Received: 4 September 2020 | Revised: 18 November 2020 | Accepted: 18 November 2020 | Published online: 5 April 2021

DOI: 10.1002/saj2.20205

Soil Science Society of America Journal

SOIL CHEMISTRY

Diffuse reflectance infrared spectroscopy estimates for soil properties using multiple partitions: Effects of the range of contents, sample size, and algorithms

Bernard Ludwig[1] | **Isabel Greenberg**[1] | **Anja Sawallisch**[1] | **Michael Vohland**[2]

[1] Department of Environmental Chemistry, Kassel University, Nordbahnhofstr. 1a, Witzenhausen 37213, Germany

[2] Geoinformatics and Remote Sensing, Institute for Geography, Leipzig University, Johannisallee 19a, Leipzig 04103, Germany

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

Assigned to Associate Editor Kang Xia

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: LU 583/19-1, VO 1509/7-1

Abstract

The RMSE of validation ($RMSE_V$) and ratio of the interquartile range to $RMSE_V$ ($RPIQ_V$) are key quality parameters in diffuse reflectance infrared (IR) spectroscopy studies, but the effects of different factors on these parameters are often not sufficiently considered. Our objectives were to reveal the effects of range of contents, sample size, data pretreatment, wavenumber region selection, and algorithms on the evaluation of IR spectra in the wavenumber range from 1,000 to 7,000 cm^{-1} (mid- and long-wave near IR) estimations. Contents of soil organic C (SOC), N, clay, and sand and pH values were determined for surface soils of an arable field in India, and IR spectra were recorded for four samples consisting of 71–263 soils. For each of the four samples, five random partitions into calibration and validation datasets were carried out, and partial least squares regression (PLSR) or support vector machine regression was performed. A plot of the $RMSE_V$ values against the interquartile ranges of measured values for the validation samples (IQR_V) indicated that the IQR_V was a key parameter for all soil properties: a sufficiently high IQR_V—which is affected by sample size and random partitioning—resulted in generally good estimation accuracies ($RPIQ_V \geq 2.70$). Optimized data pretreatment and wavenumber region selection improved estimation accuracy for SOC and pH. Support vector machine regression was superior to PLSR for the estimation of SOC, clay, and sand, but worse for pH. Overall, this study indicates that multiple partitioning of the data is essential in IR studies and suggests that $RPIQ_V$ and $RMSE_V$ need to be interpreted in the context of the respective IQR_V values.

1 | INTRODUCTION

Mid-infrared (MIR, range: 2,500–25,000 nm, 4,000–400 cm^{-1}) and visible to near-infrared (visNIR, range: 400–2,500 nm, 25,000–4,000 cm^{-1}) spectroscopy (MIRS and visNIRS, respectively) have proven to be useful techniques for the simultaneous estimation of a number of chemical and physical soil properties (Gholizadeh, Boruvka,

Abbreviations: DRIFT, diffuse reflectance infrared Fourier transform; IQR, interquartile range; IR, infrared; MIR, mid-infrared; MIRS, mid-infrared spectroscopy; NIR, near-infrared; PLSR, partial least squares regression; RPIQ, ratio of performance to interquartile distance; SOC, soil organic carbon; SVMR, support vector machine regression; visNIR, visible- and near-infrared; visNIRS, visible to near-infrared spectroscopy.

Saberioon, & Vasat, 2013; Hutengs, Ludwig, Jung, Eisele, & Vohland, 2018; Kuang et al., 2012; O'Rourke, Minasny, Holden, & McBratney, 2016; Pallottino et al., 2019; Soriano-Disla, Janik, Viscarra Rossel, Macdonald, & McLaughlin, 2014). Their application may be especially beneficial in terms of reduction of analytical costs and time in studies that require large datasets (O'Rourke, Argentati, & Holden, 2011); for example, where the focus is on spatial or temporal monitoring of soil properties.

