Received: 31 August 2022 | Revised: 9 June 2023 | Accepted: 22 June 2023

DOI: 10.1111/ejss.13394

RESEARCH ARTICLE

European Journal of Soil Science WILEY

Optimised use of data fusion and memory-based learning with an Austrian soil library for predictions with infrared data

Bernard Ludwig[1] | **Isabel Greenberg**[1] | **Michael Vohland**[2,3,4] | **Kerstin Michel**[5]

[1]Department of Environmental Chemistry, University of Kassel, Witzenhausen, Germany

[2]Geoinformatics and Remote Sensing, Institute for Geography, Leipzig University, Leipzig, Germany

[3]Remote Sensing Centre for Earth System Research, Leipzig University, Leipzig, Germany

[4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[5]Unit of Soil Ecology, Federal Research and Training Centre for Forests, Natural Hazards and Landscape (BFW), Wien, Austria

Correspondence
Bernard Ludwig, Department of Environmental Chemistry, University of Kassel, Nordbahnhofstr. 1a, 37213 Witzenhausen, Germany.
Email: bludwig@uni-kassel.de

Funding information
Deutsche Forschungsgemeinschaft

Abstract

Infrared spectroscopy in the visible to near-infrared (vis–NIR) and mid-infrared (MIR) regions is a well-established approach for the prediction of soil properties. Different data fusion and training approaches exist, and the optimal procedures are yet undefined and may depend on the heterogeneity present in the set and on the considered scale. The objectives were to test the usefulness of partial least squares regressions (PLSRs) for soil organic carbon (SOC), total carbon (C_t), total nitrogen (N_t) and pH using vis–NIR and MIR spectroscopy for an independent validation after standard calibration (use of a general PLSR model) or using memory-based learning (MBL) with and without spiking for a national spectral database. Data fusion approaches were simple concatenation of spectra, outer product analysis (OPA) and model averaging. In total, 481 soils from an Austrian forest soil archive were measured in the vis–NIR and MIR regions, and regressions were calculated. Fivefold calibration-validation approaches were carried out with a region-related split of spectra to implement independent validations with n ranging from 47 to 99 soils in different folds. MIR predictions were generally superior over vis–NIR predictions. For all properties, optimal predictions were obtained with data fusion, with OPA and spectra concatenation outperforming model averaging. The greatest robustness of performance was found for OPA and MBL with spiking with $R^2 \geq 0.77$ (N), 0.85 (SOC), 0.86 (pH) and 0.88 (C_t) in the validations of all folds. Overall, the results indicate that the combination of OPA for vis–NIR and MIR spectra with MBL and spiking has a high potential to accurately estimate properties when using large-scale soil spectral libraries as reference data. However, the reduction of cost-effectiveness using two spectrometers needs to be weighed against the potential increase in accuracy compared to a single MIR spectroscopy approach.

KEYWORDS

data fusion, independent validation, infrared spectroscopy, MBL, nitrogen, outer product analysis, pH, soil organic carbon, spiking, total carbon

1 | INTRODUCTION

Visible to near-infrared (vis–NIR, range: 400–2500 nm, 25,000–4000 cm^{-1}) spectroscopy and mid-infrared (MIR, range: 2500–25,000 nm, 4000–400 cm^{-1}) spectroscopy (vis–NIRS and MIRS, respectively) are well-established spectroscopic techniques for the simultaneous estimation of a number of chemical and physical soil properties (Gholizadeh et al., 2013; Ludwig et al., 2019, 2021; Pallottino et al., 2019; Soriano-Disla et al., 2014). These spectroscopic applications may be especially useful in studies with a focus on spatial and/or temporal monitoring of soil properties where analytical costs and measurement time should be reduced considerably (O'Rourke et al., 2011).

The reasons for accurate predictions of soil organic carbon (SOC) and total N (N_t) concentrations have been assigned to vibrations related to alkyl groups, protein amides, carboxylic acids, water-associated groups, carboxylate anions, and aromatic groups (Soriano-Disla et al., 2014). Moreover, based on experiments with different additions of wheat (*Triticum aestivum* L.) straw and clover (*Trifolium* spp.) residues to soils, Greenberg et al. (2020) summarised that predictive mechanisms for SOC and N_t appeared to be similar for vis–NIRS, but not for MIRS. A number of bands have been assigned to carbonates, especially in the MIR region (2910–2850, 2686–2460, 1850–1784, 1567–1295, 889–867, 734–719, and 719–708 cm^{-1}) and also in the NIR region (2335 nm), which makes predictions of carbonate concentrations feasible (McBride, 2022; Mirzaeitalarposhti et al., 2016; Soriano-Disla et al., 2014; Tatzber et al., 2010).

In contrast to direct estimations of SOC, total C (C_t, sum of SOC and carbonate-C) and N_t concentrations, indirect estimation of pH may be based on the presence of proton-rich clays, Al oxyhydroxide minerals and sulphides, oxidisable ammonium and organic N as amides and carboxylic acids and phenols (Leenen et al., 2019; Soriano-Disla et al., 2014). Indirect estimations are understandably a controversial topic in soil spectroscopy (McBride, 2022; Viscarra Rossel et al., 2022), since it is very likely that the calibrations used to predict soil pH by vis–NIRS or MIRS at one location will fail to predict pH at another location (McBride, 2022). In general, uncertainty tends to increase when statistical models are used under extrapolation.

A large number of soil spectral libraries (SSLs) have already been created, and their usefulness for vis–NIR and MIR predictions has been demonstrated (e.g., Viscarra Rossel & Webster, 2012; Wijewardane et al., 2018). Several studies have shown that general models derived from SSL may likely produce severely biased estimates for local-scale applications. Spiking or weighted spiking has been shown to potentially reduce bias in the estimations considerably (Guerrero et al., 2010, 2014). However, spiking did not always result in improved estimations, and Zeng et al. (2016) discussed the importance of the distribution and relationship between target set and spiked set for improving the prediction accuracy.

Highlights

- Optimal procedures for data fusion and training for SSL are yet undefined
- The combination of OPA with MBL and spiking has a high potential to estimate properties
- Besides the training and fusion effects, the IQR also affects model accuracies

Viscarra Rossel et al. (2022) summarised that nowadays SSL may be used as a source of information for building localised calibrations for specific contexts and pedologic domains rather than to create general models. Thus, an alternative approach to a general calibration model (with or without spiking) using, e.g., partial least squares regression (PLSR) is memory-based learning (MBL). For each new unit (i.e., soil), MBL extracts appropriate units from the SSL based on their spectral similarity and then calculates the regression equations individually for each new soil with the selected units (Ramirez-Lopez et al., 2013, 2022). This approach has given promising results compared with other calibration strategies with different algorithms. For instance, Jaconi et al. (2017) found for a German SSL that calibration with MBL provided more accurate SOC estimations than a general PLSR model. Similarly, Li, Li, et al. (2022) reported for a Chinese vis–NIR SSL that MBL with spiking outperformed general PLSR with spiking for the estimation of soil organic matter concentrations.

The general superiority of MIRS over vis–NIRS for predictions of spectrally active soil properties has been reported in a number of studies and has been explained by a higher information content in MIR spectra, which contain fundamental vibration bands, compared to the less intense peaks in the NIR region, which is limited to overtone and combination bands (Ng et al., 2019;

Soriano-Disla et al., 2014; Vohland et al., 2014). Nevertheless, vis–NIRS information is not entirely redundant to MIRS, and low-level [concatenation of the vis–NIR and MIR spectra or the outer product analysis (OPA) (Borràs et al., 2015)] or high-level (averaging of vis–NIR and MIR modelling results) data fusion has the potential for improved predictions. The results so far are inconclusive: Johnson et al. (2019) reported a benefit of spectra concatenation for the prediction of SOC over MIR or vis–NIR predictions for an African SSL, whereas a number of studies did not find a benefit compared to MIR predictions (e.g., Ng et al., 2019; Vohland et al., 2022). The same holds true for OPA, which corresponds to a mutual weighting of each signal (vis–NIR and MIR) by the other. Terra et al. (2019) and Xu et al. (2020) noted improved SOC prediction and soil classification, respectively, whereas Vohland et al. (2022) reported that model averaging outperformed OPA for SOC prediction. Summarising the studies above using different SSLs indicates that the benefits of spectral fusion depended on the data set. Fusing data of vis–NIRS and MIRS may create partly redundant data due to an addition of overtones and combination bands of the NIR region to the fundamental bands of the MIR range. However, a potential merit lies especially in the visible and the directly adjacent NIR region (600 to 1000 nm) because of the addition of important spectral information originated from electronic transitions in atoms (Terra et al., 2019).

Overall, the combined investigation of the training approaches (general calibration and MBL with and without spiking) and data fusion may have greatest merits for improved predictions. We use a multiple partitioning approach for independent validation of an Austrian forest soil archive and hypothesise that the combination of OPA, MBL and spiking will result in the highest prediction accuracies. Objectives were to study the effects of different training approaches (local calibration, general calibration using a national database with or without spiking and MBL with or without spiking) and data fusion methods (concatenation, OPA and model averaging) for independent predictions of SOC, C_t and N_t concentration and pH for an Austrian soil archive.

2 | MATERIALS AND METHODS

2.1 | Austrian forest soil archive: Soil sampling and chemical analyses

The Austrian Forest Soil Survey (Waldbodenzustandsinventur, WBZI) was carried out in 1987/89 in order to describe the forest soil condition of the entire federal territory on a systematic grid of 8.7 × 8.7 km, resulting in a total of 511 plots sampled (BMLFUW, 2013). From this Austrian forest soil archive, 481 plots were available for spectral measurements in this study, which focuses on the surface soils (0–10 cm, Figure 1). For each plot, site (e.g., altitude, slope, exposition and vegetation), soil (e.g., parent material, soil type and horizon) and chemical and physical soil properties were recorded.

