Soil & Environmental Health 3 (2025) 100123

Contents lists available at ScienceDirect

Soil & Environmental Health

journal homepage: www.sciencedirect.com/journal/soil-and-environmental-health

Research Paper

Modeling soil organic carbon content using mid-infrared absorbance spectra and a nonnegative MCR-ALS analysis

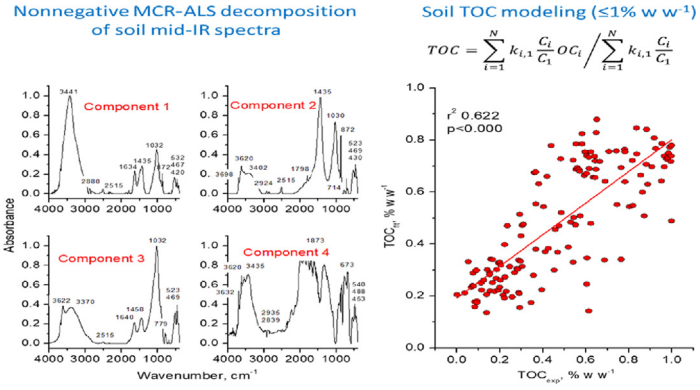
Mikhail Borisover [a,*], Marcos Lado [b], Guy J. Levy [a]

[a] Institute of Soil, Water and Environmental Sciences, Agricultural Research Organization, Volcani Institute, P.O. Box 15159, Rishon LeZion 7505101, Israel
[b] Centro de Investigaciones Científicas Avanzadas, Facultad de Ciencias, Universidade da Coruña, Spain

HIGHLIGHTS

- Soil mid-IR absorbance spectra were decomposed using the nonnegative MCR-ALS algorithm
- The identified components are characterized by concentration scores and IR spectra
- A mechanistic model is used to link scores of MCR-ALS components to soil TOC values
- Success in modeling soil TOC content depended on a threshold TOC level
- The detected threshold can help identify different types of soil organic matter

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling editor: Lena Q. Ma
Technical editor: Songlin Wu

Keywords:
Soil health
MCR-ALS algorithm
Mid-IR spectroscopy
SOM-Mineral interactions
The Beer-Lambert law
Carbon storage capacity
Dominating SOM pools
Physicochemical model
Soil organic matter (SOM)

ABSTRACT

A new approach based on mid-IR absorbance spectra is proposed for modeling total organic carbon (TOC) content in soils. This approach involves a first-time bilinear decomposition of soil mid-IR absorbance spectra using nonnegative multivariate curve resolution (MCR) with an alternating least squares (ALS) algorithm. An MCR-ALS-derived component signifies a chemically meaningful combination of soil constituents. This new mechanistic model has been developed to link the soil composition, expressed in terms of ratios of MCR-ALS-based concentration scores of the identified components, to soil TOC value. Nonnegative MCR-ALS decomposition, performed for 213 mid-IR absorbance spectra of soil samples collected in the north and south of Israel, yielded four components. Fitting the mechanistic model-derived TOC to the experimental TOC values exhibited a TOC content threshold that affected model performance. TOC content <1.0 % w w^{-1} was represented by the root mean square deviation of 0.18% with 62% of the variance being explained, whereas for larger TOC values, a sharp decline in model performance was observed. The existence of this TOC threshold in determining model performance suggested that successful TOC modeling (below 1%) could be indirect and related to IR spectral fingerprints of minerals binding soil organic matter (SOM) and forming organo-mineral complexes. Thus, a SOM fraction having weak interactions with soil minerals was poorly accounted for in some soil samples. The dependency of the model performance on soil TOC contents suggests that it might be possible to differentiate between soil samples based on

* Corresponding author.
E-mail address: vwmichel@volcani.agri.gov.il (M. Borisover).

https://doi.org/10.1016/j.seh.2024.100123
Received 20 July 2024; Received in revised form 10 November 2024; Accepted 14 November 2024
Available online 17 November 2024

their different dominating SOM pools, mineral-associated ones and those having weak interactions with minerals. Further studies, especially in soils with high SOM content, are needed to validate our findings.

1. Introduction

Soil organic matter (SOM) is well-recognized for controlling soil physical, chemical and biological properties (Baldock and Broos, 2012). The content of total organic carbon (TOC) (or organic matter) was found as the most frequently used property for characterizing soil quality, among 27 indicators (Bünemann et al., 2018). The TOC content as a soil health indicator is relevant to soil functions, practical for use and informative, however, it is not necessarily sensitive to respond detectably and quickly to land use and management changes without reflecting short-term variations (Lehmann et al., 2020). Hence, SOM fractions, such as unprotected and mineral-protected SOM, that are potentially more sensitive, were proposed to be among soil health indicators of a new generation (Lehmann et al., 2020). These fractions also play different roles in SOM turnover and carbon sequestration. Thus, SOM becomes stabilized due to its interactions with minerals and reduced accessibility (Kögel-Knabner and Rumpel, 2018; Lehmann and Kleber, 2015; Just et al., 2023) although further mineral-associated SOM may be destabilized by plant and microbial exudates via multiple pathways (Li et al., 2021). The majority of SOM in Earth's mineral soils is associated with minerals thus affecting the whole cycling of terrestrial carbon (Sokol et al., 2022). Hence, there is a fundamental interest in quantifying SOM fractions, predicting soil TOC content, and understanding the relations between SOM content and soil chemical composition.

Soil infrared (IR) spectra serve as fingerprints for soil composition due to the vibrational activities of various chemical bonds present in both soil minerals and SOM. Significant work has been done to model and predict the content of TOC and SOM fractions using soil mid- and near-IR spectra (Soriano-Disla et al., 2014; Madhavan et al., 2017; Nocita et al., 2014; Zhang et al., 2018; Baldock et al., 2018; Lussier et al., 2020; Nasonova et al., 2022; Margenot et al., 2023). Thus, based on multiple studies (including those mentioned above), the capability of mid- and near-IR spectroscopy, probing soil chemical composition, to model and predict TOC content and SOM composition, is well established (Terhoeven-Urselmans et al., 2006; Baldock et al., 2018; Ng et al., 2022). It has also been recognized that for quantifying soil TOC content, different types of soil IR spectroscopy may provide time- and cost-effective approaches that are less laborious and cheaper than conventional dry combustion analysis (Barthès and Chotte, 2021; Metzger et al., 2021; Li et al., 2022).

Partial least square (PLS) regressions (Wold et al., 2001) are commonly used for modeling and prediction of soil TOC. Barra et al. (2021) summarized predictions of multiple soil properties including TOC content, using mid-, near- and visible-near IR spectroscopies, with different multivariate techniques. Of the 90 publications examined by Barra et al. (2021), 86 papers used the PLS technique (not necessarily only) for prediction purposes. In the case of TOC predictions, the empirical PLS regression models link carbon content values to latent variables, which are linear combinations of IR absorbances determined at multiple wavenumbers. Since the PLS predictors are linear combinations of original descriptors (IR absorbances), and their loadings may be either positive or negative, the interpretation of PLS models is far from simple (Pirouz, 2010; Fritzsche et al., 2019; Xia, 2020; Lado et al., 2023).

Conceptually, a less empirical way to model TOC could be proposed if, instead of creating latent PLS variables that best fit the TOC values, the soil IR absorbance spectra would be expressed in terms of contributions from chemically meaningful independent components characterized by their spectra and specific sample-dependent concentration scores, i.e., performing so-called "mathematical chromatography" (Bro et al., 2010). Separating soil IR absorbance spectra into the contributions of independent components varying in their proportions in a sample series represents a "mixture analysis problem" and fits the family of the methods termed Multivariate Curve Resolution (MCR; de Juan and Tauler, 2021) seeking a bilinear decomposition of the dataset into component scores ("concentrations") and loadings (spectra). It should be kept in mind that this decomposition is "unsupervised" regarding the sample TOC values of interest. For further use of IR spectra decomposition to model soil TOC, a mechanistic model needs to be developed that links the soil composition expressed in terms of the component's scores to the TOC value.

Although the decomposition of IR spectra of complex (and, particularly, environmental) matrices into components' contributions is a difficult task, there are multiple studies demonstrating the potential of this analysis. The MCR methodology with alternating least squares (ALS) algorithms was applied to Fourier transform IR (FTIR) spectra to characterize acid and salt forms of humic acids interacting with lead (Gossart et al., 2003). It was also applied to decompose mid-FTIR imaging of wood fibers, yielding three components that were identified as glucans, lignin and hemicelluloses, very similar to reference spectra for these substances (Araya et al., 2019). Câmara et al. (2022) used it to recover the kerosene spectral profile from jet fuel mid- and near-IR spectra. Spectral and concentration profiles of monodentate and bidentate glyphosate complexes with ferrihydrite were identified using MCR-ALS of attenuated total reflectance (ATR) spectra (Li et al., 2023). Furthermore, Ioannidi et al. (2023) demonstrated the applicability of MCR-ALS to recover contributions from crystalline and amorphous forms of fats in chocolate using also ATR-FTIR spectra.

A decomposition of IR spectral data into the contributions of unknown components and recovering their spectral loadings and concentration profiles is also possible using a similar technique of Nonnegative/Positive Matrix Factorization (NMF/PMF, Paatero and Tapper, 1994). Fritzsche et al. (2019) decomposed mid-IR spectra of groundwater solids using PMF, providing the quantitative interpretation of PMF components (humic acid-coated goethite, kaolinite, montmorillonite, organic matter) without the *a priori* knowledge of the sample composition. Russell et al. (2009) and Takahama et al. (2011) characterized the composition of aerosols in terms of PMF components derived from mid-IR spectra. Lado et al. (2023) performed an NMF decomposition of mid-IR spectra of water-extractable organic matter obtained from soils undergoing controlled heating in an oxidizing atmosphere and demonstrated the accumulation of oxidized organic matter components. Recently, Borisover et al. (2023) examined the sensitivity of soil chemical components identified with NMF of mid-IR spectra to the treatments intended to minimize the negative consequences of treated wastewater irrigation.

Bilinear decomposition using the MCR-ALS technique does not provide generally a unique solution (Bro et al., 2010) and may lead to scale and rotation ambiguities. The former can be accounted for by scaling the data, and the latter may be minimized by introducing nonnegativity (and other) constraints, and/or its degree can be evaluated (Jaumot and Tauler, 2010). Nonnegativity constraints are naturally expected when decomposing spectra into physically reliable concentration scores and spectral loadings of components since both concentrations and absorbance spectra of components cannot be characterized by negative values (Jaumot and Tauler, 2010). Thus, the above-described potential of MCR-ALS in decomposing datasets into the component's contributions has driven our interest in applying this technique to soil mid-IR absorbance spectra.

