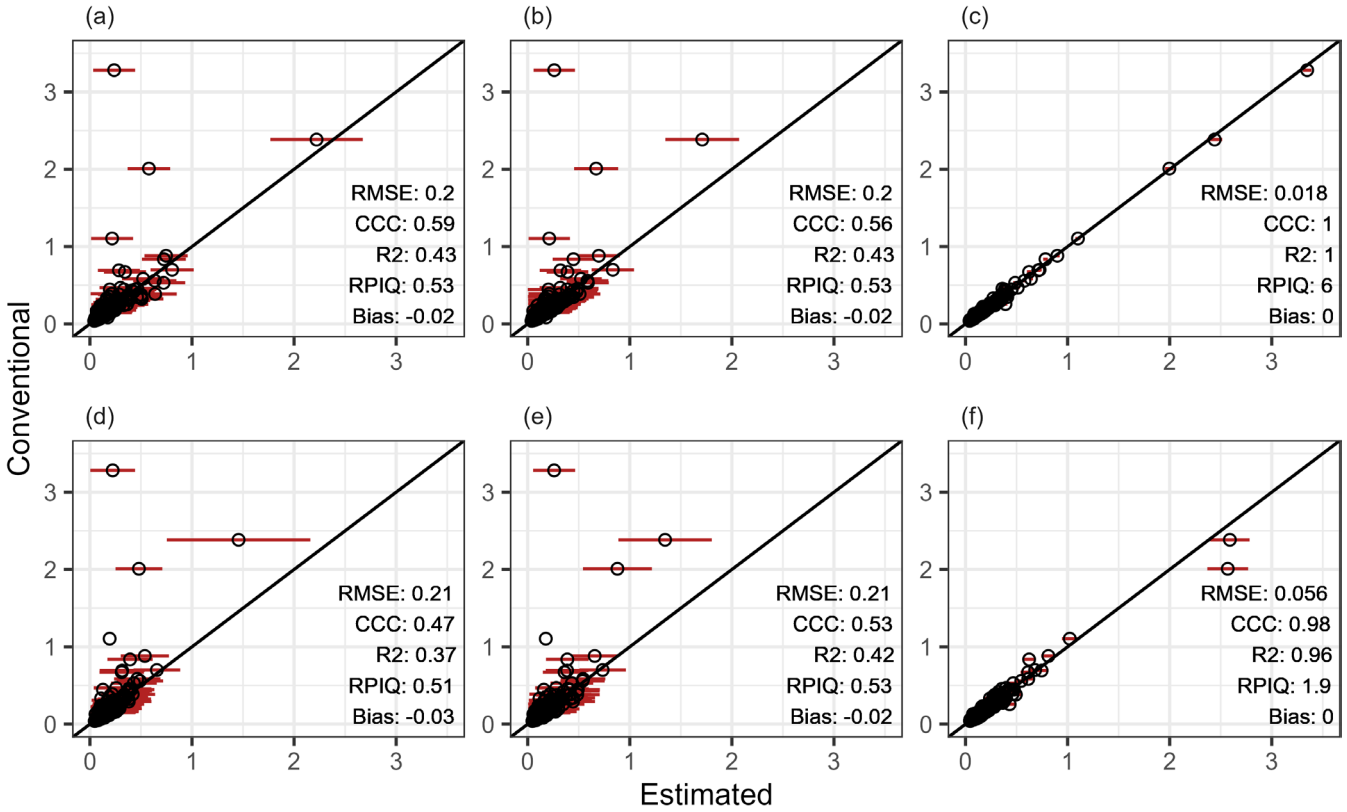


FIGURE 2 | Estimations of partial least squares regression (top-row) and random forest (bottom-row) on a validation holdout set ($n = 319$). Columns correspond to estimation methods: $(SOC/clay)_{Est}$ (a and d), $SOC_{Est}/clay_{Est}$ (b and e) and $SOC_{Est}/clay_{Con}$ (c and f). Horizontal bars (⁻) are the total estimation error variances (Equation 2).

TABLE 3 | Mean estimation metrics for SOC/clay by partial least squares regression (PLSR) and random forest (RF) regression. Metrics are for a subset of the validation set where $(SOC/clay)_{Con} < 0.125$ ($n = 144$).

Model	Estimation method	RMSE	CCC	R^2	Bias	RPIQ
PLSR	$(SOC/clay)_{Est}$	0.018	0.76	0.34	0.01	2.1
	$SOC_{Est}/clay_{Est}$	0.019	0.76	0.29	0.01	2.0
	$SOC_{Est}/clay_{Con}$	0.007	0.95	0.90	0.00	5.4
RF	$(SOC/clay)_{Est}$	0.020	0.71	0.17	0.01	1.9
	$SOC_{Est}/clay_{Est}$	0.021	0.70	0.10	0.01	1.8
	$SOC_{Est}/clay_{Con}$	0.014	0.82	0.57	0.00	2.6

TABLE 4 | Percentage of validation set observations estimated ($Index_{Est}$) to be in the same index class as when using the conventional data ($Index_{Con}$).

Method	$Index_{Con}$	n	$Index_{Est} = Index_{Con}$ (%)			
			$(SOC/clay)_{Est}$	$SOC_{Est}/clay_{Est}$	$SOC_{Est}/clay_{Con}$	RF Index
PLSR	All	319	74	75	90	—
	Very Good	175	86	88	98	—
	Good	53	45	36	68	—
	Moderate	40	55	60	82	—
	Degraded	51	78	80	90	—
RF	All	319	71	71	79	
	Very Good	175	91	90	92	94
	Good	53	43	47	40	10
	Moderate	40	45	40	62	30
	Degraded	51	53	57	86	80

concise presentation of the results, we have decided to present only the PLSR results for subsequent analyses.

3.4 | Index Class Probability Based on the Associated Error Variance

Figure 4 shows the probability of the estimated index classes given the marginal uncertainties associated with estimates (Equation 3). Ideally, the estimated index class (x-axis) would show high probability (maximum value of 1, y-axis) for observations from the corresponding conventional index class (indicated by colour). Inversely, the probability should be low for boxplots that do not match the x-axis. In some cases, there are only two boxplots as there were no misclassifications in the missing combination (e.g., samples in the Degraded index class estimated to be Very Good in Figure 1a). Overall, with a few exceptions, we see that the distribution of probabilities is most narrow and closest to 1 when there is agreement between conventional (boxplot colour) and estimated classes (x-axis label).