In this study, we used data for SOC, C_t and N_t concentrations and pH. Soils were air dried, sieved to <2 mm and ground for determinations of SOC, C_t and N_t concentrations. $CaCl_2$ (0.01 moL/L) was used for pH measurements (ÖNORM L 1083, 1989). Total N concentration was measured using Kjeldahl extraction (ÖNORM L 1082, 1989). Total C concentrations were measured via oxidation in an oxygen flux and infrared detection of the carbon dioxide generated (ÖNORM L 1080, 1989). Inorganic C was determined with the Scheibler method (ÖNORM L 1084, 1989), and SOC was calculated as the difference between C_t and inorganic C.

Soils were classified according to Austrian soil taxonomy (Nestroy et al., 2011) and consisted of 26 different soil types. We grouped these soil types into five groups A–E (Figures 1 and 2) and provided the soil types in Appendix A: "A—Fluvisols and others (28 plots, 7 soil types)"; "B—predominantly Podzols (92 plots, 3 soil types)"; "C—predominantly Cambisols (189 plots, 7 soil types)"; "D—predominantly Leptosols and Phaeozems (121 plots, 4 soil types)"; and "E—predominantly Planosols and Stagnosols (51 plots, 5 soil types).

2.2 | Spectral measurements

Dried soils were ball-milled to a particle size <0.2 mm using a Retsch MM 400 (Haan, Germany) with 10 zirconium oxide balls at 30 Hz for 5 min. Lab spectral measurements were done on two replicates per observational unit. A Foss XDS Rapid Content Analyzer (Silver Spring, MD, USA) (optical bandwidth: 8.75 nm) with 32 co-added scans was used for the vis–NIR spectra in the range of 400–2500 nm. For each measurement, approximately 10 g of soil was filled into a cell (5-cm diameter) with a quartz window. The instrument was recalibrated every ca. 30 min using an internal white reference to avoid baseline shifts. The region below 500 nm was excluded from the chemometric analysis due to instrumental artefacts (Stevens et al., 2013). A splice correction (package prospectr, Stevens & Ramirez-Lopez, 2022) with the position of the splice at 1100 nm was included to account for the detector change (Si detector from 400 to 1100 nm, PbS detector from 1100 to 2500 nm). In total, each vis–NIR spectrum consisted of 999 data points (regular increment of 2 nm). A Bruker-TENSOR 27 MIR spectrometer (Ettlingen, Germany) with an A562

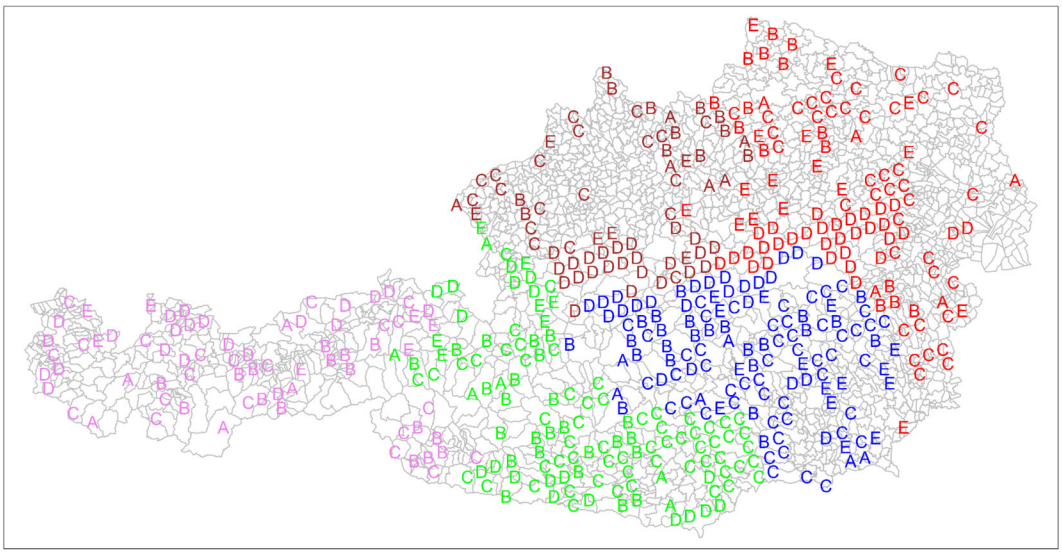

FIGURE 1 Sampling locations of the 481 soils in Austria. Different colours refer to the five regions for the fivefold calibrations-validations (red: Lower Austria and Burgenland, green: Carinthia and Salzburg, brown: Upper Austria, blue: Styria and pink: Tyrol and Vorarlberg). The letters A to E refer to groups of soil types explained in the text.

integrating sphere detector and the diffuse reflectance accessory (Ulbricht-Kugel, Ettlingen, Germany) was used for the scanning of diffuse reflectance infrared Fourier transform spectra of the soils in the range of 7000–370 to cm^{-1}. For each measurement, approximately 1.5 g of soil was filled into a cup. The instrument was calibrated every hour with a gold reference background. The spectra were measured with 200 scans at approximately 2 cm^{-1} intervals. The longwave NIR region (7000–4000 cm^{-1}) and the region <650 cm^{-1}, which has limited usefulness due to overlapping mineral and organic absorption bands (Nocita et al., 2015), were excluded from the analysis. In total, each MIR spectrum consisted of 1737 data points.

For vis–NIR and MIR spectra, the reflectance values of replicate measurements at each sampling point were averaged and converted to absorbance [$\log_{10}$(1/reflectance)] prior to the chemometric analyses.

2.3 | Formation of training and test sets for local vs. national training

In order to test the robustness and performance variance, multiple dataset partitions were created as recommended by Cawley and Talbot (2010). For this, the complete dataset of surface soils from five regions (1–states Burgenland and Lower Austrian, n = 108; 2–Carinthia and Salzburg, n = 116; 3–Upper Austria, n = 67; 4–Styria, n = 119 and 5–Tyrol and Vorarlberg, n = 71) was divided so that each of the five regions (indicated by different colours in Figure 1) served as the validation set once for each training strategy and spectral approach described below. For each of the five validation sets, 20 soils were removed in a stratified randomisation approach described below, which allowed the creation of identical independent validation sets used to test all training strategies and spectral approaches. The 20 soils removed from each region served as a local training set (strategy I) or as a spiking set (strategies III and V), as described below. Across all training strategies and spectral approaches, the number of soils remaining in the five independent validation folds was 88, 96, 47, 99 and 51 for the five regions, respectively (Table 1). Alternatively, the fivefold calibration-validations with spectra split according to the five regions for independent validations may be treated as leave-one-region-out cross validation with subsequent averaging of the results.

Five chemometric training strategies described below were tested with fivefold partitioning of the data for the original vis–NIR and MIR regions, the concatenated spectra (each spectrum consisting of 2736 data points) and the OPA (Terra et al., 2019; Vohland et al., 2022). For the latter, vis–NIR and MIR spectra were resampled to every eighth data point (125 and 218 data points for vis–NIR and MIR, respectively), resulting in each outer product consisting of 27,250 data points.

2.3.1 | Training strategy I: Local training (Cal_{local})

Local training with only small local training sets (n = 20, Cal_{local}) served as a control (and was additionally used as a spiking set in variants described below). Soils were selected for the local training with a stratified

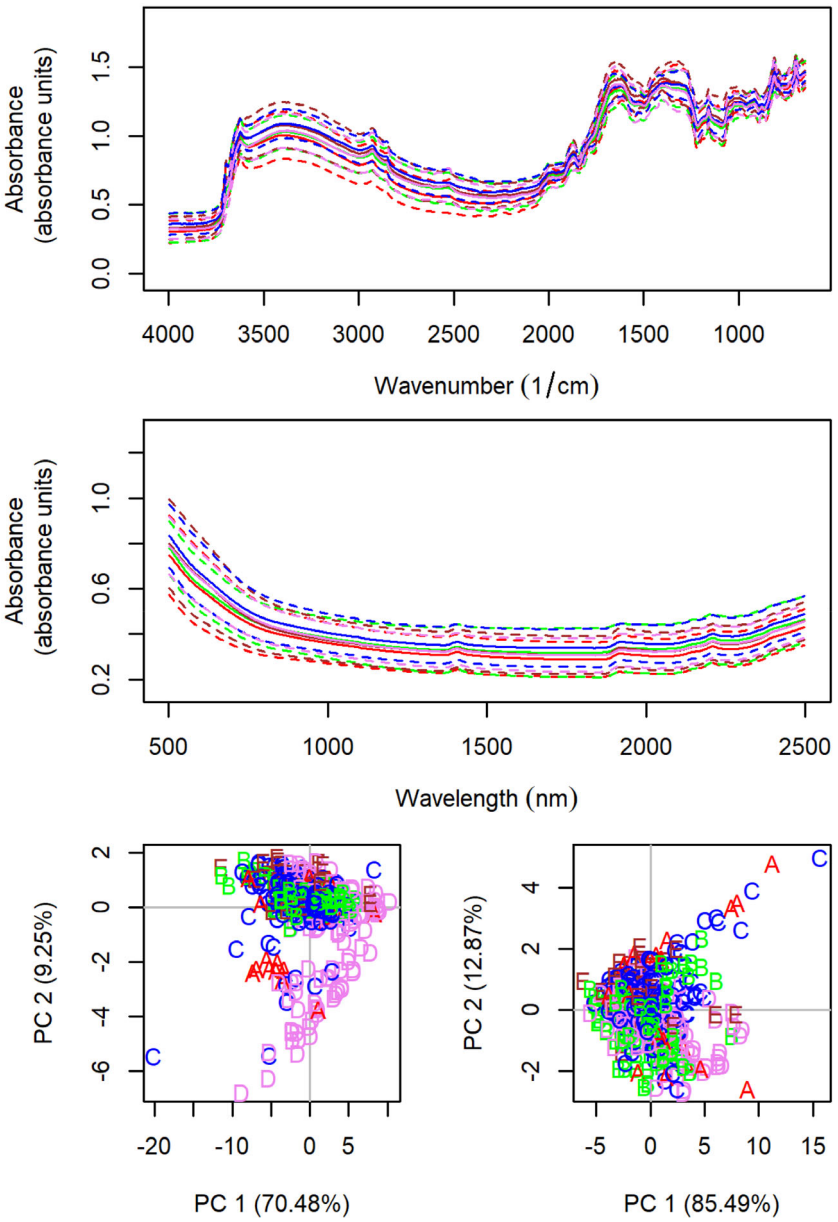

FIGURE 2 MIR (top) and vis–NIR spectra (middle) of the 481 soils: mean ± 1 standard deviation (dotted lines) for each of the five regions (colour codes: see legend of Figure 1). Scores of the principal component analysis for the 481 soils are also shown for MIRS (bottom left) and vis–NIRS (bottom right, colour codes: see legend of Figure 1; letters A to E refer to groups of soil types given in the text).

randomisation approach from each respective region by having each soil type of the respective region present at least once in the local training set. Afterwards, each local training set was completed by randomly adding soils from the respective region until n equalled 20.