Therefore, three specific objectives were pursued in this study: (*1*) to develop a mechanistic model relating differences in soil composition, identified by nonnegative bilinear decomposition (based on MCR-ALS) of mid-IR absorbance spectra, to soil TOC content, (*2*) to identify the

particular MCR-ALS-derived soil components by performing the analysis of selected soil mid-IR spectra, (*3*) to test whether these derived soil components combined within a mechanistic approach allow modeling soil TOC content. To the best of our knowledge, there are no studies in the literature that propose a physicochemical model linking soil components quantified through multivariate decomposition of mid-IR spectra and soil TOC content. Therefore, research involving these three objectives is novel and potentially could contribute to better identification of C pools and their dynamics in soils.

2. Methodologies and data sources used in model examination

2.1. Derivation of the mechanistic model

This section addresses the first objective of the research. Although soil samples represent a complex matrix for spectral analysis, the applicability of the Beer-Lambert law is generally (explicitly or implicitly) accepted in soil and mineral IR spectroscopy (Johnston and Aochi, 1996; Kaufhold et al., 2012). This is evident also in general practice to compare the absorbance of different bands or a band and a reference signal using the maximum height or the band areas (Davis et al., 1999; Margenot et al., 2023). Then, MCR analysis can be considered an extension of the Beer-Lambert law applied to a mixture of *N* identifiable components at variable wavenumbers (Jaumot et al., 2005; de Juan and Tauler, 2021). This relation between the physicochemical Beer-Lambert law and the MCR of IR spectra highlights the potential merit of this decomposition for modeling and prediction of TOC in soils; the MCR-based decomposition may bring less empirical and better interpretable characterization of soil IR spectra and their relations to soil TOC. However, it is important to recognize that MCR-ALS-derived soil components do not describe necessarily a particular chemical substance or minerals. A given component represents a set of chemical bonds and functional groups, that are active in the mid-IR range, present in a soil sample and maintain constant proportions in a soil sample series. This means that an identified MCR-ALS component reflects a chemically meaningful combination of soil constituents; a set of the MCR-ALS components characterizes soil composition.

The results of an MCR-based decomposition of absorbance spectra are (1) a number (*N*) of estimated components *i* (from 1 to *N*); (2) the spectra of the components, given as absorbance ($L_i(\bar{\nu})$, in arbitrary units) *vs* wavenumber $\bar{\nu}$; (3) sample-dependent scores C_i of each component *i* (Jaumot et al., 2005; de Juan and Tauler, 2021). The product of C_i of a component and its $L_i(\bar{\nu})$ is the absorbance $Abs_i\ (\bar{\nu})$ [$= C_iL_i(\bar{\nu})$] of the given component that contributes to the whole measured sample absorbance $Abs\ (\bar{\nu})$ at a certain $\bar{\nu}$:

$$Abs\ (\bar{\nu}) = \sum_{i=1}^{N} Abs_i(\bar{\nu}) = \sum_{i=1}^{N} C_iL_i(\bar{\nu}) \tag{1}$$

Following the Beer-Lambert law,

$$Abs_i(\bar{\nu}) = C_iL_i(\bar{\nu}) = \theta_i\varepsilon_i(\bar{\nu})d \tag{2}$$

where θ_i is a concentration of component *i* in a sample (e.g., in mass per volume), *d* is the (possibly sample-specific) optical path length, and $\varepsilon_i(\bar{\nu})$ is the absorptivity of a specific component at wavenumber $\bar{\nu}$. From Eq. (2) it follows:

$$C_i = \theta_i\varepsilon_i(\bar{\nu})d\ /\ L_i(\bar{\nu}) \tag{3}$$

For a given sample, the results of the MCR analysis may be presented as ratios of the scores C_i of the components relative to a certain one, e.g., for the first component: C_i/C_1 where $i = 1, 2, \ldots N$ (for $i = 1$, the ratio equals one). Based on Eq. (3), each such C_i/C_1 ratio is proportional to the concentration ratio θ_i/θ_1:

$$\frac{C_i}{C_1} = \frac{\theta_i}{\theta_1}\frac{\varepsilon_i(\bar{\nu})}{\varepsilon_1(\bar{\nu})}\frac{L_1(\bar{\nu})}{L_i(\bar{\nu})} \tag{4}$$

$$\frac{\theta_i}{\theta_1} = k_{i,1}\frac{C_i}{C_1} \tag{5}$$

The wavenumber-independent proportionality coefficient $k_{i,1}$ in Eq. (5) is specific for component *i* and does not change among samples (when $i = 1$, $k_{1,1} = 1$):

$$k_{i,1} = \frac{L_i(\bar{\nu})}{L_1(\bar{\nu})}\frac{\varepsilon_1(\bar{\nu})}{\varepsilon_i(\bar{\nu})} \tag{6}$$

Based on Eq. (5), the whole mass concentration of all the soil components in a sample is defined as:

$$\sum_{i=1}^{N}\theta_i = \theta_1\left(\sum_{i=1}^{N}k_{i,1}\frac{C_i}{C_1}\right) \tag{7}$$

$$\theta_1 \Big/ \sum_{i=1}^{N}\theta_i = 1 \Big/ \sum_{i=1}^{N}k_{i,1}\frac{C_i}{C_1} \tag{8}$$

The mass fraction ϕ_i of component *i* in the whole soil sample is as follows:

$$\phi_i = \theta_i \Big/ \sum_{i=1}^{N}\theta_i \tag{9}$$

From Eqs. (5), (8) and (9) it follows:

$$\phi_i = k_{i,1}\frac{C_i}{C_1} \Big/ \sum_{i=1}^{N}k_{i,1}\frac{C_i}{C_1} \tag{10}$$

Mid-IR spectra cover a variety of chemical bonds and functional groups, and it is expected that major soil components are represented, to different extents, in mid-IR spectra. Hence, one may find the total organic C content (TOC) of a soil sample by summing the organic C contents (OC_i) of each MCR component weighed by its mass fraction ϕ_i:

$$TOC = \sum_{i=1}^{N}k_{i,1}\frac{C_i}{C_1}OC_i \Big/ \sum_{i=1}^{N}k_{i,1}\frac{C_i}{C_1} \tag{11}$$

where TOC and OC_i refer to the masses of the soil sample and component *i*, respectively.

When MCR-ALS is applied to mid-IR spectra for *m* soil samples and the model suggests *N* components, there are $N \times m$ known $\frac{C_i}{C_1}$ ratios, and *m* experimentally determined the TOC values (one for each soil sample). The model (Eq. (11)) involves *N* unknown *OC* contents, one for each extracted component, and (*N*-1) unknown values of $k_{i,1}$ (since $k_{1,1} = 1$). These 2*N*-1 unknown variables may be calculated by adjusting the model to the TOC values of *m* samples. Hence, a successful fit could help in understanding whether the MCR-ALS of the soil IR spectra properly identified the soil components controlling TOC content.

To summarize, the model assumptions are the following:

1) MCR-ALS may find contributions of chemically meaningful components to soil mid-IR spectra. This means that
   - these components are not necessarily individual substances, but a given component may include diverse constituents maintaining, for different reasons, the same proportion across a sample series. Thus, such an assemblage behaves like a substance with “constant chemical composition".
   - the contributions of such components to the soil spectra follow the Beer-Lambert law.

2) All TOC-controlling components may be identified within the MCR-ALS-based decomposition of IR spectra.

2.2. The mid-IR spectra of soil samples

The soil mid-IR spectra examined in this work were collected by Nasonova et al. (2022). These spectra represented 213 soil samples obtained from Jezreel Valley and North-Western part of the Negev of Israel. Jezreel Valley and North-Western part of the Negev are two important agricultural regions characterized by different climate (Mediterranean and semiarid, respectively) and precipitation, situated in the north and south of Israel (Nasonova et al., 2022; Rinot et al., 2021). The soil samples were taken from three sites at each region and from two different depths (0–10 cm and 30–60 cm), to provide soil samples of the same type but differing in SOM quantity and (potentially) composition. Sampling was carried out at both the end of the rainy season (Spring) and at the end of the irrigation season (Autumn). The soil samples were also associated with different land uses (orchards, field crops and non-cultivated plots with natural vegetation). Variations in locations, climate, land use, sampling depth and seasons were dictated by original interest in linking soil mid-IR spectra to properties of soil extracts in a wide range of environmental/land use scenarios (Nasonova et al., 2022).

Each type of soil sample was taken from three pits providing sample triplications. The overall number of collected samples was 216 (2 regions × 3 sites × 3 land uses × 2 depths × 3 replicates × 2 sampling seasons). Only 213 mid-IR spectra were used in this work since the TOC values for 3 samples were lacking. Thus 213 mid-IR spectra were available for the modeling.

More details regarding soil types, climate differences, sample distribution and agricultural management are provided in Nasonova et al. (2022). Soil characterization involved the determination of several properties (Nasonova et al., 2022; Rinot et al., 2021). Table S1 (Supplementary information) provides a summary of data on total nitrogen, TOC, inorganic carbon (representing soil carbonates), cation exchange capacity (CEC) and texture. It should be noted that looking for relations between soil TOC contents and soil texture, CEC, carbonate contents and climate regions was out of the scope of this work which is focused on linking between a new way to decompose soil IR spectra and TOC content.

The mid-IR spectra of soil samples (air-dried at room temperature) were measured in transmission mode, in KBr pellets, using Bruker Tensor 27 FT-IR spectrometer, with a soil concentration of about 1% w w^{-1}; the wavenumber range was 4000–400 cm^{-1} with 4 cm^{-1} resolution, and 16 scans per acquisition (Nasonova et al., 2022). Spectral corrections involved atmospheric compensation and reducing the baseline slope of the final spectrum using the software OPUS 6.5. IR spectra of soil samples normalized by quartz absorbance signal (taken with permission from Elsevier, from Nasonova et al. (2022)) are exemplified in Figure S1, that shows averages of spectra of soils sampled in the north (Mediterranean) and south (semiarid) regions. For the current MCR-ALS analysis, the soil IR spectra after the "Min-Max" normalization performed with the OPUS 6.5 software were used; in this normalization, the maximal absorption is assigned the value of 2.