The general trend is that while moving down the graphs from (a) to (c), i.e. $(SOC/clay)_{Est}$ to $SOC_{Est}/clay_{Est,}$ to $SOC_{Est}/clay_{Con}$, the class estimates improve (larger counts for boxplots that match the x-axis and/or reduced counts where they do not match). Overall, probabilities for incorrect estimates had a median close to 1. However, in some cases the lower quartile of the boxplot extended to lower probability values for incorrect classes, for example the boxplots which were incorrectly estimated to be Good-Moderate with $SOC_{Est}/clay_{Est}$ (Figure 4b). Even within the $SOC_{Est}/clay_{Con}$ estimation method, we observed high probabilities for incorrect estimations (Figure 4c).

3.5 | Index Class Estimation Confidence Based on Probability

We used probability thresholds to filter estimations based on the confidence in the estimated index class. The overall logic of the procedure is as follows: (1) We estimate the index class of each sample in the validation set using PLSR. (2) Then, we use the associated error variance to quantify how certain we are of these estimated classes (expressed as probability). (3) Finally, we use the probability to determine which estimates are reliable and which need to be analysed by conventional methods. As a result, the validation set size decreases since a selection of these estimates is considered too uncertain (Figure 5, black lines). The minimum retention was just below 80% of the original validation set, considering a probability threshold of 0.99 and the $(SOC/clay)_{Est}$ method, though specific index classes had different retentions (Table 5). The filtering step based on uncertainty improved the percentage of matches in the validation set. The Good–Moderate class had lesser retention of samples, but this might be expected due to a narrower threshold range and two directions in which a sample could be misclassified.

From the proportion of estimates filtered by the number of matches (and mismatches) in the validation set as a whole, $(SOC/clay)_{Est}$ and $SOC_{Est}/clay_{Est}$ filtered 40% of mismatches (dashed blue line, Figure 5a,b). Estimations were also considered uncertain even though they matched the conventional index class (Figure 4) and were removed by filtering, of which the maximum was 17% (Figure 5b). For a probability threshold of 0.99, the ratios of filtered samples with respect to index class agreement (%mismatch/%match) were: $a = 2.5$, $b = 3.4$, $c = 2.6$ where a to c correspond to Figure 5. These ratios suggest that the filtering was informative since at least twice as many mismatches were filtered than matches (respective to their frequency in the validation set). The filtering performance for the $SOC_{Est}/clay_{Est}$ method (b) was better than that for the $(SOC/clay)_{Est}$ method (a) and also retained a higher percentage of samples (Table 5). A decrease in the percentage mismatch/match ratios for the $SOC_{Est}/clay_{Con}$ method showed that uncertainties associated with each estimate became less informative when the mean estimation metrics improved, although also a lower number of samples was filtered in total (Figure 5 and Table 5).

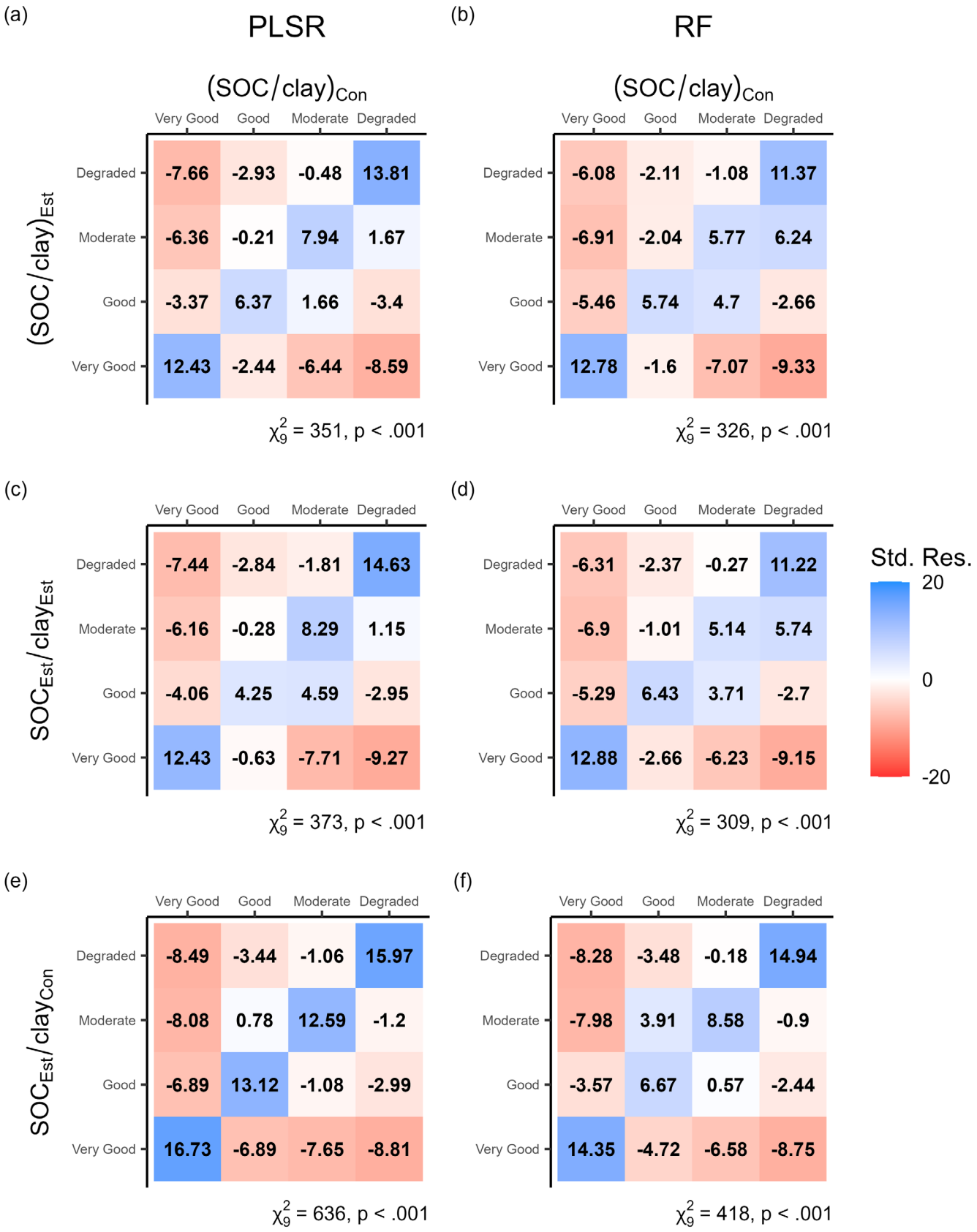


FIGURE 3 | Standardised residuals of chi-square tests for the validation set between conventional index class ($(SOC/clay)_{Con}$, *x*-axis) and estimated index class (*y*-axis) for each of the estimation methods: $(SOC/clay)_{Est}$ (a and b), $SOC_{Est}/clay_{Est}$ (c and d) and $SOC_{Est}/clay_{Con}$ (e and f). The columns correspond to the regression methods: partial least squares regression (PLSR) and random forest (RF). Positive values suggest association of estimated and conventional classes; negative values suggest a lack of association.