Analysis was performed with the statistical software R (version 4.2.0, R Core Team, 2022). The package pls (Liland et al., 2021) was used for PLSR.

Model training included leave-one-out cross validation to determine the optimal data pretreatment and number of latent variables.

The following 13 data pretreatments were tested (package prospectr): (i) use of the full spectra without manipulation, (ii–v) calculation of moving averages (over 5, 11, 17 or 23 data points) and (vi–xiii) application of the Savitzky–Golay algorithm for the reduction of noise applied with the polynomial degree (PD) set to 2, the order of the derivative (DER) ranging from 1 to 2 (with PD-DER: 2–1 or 2–2) and a window smoothing size of 5, 11, 17 or 23.

The maximum number of latent variables was set to 15, and the optimal numbers (averaged over the five folds) ranged from 2 to 4 for the local models (Tables 2a and 2b). For each fold and property, the optimal number of latent variables was determined in cross validation by considering minimisation of Akaike information criterion (AIC) (Viscarra Rossel & Behrens, 2010), calculated as

$$\mathrm{AIC} = n \times \log_e(\mathrm{RMSE}) + 2k,$$

where n is the sample size, k is the number of latent variables and RMSE is the root mean square error.

TABLE 1 Descriptive statistics for the validation samples of the fivefold validations for the four soil properties.

Property	Validation fold	Minimum	Maximum	Median	IQR	p (Shapiro–Wilk)
SOC (g kg^{-1})	1	12	380	42	63	7.8×10^{-12}
	2	10	146	50	31	2.3×10^{-5}
	3	21	380	58	65	2.8×10^{-7}
	4	14	418	67	49	8.8×10^{-13}
	5	16	258	67	57	3.3×10^{-6}
C_t (g kg^{-1})	1	12	380	42	65	1.4×10^{-11}
	2	10	205	51	36	1.8×10^{-8}
	3	21	414	58	62	2.5×10^{-7}
	4	14	418	68	54	2.1×10^{-11}
	5	16	268	69	70	1.5×10^{-5}
N_t (g kg^{-1})	1	0.7	19.0	2.6	3.4	1.5×10^{-10}
	2	0.7	11.0	2.8	1.6	9.1×10^{-8}
	3	1.0	15.7	3.0	3.5	1.2×10^{-6}
	4	0.8	14.5	3.3	2.7	4.4×10^{-9}
	5	0.8	18.8	3.4	3.5	2.8×10^{-7}
pH	1	3.2	7.3	4.2	1.7	3.2×10^{-8}
	2	3.2	7.6	4.0	0.8	2.9×10^{-11}
	3	3.0	7.2	3.9	3.0	5.5×10^{-6}
	4	2.7	7.4	3.8	0.9	8.0×10^{-11}
	5	3.0	7.5	4.5	2.9	9.5×10^{-5}

Note: Validation folds: 1–Burgenland and Lower Austria, n = 88; 2–Carinthia and Salzburg, n = 96; 3–Upper Austria, n = 47; 4–Styria, n = 99; 5–Tyrol and Vorarlberg, n = 51.
Abbreviations: C_t, total carbon; IQR, interquartile range; N_t, total nitrogen; SOC, soil organic carbon.

The RMSE is calculated as

$$\mathrm{RMSE} = \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n}},$$

where $\hat{y}_i$ is the modelled soil property and y_i is the measured soil property.

The inaccuracy determined by the RMSE consists of the mean error (ME) (or bias) and the standard deviation (SDE) (or imprecision) (Viscarra Rossel & McBratney, 1998) calculated as

$$\mathrm{RMSE}^2 = \mathrm{ME}^2 + \mathrm{SDE}^2.$$

The Nash–Sutcliffe model efficiency (Janssen & Heuberger, 1995; Ludwig et al., 2008; Yang et al., 2022), which may also be denoted as coefficient of determination (R^2), is calculated as follows:

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

where RSS is the residual sum of squares and TSS is the total sum of squares.

For a satisfactory prediction, we suggest an R^2 threshold of 0.5, analogous to the suggestion by Chang et al. (2001) for ratio of performance to deviation (RPD) values, and a threshold of 0.85 for a good prediction. However, one has to keep in mind that the usefulness of a model is always defined in its specific context.

2.3.2 | Training strategy II: Training using the national database in standard calibration-validation (Cal_{nat})

For each of the five regions (validation folds 1–5, Table 1) used in independent validations, the remaining four regions were used as the respective calibration dataset. For instance, for the validation of region 1 (Burgenland and Lower Austria, n = 88), the national dataset consisted of regions 2–5 (n = 373).

PLSR was carried out as described in 2.3.1. However, since more information was available in the

TABLE 2a Average performances (mean and standard deviation) and average factors of the PLS regressions in the training for local calibration (Cal_{local}), calibration using the national archive (Cal_{nat}) and the national archive with spiking ($Cal_{nat/sp}$). Median k-nearest neighbours are also shown for memory-based learning using the national archive (MBL_{nat}) and additionally with spiking ($MBL_{nat/sp}$).

		MIR training				Vis–NIR training			
Property	Training strategy	Average factors or median k	SDE_{CV}	ME_{CV}	R^2_{CV}	Average factors or median k	SDE_{CV}	ME_{CV}	R^2_{CV}
SOC ($g\ kg^{-1}$)	Cal_{local}	3	30.0 (10.5)	2.9 (3.0)	0.61 (0.18)	3	24.9 (8.4)	1.9 (2.1)	0.73 (0.13)
	Cal_{nat}	14	17.9 (0.6)	0.0 (0.3)	0.91 (0.01)	12	22.1 (1.0)	0.4 (0.3)	0.86 (0.01)
	MBL_{nat}	160	-	-	-	60	-	-	-
	$Cal_{nat/sp}$	16	16.9 (1.1)	0.4 (0.4)	0.91 (0.01)	12	22.0 (1.5)	0.3 (0.6)	0.86 (0.01)
	$MBL_{nat/sp}$	120	-	-	-	80	-	-	-
C_t ($g\ kg^{-1}$)	Cal_{local}	2	33.9 (13.0)	1.6 (3.0)	0.65 (0.22)	2	28.1 (8.9)	−1.9 (1.8)	0.78 (0.07)
	Cal_{nat}	1	17.3 (0.8)	−0.3 (0.4)	0.93 (0.01)	13	23.8 (1.2)	−0.2 (0.4)	0.88 (0.01)
	MBL_{nat}	120	-	-	-	100	-	-	-
	$Cal_{nat/sp}$	16	16.2 (0.8)	0.2 (0.5)	0.94 (0.01)	15	23.3 (1.4)	0.1 (0.7)	0.88 (0.01)
	$MBL_{nat/sp}$	160	-	-	-	100	-	-	-
N_t ($g\ kg^{-1}$)	Cal_{local}	2	1.7 (0.8)	0.1 (0.1)	0.49 (0.27)	3	1.5 (0.5)	0.0 (0.1)	0.62 (0.18)
	Cal_{nat}	14	1.0 (0.0)	0.0 (0.0)	0.87 (0.02)	13	1.3 (0.1)	0.0 (0.0)	0.80 (0.03)
	MBL_{nat}	80	-	-	-	100	-	-	-
	$Cal_{nat/sp}$	17	1.0 (0.1)	0.0 (0.0)	0.87 (0.02)	15	1.2 (0.1)	0.0 (0.0)	0.82 (0.01)
	$MBL_{nat/sp}$	100	-	-	-	80	-	-	-
pH ($CaCl_2$)	Cal_{local}	4	0.7 (0.3)	0.1 (0.1)	0.75 (0.18)	3	1.0 (0.2)	0.0 (0.1)	0.44 (0.26)
	Cal_{nat}	17	0.4 (0.0)	0.0 (0.0)	0.93 (0.01)	16	0.6 (0.0)	0.0 (0.0)	0.78 (0.02)
	MBL_{nat}	180	-	-	-	180	-	-	-
	$Cal_{nat/sp}$	16	0.4 (0.0)	0.0 (0.0)	0.93 (0.01)	17	0.7 (0.0)	0.0 (0.0)	0.78 (0.02)
	$MBL_{nat/sp}$	120	-	-	-	240	-	-	-

Abbreviations: ME_{CV}, mean error of cross validation; OPA, outer product analysis; R^2_{CV}, coefficient of determination (Nash–Sutcliffe model efficiency) of cross validation; SDE_{CV}, standard deviation of the error of cross validation. Other abbreviations: see Table 1.

national dataset, the maximum number of latent variables was set to 20. The optimal numbers, again determined by minimisation of AIC, ranged from 12 to 19 when averaged over the five folds (Tables 2a and 2b).