2.3. Multivariate curve resolution (MCR-ALS) of mid-IR absorbance spectra

To accomplish the second objective, the MCR-ALS analysis was performed using the MCR-ALS GUI (Graphical User Interface) 2.0 toolbox (Jaumot et al., 2015) within Matlab R2023a. The optimal number of components was selected using the singular value decomposition (SVD) algorithm. The initial estimations for component spectra and scores were made with a "purest variable" detection method (Jaumot et al., 2015). The fast nonnegative least squares algorithm was used for spectra decomposition (Bro and De Jong, 1997; Jaumot et al., 2015), with nonnegativity constraints for scores and absorbance loadings of the resolved components. The resolved spectra profiles were normalized to have equal height (Jaumot et al., 2015), which helps to fix the possible intensity ambiguities. The iteration stop criterion was set at a convergence of 0.001 (i.e., reflecting the percentage of change of standard deviation of residuals between two successive iterations). Extents of the possible rotation ambiguities were evaluated for each component using MCR-BANDS (Tauler et al., 2016), with the same constraints as in the MCR-ALS. The MCR-BAND method evaluates the contribution of a certain component to the whole signal for the mixture of N components, and when the difference between the minimal and maximal relative contributions tends to be zero, there is no remaining rotation ambiguity (Jaumot and Tauler, 2010). Typically, this difference in the MCR-ALS analyses was around 10^{-14}. The quality of optimization was characterized by (i) the standard deviation of residuals (root-mean-square deviation, RMSD); (ii) the fitting error (lack-of-fit, Lof, in %) defined as $\mathrm{Lof} = 100\sqrt{\sum_{i,j} e_{i,j}^2 / \sum_{i,j} d_{i,j}^2}$ where $d_{i,j}$ indicates absorbance at wavenumber i of a spectrum corresponding to sample j, and $e_{i,j}$ is the difference between measured and modeled absorbance values; and (iii) the explained variance, defined as $1\text{-}(\mathrm{Lof}/100)^2$ and expressed in percentage. Different optimization runs when repeated brought the same values of RMSD, Lof and the explained variance. When mid-IR spectra of MCR-ALS-identified components were compared between different decompositions, the degree of similarity was evaluated using correlations between mean-centered absorbances of these spectra (Varmuza et al., 2003). Thus, the nonnegative MCR-ALS decomposition of soil mid-IR spectra was performed without targeting the soil TOC data, being, therefore, unsupervised. Also, no *a priori* knowledge of soil chemical composition is used in the spectra decomposition although available data of actual soil components (i.e., mineral composition, particular organic substances and their classes) could potentially have some impact on the selection of the number of components.

2.4. Fitting of the principal model to soil TOC values

The third objective of the research was to test whether the developed mechanistic approach allows modeling soil TOC content, using the MCR-ALS components derived from soil mid-IR spectra. To address this objective, fitting of TOC content with Eq. (11) incorporating scores C_i of MCR-ALS-identified soil components was performed using a nonlinear estimation option in Statistica 7.0 Statsoft. It should be noted that modeling the soil TOC values using Eq. (11) does not require *per se* knowledge of mineralogical composition or identification of particular organic compounds in SOM. It depends only on specific score values of the components in Eq. (11). In the current stage of the research, no predictive steps were pursued, but only the quantitative description of the soil TOC content data was of interest. Therefore, verification of the model's predictive strength with data separation into the training/test sets was out of the research scope. For the soil TOC data modeling, the parameters of Eq. (11) were sought by the software to minimize the loss function. The minimized loss function was the sum of squared deviations between experimental and fitted values, introducing penalties when the values of adjustable parameters were negative or exceeded 100. In this way, only physically meaningful positive values were assigned to the model parameters. The penalties for the values exceeding 100 helped to eliminate physically impossible organic C contents of the MCR-ALS components (OC_i) given in %, and strong dominance of some $\frac{C_i}{C_1}$ ratios in the model (Eq. (11)). Hence, when the optimal values of the model parameters were found, the penalty did not contribute to a final loss function. Fitting was repeated with 30 randomly generated combinations of initial values of the adjustable parameters, varying from 0 to 100, to ensure that a stable minimum of the loss function was reached. The Hooke-Jeeves pattern move algorithm providing a fast optimization was used for model fitting. Table S2 shows the results of 10 optimizations starting from different initial values of adjustable parameters, thus illustrating the stability of the optimization.

3. Results and discussion

3.1. Decomposition of soil mid-IR spectra

The preliminary examination of 213 IR absorbance spectra of the soil samples using the SVD plot (Jaumot et al., 2015) suggested that four components explain the variability of the data. This model explained 99.7% of the variance, with a fitting error (Lof) of 5.88% and a standard deviation of residuals of 0.037. When varying the number of the components from 3 to 6 and plotting RMSD and Lof against the number of the components (Fig. S2), there was a linear trend covering changes from 4 to 6 components. For a three-component decomposition, elevated RMSD and Lof values, distinctly deviating from this linear association, were found. Hence, four components were used for further analysis.

Mid-IR spectral loadings of the four components are presented in Fig. 1, whereas the positions and assignments of absorbance maxima are described in Table 1. Components 1, 2 and 3 differ in a specific combination of absorbance peaks associated with inorganic constituents such as clay minerals, possibly, iron-containing phases (such as oxides and oxyhydroxides), and organic matter. It is worth noting that among components 1–3 (Fig. 1), component 2 is characterized by having the strongest absorbance at 1435 cm^{-1}, which together with distinct absorbance at 2515, and sharp peaks at 872 and 714 cm^{-1}, indicates that this component represents carbonate-rich materials. All the soil samples studied contained inorganic carbon commonly associated with carbonates (the illustrative data are provided in the overview of soil samples in Supplementary information, Table S1). When looking at the absorbance at 1032 cm^{-1} and bands >3300 cm^{-1}, component 3 seems to be enriched as compared with component 1 by chemical bonds absorbing at 1032 cm^{-1}. This suggests the enrichment of component 3 as compared with component 1 with aluminosilicate Si–O bonds and the organic C–O/C–O–C groups relative to the presence of diverse inorganic and organic OH groups (Table 1). This further implies that components 1 and 3 are essentially different organo-mineral complexes whereas component 3 seems to be more hydrophobic (with a lesser relative content of hydrophilic OH groups) and possibly enriched by SOM. The fourth component is characterized by absorbance bands similar to those found for components 1–3 above 3000 and below 600 cm^{-1}. However, in the 1000-2000 cm^{-1} range, the spectrum is complicated and exhibits more "noisy" shapes, in particular, for absorbance in the 1500-2000 cm^{-1} range. It is difficult to know why this less regular shape of absorbance appears here, and whether it is caused by the ability of the MCR-ALS method to separate the contributions from the IR-absorbing components, spectral noise, or other non-accounted effects which may appear when using KBr pellets, e.g., light scatter (Tesfamichael et al., 2001). Less regular absorbance profiles in the 1000-1800 cm^{-1} range were found by Fritzsche et al. (2019) who examined mid-FTIR spectra of ground water-derived solids in KBr pellets and applied PMF, an MCR technique,

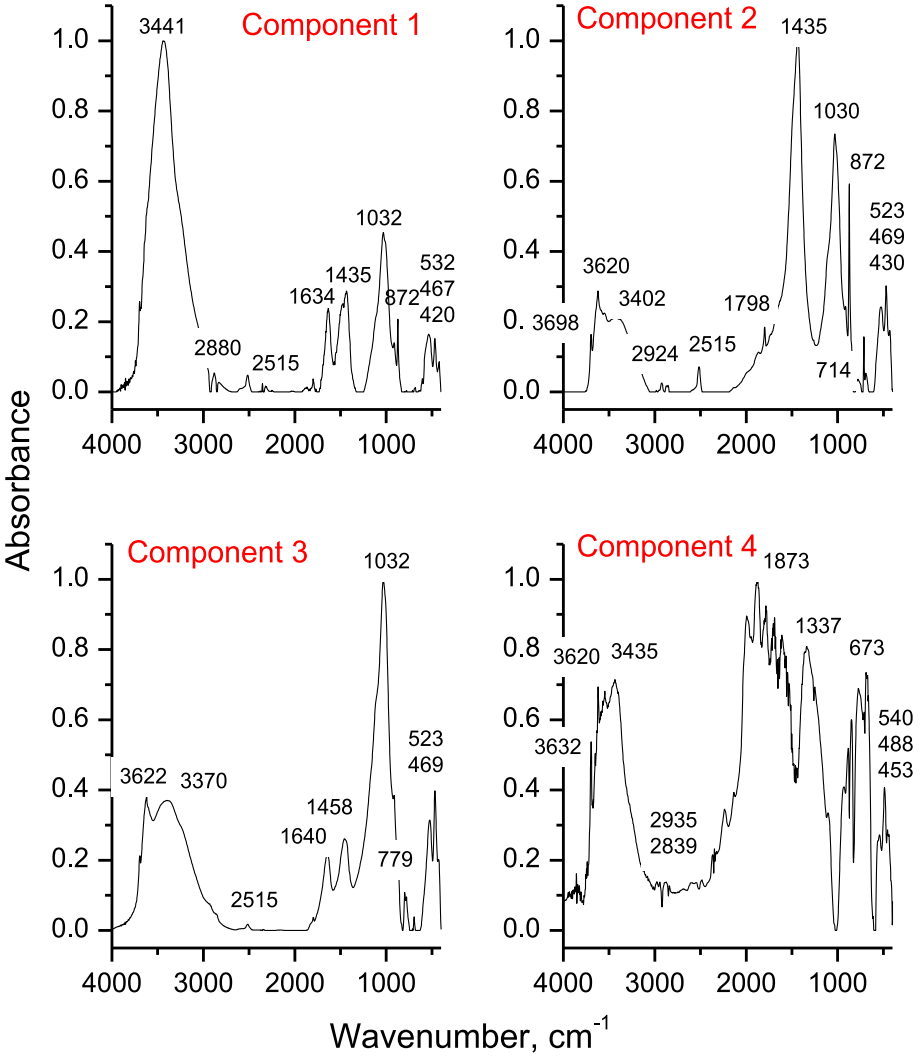


Fig. 1. MCR-ALS identified mid-IR spectra of the four soil components that contribute to the whole soil mid-IR spectra. Spectra are normalized to unit absorbance at its maximal value.