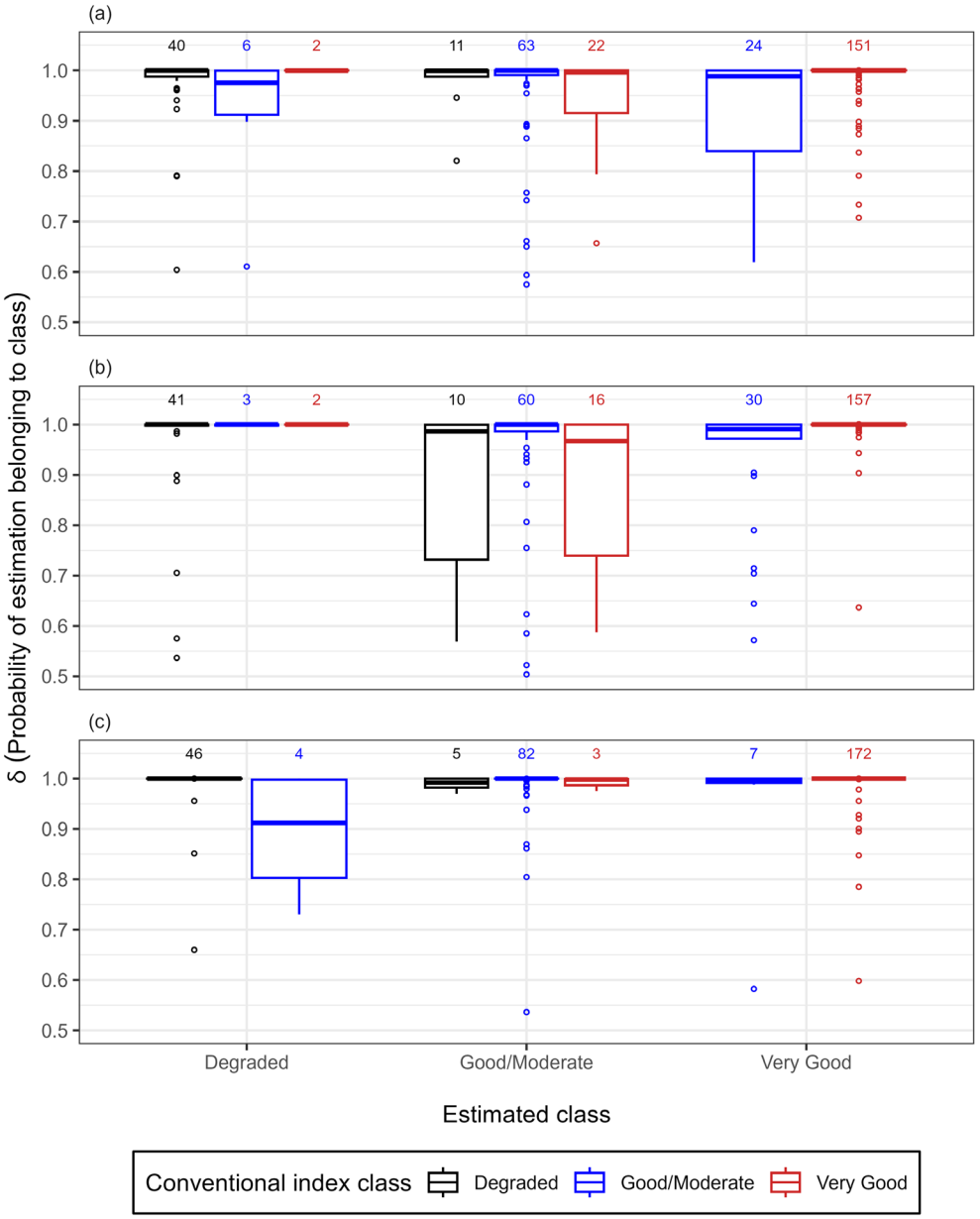


FIGURE 4 | Boxplots of the estimated index class probability (δ, Equation 3) as a function of the estimated index class, by different estimation method: $(SOC/clay)_{Est}$ (a); $SOC_{Est}/clay_{Est}$ (b) and $SOC_{Est}/clay_{Con}$ (c). Boxplots have been grouped and coloured by the index class based on conventionally measured data, with the corresponding count displayed above.

4 | Discussion

4.1 | Estimating SOC and Clay Concentrations

Estimates of SOC concentration by the conventional and MIRS methods both rely on the soil chemical composition. For clay however, the conventional method measures particle size fractions whereas MIRS measures clay mineralogy. Hence, the two methods gave comparable estimates of SOC concentration but larger discrepancy for clay concentration. Our SOC estimates and their uncertainties are in line with previous studies using MIRS. Soriano-Disla et al. (2014) reported a median R^2 of 0.93 for SOC by the two methods across 20 studies, and $R^2 = 0.80$ for clay across 14 studies. Published RMSE values for SOC tended to be higher than our PLSR SOC predictions, whereas they were comparable with RF SOC predictions (e.g., 6.4–11.9 g kg^{-1}, Dangal et al. 2019; 2–6 g kg^{-1}, Clairotte et al. 2016; 6.3–7.2 g kg^{-1}, Grinand et al. 2012). Our RMSE values for clay concentration tended to be higher than published values (e.g., 24.7–39.5 g kg^{-1}, Dangal et al. 2019; 28.3–42.1 g kg^{-1}, Sanderman et al. 2020; 84.5 g kg^{-1}, Terra et al. 2015).

Clay-sized minerals have distinct absorption features in the MIR region, so clay concentration is generally accurately estimated by MIRS (Madejová 2003; Tinti et al. 2015; Huang et al. 2025). However, estimates of clay concentration by particle size analysis will miss clay minerals in aggregates exceeding 2 µm diameter and will include non-clay minerals such as quartz and carbonate (as for example in Figure S1). The carbonate peak at 2510 cm^{-1} (Comstock et al. 2019) is distinctive in our samples (Figure S4) and seems to have some significance for the clay and SOC/clay estimations.