2.3.3 | Training strategy III: Training using the national database and MBL (MBL_{nat})

Training strategy III used the same regions 1–5 in the independent fivefold validations (Table 1) and the same respective national training sets as strategy II described above with the optimal data pretreatments for the respective properties and regions obtained in strategy II. However, in contrast to the calibration-validation-runs described above, MBL does not derive general or global models for each of the five folds. In contrast, it derives a local model for each soil of each fold separately. In the first step for each soil, MBL seeks a sequence of k-nearest neighbours of the respective soil from the reference library (i.e., the national database) based on the Mahalanobis distances in the principal component space [the default setting of the mbl() command of the resemble package (Ramirez-Lopez et al., 2022)]. A local model is then fitted using different approaches. In this study, a weighted average PLSR model (Shenk et al., 1997) was used. Briefly, the weighted average PLSR approach uses multiple models generated by considering various numbers of pls components (i.e., between a minimum of 3 and a maximum number of 20 pls components in this study). At each local partition, the final predicted value is a weighted average of all the predicted values generated by the multiple pls models (Ramirez-Lopez et al., 2022; Shenk et al., 1997). In this study, the sequence of k-nearest

TABLE 2b Average performances (mean and standard deviation) and average factors of the PLS regressions in the training for local calibration (Cal_{local}), calibration using the national archive (Cal_{nat}) and the national archive with spiking ($Cal_{nat/sp}$). Median k-nearest neighbours are also shown for memory-based learning using the national archive (MBL_{nat}) and additionally with spiking ($MBL_{nat/sp}$).

		Vis–NIR-MIR concatenation				OPA vis–NIR-MIR training			
Property	Training strategy	Average factors or median k	SDE_{CV}	ME_{CV}	R^2_{CV}	Average factors or median k	SDE_{CV}	ME_{CV}	R^2_{CV}
SOC (g kg^{-1})	Cal_{local}	3	29.2 (11.7)	1.4 (4.6)	0.64 (0.19)	4	19.0 (7.1)	2.4 (3.6)	0.82 (0.11)
	Cal_{nat}	16	17.1 (0.5)	−0.1 (0.3)	0.92 (0.01)	14	14.2 (0.5)	0.7 (0.5)	0.94 (0.00)
	MBL_{nat}	100	-	-	-	300	-	-	-
	$Cal_{nat/sp}$	18	16.0 (1.0)	0.4 (0.2)	0.92 (0.01)	15	13.1 (0.6)	0.5 (0.4)	0.95 (0.01)
	$MBL_{nat/sp}$	140	-	-	-	240	-	-	-
C_t (g kg^{-1})	Cal_{local}	3	30.2 (12.0)	1.4 (3.9)	0.72 (0.19)	4	18.3 (9.5)	2.2 (4.1)	0.88 (0.11)
	Cal_{nat}	17	16.7 (0.8)	−0.3 (0.4)	0.94 (0.01)	13	13.1 (1.2)	1.0 (0.3)	0.96 (0.01)
	MBL_{nat}	100	-	-	-	120	-	-	-
	$Cal_{nat/sp}$	18	15.5 (0.9)	0.4 (0.4)	0.95 (0.01)	15	12.4 (0.7)	0.5 (0.4)	0.97 (0.01)
	$MBL_{nat/sp}$	80	-	-	-	100	-	-	-
N_t (g kg^{-1})	Cal_{local}	2	1.7 (0.8)	0.1 (0.2)	0.51 (0.31)	3	1.3 (0.5)	0.1 (0.3)	0.73 (0.13)
	Cal_{nat}	15	1.0 (0.0)	0.0 (0.0)	0.88 (0.01)	17	0.9 (0.0)	0.0 (0.0)	0.91 (0.01)
	MBL_{nat}	120	-	-	-	200	-	-	-
	$Cal_{nat/sp}$	17	1.0 (0.1)	0.0 (0.0)	0.89 (0.01)	20	0.8 (0.0)	0.0 (0.0)	0.92 (0.01)
	$MBL_{nat/sp}$	100	-	-	-	180	-	-	-
pH ($CaCl_2$)	Cal_{local}	3	0.8 (0.3)	0.1 (0.1)	0.70 (0.18)	4	0.8 (0.3)	0.0 (0.1)	0.70 (0.12)
	Cal_{nat}	19	0.4 (0.0)	0.0 (0.0)	0.92 (0.01)	18	0.4 (0.0)	0.0 (0.0)	0.92 (0.01)
	MBL_{nat}	100	-	-	-	220	-	-	-
	$Cal_{nat/sp}$	19	0.4 (0.0)	0.0 (0.0)	0.93 (0.01)	19	0.4 (0.0)	0.0 (0.0)	0.92 (0.01)
	$MBL_{nat/sp}$	100	-	-	-	240	-	-	-

Abbreviations: see Tables 1 and 2a.

neighbours to be tested ranged from 40 to 340 in steps of 20. Tables 2a and 2b show the median numbers of k of the respective five folds for the four properties, which ranged from 60 to 300.

2.3.4 | Training strategy IV: Training using the national database with spiking in standard calibration-validation ($Cal_{nat/sp}$)

Training strategy IV used the same regions 1–5 in the independent fivefold validations (Table 1) and the same respective national training sets as described above. The calibration was carried out as described for training strategy II above, except that for each of the five folds, the 20 soils (described in training strategy I) were used additionally in the calibration sets, and each spiking soil was weighted fivefold.

2.3.5 | Training strategy V: Training using the national database with spiking and MBL ($MBL_{nat/sp}$)

Training strategy V also used the regions 1–5 in the fivefold validation approach (Table 1) and the same respective national training sets as described above. The training was carried out as described for training strategy IV with the optimal data pretreatments for the respective properties and regions obtained in strategy IV, except that for each of the five folds, the 20 soils were used additionally in the training set as spiking soils.

2.4 | Additional analyses

Principal component analyses (PCAs) were carried out for the vis–NIR and MIR regions (Figure 2) using centred

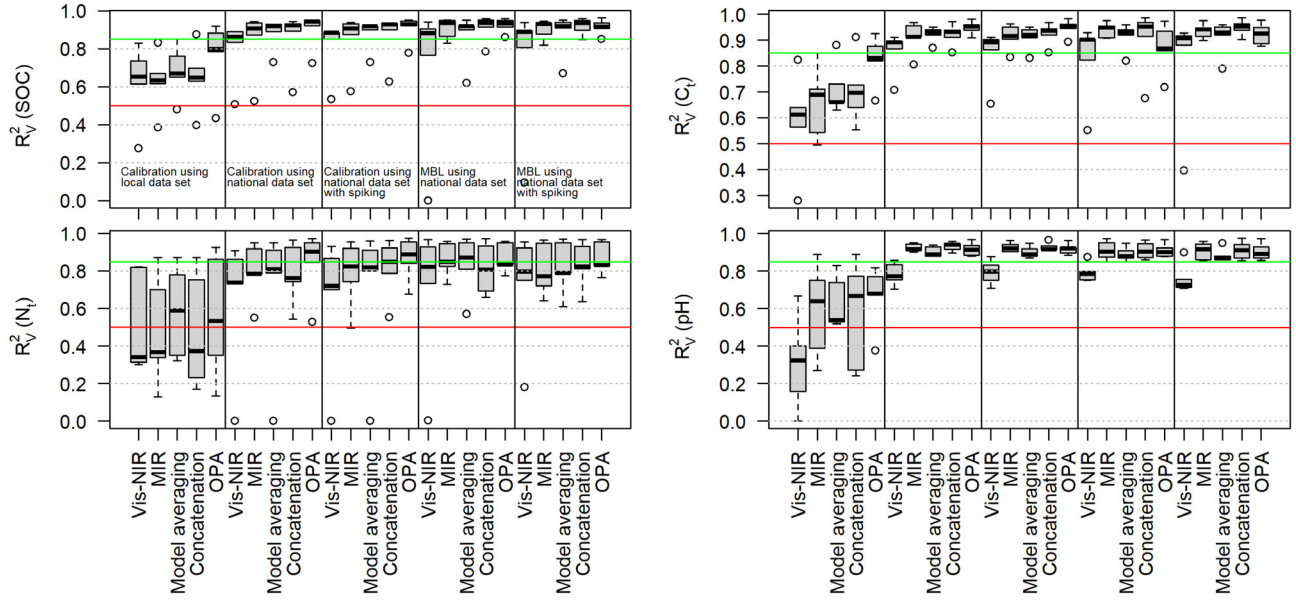

FIGURE 3 Boxplots of R^2_V (coefficient of determination, Nash–Sutcliffe model efficiency) of validation values for the four soil properties using the vis–NIR or MIR spectra or data fusion approaches [spectra concatenation, outer product analysis (OPA) and model averaging] for the five training variants. Green and red horizontal lines indicate thresholds for good ($R^2_V \geq 0.85$) and satisfactory ($R^2_V \geq 0.5$) validation results, respectively. Negative R^2_V values were set to zero.

spectral data and the variance–covariance matrix. Sampling points in Austria (Figure 1) were plotted using the packages sf (Pebesma, 2018) and tmap (Tennekes, 2018). Spearman rank correlation coefficients were calculated to study the relationships between R^2_V and IQR values.

3 | RESULTS

3.1 | Chemical and spectral soil properties

The ranges and IQR of the soil properties differed between the five regions, which were separated according to states and allowed five independent validations. For SOC, IQR values were almost identical with a ratio of 1.0 between regions 1 (Burgenland and Lower Austria) and 3 (Upper Austria) but differed by a factor of 2.1 between regions 2 (Carinthia and Salzburg) and 3 (Table 1).

MIR and vis–NIR spectra of the different regions are quite similar (Figure 2 shows mean values of absorbances ±1 standard deviation for each region). For MIRS, absorption was pronounced between 3600 and 3700 cm^{-1} (related to clay minerals kaolinite, smectite and illite), around 2000 and 1880 cm^{-1} (likely indicating quartz content), around 2950 and 2870 cm^{-1} (indicating aliphatic CH) and around 1600 cm^{-1} (indicating aromatic compounds). In contrast, vis–NIR spectra showed much less pronounced features.