The usefulness of MIRS and visNIRS for accurate estimations of soil organic C (SOC) and total N content is well established (Baldock, Hawke, Sanderman, & Macdonald, 2013; Bellon-Maurel & McBratney, 2011; Reeves, 2010), and important vibrations are those related to alkyl groups, protein amides, carboxylic acids, water-associated groups, carboxylate anions, and aromatic groups (Soriano-Disla et al., 2014). For SOC and N, a marked overlap in the MIR band assignments have been reported (Vohland, Ludwig, Thiele-Bruhn, & Ludwig, 2014). Overall, it cannot be ruled out that N contents are estimated to some extent indirectly in the spectral range under investigation. In a laboratory spectroscopy study with different additions of wheat (*Triticum aestivum* L.) straw and clover (*Trifolium* spp.) residues to soils, Greenberg, Linsler, Vohland, and Ludwig (2020) summarized that based on the loadings of partial least squares regression (PLSR) components, the predictive mechanisms for SOC and N appear to be quite similar for visNIRS, but the important wavelengths were less aligned for MIRS.

Important vibrations for the accurate estimation of clay and sand are those associated with kaolinite (3,690–3,620 cm^{-1}), smectite and illite (3,630–3,620 and 3,400–3,300 cm^{-1}, respectively), and quartz (1,100–1,000 cm^{-1}; Soriano-Disla et al., 2014).

The RMSE of validation ($RMSE_V$), ratio of SD to $RMSE_V$ (RPD_V), and ratio of the interquartile range (IQR) to $RMSE_V$ ($RPIQ_V$) are key quality parameters in infrared (IR) studies to describe the performance (accuracy and precision) of a spectroscopic model (Bellon-Maurel, Fernandez-Ahumada, Palagos, Roger, & McBratney, 2010; Soriano-Disla et al., 2014). The $RMSE_V$ (consisting of bias and imprecision) is commonly used to describe the predictive ability in IR studies and appears as an averaged error recorded on the validation sample according to Bellon-Maurel et al. (2010). The RPD_V is used to compare the model performance across datasets (Soriano-Disla et al., 2014) but is appropriate only for normally distributed soil properties. Bellon-Maurel et al. (2010) thus suggested calculation of $RPIQ_V$ values instead. The IQR in the numerator gives the range that accounts for 50% of the population around the median, and thus the $RPIQ_V$ better describes the spread of the population for typical cases of non-normally distributed soil properties (Bellon-Maurel et al., 2010). The factors affecting $RMSE_V$ and $RPIQ_V$ (which has the $RMSE_V$ in the denominator) are manifold for spectrally active soil properties, and the relationships are not always simple. The $RMSE_V$ and (thus also) $RPIQ_V$ are instrument related (signal to noise ratios differ between instruments), soil property related (see, e.g., Kuang et al., 2012; Soriano-Disla et al., 2014), and dependent on the soil treatment prior to the scanning (e.g., Hutengs, Ludwig, Oertel, Seidel, and Vohland [2019] found $RMSE_V$ to be greatly affected by drying and grinding). Furthermore, Stenberg, Viscarra Rossel, Mouazen, and Wetterlind (2010) showed a positive approximately linear relationship between RMSE and SD for SOC predictions, demonstrating that RMSE is also affected by data variability. Moreover, they also noted that RMSE for SOC may be dependent on the clay contents.

Additionally, a number of studies have focused on optimizing several chemometric factors to achieve high RPIQ and low RMSE values, including the algorithms applied, mathematical pretreatments, and the use of specific wavenumbers or wavenumber regions (Dotto, Dalmolin, Grunwald, ten Caten, & Filho, 2017; Ludwig, Murugan, Parama, & Vohland, 2019). Unfortunately, results have been variable for different properties and in the different studies. Regarding the algorithm, it is well established in spectroscopy that the linear approach PLSR is superior to multiple linear regression, since it appropriately handles the multicollinearity of spectral data (Wehrens, 2011). Several authors suggested that nonlinear approaches, such as support vector machine regression (SVMR) with a radial kernel, may be more appropriate to analyze spectral data, especially for datasets containing a high level of nonlinearities (Xu, Lu, Baldea, Edgar, & Nixon., 2018). For instance, Deiss, Margenot, Culman, and Demyan (2020) reported in a MIR study that SVMR outperformed PLSR for estimations of clay, sand, pH, SOC, and permanganate oxidizable C in calibration and validation samples and emphasized that tuning of the hyperparameters is important. In a NIR study on potentially toxic elements in soils, it was suggested that SVMR is the best solution for handling the calibration (Gholizadeh et al., 2015). In contrast, Nawar, Buddenbaum, Hill, Kozak, and Mouazen

Core Ideas

- Multiple partitioning of the data is essential in infrared studies.
- The interquartile range (IQR) of measured data was a key parameter affecting the evaluation.
- A sufficiently high IQR resulted in generally good estimation accuracies in this field study.
- SVMR was slightly superior to PLSR for the estimation of SOC, clay, and sand.
- Optimum pretreatments and wavenumber region selection were useful for SOC and pH estimations.