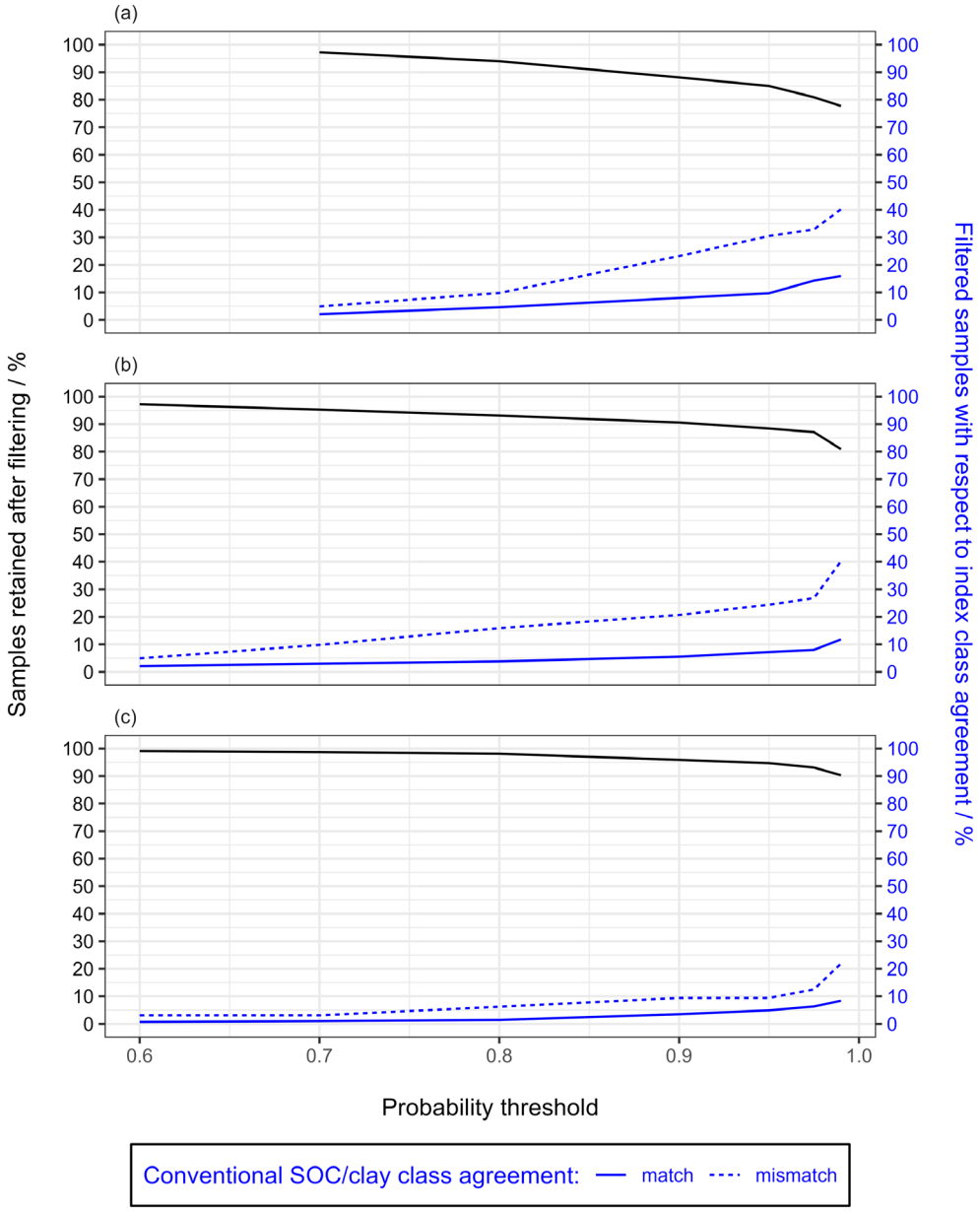


FIGURE 5 | Percentage of samples from the validation set retained after filtering based on the estimated SOC/clay class probability threshold (left *y*-axis, black line) and the filtered samples as a percentage of the number of samples with their respective index class agreement status (right *y*-axis, blue lines). Rows correspond to index estimation methods: $(SOC/clay)_{Est}$ (a), $SOC_{Est}/clay_{Est}$ (b), and $SOC_{Est}/clay_{Con}$ (c).

TABLE 5 | Percentage of validation set observations estimated ($Index_{Est}$) to be in the same index class as when using the conventional data ($Index_{Con}$).

			$Index_{Est} = Index_{Con}$ (%)		
$(SOC/clay)_{Con}$	*n*	% retained	$(SOC/clay)_{Est}$	$SOC_{Est}/clay_{Est}$	$SOC_{Est}/clay_{Con}$
All	319	78, 81, 90	86 (+12)	88 (+13)	96 (+6)
Very Good	175	85, 90, 94	91 (+5)	94 (+6)	99 (+1)
Good–Moderate	93	67, 67, 83	77 (+9)	71 (+1)	91 (+3)
Degraded	51	73, 75, 90	81 (+3)	90 (+10)	94 (+4)

Note: Numbers in parentheses are the difference in the value between pre- (Table 4) and post-filtering based on a 0.99 probability (Equation 3). As the marginal uncertainties differ between the $(SOC/clay)_{Est}$, $SOC_{Est}/clay_{Est}$, and $SOC_{Est}/clay_{Con}$ methods, different percentages of the original number of samples in the validation set (*n*) were retained (listed in column order left to right, % retained).

In the literature, X-ray diffraction (XRD) spectra have been shown to provide additional explanatory power for estimating particle size (Chen et al. 2021) and mineralogy (Huang et al. 2025) compared to infrared spectroscopy on its own. Calibrations between XRD and MIR spectra are a significant advance in soil analysis, providing a more direct measurement of mineral composition than particle size alone. This has multiple benefits through a better alignment in what is being measured (chemical bond vibrations with MIR) and the calibrated property (crystalline to amorphous phase arrangements of atoms). Furthermore, soils with the same particle size can have different biogeochemical properties, for example in the sorption behaviour and physical occlusion of carbon (Six et al. 2002; Feng et al. 2013), so accounting for mineralogy provides added value to the measurement. The ability to derive multiple soil property values from a single MIR spectrum also reduces the separation of measurements for interpretation of results. That is, with conventional laboratory methods, SOC and soil particle size analysis would be conducted using separate sample portions with different amounts required and represented.

4.2 | Estimating the SOC/Clay Indices

The relative weighting of clay concentration in the index (Poeplau and Don 2023) does mean that errors in this will have a greater impact than SOC on index estimation. However, we observed index classification rates of >80% for the highest and lowest classes in the proposed three-class index, which is relevant for more cost-effective analysis in the context of the Soil Monitoring and Resilience Directive, where the 1/13 threshold has been set as an indicator of soil organic matter status (COM 2025/2360). However, a regional adaptation of the 1/13 threshold has been advocated (Bernardini et al. 2025) and, once this methodology is established, future work could assess how this coefficient affects SOC/clay class prediction from MIRS.