PCA for MIRS showed a large overlap for the five regions (indicated by different colours, Figure 2), which suggests that independent validations for each of the five regions may be successful for the spectrally active properties SOC, C_t and N_t. However, exceptions included one soil from region 4 (Styria) on PC 1 and several from regions 1 (Burgenland and Lower Austria), 4 (Styria) and 5 (Tyrol and Vorarlberg) on PC 2. For these soils, predictions may be less accurate for spectrally active properties. The five regions were also similar with respect to soil types: regions 3 to 5 comprised 15 different soil types and regions 1 and 2 had 19 and 17 different soil types, respectively. For vis–NIRS, PCA indicated some deviations from the average spectra especially for regions 1 and 4 (Figure 2).

3.2 | Performance of the spectral approaches and training strategies in the fivefold validations

Figure 3 shows summarising information on the performances (measured as R^2_V) in the fivefold independent validations for the different spectral approaches and training strategies. For all the four soil properties, the calibrations using a local data set with n = 20 showed generally poor performances. However, OPA performed better than all other spectral approaches already for the local calibrations for C_t (Figure 3).

TABLE 3a Average performances (mean and standard deviation) of the PLS regressions in the validation for local calibration (Cal_{local}), calibration using the national archive (Cal_{nat}) and the national archive with spiking ($Cal_{nat/sp}$). Average performances are also shown for memory-based learning using the national archive (MBL_{nat}) and additionally with spiking ($MBL_{nat/sp}$).

		MIR validation			Vis–NIR validation		
Property	**Training strategy**	**SDE_V**	**ME_V**	**R^2_V**	**SDE_V**	**ME_V**	**R^2_V**
SOC (g kg^{-1})	Cal_{local}	31.5 (9.4)	2.0 (9.2)	0.63 (0.16)	31.5 (7.5)	2.6 (9.3)	0.62 (0.21)
	Cal_{nat}	18.4 (1.7)	0.1 (3.3)	0.84 (0.18)	22.4 (3.1)	−0.4 (3.2)	0.80 (0.16)
	MBL_{nat}	16.1 (4.7)	0.3 (2.8)	0.91 (0.06)	22.5 (4.8)	0.8 (7.8)	0.67 (0.45)
	$Cal_{nat/sp}$	18.4 (1.9)	−0.3 (3.6)	0.85 (0.15)	21.8 (3.7)	0.1 (4.8)	0.81 (0.15)
	$MBL_{nat/sp}$	16.5 (4.3)	0.1 (3.0)	0.90 (0.05)	21.7 (3.7)	0.2 (7.0)	0.73 (0.36)
C_t (g kg^{-1})	Cal_{local}	35.5 (9.0)	−0.2 (10.0)	0.66 (0.14)	39.2 (7.4)	1.6 (11.8)	0.58 (0.20)
	Cal_{nat}	17.2 (1.5)	0.0 (3.8)	0.91 (0.06)	23.1 (2.9)	−1.6 (4.5)	0.85 (0.08)
	MBL_{nat}	15.1 (4.0)	−0.4 (3.3)	0.94 (0.03)	23.8 (2.8)	0.6 (4.5)	0.82 (0.16)
	$Cal_{nat/sp}$	17.4 (2.0)	−0.9 (3.7)	0.91 (0.05)	23.3 (2.2)	−1.2 (4.8)	0.84 (0.11)
	$MBL_{nat/sp}$	15.6 (4.2)	−0.6 (3.4)	0.94 (0.03)	23.5 (3.9)	−0.5 (3.3)	0.81 (0.23)
N_t (g kg^{-1})	Cal_{local}	1.8 (0.4)	0.1 (0.6)	0.48 (0.30)	1.7 (0.4)	0.1 (0.6)	0.52 (0.27)
	Cal_{nat}	1.1 (0.2)	0.0 (0.2)	0.80 (0.16)	1.9 (1.4)	−0.2 (0.4)	−0.70 (3.37)
	MBL_{nat}	0.9 (0.2)	0.0 (0.1)	0.86 (0.09)	1.2 (0.4)	0.0 (0.3)	0.69 (0.39)
	$Cal_{nat/sp}$	1.1 (0.2)	0.0 (0.1)	0.79 (0.18)	1.7 (1.0)	−0.1 (0.3)	−0.11 (2.05)
	$MBL_{nat/sp}$	1.1 (0.3)	0.0 (0.1)	0.81 (0.14)	1.2 (0.3)	0.0 (0.3)	0.72 (0.31)
pH ($CaCl_2$)	Cal_{local}	0.8 (0.3)	0.0 (0.3)	0.59 (0.26)	1.1 (0.2)	−0.2 (0.1)	0.24 (0.37)
	Cal_{nat}	0.3 (0.0)	0.0 (0.1)	0.93 (0.02)	0.6 (0.0)	0.0 (0.1)	0.79 (0.06)
	MBL_{nat}	0.4 (0.1)	0.0 (0.1)	0.91 (0.05)	0.6 (0.1)	0.0 (0.1)	0.79 (0.05)
	$Cal_{nat/sp}$	0.3 (0.1)	0.0 (0.1)	0.93 (0.03)	0.6 (0.0)	0.0 (0.1)	0.79 (0.07)
	$MBL_{nat/sp}$	0.4 (0.1)	0.0 (0.1)	0.91 (0.05)	0.6 (0.1)	0.0 (0.1)	0.76 (0.08)

Abbreviations: ME_V, mean error of validation; R^2_V, coefficient of determination (Nash–Sutcliffe model efficiency) of validation; SDE_V, standard deviation of the error of validation. Other abbreviations: see Table 1.

For the training strategies which included the national database (strategies II to V), MIRS generally performed better than vis–NIRS for all the four properties. High-level data fusion by model averaging had generally similar performance compared with MIR alone for SOC, C_t and pH. However, for N_t, model averaging performed poorly in strategies II and IV compared with MIR alone. The low-level fusion approaches had generally similar or better performance compared with use of MIR alone. Also, MBL (strategies III and V) was superior over the general PLSR models (strategies II and IV) in several cases. The effect of spiking was variable (Figure 3, Tables 3a and 3b). Specifically, optimal validation performances based on mean R^2_V values (Tables 3a and 3b) and highest minimum R^2_V values of the five folds (Figure 3) were obtained for SOC and N_t by OPA of vis–NIRS and MIRS and MBL_{nat} (strategy III, SOC: mean $R^2_V = 0.92$ and highest minimum $R^2_V = 0.86$, N_t: mean $R^2_V = 0.87$ and highest minimum $R^2_V = 0.78$), for C_t by OPA and Cal_{nat} (strategy II: mean $R^2_V = 0.95$ and highest minimum $R^2_V = 0.91$) and for pH by spectra concatenation and $Cal_{nat/sp}$ (strategy IV: mean $R^2_V = 0.93$ and highest minimum $R^2_V = 0.91$).

The greatest robustness of performance over all the four properties was found for OPA and $MBL_{nat/sp}$ with $R^2 \geq 0.77$ (N_t), 0.85 (SOC), 0.86 (pH) and 0.88 (C_t) in the validations of all folds (Figure 4). For SOC, all predictions were good ($R^2_V \geq 0.85$), and only few larger deviations between measured and modelled data were noted (one outlier in regions 1, 3 and 4 each). For C_t, agreements were even better, with R^2_V values ranging from 0.88 to 0.98 and only one outlier in regions 3 and 4. For N_t and pH, validations were at least satisfactory ($0.77 \leq R^2_V \leq 0.97$) for N_t and good for pH ($0.86 \leq R^2_V \leq 0.98$, Figure 4). For N_t, region 2, which had the worst performance, was the one with the smallest IQR values (Figure 4, Table 1).

TABLE 3b Average performances (mean and standard deviation) of the PLS regressions in the validation for local calibration (Cal_{local}), calibration using the national archive (Cal_{nat}) and the national archive with spiking ($Cal_{nat/sp}$). Average performances are also shown for memory-based learning using the national archive (MBL_{nat}) and additionally with spiking ($MBL_{nat/sp}$).