Table 1
Positions of mid-IR absorbance maxima (cm^{-1}), with indicating a number of a specific MCR component (in the parentheses) and their assignments.

Wavenumbers of positions of maxima (or intervals), cm^{-1}	Assignment
3698(2), 3620(2), 3622(3), 3632(4)	O–H stretching in kaolinite[a]
3441(1), 3402(2), 3370(3), 3435(4)	O–H stretching in phyllosilicates[b], water, carboxyl and hydroxyl groups[c]
2880(1), 2924(2), 2935–2929(4)	Symmetric and asymmetric C–H stretching of methyl and methylene groups[d]
2515 (1–3)	Vibrations of CO_3^{2-} in calcite[e]
1798(2), 1873(4)	Overtones and combination bands in quartz and silicates[f]
1634(1), 1640(3)	Asymmetric stretching of COO^-[g]; amide I (C=O stretching) and II (NH and NH_2 bending) bands[h]; OH-bending of hydration water in phyllosilicates[i]; H-bonded C=O[j]; the conjugated C=C stretching in ketones, carboxylic acids and amides[k];
1435(1,2), 1458(3)	Vibrations of CO_3^{2-} in calcite[l]; C–H bending of methyl and methylene groups[h]
1337 (4)	Symmetric stretching vibrations of COO^-[g]
1032 (1,3), 1030(2)	Si–O stretching in aluminosilicates[b]; the C–O/C–O–C vibrations in alcohols, phenols, carboxylic acids, polysaccharides, carbohydrates[d,m]
714, 872–873 (1,2,4)	Vibrations of CO_3^{2-} in calcite[e], iron oxides and oxyhydroxides[n]
523-540(1–4), 467–488(1-4), 420–453 (1,2,4)	Bending of O–Si–O bonds; vibrations of AlO_6-octahedral groups[o]; Si–O stretching[b]

[a] McKissock et al., 2003; Nguyen et al. (1991); Ravisankar et al. (2011); Simkovic et al. (2008).
[b] Margenot et al., 2017.
[c] Ouatmane et al., 2000; Ellerbrock and Gerke (2004); Simkovic et al. (2008).
[d] Silverstein & Webster (1997); Ellerbrock and Gerke (2004); Simkovic et al. (2008).
[e] Nguyen et al., 1991; Ravisankar et al. (2011); Tinti et al. (2015).
[f] Nguyen et al., 1991; Calderón et al. (2011).
[g] Silverstein & Webster (1997); Hay and Myneni (2007).
[h] Silverstein & Webster (1997).
[i] Spaccini et al., 2001.
[j] Schnitzer, 1978.
[k] Silverstein & Webster (1997); Ellerbrock and Gerke (2004); Ellerbrock and Kaiser (2005); Simkovic et al. (2008).
[l] Huang & Kerr (1960); Gunasekaran et al. (2006).
[m] Rumpel et al., 2001; Smidt and Schwanninger (2005); Tarchitzky et al. (2007).
[n] Soriano-Disla et al., 2014.
[o] Fernández-Carrasco et al., 2012.

for the decomposition of spectra. Of the seven IR light-absorbing components identified by Fritzsche et al. (2019), two components, termed "organic matter-like", exhibited a more complex shape than the other ones. Increasing the number of components in the MCR- ALS decomposition seemed to lead to the appearance of noisy shapes also in spectra of other constituents (data not shown) which supported our decision to use just four components for further TOC modeling.

In summary, for the goal of soil TOC modeling, use of the MCR-ALS methodology enabled the identification of four components (and obtaining their concentration scores). These four identified components were capable of providing the reasonable description of IR spectroscopic signatures of soil composition.

3.2. Modeling the TOC contents of the whole sample set

The four MCR-ALS components contributing to soil mid-IR spectra were incorporated into Eq. (11) to describe the TOC contents of all the soil samples. Table 2 summarizes the statistics of this description and the optimized values of the adjustable parameters. Only a small portion (<27%) of the TOC variance of the whole dataset was explained, and the RMSD value was quite large (0.54% w w^{-1}). For comparison, an earlier analysis by Nasonova et al. (2022) using the PLS regression with the same dataset allowed to model the TOC values with an explained variance as high as 76%. The RMSD based on the PLS regression modeling was 0.36% w w^{-1} (obtained from the statistical measures reported by Nasonova et al. (2022) and original experimental data). Hence, when modeling TOC contents of the whole dataset, the commonly used PLS regression was more successful as compared with the current method based on the MCR decomposition. It is worth noting that for this dataset the variance explained by the PLS regression was shown to be well within the explained variance observed in a series of publications (i.e., 52–96%; detailed in Nasonova et al., 2022).

To visualize the quality of modeling, the fitted TOC values were plotted against the measured TOC values in Fig. 2a. Inspection of the data distribution in Fig. 2a suggests that the MCR-ALS-based model fails in soil samples with elevated TOC contents. This failure is demonstrated when plotting the differences between the fitted and experimental TOC contents against the experimental data (Fig. 2b). Such a tendentious distribution of residuals is a clear indication of a systematic failure, i.e., lack-of-fit, for a model $TOC_{fit} = TOC_{exp}$ (Draper and Smith, 1968).

The clear trend of increasing deviations between experimental and fitted TOC values with increasing TOC content in the soil samples deserves special attention. It suggests that in the samples with elevated TOC contents, a portion of the SOM has weaker relations with measured soil mid-IR spectra. One qualitative understanding of this apparently paradoxical conclusion may result from the existence of SOM fractions in soil, that interact differently with soil minerals, including, for example, mineral-associated SOM as well as organic matter having less interactions with minerals (e.g., particulate organic matter composed of residuals from plants and microbes, black carbon). Soil inorganic components have dominant contributions to soil mid-IR spectra, and therefore, the MCR-ALS decomposition intending to provide the best fit of IR spectra accounts foremost for the mineral spectral signatures. Hence, if these mineral components could be characterized by certain *OC*-holding capacity (as is implied in Eq. (11)), the content of mineral-bound organic matter in soil samples with lower TOC content would be better modeled as compared to the case of elevated soil TOC contents. In the latter case, the relation between minerals (and their pronounced IR spectral fingerprints) and TOC becomes weaker, and, therefore, the fitting of the TOC values becomes less successful. This explanation proposes that the ability of soil mid-IR spectra to model TOC content using Eq. (11) was rather indirect, at least, in part, and strongly affected by the association between SOM and minerals.

Considering that the fourth component has a less regular spectral profile that could be associated with spectral artifacts (Fig. 1; section 3.1) and the fitted model showed zero *OC* content in this component (Table 2), we have examined, therefore, the model performance when the fourth component was omitted, i.e., using only three components of the four-component model. The model parameters are summarized also in Table 2, where it can be observed that the model performance, i.e., the RMSD value and the variance explained, did not depend on whether four or three components were incorporated into Eq. (11); this is a proof that the fourth component had no impact on the TOC modeling for the whole dataset.

Table 2
Parameters of models linking the MCR-ALS-derived scores of the mid-IR-active components to the whole soil TOC content (Eq. (11)). The $k_{i,1}$ and OC_i values are followed with their standard errors (in parentheses); the p value if less than 0.1 is specified in the footnotes.

	Component *i*					
	1	2	3	4	RMSD, % w w^{-1}	Variance explained, %
The whole dataset: Four components					0.539	27
$k_{i,1}$ (unitless)	1 (–)	1.50 (3.16)	1.44 (4.83)	0.72 (5.8)		
OC_i, % w w^{-1}	0.16 (0.74)	0.00 (0.32)	1.49 (0.86)[a]	0.00 (4.5)		
The whole dataset: Three components (of the four-component model)					0.540	27
$k_{i,1}$ (unitless)	1 (–)	1.34 (0.92)	1.13 (0.78)	–		
OC_i, % w w^{-1}	0.12 (0.28)	0.00 (0.22)	1.55 (0.21)[b]	–		
The "low" dataset[c]: Four components					0.175	62
$k_{i,1}$ (unitless)	1 (–)	1.73 (2.17)	1.41 (1.72)	4.25 (5.82)		
OC_i, % w w^{-1}	0.31 (0.16)[d]	0.15 (0.08) [a]	0.90 (0.08)[b]	0.00 (0.14)		
The "low" dataset: Three components (of the four-component model)					0.181	60
$k_{i,1}$ (unitless)	1 (–)	1.07 (0.55)	0.70 (0.30)[d]	–		
OC_i, % w w^{-1}	0.16 (0.08)[d]	0.09 (0.07)	0.94 (0.06)[b]	–		
The "high" dataset[e]: Four components					0.521	13
$k_{i,1}$ (unitless)[f]	1 (–)	1.47	3.79	0.05		
OC_i, % w w^{-1} [f]	2.22	0.00	1.30	70.5		

[a] p<0.10.
[b] p < 0.000.
[c] "low": TOC<1% w w^{-1}.
[d] p < 0.05.
[e] "high": TOC>1% w w^{-1}.
[f] It is not possible to provide standard errors for determinable parameters since the matrix is ill-conditioned.

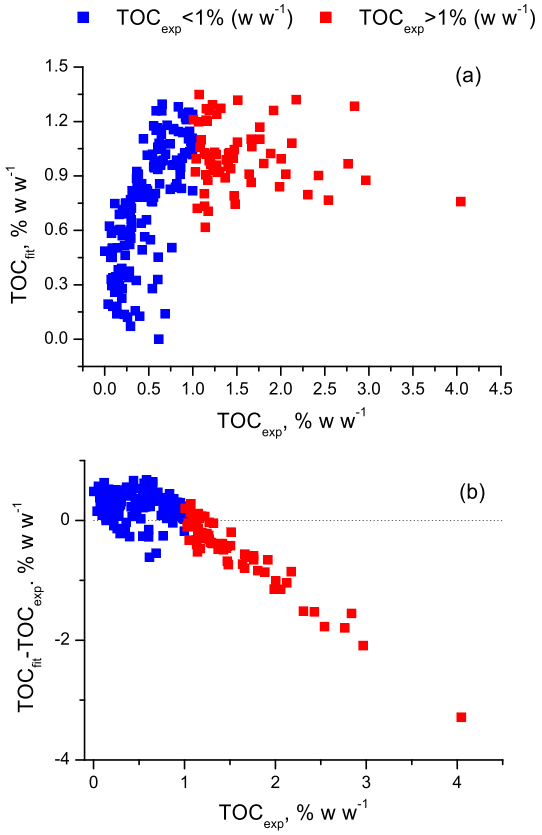


Fig. 2. (a) TOC contents of soils approximated by using Eq. (11) (TOC_{fit}) are plotted against experimental values (TOC_{exp}). (b) Differences between TOC_{fit} and TOC_{exp} are plotted against TOC_{exp}. The dotted line represents zero differences between TOC_{fit} and TOC_{exp}.