Regarding previous work on spectral methods for estimating a combined SOC and clay metric, the analysis of Hermansen et al. (2016) had similar aims, using vis–NIRS spectra to predict clay and SOC to compare with a clay/SOC ratio (clay/SOC = 10 based on Dexter et al. 2008). They used a dataset containing eight agricultural fields with 24–95 samples per site and showed excellent results for SOC and clay (as a particle-size fraction) estimations. They achieved an RMSE of 0.64 for direct estimations of clay/SOC. Clay/SOC estimation with our dataset performed very similarly to the SOC/clay estimation with PLSR analysis (Sections 3.2–3.5), considering the difference in the reciprocal scale for quantitative performance (work not presented). The high performance of the Hermansen et al. (2016) dataset compared to ours is likely due to the high local sample density compared to the data we used (minimum 5 km between points). We therefore postulate that the 5 km sampling grid of the NSI could impede accurate and precise predictions at the field scale. Furthermore, with higher sampling densities, the inconsistency between particle size and MIR spectra may be masked through greater representation of mineralogy and clay concentration combinations.

Therefore, whilst our methodology is relevant to the application of soil spectroscopy at any scale, the use of the NSI at field scale might improve through 'spiking' with local samples. Spiking a library involves using a relatively small number of samples from a site with conventional data to increase the representativeness of a particular area in the model and has shown improvements for using large spectral libraries at smaller geographical scales (Brown 2007; Guerrero et al. 2010; Viscarra Rossel et al. 2009; Wetterlind and Stenberg 2010). Smaller subsets of national spectral libraries could also be an option which can be complemented by spiking. This has been shown to improve local-scale estimations, providing that analytical methodologies of the spectral library and spiking dataset are equal (Seidel et al. 2019; Breure, Prout, et al. 2022b).

Filtering index class estimates based on the probability that they were correctly assigned provides a measure of quality control. Largely, the removal of these samples improved the percentages correctly assigned, although different estimation methods also resulted in a varying number of samples filtered. With the dataset we used, and the strictest threshold, approximately 20% of samples would be considered for reanalysis. The filtering rate reduced to 10% with $SOC_{Est}/clay_{Con}$ which represents a theoretical maximum estimation performance for clay. By filtering samples in this way, proportionally more misclassified samples were filtered out than those which were correct across methods (Figure 5). After conventional analyses, these might be fed into the model, potentially improving estimations. These findings underline the potential value of uncertainty estimates and suggest that future research on improved quantification of estimation uncertainty is warranted.

4.3 | Effects of Different Methods to Estimate SOC/Clay From MIRS

As the following may have impacts for practitioners, and there are no practical reasons to use one instead of the other, we tested whether to estimate the index value directly ($(SOC/clay)_{Est}$) or calculate from the component estimates of SOC and clay ($SOC_{Est}/clay_{Est}$).

Overall, there was not a large a difference in the statistics between the two estimation methods when estimating the numeric index value. However, it may be more transparent to estimate the separate variables (SOC and clay) than their derived value. We had thought that estimating the index value directly might be better estimated due to the single model step, but this was not the case. We also thought the choice of model might have led to better predictions, but the RF method did not lead to improvements. This could be because PLSR successfully captured the multi-collinearity between the spectra and soil properties in the latent variables. Another reason to use RF would be that, as a classifier, it can provide probabilities. However, these probabilities are affected by class frequency in the calibration set and would need to be corrected (Niculescu-Mizil and Caruana 2005).

The differences in categorical index class estimation showed that $SOC_{Est}/clay_{Est}$ for the most part achieved higher agreement than $(SOC/clay)_{Est}$ (Table 5). After filtering based on probability that a sample was in the correct class, there were higher percentages of samples retained and higher estimation agreement percentages for $SOC_{Est}/clay_{Est}$ than for $(SOC/clay)_{Est}$ (the Good–Moderate

class being an exception). These findings suggest that error propagation of the component estimates of SOC and clay led to more informative uncertainties. For these reasons, our results indicate that the $SOC_{Est}/clay_{Est}$ method is the preferred method to estimate the SOC/clay index.

4.4 | Practical Implications and Future Work

From an applied soil monitoring perspective, further research should compare the estimation of a given index class accounting for uncertainty of both conventional methods as well as MIRS. A level of measurement variation occurs in conventional laboratory methods although this variation is seldom reported (Pleijsier 1989; Batjes et al. 2020; van Leeuwen et al. 2022). Sanderman et al. (2021) showed that a large US spectral library could be used to achieve monitoring of SOC comparable to conventional lab techniques for several long-term experiments in the USA. Furthermore, relative performance between models changed when European-scale SOC predictions were considered in the context of the laboratory error (Breure et al. 2026).

Our results also illustrate the advantages of MIRS for routine analysis of soil mineral composition. Compared with particle-size analysis, MIR spectra provide a more accurate representation of the organo-mineral interactions that are the basis of the SOC/clay index (Six et al. 2002; Dexter et al. 2008; Schjønning et al. 2012). Calibration of MIRS to mineralogy data (e.g., from XRD) would allow for the characterisation of mineral activity in large datasets, such as the NSI, while capturing other spectrally-active properties. This would also add value to existing spectral libraries, e.g., the Open Soil Spectral Library (OSSL; Safanelli et al. 2025).

We showed that the probability based on the associated error variance has practical value to determine how certain we are of a SOC/clay class estimation. Further work could assess how uncertainty in MIRS estimates affects the cost-effectiveness of using soil spectroscopy to predict SOC/clay indices under different conditions (Li et al. 2022). For example, for our case, 20% of validation samples were filtered based on their uncertainty and we could assume that MIRS are 1/5 of the conventional laboratory costs (Li et al. 2022). This means that these 20% of the validation set are analysed by both MIRS and subsequently by conventional analysis, whereas 80% are by MIRS only. The cost savings would be 60% compared to analysing the validation set only by conventional laboratory methods. However, this disregards the costs of building the spectral library. Also, when samples are selected for re-analysis, the costs need to be weighed against improvements to calibration models that such samples can provide. Furthermore, misclassifications might come with associated costs as well (Breure, Haefele, et al. 2022a). Overall, in an applied MIRS context, we show that the associated uncertainty in estimates is an important criterion compared to improved prediction metrics.

Lastly, decision-making based on soil monitoring data is often based on categorical classes calculated from continuous values. Taking again the Soil Monitoring Resilience Directive as an example, for each soil property there are proposed thresholds to assess whether a soil is degraded. For example, <4 deci-Siemens m^{-1} to determine whether a soil is affected by salinisation when analysed by the saturated soil paste extract (eEC) method (COM 2025/2360). Thus, research on comparing conventional and MIRS for index class estimation, together with their associated uncertainties, can also be generalised to other categorical soil indicator metrics.