(2016) reported, also for a NIR study, cases where PLSR outperformed SVMR in the validation. Explanations in cases of superior performances of PLSR over SVMR may refer to a tendency of SVMR to produce overfitted models (Grunwald, Yu, & Xiong, 2018). A number of different spectral preprocessing approaches has also been favored by different research groups for use in conjunction with certain algorithms for the estimation of specific soil properties (see, e.g., the overview provided by Bellon-Maurel and McBratney [2011] for SOC estimation).

Several studies indicated that the factors given above might need to be studied in the context of the variability present in the sample. For instance, Clingensmith, Grunwald, and Wani (2019) found application of subsetting strategies to increase the variability of soil parameters and/or spectra in the calibration dataset had a greater influence on model performance than the algorithm applied. For the field of soil C mapping, Somarathna, Minasny, and Malone (2017) summarized that their results showed that the accuracy of spatial prediction of soil C was more sensitive to training sample size compared to the model type used. Recently, IR results on the relationship between variability and optimal algorithm showed that the usefulness of SVMR over PLSR generally decreased with decreasing sample size used for the calibration (thus decreasing the information provided), and PLSR partly outperformed SVMR in the validation (Ludwig et al., 2019). In a visNIR study estimating total C with 216 soils from an Italian forested area, the required minimum number of soils in the calibration sample was 72 for SVMR and 130 for PLSR in order to have nonsignificant differences in the validation sample $RMSE_V$ compared with the best model (SVMR using the full calibration sample size of 144) (Lucà et al., 2017). For the estimation of clay contents and exchangeable Ca^{2+}, Ramirez-Lopez et al. (2014) reported that for all the sampling algorithms in both datasets and for calibration sample sizes <200, a general trend was observed in which the training RMSE, the normalized training RMSE, and the $RMSE_V$ decreased considerably as the calibration sample size increased. The generalizability of these findings with respect to the various soil properties, scales (field, regional, or global), and locations, however, remains unresolved.

Several problems and contradictions reported above have also been encountered in scientific fields other than spectroscopy. Cawley and Talbot (2010) emphasized that overfitting and variability are key issues for machine learning algorithms. If algorithms have different numbers of parameters to be optimized, they might perform differently, especially with a limited supply of available data. In order to advance knowledge, Cawley and Talbot (2010) recommended investment in sufficient processor time in order to evaluate performance over a wide range of datasets, using multiple randomized partitions of the available data, with model selection performed separately in each trial.

Objectives were thus to quantify the effects of range of contents, sample size, spectral preprocessing, wavenumber region selection, and algorithms on the accuracies of IR estimates in the wavenumber range from 1,000 to 7,000 cm^{-1} (mid- and long-wave NIR) for SOC, N, pH, clay, and sand using multiple partitions of surface soils from an arable field in Bangalore (India). We hypothesized that a relationship between algorithm performance and sample size exists, with increasing sample size resulting in (a) higher robustness (i.e., lower variability of estimation accuracies) across model partitions of respective algorithms and (b) more pronounced benefits of SVMR over PLSR across the studied soil properties, as reported by Ludwig et al. (2019).

2 | MATERIALS AND METHODS

2.1 | Soils

Soils from an arable field trial in Bangalore were investigated. The soils of this trial were red sandy soils (Kandic Paleustalfs or Dystric Nitisols) derived from granite rock. Two sampling campaigns were carried out. In July 2016, after the establishment of the trial, 144 surface soils (0–5 cm), hereafter referred to as sample S_{144}, were sampled on a 79-m × 41-m grid consisting of 16 rows and nine columns, resulting in average distance of 4.7 m between sample units (Ludwig, Murugan, Parama, & Vohland, 2018, 2019; see map in Moeckel et al., 2018) (Table 1). In December 2016, the sampling plan had to be revised to work around other measurements being collected simultaneously on the experimental field. For this reason, the sampled region of the field expanded on three sides to a grid size of 90 m × 60 m. Sample S_{119} was collected on a grid consisting of 12 rows and 10 columns (resulting in 119 surface soils as one of the 120 soils was lost), with an average distance of 6.7 m between sample units. A random subsample S_{71} was taken from S_{119} in order to study the performance of the algorithms for a small sample. Sample S_{263} consisted of the combined samples S_{144} and S_{119} (Table 1).