Using three or four components in the four-component model (Eq. (11)) had no significant effect also on the values of model parameters (Table 2). Component 3 might be of particular interest as its computed *OC* content is determined as a finite non-zero value, thus making it a major component that controls soil TOC content.

3.3. Separating the soil sample set in the subsets with different levels of TOC contents

In light of the different behavior of the model at different soil TOC ranges (Fig. 2b), the whole dataset was divided into two subsets: one containing samples with TOC<1%, w w^{-1} (termed "low"; 142 IR spectra) and one containing samples with TOC>1% w w^{-1} (termed "high"; 71 IR spectra). This 1% w w^{-1} threshold was selected because at this TOC content, the difference between the fitted and experimental TOC contents approaches zero (Fig. 2b). Both subsets were fitted, maintaining the same number (four) of the MRC-ALS components as proposed in the examination of the whole dataset.

The fitted model to the "low" subset explained 99.7% of the variance and had Lof of 5.87%, and a standard deviation of residuals of 0.039. The "high" subset optimization resulted in 99.8% of the variance explained, Lof of 4.97%, and the standard deviation of residuals of 0.029. Fig. 3a presents the four components from the decomposition of the whole dataset, and the "low" and "high" datasets. It is seen in Fig. 3a that the spectra of components 1–3 obtained in the whole dataset and the "low" subset essentially coincide, except for some small differences in the spectra of the third component at wavenumbers above 3000 cm^{-1}. This coincidence is supported by a direct examination of determination coefficients calculated for correlations between mean-centered absorbances of the spectra under comparison (Varmuza et al., 2003), shown in Table 3. For the first three components, the r^2 values describing correlations between spectra obtained from the whole dataset and its "low" part are $\geq$ 0.99. Even for the fourth component, the correlation is high, with r^2 = 0.914. However, when examining the spectra of the four identified components in the "high" subset, they show clear differences from the spectra associated with the whole dataset. These differences are reflected in r^2 values lower than 0.9 (Table 3). This suggests that the mid-IR spectra of the components extracted from the whole dataset were characteristic foremost for the "low" dataset which could be due to the imbalance in the number of samples in the "low" and "high" datasets (142 and 71, respectively).

3.4. Modeling the TOC contents of sample subsets

We applied Eq. (11) to model the TOC content of soil samples belonging to the two subsets, "low" and "high", based on the concentration scores of the four components determined separately for each of

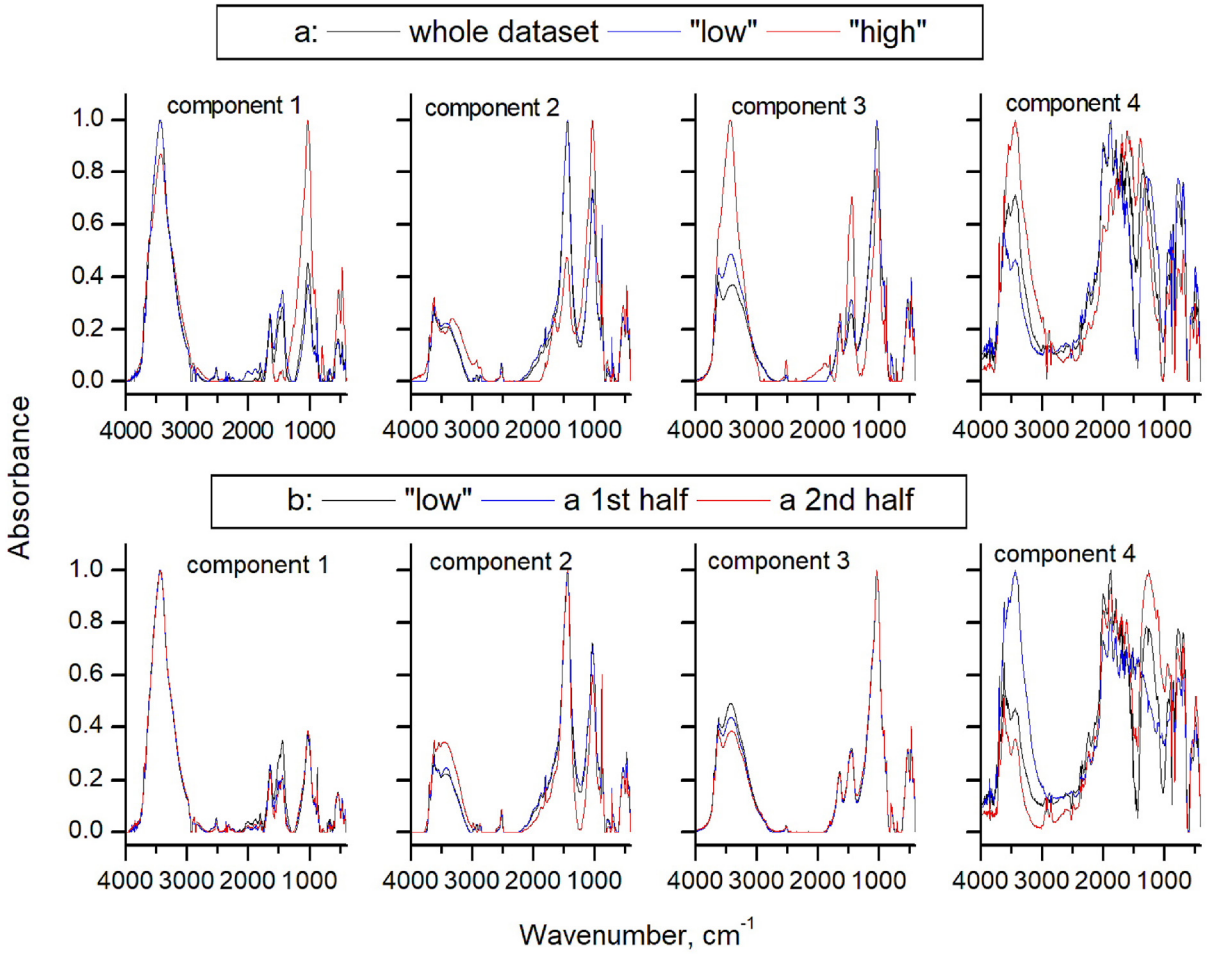


Fig. 3. (a) Mid-IR spectra of the four MCR-identified soil components in the whole dataset and its two subsets, including the samples with TOC content below 1% (w w^{-1}; "low") and above ("high"). (b) Mid-IR spectra of the four MCR-identified soil components in the "low" subset and in its two randomly obtained halves. Spectra are normalized to unit absorbance at its maximal value.

Table 3
Correlations between the component's spectra identified in different datasets.

Datasets of spectra decomposed into components' contributions, under comparison	Component (as depicted in Figs. 1 and 3)	r^2
Whole dataset *vs* "low" dataset	1	0.995
	2	0.999
	3	0.990
	4	0.914
Whole dataset *vs* "high" dataset	1	0.863
	2	0.813
	3	0.759
	4	0.864
The "low" dataset *vs* its 1st half	1	0.995
	2	0.999
	3	0.999
	4	0.707
The "low" dataset *vs* its 2nd half	1	0.992
	2	0.954
	3	0.992
	4	0.909
The "low" dataset: 1st half *vs* 2nd half	1	0.999
	2	0.948
	3	0.997
	4	0.621

them. The results of this modeling, including the values of adjustable parameters, standard errors, RMSD, and explained variance are provided in Table 2. Fig. 4 shows the fitted TOC values plotted against the experimental ones, for each of two soil sample subsets. The separation of the whole sample set led to an essential improvement in modeling soil TOC<1% (w w^{-1}), with an explained variance of 62% and RMSD of 0.175% w w^{-1} (Table 2) which is comparable with commonly explained variance when modeling TOC using PLS regressions (e.g., 52–96%;

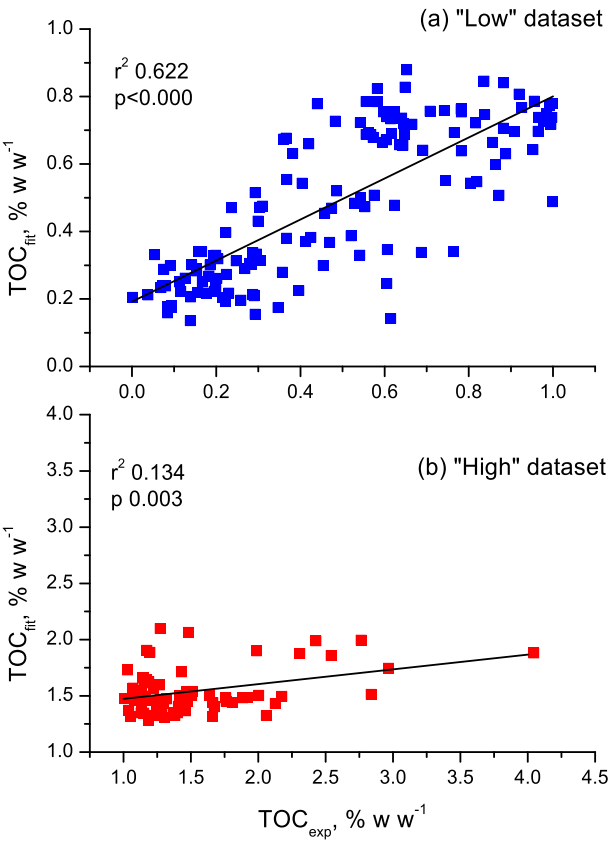


Fig. 4. TOC contents of soils approximated using Eq. (11) (TOC_{fit}) are plotted against experimental values (TOC_{exp}): (a) the soil subset with TOC<1% w w^{-1} ("low"); (b) the soil subset with TOC>1% w w^{-1} ("high"). The straight lines represent the linear regressions.