		Vis–NIR-MIR concatenation validation			OPA vis–NIR-MIR validation			Vis–NIR and MIR model averaging validation		
Property	**Training strategy**	**SDE_V**	**ME_V**	**R^2_V**	**SDE_V**	**ME_V**	**R^2_V**	**SDE_V**	**ME_V**	**R^2_V**
SOC (g kg^{-1})	Cal_{local}	30.0 (8.9)	0.9 (9.4)	0.65 (0.17)	23.3 (5.7)	2.2 (7.6)	0.76 (0.19)	29.3 (9.1)	2.3 (9.0)	0.68 (0.14)
	Cal_{nat}	17.8 (2.2)	−0.6 (2.9)	0.85 (0.16)	15.4 (2.6)	0.5 (3.3)	0.90 (0.10)	17.9 (3.5)	0.0 (3.1)	0.88 (0.09)
	MBL_{nat}	15.2 (3.5)	0.2 (2.1)	0.91 (0.07)	14.8 (4.3)	0.4 (2.4)	0.92 (0.04)	17.2 (2.9)	0.5 (3.6)	0.86 (0.14)
	$Cal_{nat/sp}$	17.6 (2.9)	−0.5 (3.3)	0.86 (0.13)	15.4 (2.2)	−0.1 (2.5)	0.91 (0.07)	17.9 (3.7)	0.1 (4.0)	0.88 (0.08)
	$MBL_{nat/sp}$	15.1 (4.1)	0.5 (2.2)	0.92 (0.05)	15.3 (3.7)	0.6 (2.9)	0.92 (0.04)	16.5 (3.7)	0.1 (3.4)	0.88 (0.12)
C_t (g kg^{-1})	Cal_{local}	32.8 (10.5)	1.7 (9.3)	0.71 (0.13)	25.1 (5.0)	0.8 (7.5)	0.82 (0.10)	33.6 (8.9)	1.3 (9.5)	0.71 (0.10)
	Cal_{nat}	16.8 (3.0)	−1.0 (3.0)	0.92 (0.04)	14.0 (3.2)	−0.1 (1.9)	0.95 (0.03)	17.6 (2.9)	−0.7 (3.8)	0.92 (0.03)
	MBL_{nat}	15.9 (5.2)	−0.4 (1.9)	0.90 (0.13)	20.2 (4.0)	−0.5 (3.8)	0.87 (0.10)	17.5 (2.4)	0.1 (2.1)	0.91 (0.05)
	$Cal_{nat/sp}$	16.6 (2.5)	−0.8 (2.6)	0.92 (0.04)	13.7 (2.8)	−0.4 (1.3)	0.95 (0.03)	18.3 (2.4)	−0.8 (2.9)	0.91 (0.05)
	$MBL_{nat/sp}$	13.5 (3.8)	−0.4 (1.6)	0.95 (0.03)	16.8 (5.4)	0.5 (4.0)	0.92 (0.04)	17.3 (2.8)	−0.6 (2.2)	0.91 (0.07)
N_t (g kg^{-1})	Cal_{local}	1.8 (0.4)	0.0 (0.5)	0.48 (0.31)	1.5 (0.4)	0.1 (0.5)	0.56 (0.34)	1.6 (0.5)	0.1 (0.6)	0.58 (0.25)
	Cal_{nat}	1.1 (0.3)	0.0 (0.1)	0.79 (0.17)	0.9 (0.2)	0.0 (0.1)	0.84 (0.18)	1.4 (0.6)	−0.1 (0.2)	0.45 (0.92)
	MBL_{nat}	1.0 (0.3)	0.0 (0.1)	0.81 (0.14)	0.9 (0.2)	0.0 (0.2)	0.87 (0.08)	0.9 (0.3)	0.0 (0.2)	0.83 (0.16)
	$Cal_{nat/sp}$	1.0 (0.2)	0.0 (0.2)	0.81 (0.16)	0.8 (0.2)	0.0 (0.2)	0.87 (0.12)	1.2 (0.5)	0.0 (0.2)	0.61 (0.59)
	$MBL_{nat/sp}$	1.0 (0.2)	0.0 (0.1)	0.83 (0.13)	0.9 (0.2)	0.0 (0.2)	0.87 (0.09)	1.0 (0.3)	0.0 (0.2)	0.82 (0.15)
pH ($CaCl_2$)	Cal_{local}	0.8 (0.3)	0.0 (0.3)	0.57 (0.30)	0.7 (0.2)	0.1 (0.1)	0.66 (0.17)	0.8 (0.2)	−0.1 (0.2)	0.63 (0.14)
	Cal_{nat}	0.3 (0.0)	0.0 (0.1)	0.93 (0.02)	0.4 (0.1)	0.0 (0.1)	0.92 (0.04)	0.4 (0.0)	0.0 (0.1)	0.90 (0.03)
	MBL_{nat}	0.4 (0.1)	0.0 (0.1)	0.91 (0.05)	0.4 (0.1)	0.0 (0.1)	0.91 (0.04)	0.4 (0.1)	0.0 (0.1)	0.89 (0.04)
	$Cal_{nat/sp}$	0.3 (0.0)	0.0 (0.1)	0.93 (0.02)	0.4 (0.1)	0.0 (0.1)	0.92 (0.03)	0.4 (0.0)	0.0 (0.1)	0.90 (0.03)
	$MBL_{nat/sp}$	0.4 (0.1)	0.0 (0.1)	0.91 (0.05)	0.4 (0.1)	0.0 (0.1)	0.91 (0.05)	0.4 (0.1)	0.0 (0.1)	0.88 (0.04)

Abbreviations: see Tables 1 and 3a.

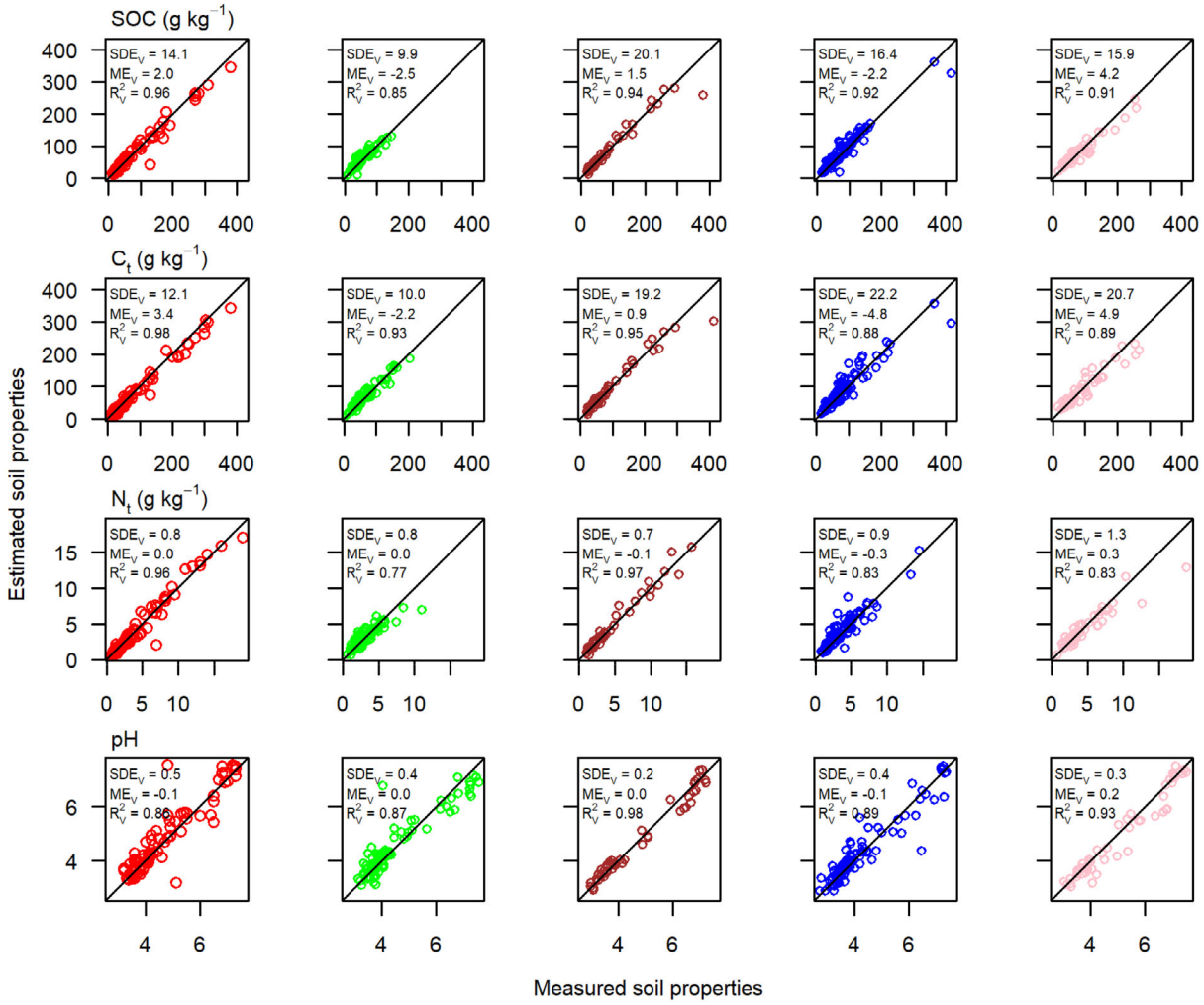

FIGURE 4 Validation results for the outer product analysis vis–NIR-MIR validation with memory-based learning for the five folds using the national dataset and spiking (red: Lower Austria and Burgenland, green: Carinthia and Salzburg, brown: Upper Austria, blue: Styria and pink: Tyrol and Vorarlberg).

3.3 | Relationships between R^2, RMSE and IQR

For all the four properties, R^2_V values in the different folds showed a marked scatter (Figure 3). This scatter is related to the variation of the response variable (TSS), which is related to the IQR, and RMSE (consisting of bias and imprecision). Figure 5 shows the R^2_V values against IQR and $RMSE_V$ values for SOC for the training strategies II ($cal_{nat/sp}$) and IV ($MBL_{nat/sp}$) for MIRS (open symbols) and OPA (closed symbols) as examples. For SOC, the generally smaller $RMSE_V$ values for OPA indicate the benefit of this data fusion approach. The benefits of MBL are especially visible for small IQR values, which overall resulted in good [SOC (Figure 5), C_t and pH] and satisfactory (N_t) predictions for all five folds in the respective validations.

The smallest R^2_V values were obtained for region 2, the region with the smallest IQR and thus smallest variation of the contents of response variable (Figure 5). The Spearman rank correlation coefficient between R^2_V and IQR for the SOC example was 0.86, which shows the profound effect of the variation of the contents on the R^2_V values.

4 | DISCUSSION

4.1 | Performance of the spectral approaches and training strategies in the fivefold validations

Calibrations with local datasets of n = 20 did not result in convincing validation results for the different spectral approaches. These results indicate that such a small set does not include sufficient spectral variation for successful validations at this geographical scale, despite the

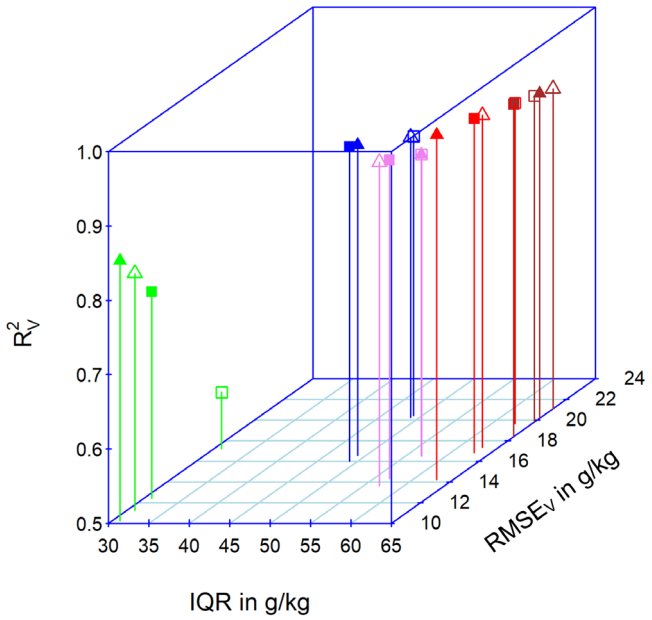

FIGURE 5 R^2_V (coefficient of determination, Nash–Sutcliffe model efficiency) of validation values against IQR (interquartile range) and $RMSE_V$ (root mean square error of validation) errors for MIRS (open symbols) and outer product analysis (closed symbols) with the variants calibration (squares) and memory-based learning (MBL) (triangles) using the national dataset and spiking (colour codes: see legend of Figure 1) for soil organic carbon (SOC).

stratified random approach to selecting diverse soil types for the training set. The Kennard-Stone algorithm was also tested to select a spectrally diverse subset for training, but performance was likewise generally poor. However, local calibrations with small samples sizes, e.g., n = 20, have been successful for SOC and N at the field scale (Debaene et al., 2014; Greenberg et al., 2022) and 5 km^2 scale (Guerrero et al., 2010), suggesting that the performance of small local calibrations is related to the heterogeneity of the target site. In the present study, local calibrations showed good performances for all the four properties for at least one validation fold of one approach and region (Figure 3) but poor performances for several other validation folds, indicating that the use of different validation sets defined by multiple partitioning is essential for conclusive evaluations.