5 | Conclusions

MIRS estimations of the SOC/clay index suggest that a practical framework to identify soils in need of improvement or protection could be based on a modified three-class index. The three classes are Very Good, Good–Moderate and Degraded where the middle threshold separating Good and Moderate classes (SOC/clay = 1/10) is removed. There did not seem to be a net benefit to directly estimating the index compared to computing the value from the component estimates of SOC and clay concentrations. We could correctly classify at least 80% of the Very Good and Degraded soils (with approximately 70% for Good–Moderate). The estimation error for clay concentration seemed to be the driver of the incorrect estimations. This may be because of the discrepancy in calibrating clay concentration measured by particle size analysis to the MIR spectra, which reflect the chemistry of a sample. Calibration of MIRS to soil mineralogy would allow the additional information provided by MIRS to be exploited in an improved SOC/clay index which better reflects the SOC–mineral interactions affecting SOC dynamics and soil physical properties. We evaluated to what extent marginal uncertainties could be used to filter out samples that did not match the conventional index class. Filtering suggested re-measurement of approximately 10%–20% of the validation set. Relative to their frequency in the validation set, at least twice as many misclassified samples compared to correctly classified samples were filtered. Having filtered the validation set, the classification rates were improved. The filtering was more informative when uncertainty was calculated from the component estimates of SOC and clay. Our methodological framework is generalisable to other categorical soil indicator metrics to be predicted from infrared spectra.

Author Contributions

Jonah M. Prout: conceptualization, methodology, software, data curation, investigation, validation, formal analysis, visualization, writing – original draft, writing – review and editing. **Timo S. Breure:** conceptualization, methodology, software, investigation, validation, formal analysis, visualization, writing – original draft, writing – review and editing. **Keith D. Shepherd:** conceptualization, methodology, supervision, funding acquisition, writing – review and editing. **Gifty E. Acquah:** conceptualization, methodology, investigation, writing – review and editing. **Franklin D. Hewlett:** methodology, investigation. **Steve P. McGrath:** conceptualization, supervision, funding acquisition, project administration, writing – review and editing. **Guy J. D. Kirk:** conceptualization, supervision, funding acquisition, project administration, resources, writing – review and editing. **Stephan M. Haefele:** conceptualization, supervision, funding acquisition, project administration, resources, writing – review and editing.

Acknowledgements

Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom. We acknowledge support from the Growing Health

[BB/X010953/1; BBS/E/RH/230003B; BBS/E/RH/230003C] and Soils to Nutrition [BBS/E/C/000I0310] Institute Strategic Programmes.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data originated from Cranfield University, Land Information System LandIS Portal National Soil Inventory as Public Data 2026, subject to the LandIS Portal Open Licence.

References

Avery, B. W., and C. L. Bascomb. 1974. *Soil Survey Laboratory Methods*. Rothamsted Experimental Station, Technical Monograph (Soil Survey of England and Wales) no. 6.

Batjes, N. H., E. Ribeiro, and A. Van Oostrum. 2020. "Standardised Soil Profile Data to Support Global Mapping and Modelling (WoSIS Snapshot 2019)." *Earth System Science Data* 12, no. 1: 299–320.

Bellon-Maurel, V., E. Fernandez-Ahumada, P. Palagos, J.-M. Roger, and A. McBratney. 2010. "Critical Review of Chemometric Indicators Commonly Used for Assessing the Quality of the Prediction of Soil Attributes by NIR Spectroscopy." *Trends in Analytical Chemistry* 29: 1073–1081.

Bernardini, L. G., E. Bruni, E. Izquierdo-Verdiguier, E. Smit, and C. Rosinger. 2025. "Letter to the Editor Regarding Feeney et al., 2024, Benchmarking Soil Organic Carbon (SOC) Concentration Provides More Robust Soil Health Assessment Than the SOC/Clay Ratio at European Scale." *Science of the Total Environment* 976: 179278.

Breure, T. S., S. M. Haefele, J. A. Hannam, et al. 2022a. "A Loss Function to Evaluate Agricultural Decision-Making Under Uncertainty: A Case Study of Soil Spectroscopy." *Precision Agriculture* 23, no. 4: 1333–1353.

Breure, T. S., A. Jones, and P. Panagos. 2026. "Evaluating Visible Near-Infrared Spectroscopy in Context of a Repeated Sampling Survey Across the European Union." *Geoderma* 465: 117647.

Breure, T. S., J. M. Prout, S. M. Haefele, et al. 2022b. "Comparing the Effect of Different Sample Conditions and Spectral Libraries on the Prediction Accuracy of Soil Properties From Near- and Mid-Infrared Spectra at the Field-Scale." *Soil and Tillage Research* 215: 105196.

Brown, D. J. 2007. "Using a Global VNIR Soil-Spectral Library for Local Soil Characterization and Landscape Modeling in a 2nd-Order Uganda Watershed." *Geoderma* 140: 444–453.

Brown, D. J., K. D. Shepherd, M. G. Walsh, M. Dewayne Mays, and T. G. Reinsch. 2006. "Global Soil Characterization With VNIR Diffuse Reflectance Spectroscopy." *Geoderma* 132, no. 3–4: 273–290.

Chen, Y., S. Gao, E. J. Jones, and B. Singh. 2021. "Prediction of Soil Clay Content and Cation Exchange Capacity Using Visible Near-Infrared Spectroscopy, Portable X-Ray Fluorescence, and X-Ray Diffraction Techniques." *Environmental Science & Technology* 55: 4629–4637.

Clairotte, M., C. Grinand, E. Kouakoua, et al. 2016. "National Calibration of Soil Organic Carbon Concentration Using Diffuse Infrared Reflectance Spectroscopy." *Geoderma* 276: 41–52.

COM. 2025. "Directive (EU) 2025/2360 of the European Parliament and of the Council of 12 November 2025 on Soil Monitoring and Resilience (Soil Monitoring Law)". Accessed March 25, 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202502360.

Comstock, J. P., S. R. Sherpa, R. Ferguson, et al. 2019. "Carbonate Determination in Soils by Mid-IR Spectroscopy With Regional and Continental Scale Models." *PLoS One* 14, no. 2: e0210235. https://doi.org/10.1371/journal.pone.0210235.

Dangal, S. R. S., J. Sanderman, S. Wills, and L. Ramirez-Lopez. 2019. "Accurate and Precise Prediction of Soil Properties From a Large Mid-Infrared Spectral Library." *Soil Systems* 3, no. 1: 1–23.

Dexter, A. R., G. Richard, D. Arrouays, E. A. Czyz, C. Jolivet, and O. Duval. 2008. "Complexed Organic Matter Controls Soil Physical Properties." *Geoderma* 144: 620–627.