Nduwamungu et al., 2009; Baldock et al., 2013; Soriano-Disla et al., 2014; Matamala et al., 2019; Gomez et al., 2020). However, note that interpretations of PLS regressions are not obvious (Pirouz, 2010; Fritzsche et al., 2019; Xia, 2020; Lado et al., 2023). In contrast, the MCR-ALS-based descriptors are associated with interpretable sets of IR-active chemical bonds and functional groups, where changes in the concentration scores of MCR-ALS components reflect changes in sample chemical composition.

Two MCR-ALS components, the first and the third, are characterized by the largest OC_i values and contribute distinctly to the soil TOC in the samples belonging to the "low" dataset modeled with four components (Table 2). Once again, as in the whole dataset, component 4 did not contribute to TOC. Interestingly, the computed *OC* content associated with the third component (0.90 ± 0.08% w w^{-1}) is essentially three times larger than that found for the first component (0.31 ± 0.16% w w^{-1}). This is in agreement with an earlier conclusion that components 1 and 3 are different organo-mineral complexes whereas component 3 seems to possibly be enriched by SOM (section 3.1). Further proof of the robustness of the IR spectra decomposition using MCR-ALS emerged from a split-half analysis which involved the separation of the "low" subset into two randomly selected halves. Each half was once again decomposed into components using MCR-ALS, with the variance explained being 99.6 and 99.7%, Lof being 6.12 and 5.18%, and the standard deviation of residuals being 0.040 and 0.034, for the first and second half, respectively. Fig. 3b compares the spectral loadings of the components obtained during the decomposition of mid-IR spectra of the "low" subset and its two random halves. The spectral loadings of the first three components are very close in these three datasets. This similarity is evidenced also by strong correlations between spectra in paired comparisons (high r^2, Table 3). Thus, the identification of these components did not depend substantially on the specific dataset as they have been found in all four datasets: the whole one, a "low" subset, and the two halves of the "low" subset (Fig. 3a and b).

In the above comparison (Fig. 3b), the fourth component behaved differently; its spectral loading depended on the specific dataset, and the correlations of its spectral loading between different datasets were much weaker (Table 3). As mentioned earlier regarding the analysis of the whole dataset, the role of this fourth component in modeling the TOC content seems negligible. Indeed, when omitting the fourth component from the fitting of Eq. (11) for the "low" dataset, TOC modeling yielded practically the same RMSD as with the four components, a very similar percentage of explained variance, and well-comparable values of the adjusted model parameters, including detectable non-zero *OC* contents of components 1 and 3 (Table 2).

In contrast to the modeling of the "low" subset, the dataset with TOC exceeding 1% w w^{-1} threshold, was poorly modeled, with RMSD 0.521% w w^{-1} (Table 2) showing that the model (Eq. (11)) cannot accurately represent elevated contents of TOC. In this poorly modeled case, Statistica 7.0 Statsoft did not allow to obtain the standard errors of parameters, reporting the ill-conditioned data matrix.

3.5. *Feasibility of the model, its assumptions and perspectives – discussion*

Recalling our third objective, which was to test whether the MCR-ALS-derived soil components allow modeling soil TOC content, we noted that the modeling success depended on a specific range of TOC content. As explained in section 2.1, there are two central assumptions in the approach we used. One is that it is possible to identify chemically meaningful components explaining soil mid-IR spectra using nonnegative MCR-ALS decomposition. This assumption seems to be valid considering the high extent of the IR absorbance variance explained by the nonnegative MCR-ALS components in the whole dataset (99.7%), its subsets (99.7% and 99.8%, for the "low" and "high" subsets, respectively), and the robustness of decomposition (as discussed above, section 3.4). It is worth keeping in mind that the MCR-ALS–derived components are not individual substances but each one may include different organic molecules and inorganic constituents, such as clay minerals, carbonates, metal oxides, oxyhydroxides and others. In an MCR-ALS component, the individual chemical materials maintain constant proportions in the samples studied. A high degree of the absorbance variance explained indicates that if above-mentioned chemical materials contribute meaningfully to the measured IR spectra, they are also represented by MCR-ALS components.

The second assumption was that it is possible to identify all TOC-controlling components using the MCR-ALS-based decomposition of IR spectra. The results of the work showed that this assumption was violated when soil TOC content exceeded 1% w w^{-1}. The existence of the TOC threshold resulted in a sharp loss of capability of the MCR-ALS components to model soil TOC at >1% w w^{-1}. At elevated TOC contents the MCR-ALS-based decomposition of soil mid-IR spectra, although explaining more than 99.7% of the spectral variance, did not succeed in quantifying all TOC content-controlling components. The threshold determined for the current particular dataset is not necessarily applicable to other soil samples.

It is hardly plausible that there are SOM pools that are not active in mid-IR spectra. Hence, the failure of TOC modeling at elevated TOC contents suggests that some important soil constituents controlling TOC content are masked in soil mid-IR spectra by others. A natural explanation is that inorganic constituents (e.g., clay minerals) dominate in the MCR-ALS decomposition of soil mid-IR spectra. Hence, modeling of the soil TOC content by MCR-ALS components below 1% w w^{-1} was probably indirect, at least, in part. It can be strongly linked to minerals present that variously bind SOM and form organo-mineral complexes. A loosely mineral-bound SOM is probably less mirrored by mineral IR spectral fingerprints and, therefore, its contribution to TOC is only poorly accounted for. Thus, the 1% w w^{-1} threshold suggests that above this

value, in our studied soils there is a significant part of SOM that is not linked tightly to soil minerals. Hence, MCR-ALS components identified in mid-IR spectra and dominated by the minerals could not account (most probably implicitly) for this SOM fraction.

The above mentioned explanation should yet be considered as a hypothesis. Nevertheless, the existence of a threshold in the TOC modeling efficacy using MCR-ALS components extracted from soil IR spectra is of interest. This may provide a way to distinguish soil samples with dominance of different types of SOM, from those ones "tightly" bound to minerals (and therefore, their contribution to TOC is correlated with mineral IR spectral fingerprints), to those rather loosely bound to minerals or particulate SOM. The SOM fraction "tightly bound to minerals" might be related to (not necessarily coincide with) the mineral-associated fraction of SOM determined by common methods (Yu et al., 2022). Thus, the TOC "threshold", if identified, can be informative when associations between minerals and SOM become weaker, or a mineral's capacity to interact with SOM becomes, to a certain extent, exhausted. This phenomenon is of interest for better understanding the SOM fractions that control carbon storage and stabilization (Kögel-Knabner and Rumpel, 2018; Lehmann and Kleber, 2015). Hence, it could be a future task to apply the nonnegative MCR-ALS decomposition of mid-IR spectra and to perform TOC modeling in whole soil samples and their fractions differing by contributions of mineral-associated SOM, and to examine whether there are links between the model efficacy to describe TOC content and the mineral-associated SOM fraction in the whole TOC. This type of work is currently on its way.

The major current limitation of the proposed approach to model soil TOC is seen as the lack of applicability to soils widely differing in TOC. Hence, the further expected (and important) question is whether (and how) this approach can be successfully extended to diverse soils differing in TOC content and mineral composition. As follows from the above discussion, the problem in TOC modeling may result when (i) TOC is contributed by portions of SOM that are not associated sufficiently strongly with mineral surfaces, so that it could be accounted for indirectly through mineral contributions to IR spectra, and (ii) the SOM contribution *per se* to IR spectra of soils and MCR-ALS components is not significant. Therefore, in its current form, our approach is expected to be applicable for SOM-poor soils, yet rich in minerals that provide significant interfaces for interacting with SOM, or soils rich by SOM contributing more distinctly to measured IR spectra. TOC content in soils "intermediately" enriched by SOM will be less successfully modeled using MCR-ALS components derived from IR spectra (it is not possible yet to provide strict limits for "rich", "poor" or "intermediate" cases). It is also expected that in soils rich in SOM, better fingerprinting organic matter in IR spectra should lead to an improved modeling of the TOC content, but an identification of a SOM pool stronger interacting with minerals may become more difficult.

A possible way to improve the efficacy of the described approach that explores MCR-ALS decomposition of mid-IR spectra and to make it more universal for modeling TOC contents in soils is to use those regions in soil mid-IR spectra that are largely contributed by molecular vibrations of SOM constituents. The latter will make SOM contributions to MCR-ALS-derived components more influential. However, an effective solution that can essentially improve the ability of Eq. (11) (and nonnegative MCR-ALS components derived from soil IR spectra) to model soil TOC could be based on consideration of a principal difference between the common use of the PLS regressions and the approach proposed in this paper. From the beginning, PLS regressions are forced to fit the goal that is soil TOC content. The MCR-ALS decomposition is not related *per se* to the soil TOC modeling which is performed with Eq. (11). Hence, the additional constraint may be introduced into the MCR-ALS decomposition of soil mid-IR spectra. This constraint is the requirement of best fit for soil TOC values (through Eq. (11)). The soil TOC best-fit constraint is supposed to make decomposition "supervised" and less flexible, but the obtained components will be forced to fit TOC values better.

4. Conclusions

A bilinear nonnegative MCR-ALS decomposition of soil mid-IR absorbance spectra provides a way to identify meaningful sets of chemical bonds and functional groups maintaining constant proportions among soil samples. Decomposed spectral components representing identifiable soil components can be linked to multiple soil properties, and, foremost to soil TOC contents, thus providing a basis for mechanistic physicochemical models.

In our particular soil sample set, the performance of the developed mechanistic model in representing soil TOC values depended on a specific TOC range and revealed a threshold TOC in model efficacy. The appearance of this threshold upon the increase in soil TOC content may indicate the significance of SOM which could be particulate or weakly associated with minerals. Identifying such a threshold TOC for model performance could be of great interest considering the commonly accepted importance of mineral-associated SOM for soil carbon storage and sequestration.

This work is the first attempt to perform a nonnegative MCR-ALS decomposition of soil mid-IR spectra, and in particular, to link this decomposition to soil TOC. Certainly, more work is needed to extend and verify the applicability of the proposed approach to soils with variable mineral composition and SOM content and examine the relations between the model capability to represent soil TOC and the distribution of SOM between the fractions differing in their association with minerals.

CRediT authorship contribution statement

Mikhail Borisover: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Marcos Lado:** Writing – review & editing, Visualization, Methodology, Formal analysis, Conceptualization. **Guy J. Levy:** Writing – review & editing, Visualization, Formal analysis, Conceptualization.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.seh.2024.100123.