2.2 | Laboratory analyses and IR spectroscopy

Soils were air dried, sieved to <2 mm, ground, and stored in plastic vials at room temperature before analysis. Contents of total C (interpreted as SOC because carbonates were absent) and N were analyzed by dry combustion (Elementar Vario El, Heraeus). The pH was determined according to DIN ISO 10390 (2005) with 10 g field-moist soil (i.e., at 50% of its water holding capacity) in 25 ml of 0.01 M $CaCl_2$. Soil texture was determined with the pipet method (DIN ISO 11277, 2002). Table 1 shows descriptive statistics of these properties.

TABLE 1 Descriptive statistics and results of the Shapiro–Wilk tests for the soil organic C (SOC) content, N content, pH, and soil texture of samples containing 71, 119, 144, and 263 soils (S_{71}, S_{119}, S_{144}, and S_{263}, respectively)

Property	Sample	Min.	Max.	Median	Interquartile range	*p* (Shapiro–Wilk)
SOC, g kg^{-1}	S_{263}	3.0	14.4	10.0	1.61	1.3×10^{-9}
	S_{144}	8.0	13.1	10.6	1.09	.12
	S_{119}	3.0	14.4	9.3	1.52	2.3×10^{-6}
	S_{71}	3.3	11.3	9.2	1.41	5.9×10^{-6}
N, g kg^{-1}	S_{263}	0.29	1.2	0.90	0.13	4.5×10^{-10}
	S_{144}	0.72	1.2	0.93	0.083	.17
	S_{119}	0.29	1.2	0.84	0.16	4.8×10^{-4}
	S_{71}	0.33	1.2	0.83	0.15	1.5×10^{-3}
pH	S_{263}	4.75	7.10	5.23	0.58	8.8×10^{-15}
	S_{144}	4.86	6.38	5.11	0.19	1.1×10^{-12}
	S_{119}	4.75	7.10	5.70	0.80	.094
	S_{71}	4.75	6.77	5.64	0.69	.53
Clay, %	S_{263}	7.88	33.69	27.68	5.35	1.2×10^{-11}
	S_{144}	22.61	33.69	29.59	3.34	.027
	S_{119}	7.88	31.18	24.65	5.02	7.3×10^{-7}
	S_{71}	10.42	30.85	24.62	4.87	8.9×10^{-4}
Sand, %	S_{263}	49.60	87.85	58.54	4.19	$<2.2 \times 10^{-16}$
	S_{144}	52.67	64.26	57.53	2.43	.0013
	S_{119}	49.60	87.85	61.02	5.09	1.2×10^{-10}
	S_{71}	54.41	82.58	61.50	3.62	2.6×10^{-7}

Soils were ground to <0.2 mm and stored in a desiccator until the diffuse reflectance IR Fourier transform (DRIFT) measurements. The DRIFT spectra of the soils (~1.5 g) were recorded using a Bruker-TENSOR 27 IR spectrometer with a diffuse-reflectance accessory (Ulbricht-Kugel) in the MIR region from 370 to 4,000 cm^{-1}, as well as the long-wave part of NIR from 4,000 to 7,000 cm^{-1} (1,430–2,500 nm). No KBr was added to the soils. Each spectrum was recorded at ~2-cm^{-1} intervals with 200 scans, which resulted in 3,438 data points per spectrum. For each soil, the average of two measurements was taken and the reflectance spectra were transformed to absorbance spectra [log(1/reflectance)] (Figure 1).