Feng, W., A. F. Plante, and J. Six. 2013. "Improving Estimates of Maximal Organic Carbon Stabilization by Fine Soil Particles." *Biogeochemistry* 112: 81–93.

Grinand, C., B. G. Barthès, D. Brunet, et al. 2012. "Prediction of Soil Organic and Inorganic Carbon Contents at a National Scale (France) Using Mid-Infrared Reflectance Spectroscopy (MIRS)." *European Journal of Soil Science* 63: 141–151.

Guerrero, C., R. Zornoza, I. Gómez, and J. Mataix-Beneyto. 2010. "Spiking of NIR Regional Models Using Samples From Target Sites: Effect of Model Size on Prediction Accuracy." *Geoderma* 158: 66–77.

Hallett, S. H., R. Sakrabani, C. A. Keay, and J. A. Hannam. 2017. "Developments in Land Information Systems: Case Studies in Land Resource Management Capabilities and Options." *Soil Use and Management* 33, no. 4: 514–529. https://doi.org/10.1111/sum.12380.

Hastie, T., J. Friedman, and R. Tibshirani. 2013. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." In *10th Printing With Corrections*. Springer Paragraph 7.10.

Hermansen, C., M. Knadel, P. Moldrup, M. H. Greve, R. Gislum, and L. W. De Jonge. 2016. "Visible-Near-Infrared Spectroscopy Can Predict the Clay/Organic Carbon and Mineral Fines/Organic Carbon Ratios." *Soil Science Society of America Journal* 80: 1486–1495.

Huang, Y. C., W. Ng, B. Minasny, and A. B. McBratney. 2025. "Characterising and Quantifying Soil Clay-Sized Minerals Using Mid-Infrared Spectroscopy." *Soil and Tillage Research* 252: 106590.

Kalembasa, S. J., and D. S. Jenkinson. 1973. "A Comparative Study of Titrimetric and Gravimetric Methods for the Determination of Organic Carbon in Soil." *Journal of the Science of Food and Agriculture* 24: 1085–1090.

Li, S., R. A. Viscarra Rossel, and R. Webster. 2022. "The Cost-Effectiveness of Reflectance Spectroscopy for Estimating Soil Organic Carbon." *European Journal of Soil Science* 73: e13202.

Liland, K. H., B. H. Mevik, and R. Wehrens. 2022. "PLS: Partial Least Squares and Principal Component Regression". R Package Version 2.8-1. https://CRAN.R-project.org/package=pls.

Lin, L. I. 1989. "A Concordance Correlation Coefficient to Evaluate Reproducibility." *Biometrics* 45: 255–268.

Madejová, J. J. V. S. 2003. "FTIR Techniques in Clay Mineral Studies." *Vibrational Spectroscopy* 31, no. 1: 1–10.

McCarty, G. W., J. B. Reeves III, V. B. Reeves, R. F. Follett, and J. M. Kimble. 2002. "Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurements." *Soil Science Society of America Journal* 66: 640–646.

McGrath, S. P., and P. J. Loveland. 1992. *The Soil Geochemical Atlas of England and Wales*. Blackie.

Nguyen, T. T., L. J. Janik, and M. Raupach. 1991. "Diffuse Reflectance Infrared Fourier Transform (DRIFT) Spectroscopy in Soil Studies." *Soil Research* 29, no. 1: 49–67.

Niculescu-Mizil, A., and R. Caruana. 2005. "Predicting Good Probabilities With Supervised Learning." In *Proceedings of the 22nd International Conference on Machine Learning*, 625–632. Association for Computing Machinery.

Nocita, M., A. Stevens, B. van Wesemael, et al. 2015. "Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring." *Advances in Agronomy* 132: 139–159.

Pleijsier, K. 1989. "Variability in Soil Data." In *Land Qualities in Space and Time*, edited by J. Bouma and A. K. Bregt, 89–98. PUDOC.

Poeplau, C., and A. Don. 2023. "A Simple Soil Organic Carbon Level Metric Beyond the Organic Carbon-to-Clay Ratio." *Soil Use and Management* 39, no. 3: 1057–1067. https://doi.org/10.1111/sum.12921.

Prout, J. M., K. D. Shepherd, S. P. McGrath, G. J. D. Kirk, and S. M. Haefele. 2021. "What Is a Good Level of Soil Organic Matter? An Index Based on Organic Carbon to Clay Ratio." *European Journal of Soil Science* 72: 2493–2503. https://doi.org/10.1111/ejss.13012.

Prout, J. M., K. D. Shepherd, S. P. McGrath, G. J. Kirk, K. L. Hassall, and S. M. Haefele. 2022. "Changes in Organic Carbon to Clay Ratios in Different Soils and Land Uses in England and Wales Over Time." *Scientific Reports* 12: 5162.

R Core Team. 2021. "R: A Language and Environment for Statistical Computing." In *R Foundation for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rabot, E., N. P. Saby, M. P. Martin, et al. 2024. "Relevance of the Organic Carbon to Clay Ratio as a National Soil Health Indicator." *Geoderma* 443: 116829.

Ramirez-Lopez, L., A. Stevens, R. Viscarra Rossel, C. Lobsey, A. Wadoux, and T. Breure. 2022. "Resemble: Regression and Similarity Evaluation for Memory-Based Learning in Spectral Chemometrics". R package Vignette R package version 2.2.1. Accessed September 11, 2023. https://CRAN.R-project.org/package=resemble.

Reeves, J. B. 2010. "Near- Versus Mid-Infrared Diffuse Reflectance Spectroscopy for Soil Analysis Emphasizing Carbon and Laboratory Versus On-Site Analysis: Where Are We and What Needs to Be Done?" *Geoderma* 158: 3–14.

Safanelli, J. L., T. Hengl, L. L. Parente, et al. 2025. "Open Soil Spectral Library (OSSL): Building Reproducible Soil Calibration Models Through Open Development and Community Engagement." *PLoS One* 20, no. 1: e0296545.

Sanderman, J., K. Savage, and S. R. S. Dangal. 2020. "Mid-Infrared Spectroscopy for Prediction of Soil Health Indicators in the United States." *Soil Science Society of America Journal* 84, no. 1: 251–261.

Sanderman, J., K. Savage, S. R. S. Dangal, et al. 2021. "Can Agricultural Management Induced Changes in Soil Organic Carbon Be Detected Using Mid-Infrared Spectroscopy?" *Remote Sensing* 13: 1–18.

Sauzet, O., A. Johannes, C. Deluz, et al. 2024. "The Organic Carbon-To-Clay Ratio as an Indicator of Soil Structure Vulnerability, A Metric Focused on the Condition of Soil Structure." *Soil Use and Management* 40: e13060.