For the local calibrations with limited spectral information, the greater potential of MIRS over vis–NIRS is obvious (except for N_t), but OPA already showed its potential for the property C_t, indicating that vis–NIRS was not entirely redundant to MIRS and provided complementary information helpful to improve estimation accuracies as has been pointed out by Terra et al. (2019) for the region of 600 to 1000 nm.

Training strategy II to V indicated a general superiority of MIRS over vis–NIRS for all the four properties [with n ranging from 362 to 414 for Cal_{nat} (strategy II) and MBL_{nat} (strategy IV)]. This superiority is in line with other studies and can be traced back to more pronounced information in the MIR spectra, consisting of fundamental vibrations (e.g., Ng et al., 2019; Soriano-Disla et al., 2014; Vohland et al., 2014).

In our study, low-level data fusion generally outperformed MIRS, with slightly greater benefits of OPA (for SOC, C_t and N_t) over model concatenation (for pH) based on mean R^2_V values (Tables 3a and 3b) and highest minimum R^2_V values of the five folds (Figure 3). In contrast, model averaging was not as successful, which is not surprising considering the superior performance of MIRS over vis–NIRS. The very good potential of OPA—also reported by Terra et al. (2019) and Xu et al. (2020)—has been assigned to the mutual weighting of each signal (vis–NIRS and MIRS) by the other, with maximum products when the intensities are simultaneously high in the two domains (Barros et al., 2008). In contrast, Vohland et al. (2022) did not find a benefit of OPA for SOC prediction, where model averaging outperformed OPA. The latter study, however, was based on a smaller data set of 186 spectra, and the differences between vis–NIR and MIR prediction accuracies were not as pronounced as in this study, resulting in more reliable vis–NIR estimates for model averaging in that study. From a practical point of view with respect to cost-effectiveness of reflectance spectroscopy (Li, Viscarra Rossel, & Webster, 2022), however, it has to be noted that fusion of vis–NIRS and MIRS increases the cost of spectroscopy and thus decreases the cost-effectiveness compared with a stand-alone MIRS approach.

The combination of OPA and MBL with spiking achieved the greatest robustness of performance over all the four properties. Prediction accuracies were good ($R^2_V \geq 0.85$ for all five validation folds) for SOC, C_t and pH, whereas for N_t predictions, accuracies were at least satisfactory ($0.77 \leq R^2 \leq 0.97$ for the five validation folds). The usefulness of MBL for any validation set will depend largely on the representativeness of the considered SSL and the defined validation set. In addition, dissimilarity methods used to find spectral neighbours will determine the predictive success of MBL models (Summerauer et al., 2021).

4.2 | Relationships between R^2, RMSE and IQR

Besides the training and data fusion effects discussed above, the interquartile distance (which is related to the variation of the contents of the response variable) also affected the R^2_V values markedly (the Spearman rank correlation coefficient between R^2_V and IQR was 0.86 for

SOC for MIRS and OPA for training strategies II and IV, Figure 5). This indicates the need to focus on both contributions for quality assessment in spectroscopy: factors that affect $RMSE_V$, SDE_V and ME_V, such as instrumental specification, spectral approach, training approach and soil mineral composition, must be considered in conjunction with the range of contents, which affects R^2_V values, RPD values and ratios of performance to interquartile distance (RPIQ).

The results suggest that by selecting only data sets (and folds) with large standard deviations and/or IQR, any model for spectrally active properties can be pushed to become very successful when rated only by R^2_V values, RPD or RPIQ values and that a focus on error measures such as RMSE, SDE and ME is equally important.

AUTHOR CONTRIBUTIONS

Bernard Ludwig: Conceptualization; methodology; software; data curation; supervision; formal analysis; validation; investigation; funding acquisition; writing – original draft; visualization; project administration; resources. **Isabel Greenberg:** Methodology; investigation; visualization; writing – review and editing. **Michael Vohland:** Writing – review and editing; funding acquisition; validation; methodology. **Kerstin Michel:** Writing – review and editing; resources; data curation; conceptualization; investigation; methodology; formal analysis.

ACKNOWLEDGEMENTS

The map of Austria was created using data provided by Statistik Austria: data.statistik.gv.at. This project was supported by a grant from the German Research Foundation (DFG, LU 583/19-1, VO 1509/7-1). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Bernard Ludwig https://orcid.org/0000-0001-8900-6190
Isabel Greenberg https://orcid.org/0000-0002-4762-8474
Michael Vohland https://orcid.org/0000-0002-6048-1163
Kerstin Michel https://orcid.org/0000-0001-6139-6596

REFERENCES

Barros, A. S., Pinto, R., Jouan-Rimbaud Bouveresse, D., & Rutledge, D. N. (2008). Principal component transform outer product analysis in the PCA context. *Chemometrics and Intelligent Laboratory Systems*, *93*, 43–48. https://doi.org/10.1016/j.chemolab.2008.03.009

BMLFUW—(ÖSTERREICHISCHES) BUNDESMINISTERIUM FÜR LAND- UND FORSTWIRTSCHAFT, UMWELT UND WASSERWIRTSCHAFT. (2013). Bodenfunktionsbewertung: Methodische Umsetzung der ÖNORM L 1076. Wien. 105 pp.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment—A review. *Analytica Chimica Acta*, *891*, 1–14. https://doi.org/10.1016/j.aca.2015.04.042

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107. https://doi.org/10.5555/1756006.1859921

Chang, C. W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R., Jr. (2001). Near infrared reflectance spectroscopy: Principal components regression analyses of soil properties. *Soil Science Society of America Journal*, *65*, 480–490. https://doi.org/10.2136/sssaj2001.652480x

Debaene, G., Niedzwiecki, J., Pecio, A., & Zurek, A. (2014). Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. *Geoderma*, *214–215*, 114–125. https://doi.org/10.1016/j.geoderma.2013.09.022

Gholizadeh, A., Boruvka, L., Saberioon, M., & Vasat, R. (2013). Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. *Applied Spectroscopy*, *67*, 1349–1362. https://doi.org/10.1366/13-07288

Greenberg, I., Linsler, D., Vohland, M., & Ludwig, B. (2020). Robustness of visible-near infrared and mid-infrared spectroscopic models to changes in the quantity and quality of crop residues in soil. *Soil Science Society of America Journal*, *84*, 963–977. https://doi.org/10.1002/saj2.20067

Greenberg, I., Seidel, M., Vohland, M., Koch, H.-J., & Ludwig, B. (2022). Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies. *Geoderma*, *409*, 115614. https://doi.org/10.1016/j.geoderma.2021.115614

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., Zornoza, R., Ruiz-Sinoga, J. D., & Kuang, B. (2014). Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset. *European Journal of Soil Science*, *65*, 248–263.

Guerrero, C., Zornoza, R., Gomez, I., & Mataix-Beneyto, J. (2010). Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma*, *158*, 66–77.

Jaconi, A., Don, A., & Freibauer, A. (2017). Prediction of soil organic carbon at the country scale: Stratification strategies for near-infrared data. *European Journal of Soil Science*, *68*, 919–929. https://doi.org/10.1111/ejss.12485

Janssen, P., & Heuberger, P. (1995). Calibration of process-oriented models. *Ecological Modelling*, *83*, 55–66. https://doi.org/10.1016/0304-3800(95)00084-9

Johnson, J.-M., Vandamme, E., Senthilkumar, K., Sila, A., Shepherd, K. D., & Saito, K. (2019). Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-Saharan Africa.

Geoderma, *354*, 113840. https://doi.org/10.1016/j.geoderma.2019.06.043

Leenen, M., Welp, G., Gebbers, R., & Pätzold, S. (2019). Rapid determination of lime requirement by mid-infrared spectroscopy: A promising approach for precision agriculture. *Journal of Plant Nutrition and Soil Science*, *182*, 953–963. https://doi.org/10.1002/jpln.201800670

Li, H., Li, Y., Yang, M., Chen, S., & Shi, Z. (2022). Strategies for efficient estimation of soil organic content at the local scale based on a national spectral database. *Land Degradation & Development*, *33*, 1649–1661. https://doi.org/10.1002/ldr.4223

Li, S., Viscarra Rossel, R. A., & Webster, R. (2022). The cost-effectiveness of reflectance spectroscopy for estimating soil organic carbon. *European Journal of Soil Science*, *73*, e13202. https://doi.org/10.1111/ejss.13202

Liland, K., Mevik, B., & Wehrens, R. (2021). pls: Partial Least Squares and Principal Component Regression. R package version 2.8-0. https://CRAN.R-project.org/package=pls

Ludwig, B., Greenberg, I., Sawallisch, A., & Vohland, M. (2021). Diffuse reflectance infrared spectroscopy estimates for soil properties using multiple partitions: Effects of the range of contents, sample size, and algorithms. *Soil Science Society of America Journal*, *85*, 546–559. https://doi.org/10.1002/saj2.20205

Ludwig, B., Murugan, R., Parama, V. R. R., & Vohland, M. (2019). Accuracy of estimating soil properties with mid-infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. *Soil Science Society of America Journal*, *83*, 1542–1552. https://doi.org/10.2136/sssaj2018.11.0413

Ludwig, B., Nitschke, R., Terhoeven-Urselmans, T., Michel, K., & Flessa, H. (2008). Use of mid-infrared spectroscopy in the diffuse reflectance mode for the prediction of the composition of organic matter in soil and litter. *Journal of Plant Nutrition and Soil Science*, *171*, 384–391.