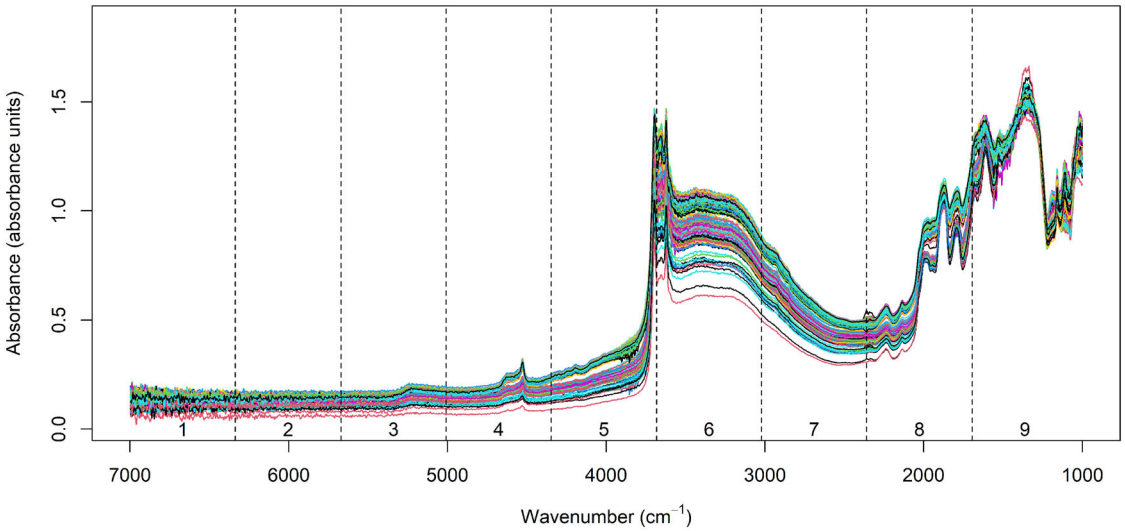

FIGURE 1 Mid-infrared absorbance spectra (4,000–1,000 cm^{-1}) for 263 soils. Parts of the near-infrared region (7,000–4,000 cm^{-1}) are also shown. Dashed lines separate the nine wavenumber regions used for chemometric modeling

2.3 | Preprocessing and wavenumber region selection

We used the statistical software R versions 3.5.3 (PLSR calculations on a PC) and 3.6.0 (SVMR calculations on a LINUX computing cluster) (R Core Team, 2019) for the mathematical treatments of the spectra. The prospectr package (miscellaneous functions for processing and sample selection of diffuse reflectance data; Stevens & Ramirez-Lopez, 2020) was used for spectral preprocessing. The region <1,030 cm^{-1} was removed from the spectra to decrease noise and the remaining region contained 3,094 data points.

For each property, 18,396 PLSR models were calculated in order to test the usefulness of pre-processing and wavenumber region selection and are assigned to variants A (1 regression model), B (510 models), C (35 models), and D (17,850 models):

- Variant A: control variant consisting of all 3,094 data points (i.e., no wavenumber region selection and no preprocessing).
- Variant B: use of optimized wavenumber regions with an automatic selection approach and without preprocessing. The spectra were split into nine wavenumber regions (Ludwig et al., 2019) using R in a similar way as implemented in the Bruker OPUS Quant chemometric software, ranging from 6,998 to 6,336 (Region 1), 6,336 to 5,671 (Region 2), 5,671 to 5,009 (Region 3), 5,009 to 4,346 (Region 4), 4,346 to 3,682 (Region 5), 3,682 to 3,021 (Region 6), 3,021 to 2,359 (Region 7), 2,359 to 1,694 (Region 8), and 1,694 to 1,030 cm^{-1} (Region 9) (Figure 1). The automatic testing of all wavenumber region combinations resulted in 510 (2^9 regions − 2) tested PLS regression models per property for variant B. The implemented automatic selection of spectral ranges has some similarities to interval PLS (iPLS), where the focus is also on use of important spectral regions and the removal of interference from other regions (Zou, Zhao, Povey, Holmes, & Mao, 2010).
- Variant C: use of optimized preprocessing without wavenumber region selection. Preprocessing methods included calculation of moving averages (calculated over 17 data points), resampling (keeping every second data point because of collinearity), and the use of the Savitzky–Golay algorithm for the reduction of noise. The PLSR models calculated with the original absorbance spectra (i.e., without the Savitzky–Golay algorithm) included three variants: use of moving averages, use of moving averages with resampling, and use of resampling without moving averages. The Savitzky–Golay algorithm was applied to the full spectra (with and without calculated moving averages and with or without resampling) as follows: the polynomial (PG) degree was set to 2, the order of derivative (DER) ranged from 1 to 2 (PG–DER: 2–1, 2–2), and the window size for smoothing was set to 5, 11, 17, or 23. Thus, variant C consisted of $3 + 2 \times 2 \times 2 \times 4 = 35$ PLSR models per property.
- Variant D: use of optimized preprocessing and wavenumber region selection. Thus, the combined use of optimized preprocessing (35 PLSR models per property) and optimized wavenumber region selection (2^9 regions − 2) resulted in 17,850 models that were tested per property.