Schjønning, P., L. W. de Jonge, L. J. Munkholm, P. Moldrup, B. T. Christensen, and J. E. Olesen. 2012. "Clay Dispersibility and Soil Friability—Testing the Soil Clay-to-Carbon Saturation Concept." *Vadose Zone Journal* 11: 1. https://doi.org/10.2136/vzj2011.0067.

Seidel, M., C. Hutengs, B. Ludwig, S. Thiele-Bruhn, and M. Vohland. 2019. "Strategies for the Efficient Estimation of Soil Organic Carbon at the Field Scale With Vis-NIR Spectroscopy: Spectral Libraries and Spiking vs. Local Calibrations." *Geoderma* 354: 113856.

Six, J., R. T. Conant, E. A. Paul, and K. Paustian. 2002. "Stabilization Mechanisms of Soil Organic Matter: Implications for C-Saturation of Soils." *Plant and Soil* 241: 155–176.

Soriano-Disla, J. M., L. J. Janik, R. A. Viscarra Rossel, L. M. Macdonald, and M. J. McLaughlin. 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., M. Nocita, G. Tóth, L. Montanarella, and B. van Wesemael. 2013. "Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy." *PLoS One* 8, no. 6: e66409.

Stevens, A., and L. Ramirez-Lopez. 2022. "An Introduction to the Prospectr Package". R package version 0.2.6. Accessed September 11, 2023. https://cran.r-project.org/package=prospectr.

Stevenson, M., T. Nunes, C. Heuer, et al. 2023. "epiR: Tools for the Analysis of Epidemiological Data". R package version 2.0.57. Accessed September 11, 2023. https://CRAN.R-project.org/packag=epiR.

Terra, F. S., J. A. Demattê, and R. A. V. Rossel. 2015. "Spectral Libraries for Quantitative Analyses of Tropical Brazilian Soils: Comparing Vis–NIR and Mid-IR Reflectance Data." *Geoderma* 255: 81–93.

Thomas, C. L., J. Hernandez-Allica, S. J. Dunham, S. P. McGrath, and S. M. Haefele. 2021. "A Comparison of Soil Texture Measurements Using Mid-Infrared Spectroscopy (MIRS) and Laser Diffraction Analysis (LDA) in Diverse Soils." *Scientific Reports* 11, no. 1: 16. https://doi.org/10.1038/s41598-020-79618-y.

Tinti, A., V. Tugnoli, S. Bonora, and O. Francioso. 2015. "Recent Applications of Vibrational Mid-Infrared (IR) Spectroscopy for Studying Soil Components: A Review." *Journal of Central European Agriculture* 16, no. 1: 1–22.

van Leeuwen, C. C., V. L. Mulder, N. H. Batjes, and G. B. Heuvelink. 2022. "Statistical Modelling of Measurement Error in Wet Chemistry Soil Data." *European Journal of Soil Science* 73: e13137.

Viscarra Rossel, R. A., T. Behrens, E. Ben-Dor, et al. 2016. "A Global Spectral Library to Characterize the World's Soil." *Earth-Science Reviews* 155: 198–230.

Viscarra Rossel, R. A., S. R. Cattle, A. Ortega, and Y. Fouad. 2009. "In Situ Measurements of Soil Colour, Mineral Composition and Clay Content by Vis-NIR Spectroscopy." *Geoderma* 150: 253–266.

Viscarra Rossel, R. A., D. J. J. Walvoort, A. B. McBratney, L. J. Janik, and J. O. Skjemstad. 2006. "Visible, Near Infrared, Mid Infrared or Combined Diffuse Reflectance Spectroscopy for Simultaneous Assessment of Various Soil Properties." *Geoderma* 131: 59–75.

Wehrens, R., H. Putter, and L. M. Buydens. 2000. "The Bootstrap: A Tutorial." *Chemometrics and Intelligent Laboratory Systems* 54: 35–52.

Wetterlind, J., and B. Stenberg. 2010. "Near-Infrared Spectroscopy for Within-Field Soil Characterization: Small Local Calibrations Compared With National Libraries Spiked With Local Samples." *European Journal of Soil Science* 61: 823–843.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., M. Averick, J. Bryan, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4, no. 43: 1686. https://doi.org/10.21105/joss.01686.

Wright, M. N., and A. Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77, no. 1: 1–17. https://doi.org/10.18637/jss.v077.i01.

Zhang, Y. 2014. *Quantification of Prediction Uncertainty for Principal Components Regression and Partial Least Squares Regression*. University College London.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Loadings of the pre-processed spectra for each of SOC (top), clay (middle) and SOC/clay (bottom) from the partial least squares regression analysis. **Figure S2:** Standardised residuals of chi-square tests for the validation set between reference index class ($(SOC/clay)_{Con}$, *x*-axis) and estimated index class ($Index_{Est}$, *y*-axis) for

each of the estimation methods: $(SOC/clay)_{Est}$ (a and b), $SOC_{Est}/clay_{Est}$ (c and d), $SOC_{Est}/clay_{Con}$ (e and f). Columns correspond to the calibration regression methods used: partial least squares regression (PLSR) and random forest (RF) regression. Combining the Good and Moderate SOC/clay index classes improved the sensitivity and specificity. Positive values suggest association of estimated and reference classes, negative values suggest a lack of association. **Figure S3:** Standardised residuals of chi-square tests for the validation set between reference index class ($(SOC/clay)_{Con}$, *x*-axis) and estimated index class ($Index_{Est}$, *y*-axis) for each a random forest (RF) regression directly on the SOC/clay index class. Positive values suggest association of estimated and reference classes, negative values suggest a lack of association. **Figure S4:** Loadings from the partial least squares regression analysis for inorganic carbon (IC). The subset of data differed from the subset used for the main analysis of SOC/clay due to the presence or absence of inorganic carbon measurements, however selection criteria of clay $>0\,g\,kg^{-1}$ and soil organic carbon $\leq 12.0\,g\,kg^{-1}$ were applied as for the main dataset.



https://dspace.org/

Midinfrared spectroscopy to estimate ratios of soil organic carbon to clay

Prout, Jonah M.; Breure, Timo S.; Shepherd, Keith D.; Acquah, Gifty E.; Hewlett, Franklin D. ...

2026-06
Wiley
https://doi.org/10.1111/ejss.70353
Prout JM, Breure TS, Shepherd KD, et al., (2026) Midinfrared spectroscopy to estimate ratios of soil organic carbon to clay. European Journal of Soil Science, Volume 77, Issue 3, May-June 2026, Article number e70353

Downloaded from a DSpace repository. https://dspace.org/