McBride, M. B. (2022). Estimating soil chemical properties by diffuse reflectance spectroscopy: Promise versus reality. *European Journal of Soil Science*, *73*, e13192. https://doi.org/10.1111/ejss.13192

Mirzaeitalarposhti, R., Demyan, M. S., Rasche, F., Cadisch, G., & Müller, T. (2016). Overcoming carbonate interference on labile soil organic matter peaks for midDRIFTS analysis. *Soil Biology and Biochemistry*, *99*, 150–157. https://doi.org/10.1016/j.soilbio.2016.05.010

Nestroy, O., Aust, G., Blum, W. E. H., Englisch, M., Hager, H., Herzberger, E., Kilian, W., Nelhiebl, P., Ortner, G., Pecina, E., Pehamberger, A., Schneider, W., & Wagner, J. (2011). Systematische Gliederung der Böden Österreichs. Österreichische Bodensystematik 2000 in der revidierten Fassung von 2011. *Mitteilungen der Österreichischen Bodenkundlichen Gesellschaft, 79*, Wien.

Ng, W., Minasny, B., Montazerolghaeum, M., Padarian, J., Ferguson, R., Bailey, S., & McBratney, A. B. (2019). Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*, *352*, 251–267. https://doi.org/10.1016/j.geoderma.2019.06.016

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. B., Brown, D. J., Clairotte, M., Csorba, A., & Dardenne, P. (2015). Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in Agronomy*, *132*, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002

ÖNORM L 1080. (1989). Chemische Bodenuntersuchungen; Bestimmung des organischen Kohlenstoffs und des Humusgehalts durch trockene Verbrennung unter Berücksichtigung der Carbonate und des elementaren Kohlenstoffs. Österr. Normungsinstitut, Wien.

ÖNORM L 1082. (1989). Chemische Bodenuntersuchungen; Bestimmung von Gesamtstickstoff. Österr. Normungsinstitut, Wien.

ÖNORM L 1083. (1989). Chemische Bodenuntersuchungen; Bestimmung der Acidität. Österr. Normungsinstitut, Wien.

ÖNORM L 1084. (1989). Chemische Bodenuntersuchungen—Bestimmung von Carbonat unter Berücksichtigung von Luftdruck und Temperatur. Österr. Normungsinstitut, Wien.

O'Rourke, S. M., Argentati, I., & Holden, N. M. (2011). The effect of region of interest size on model calibration for soil organic carbon prediction from hyperspectral images of prepared soils. *Journal of Near Infrared Spectroscopy*, *19*, 161–170. https://doi.org/10.1255/jnirs.930

Pallottino, F., Antonucci, F., Costa, C., Bisagli, C., Figorilli, S., & Menesatti, P. (2019). Optoelectronic proximal sensing vehicle-mounted technologies in precision agriculture: A review. *Computers and Electronics in Agriculture*, *162*, 859–873. https://doi.org/10.1016/j.compag.2019.05.034

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*, 439–446. https://doi.org/10.32614/RJ-2018-009

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

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil Vis–NIR spectra of complex datasets. *Geoderma*, *195–196*, 268–279. https://doi.org/10.1016/j.geoderma.2012.12.014

Ramirez-Lopez, L., Stevens, A., Viscarra Rossel, R., Lobsey, C., Wadoux, A., & Breure, T. (2022). resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics. R package Vignette R package version 2.1.2. https://CRAN.R-project.org/package=resemble

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of near Infrared Spectroscopy*, *5*, 223–232. https://doi.org/10.1255/jnirs.115

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

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PLoS One*, *8*, e66409. https://doi.org/10.1371/journal.pone.0066409

Stevens, A., & Ramirez-Lopez, L. (2022). An introduction to the prospectr package. R package Vignette R package version 0.2.6.

https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr-intro.pdf

Summerauer, L., Baumann, P., Ramirez-Lopez, L., Barthel, M., Bauters, M., Bukombe, B., Reichenbach, M., Boeckx, P., Kearsley, E., Van Oost, K., Vanlauwe, B., Chiragaga, D., Heri-Kazi, A. B., Moonen, P., Sila, A., Shepherd, K., Mujinya, B. B., Van Ranst, E., Baert, G., … Six, J. (2021). The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis. *The Soil*, *7*, 693–715. https://doi.org/10.5194/soil-7-693-2021

Tatzber, M., Mutsch, F., Mentler, A., Leitgeb, E., Englisch, M., & Gerzabek, M. H. (2010). Determination of organic and inorganic carbon in forest soil samples by mid-infrared spectroscopy and partial least squares regression. *Applied Spectroscopy*, *64*, 1167–1175. https://doi.org/10.1366/000370210792973460

Tennekes, M. (2018). Tmap: Thematic maps in R. *Journal of Statistical Software*, *84*, 1–39. https://doi.org/10.18637/jss.v084.i06

Terra, F. S., Viscarra Rossel, R., & Demattê, J. A. M. (2019). Spectral fusion by outer product analysis (OPA) to improve predictions of soil organic C. *Geoderma*, *335*, 35–46. https://doi.org/10.1016/j.geoderma.2018.08.005

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., & Shen, Z. (2022). Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. *European Journal of Soil Science*, *73*, e13271. https://doi.org/10.1111/ejss.13271

Viscarra Rossel, R. A., & McBratney, A. B. (1998). Soil chemical analytical accuracy and costs: Implications from precision agriculture. *Australian Journal of Experimental Agriculture*, *38*, 765–775.

Viscarra Rossel, R. A., & Webster, R. (2012). Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *European Journal of Soil Science*, *63*, 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x

Vohland, M., Ludwig, B., Seidel, M., & Hutengs, C. (2022). Quantification of soil organic carbon at regional scale: Benefits of fusing Vis–NIR and MIR diffuse reflectance data are greater for in situ than for laboratory-based modelling approaches. *Geoderma*, *405*, 115426. https://doi.org/10.1016/j.geoderma.2021.115426

Vohland, M., Ludwig, M., Thiele-Bruhn, S., & Ludwig, B. (2014). Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. *Geoderma*, *223–225*, 88–96. https://doi.org/10.1016/j.geoderma.2014.01.013

Wijewardane, N. K., Ge, Y. F., Wills, S., & Libohova, Z. (2018). Predicting physical and chemical properties of us soils with a mid-infrared reflectance spectral library. *Soil Science Society of America Journal*, *82*, 722–731. https://doi.org/10.2136/sssaj2017.10.0361

Xu, H., Xu, D., Chen, S., Ma, W., & Shi, Z. (2020). Rapid determination of soil class based on visible-near infrared, mid-infrared spectroscopy and data fusion. *Remote Sensing*, *12*, 1512. https://doi.org/10.3390/rs12091512

Yang, Y., Shen, Z., Bissett, A., & Viscarra Rossel, R. A. (2022). Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions. *The Soil*, *8*, 223–235. https://doi.org/10.5194/soil-8-223-2022

Zeng, R., Zhao, Y. G., Wu, D. W., Wei, C. L., & Zhang, G. L. (2016). Comparison of different strategies for predicting soil organic matter of a local site from a regional Vis–NIR soil spectral library. In G. L. Zhang, D. Brus, F. Liu, X. D. Song, & P. Lagacherie (Eds.), *Digital soil mapping across paradigms, scales and boundaries* (Springer Environmental Science and Engineering). Springer. https://doi.org/10.1007/978-981-10-0415-5_26

How to cite this article: Ludwig, B., Greenberg, I., Vohland, M., & Michel, K. (2023). Optimised use of data fusion and memory-based learning with an Austrian soil library for predictions with infrared data. *European Journal of Soil Science*, *74*(4), e13394. https://doi.org/10.1111/ejss.13394

APPENDIX A

A.1 | Classification of soils according to Austrian soil taxonomy (Nestroy et al., 2011).

Group "A—Fluvisols and others (number of plots: 28)": Anmoor (2), Brauner Auboden (6), Grundwassergley (6), Kunstböden (1), Rohauboden, Grauer Auboden (2), Rohböden und Ranker (9), Schwemmböden und Bachauböden (2).

Group "B—predominantly Podzols (92)": Klimabedingter Podsol (11), Semipodsol auf Kristallin (76), Substratbedingter Podsol (5).

Group "C—predominantly Cambisols (189)": Bindige Braunerde auf Moränen, Geschiebe, Staublehm und tonhältigem Ausgangsmaterial allgemein (auch Werfener Schichten) (21), Braunerde aus Löß (3), Braunerde und Hangkolluvien auf ärmerem Kristallin (86), Braunerde und Kolluvien auf basenreichem Kristallin und kalkbeeinflusste Braunerde (60), Leichte Braunerde und podsolige Braunerde auf Lockersedimenten (14), Parabraunerde (2), Silikatischer Braunlehm, Rotlehm (3).

Group "D—predominantly Leptosols and Phaeozems (121)": Mischböden aus Rendsina und Terra fusca (28), Pararendsina (3), Rendsina und Rohböden auf Kalk (40), Terra fusca und Kalksteinlehm allgemein (50).

Group "E—predominantly Planosols and Stagnosols (51)": Hangpseudogley und Hanggley (19), Pseudogley auf Flysch, Werfener Schichten, Fleckenmergel und anderem tonhaltigem, festem Grundgestein (11), Pseudogley auf Lockersedimenten (17), Pseudogley auf Löß (1), Stagnogley (3).