References

Araya, J.A., Carneiro, R.L., Freer, J., Neira, J.Y., Castillo, R. del P., 2019. Fourier transform infrared imaging and quantitative analysis of pre-treated wood fibers: a comparison between partial least squares and multivariate curve resolution with alternating least squares methods in a case study. Chemometr. Intell. Lab. Syst. 195. https://doi.org/10.1016/j.chemolab.2019.103890.

Baldock, J.A., Beare, M.H., Curtin, D., Hawke, B., 2018. Stocks, composition and vulnerability to loss of soil organic carbon predicted using mid-infrared spectroscopy. Soil Res. 56, 468–480. https://doi.org/10.1071/SR17221.

Baldock, J.A., Broos, K., 2012. Soil organic matter. In: Huang, P.M., Li, Y., Sumner, M.E. (Eds.), Hanbook of Soil Sciences. Properties and Processes. CRC Press, Taylor and Francis Group, pp. 11–52.

Baldock, J.A., Hawke, B., Sanderman, J., MacDonald, L.M., 2013. Predicting contents of carbon and its component fractions in Australian soils from diffuse reflectance mid-infrared spectra. Soil Res. 51 (7–8), 577–595. https://doi.org/10.1071/SR13077.

Barra, I., Haefele, S.M., Sakrabani, R., Kebede, F., 2021. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: recent advances–A review. TrAC, Trends Anal. Chem. 135, 116166. https://doi.org/10.1016/j.trac.2020.116166.

Barthès, B.G., Chotte, J.L., 2021. Infrared spectroscopy approaches support soil organic carbon estimations to evaluate land degradation. Land Degrad. Dev. 32 (1), 310–322. https://doi.org/10.1002/ldr.3718.

Borisover, M., Bar-Tal, A., Bukhanovsky, N., Berezkin, A., Lado, M., Levy, G.J., 2023. Optical properties of water-extractable organic matter as indicators for soil organic

matter response to irrigation water quality and management. SSRN Electron. J. https://doi.org/10.2139/ssrn.4491036.
Bro, R., De Jong, S., 1997. A fast non-negativity-constrained least squares algorithm. J. Chemometr. 11 (5), 393–401. https://doi.org/10.1002/(SICI)1099-128X(199709/10)11:5<393::AID-CEM483>3.0.CO;2-L.
Bro, R., Viereck, N., Toft, M., Toft, H., Hansen, P.I., Engelsen, S.B., 2010. Mathematical chromatography solves the cocktail party effect in mixtures using 2D spectra and PARAFAC. TrAC, Trends Anal. Chem. 29 (4), 281–284. https://doi.org/10.1016/j.trac.2010.01.008.
Bünemann, E.K., Bongiorno, G., Bai, Z., Creamer, R.E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T.W., Mäder, P., Pulleman, M., Sukkel, W., van Groenigen, J.W., Brussaard, L., 2018. Soil quality – a critical review. Soil Biol. Biochem. 120, 105–125. https://doi.org/10.1016/j.soilbio.2018.01.030.
Calderón, F.J., Reeves, J.B., Collins, H.P., Paul, E.A., 2011. Chemical differences in soil organic matter fractions determined by diffuse-reflectance mid-infrared spectroscopy. Soil Sci. Soc. Am. J. 75 (2), 568–579. https://doi.org/10.2136/sssaj2009.0375.
Câmara, A.B.F., da Silva, W.J.O., Moura, H.O.M.A., Silva, N.K.N., de Lima, K.M.G., de Carvalho, L.S., 2022. Multivariate strategy for identifying and quantifying jet fuel contaminants by MCR-ALS/PLS models coupled to combined MIR/NIR spectra. Anal. Bioanal. Chem. 414 (27), 7897–7909. https://doi.org/10.1007/s00216-022-04324-9.
Davis, W.M., Erickson, C.L., Johnston, C.T., Delfino, J.J., Porter, J.E., 1999. Quantitative Fourier transform infrared spectroscopic investigation of humic substance functional group composition. Chemosphere 38 (12), 2913–2928. https://doi.org/10.1016/S0045-6535(98)00486-X.
de Juan, A., Tauler, R., 2021. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – a review. Anal. Chim. Acta 1145, 59–78. https://doi.org/10.1016/j.aca.2020.10.051.
Draper, N.R., Smith, H., 1968. Applied Regression Analysis. John Wiley and Sons, Inc.
Ellerbrock, R.H., Gerke, H.H., 2004. Characterizing organic matter of soil aggregate coatings and biopores by Fourier transform infrared spectroscopy. Eur. J. Soil Sci. 55 (2), 219–228. https://doi.org/10.1046/j.1365-2389.2004.00593.x.
Ellerbrock, R.H., Kaiser, M., 2005. Stability and composition of different soluble soil organic matter fractions - evidence from δ13C and FTIR signatures. Geoderma 128 (1–2), 28–37. https://doi.org/10.1016/j.geoderma.2004.12.025.
Fernández-Carrasco, L., Torrens-Martín, D., Morales, L.M., Martínez-Ramírez, S., 2012. Infrared spectroscopy in the analysis of building and construction materials. Infrared Spectroscopy - Materials Science, Engineering and Technology. Intech. https://doi.org/10.5772/36186.
Fritzsche, A., Ritschel, T., Schneider, L., Totsche, K.U., 2019. Identification and quantification of single constituents in groundwater with Fourier-transform infrared spectroscopy and Positive Matrix Factorization. Vib. Spectrosc. 100, 152–158. https://doi.org/10.1016/j.vibspec.2018.09.008.
Gomez, C., Chevallier, T., Moulin, P., Bouferra, I., Hmaidi, K., Arrouays, D., Jolivet, C., Barthès, B.G., 2020. Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library. Geoderma 375, 114469. https://doi.org/10.1016/j.geoderma.2020.114469.
Gossart, P., Semmoud, A., Ruckebusch, C., Huvenne, J.P., 2003. Multivariate curve resolution applied to Fourier transform infrared spectra of macromolecules: structural characterisation of the acid form and the salt form of humic acids in interaction with lead. Anal. Chim. Acta 477 (2), 201–209. https://doi.org/10.1016/S0003-2670(02)01415-0.
Gunasekaran, S., Anbalagan, G., Pandi, S., 2006. Raman and infrared spectra of carbonates of calcite structure. J. Raman Spectrosc. 37 (9), 892–899. https://doi.org/10.1002/jrs.1518.
Hay, M.B., Myneni, S.C.B., 2007. Structural environments of carboxyl groups in natural organic molecules from terrestrial systems. Part 1: infrared spectroscopy. Geochem. Cosmochim. Acta 71 (14), 3518–3532. https://doi.org/10.1016/j.gca.2007.03.038.
Huang, C.K., Kerr, P.F., 1960. Infrared study of carbonate minerals. Am. Mineral. 45 (3–4), 311–324.
Ioannidi, E., Aarøe, E., de Juan, A., Risbo, J., van den Berg, F.W.J., 2023. Modeling changes in chocolate during production and storage by ATR-FT-IR spectroscopy and MCR-ALS hybrid soft and hard modeling. Chemometr. Intell. Lab. Syst. 233, 104735. https://doi.org/10.1016/j.chemolab.2022.104735.
Jaumot, J., de Juan, A., Tauler, R., 2015. MCR-ALS GUI 2.0: new features and applications. Chemometr. Intell. Lab. Syst. 140, 1–12. https://doi.org/10.1016/j.chemolab.2014.10.003.
Jaumot, J., Gargallo, R., De Juan, A., Tauler, R., 2005. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. Chemometr. Intell. Lab. Syst. 76 (1), 101–110. https://doi.org/10.1016/j.chemolab.2004.12.007.
Jaumot, J., Tauler, R., 2010. MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution. Chemometr. Intell. Lab. Syst. 103 (2), 96–107. https://doi.org/10.1016/j.chemolab.2010.05.020.
Johnston, C.T., Aochi, Y.O., 1996. Fourier transform infrared and Raman spectroscopy. Chapter 10. In: Sparks, A.L., et al. (Eds.), Methods of Soil Analysis, Part 3: Chemical Methods. Soil Science Society of America and American Society of Agronomy, pp. 269–321. https://doi.org/10.2136/sssabookser5.3.c10.
Just, C., Armbruster, M., Barkusky, D., Baumecker, M., Diepolder, M., Döring, T.F., Heigl, L., Honermeier, B., Jate, M., Merbach, I., Rusch, C., Schubert, D., Schulz, F., Schweitzer, K., Seidel, S., Sommer, M., Spiegel, H., Thumm, U., Urbatzka, P., et al., 2023. Soil organic carbon sequestration in agricultural long-term field experiments as derived from particulate and mineral-associated organic matter. Geoderma 434, 116472. https://doi.org/10.1016/j.geoderma.2023.116472.
Kaufhold, S., Hein, M., Dohrmann, R., Ufer, K., 2012. Quantification of the mineralogical composition of clays using FTIR spectroscopy. Vib. Spectrosc. 59, 29–39. https://doi.org/10.1016/j.vibspec.2011.12.012.
Kögel-Knabner, I., Rumpel, C., 2018. Advances in molecular approaches for understanding soil organic matter composition, origin, and turnover: a historical overview. Adv. Agron. 149, 1–48. https://doi.org/10.1016/bs.agron.2018.01.003.
Lado, M., Sayegh, J., Gia Gadňay, A., Ben-Hur, M., Borisover, M., 2023. Heat-induced changes in soil water-extractable organic matter characterized using fluorescence and FTIR spectroscopies coupled with dimensionality reduction methods. Geoderma 430, 116347. https://doi.org/10.1016/j.geoderma.2023.116347.
Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nat. Rev. Earth Environ. 1 (10), 544–553. https://doi.org/10.1038/s43017-020-0080-8.
Lehmann, J., Kleber, M., 2015. The contentious nature of soil organic matter. Nature 528 (7580), 60–68. https://doi.org/10.1038/nature16069.
Li, H., Bolscher, T., Winnick, M., Tfaily, M.M., Cardon, Z.G., Keiluweit, M., 2021. Simple plant and microbial exudates destabilize mineral-associated organic matter via multiple pathways. Environ. Sci. Technol. 55 (5), 3389–3398. https://doi.org/10.1021/acs.est.0c04592.
Li, S., Viscarra Rossel, R.A., Webster, R., 2022. The cost-effectiveness of reflectance spectroscopy for estimating soil organic carbon. Eur. J. Soil Sci. 73 (1). https://doi.org/10.1111/ejss.13202.
Li, X., Yang, P., Zhao, W., Guo, F., Jaisi, D.P., Mi, S., Ma, H., Lin, B., Feng, X., Tan, W., Wang, X., 2023. Adsorption mechanisms of glyphosate on ferrihydrite: effects of Al substitution and aggregation state. Environ. Sci. Technol. 57 (38), 14384–14395. https://doi.org/10.1021/acs.est.3c04727.
Lussier, J.M., Krzic, M., Smukler, S.M., Neufeld, K.R., Chizen, C.J., Bomke, A.A., 2020. Labile soil carbon fractions as indicators of soil quality improvement under short-term grassland set-aside. Soil Res. 58 (4), 364–370. https://doi.org/10.1071/SR19180.
Madhavan, D.B., Baldock, J.A., Read, Z.J., Murphy, S.C., Cunningham, S.C., Perring, M.P., Herrmann, T., Lewis, T., Cavagnaro, T.R., England, J.R., Paul, K.I., Weston, C.J., Baker, T.G., 2017. Rapid prediction of particulate, humus and resistant fractions of soil organic carbon in reforested lands using infrared spectroscopy. J. Environ. Manag. 193, 290–299. https://doi.org/10.1016/j.jenvman.2017.02.013.
Margenot, A.J., Calderón, F.J., Goyne, K.W., Mukome, F.N.D., Parikh, S.J., 2017. IR spectroscopy, soil analysis applications. In: Lindon, J.C., Tranter, G.E., Koppenaal, D.W. (Eds.), Encyclopedia of Spectroscopy and Spectrometry, pp. 448–454. https://doi.org/10.1016/B978-0-12-409547-2.12170-5.
Margenot, A.J., Parikh, S.J., Calderón, F.J., 2023. Fourier-transform infrared spectroscopy for soil organic matter analysis. Soil Sci. Soc. Am. J. 87 (6), 1503–1528. https://doi.org/10.1002/saj2.20583.
Matamala, R., Jastrow, J.D., Calderón, F.J., Liang, C., Fan, Z., Michaelson, G.J., Ping, C.-L., 2019. Predicting the decomposability of arctic tundra soil organic matter with mid infrared spectroscopy. Soil Biol. Biochem. 129, 1–12. https://doi.org/10.1016/j.soilbio.2018.10.014.
McKissock, I., Gilkes, R.J., Van Bronswijk, W., 2003. The relationship of soil water repellency to aliphatic C and kaolin measured using DRIFT. Aust. J. Soil Res. 41 (2), 251–265. https://doi.org/10.1071/SR01091.
Metzger, K., Zhang, C., Daly, K., 2021. From benchtop to handheld MIR for soil analysis: predicting lime requirement and organic matter in agricultural soils. Biosyst. Eng. 204, 257–269. https://doi.org/10.1016/j.biosystemseng.2021.01.025.
Nasonova, A., Levy, G.J., Rinot, O., Eshel, G., Borisover, M., 2022. Organic matter in aqueous soil extracts: prediction of compositional attributes from bulk soil mid-IR spectra using partial least square regressions. Geoderma 411, 115678. https://doi.org/10.1016/j.geoderma.2021.115678.
Nduwamungu, C., Ziadi, N., Tremblay, G.F., Parent, L.-É., 2009. Near-infrared reflectance spectroscopy prediction of soil properties: effects of sample cups and preparation. Soil Sci. Soc. Am. J. 73 (6), 1896–1903. https://doi.org/10.2136/sssaj2008.0213.
Ng, W., Minasny, B., Jeon, S.H., McBratney, A., 2022. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. Soil Security 6, 100043. https://doi.org/10.1016/j.soisec.2022.100043.
Nguyen, T.T., Janik, L.J., Raupach, M., 1991. Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. Aust. J. Soil Res. 29 (1), 49–67. https://doi.org/10.1071/SR9910049.
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 Biol. Biochem. 68, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022.
Ouatmane, A., Provenzano, M.R., Hafidi, M., Senesi, N., 2000. Compost maturity assessment using calorimetry, spectroscopy and chemical analysis. Compost Sci. Util. 8 (2), 124–134. https://doi.org/10.1080/1065657X.2000.10701758.
Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5 (2), 111–126. https://doi.org/10.1002/env.3170050203.
Pirouz, D.M., 2010. An overview of partial least squares. SSRN Electron. J. https://doi.org/10.2139/ssrn.1631359.
Ravisankar, R., Chandrasekaran, A., Kalaiarsi, S., Eswaran, P., Rajashekhar, C., Vanasundari, K., Athavale, A., 2011. Mineral analysis in beach rocks of Andaman Island, India by spectroscopic techniques. Arch. Appl. Sci. Res. 3 (3), 77–84. http://scholarsresearchlibrary.com/aasr-vol3-iss3/AASR-2011-3-3-77-84.pdf.
Rinot, O., Borisover, M., Levy, G.J., Eshel, G., 2021. Fluorescence spectroscopy: a sensitive tool for identifying land-use and climatic region effects on the characteristics of water-extractable soil organic matter. Ecol. Indicat. 121, 107103. https://doi.org/10.1016/j.ecolind.2020.107103.
Rumpel, C., Janik, L.J., Skjemstad, J.O., Kögel-Knabner, I., 2001. Quantification of carbon derived from lignite in soils using mid-infrared spectroscopy and partial least squares. Org. Geochem. 32 (6), 831–839. https://doi.org/10.1016/S0146-6380(01)00029-8.

Russell, L.M., Takahama, S., Liu, S., Hawkins, L.N., Covert, D.S., Quinn, P.K., Bates, T.S., 2009. Oxygenated fraction and mass of organic aerosol from direct emission and atmospheric processing measured on the R/V Ronald Brown during TEXAQS/GoMACCS 2006. J. Geophys. Res. Atmos. 114 (D7). https://doi.org/10.1029/2008JD011275.

Schnitzer, M., 1978. Humic substances: chemistry and reactions. In: Schnitzer, M., Khan, S.U. (Eds.), Soil Organic Matter. Elsevier Scientific Publishing Co., New York l–64.

Silverstein, R.M., Webster, F.X., 1997. Spectrometric Identification of Organic Compounds, sixth ed. John Wiley & Sons, Inc.

Simkovic, I., Dlapa, P., Doerr, S.H., Mataix-Solera, J., Sasinkova, V., 2008. Thermal destruction of soil water repellency and associated changes to soil organic matter as observed by FTIR spectroscopy. Catena 74 (3), 205–211. https://doi.org/10.1016/j.catena.2008.03.003.

Smidt, E., Schwanninger, M., 2005. Characterization of waste materials using FTIR spectroscopy: process monitoring and quality assessment. Spectrosc. Lett. 38 (3), 247–270. https://doi.org/10.1081/SL-200042310.

Sokol, N.W., Whalen, E.D., Jilling, A., Kallenbach, C., Pett-Ridge, J., Georgiou, K., 2022. Global distribution, formation and fate of mineral-associated soil organic matter under a changing climate: a trait-based perspective. Funct. Ecol. 36 (6), 1411–1429. https://doi.org/10.1111/1365-2435.14040.

Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49 (2), 139–186. https://doi.org/10.1080/05704928.2013.811081.

Spaccini, R., Piccolo, A., Haberhauer, G., Stemmer, M., Gerzabek, M.H., 2001. Decomposition of maize straw in three European soils as revealed by DRIFT spectra of soil particle fractions. Geoderma 99 (3–4), 245–260. https://doi.org/10.1016/S0016-7061(00)00073-2.

Takahama, S., Schwartz, R.E., Russell, L.M., MacDonald, A.M., Sharma, S., Leaitch, W.R., 2011. Organic functional groups in aerosol particles from burning and non-burning forest emissions at a high-elevation mountain site. Atmos. Chem. Phys. 11 (13), 6367–6386. https://doi.org/10.5194/acp-11-6367-2011.

Tarchitzky, J., Lerner, O., Shani, U., Arye, G., Lowengart-Aycicegi, A., Brener, A., Chen, Y., 2007. Water distribution pattern in treated wastewater irrigated soils: hydrophobicity effect. Eur. J. Soil Sci. 58 (3), 573–588. https://doi.org/10.1111/j.1365-2389.2006.00845.x.

Tauler, R., de Juan, A., Jaumot, J., 2016. Multivariate curve resolution homepage. http://www.mcrals.info/.

Terhoeven-Urselmans, T., Michel, K., Helfrich, M., Flessa, H., Ludwig, B., 2006. Near-infrared spectroscopy can predict the composition of organic matter in soil and litter. J. Plant Nutr. Soil Sci. 169 (2), 168–174. https://doi.org/10.1002/jpln.200521712.

Tesfamichael, T., Hoel, A., Niklasson, G.A., Wäckelgård, E., Gunde, M.K., Orel, Z.C., 2001. Optical characterization method for black pigments applied to solar-selective absorbing paints. Appl. Opt. 40 (10), 1672–1681. https://doi.org/10.1364/ao.40.001672.

Tinti, A., Tugnoli, V., Bonora, S., Francioso, O., 2015. Recent applications of vibrational mid-infrared (IR) spectroscopy for studying soil components: a review. J. Cent. Eur. Agric. 16 (1), 1–22. https://doi.org/10.5513/JCEA01/16.1.1535.

Varmuza, K., Karlovits, M., Demuth, W., 2003. Spectral similarity versus structural similarity: infrared spectroscopy. Anal. Chim. Acta 490 (1–2), 313–324. https://doi.org/10.1016/S0003-2670(03)00668-8.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. 58 (2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

Xia, Y., 2020. Correlation and association analyses in microbiome study integrating multiomics in health and disease. Progress in Molecular Biology and Translational Science 171, 309–491. https://doi.org/10.1016/bs.pmbts.2020.04.003.

Yu, W., Huang, W., Weintraub-Leff, S.R., Hall, S.J., 2022. Where and why do particulate organic matter (POM) and mineral-associated organic matter (MAOM) differ among diverse soils? Soil Biol. Biochem. 172, 108756. https://doi.org/10.1016/j.soilbio.2022.108756.

Zhang, L., Yang, X., Drury, C., Chantigny, M., Gregorich, E., Miller, J., Bittman, S., Reynolds, D., Yang, J., 2018. Infrared spectroscopy prediction of organic carbon and total nitrogen in soil and particulate organic matter from diverse Canadian agricultural regions. Can. J. Soil Sci. 98 (1), 77–90. https://doi.org/10.1139/cjss-2017-0070.