2.4 | Chemometric approaches: PLSR and SVMR

The PLSR models were calculated using the kernel algorithm (Wehrens, 2011) provided in the pls package (partial least squares and principal component regression; Mevik, Wehrens, & Liland, 2019). No outlier elimination was carried out. The maximum number of factors was set to 10 in all cases. Leave-one-out (LOO) cross-validation was used to identify the best PLSR model for each model variant A to D—defined as the model that achieved the highest RPIQ in the cross-validations ($RPIQ_{CV}$)—with an optimum number of latent variables (Table 2). These optimal models were then tested using independent validations. Additional tests using 10-fold cross-validations instead of LOO cross-validations did not result in improved validation accuracies (data not shown).

The SVMR models with a radial kernel were calculated using the caret package (classification and regression training; Kuhn et al., 2019), which calls the kernlab package (Kernel-Based Machine Learning Laboratory; Karatzoglou, Smola, & Hornik, 2018). The cost function C and the smoothing parameter sigma were optimized in the tuning process by grid searches: C was set to 2^n, with n varying from 0 to 15 by a step of 1 and sigma to 2^{-n} with n varying between 25 and 0 by a step of 5. Since the result of each SVMR model depends on its initialization, 100 different initializations were tested for each calculation using the set.seed(i) command, with i varying from 1 to 100. The SVMR models were calculated for variant A (full spectra without pretreatments) and variants B–D with the optimized pretreatments and region selections obtained in the respective PLSR optimizations. Thus, 400 SVMR models were calculated for each property in each partitioning step discussed below. Ten-fold cross-validations were carried out to identify the models which maximized $RPIQ_{CV}$ for each model variant A–D, and these optimal models were then tested using independent validation discussed below.

2.5 | Performance measures

The RPIQ (i.e., IQR of laboratory results divided by the RMSE of cross-validation [$RMSE_{CV}$] or RMSE of validation [$RMSE_V$]) (Bellon-Maurel et al., 2010) was used as a quality

TABLE 2 Parameterization of the partial least squares regression (PLSR) and support vector machine regression (SVMR, with a radial kernel) models resulting from five partitions of the samples for prediction of soil organic C (SOC) content, N content, pH, and soil texture for variants A (original full spectrum absorption data) and D (optimal preprocessing approach and automatic selection of wavenumber ranges) for the samples containing 71, 119, 144, and 263 soils (S_{71}, S_{119}, S_{144}, and S_{263}, respectively)

					PLSR	SVMR	
Property	Sample	Variant	Approach and math treatment[a]	Wavenumber regions[b]	Factors	Cost	Sigma
SOC	S_{263}	(A)	*n*–*n*–0–0–0	1–9	7–9	8/12/13	$2^{-15}/2^{-20}$
		(D)	17–*n*/*y*–0/2–0/1–0/23	4, 5, 7, 8	10	8/10/13	2^{-15}
	S_{144}	(A)	*n*–*n*–0–0–0	1–9	6–10	5/10/12	$2^{-15}/2^{-20}$
		(D)	17–*n*–0–0–0	3–5, 7–9	9–10	13/14/15	2^{-20}
	S_{119}	(A)	*n*–*n*–0–0–0	1–9	6–10	10/11/15	$2^{-20}/2^{-25}$
		(D)	*n*/17–*y*–0/2–0/1–0/17/23	4, 5, 7, 8	10	7/8/11/13/15	$2^{-15}/2^{-20}$
	S_{71}	(A)	*n*–*n*–0–0–0	1–9	7–9	11/15	2^{-25}
		(D)	*n*/17–*n*/*y*–2–1–17/23	2, 4–8	4–10	11/12/15	$2^{-20}/2^{-25}$
N	S_{263}	(A)	*n*–*n*–0–0–0	1–9	7–8	7/9/12	$2^{-15}/2^{-20}$
		(D)	*n*–*y*–0/2–0/1–0/23	1–9	7–10	7/9/12	$2^{-15}/2^{-20}$
	S_{144}	(A)	*n*–*n*–0–0–0	1–9	6–10	10/11/12	2^{-20}
		(D)	*n*–*y*–0–0–0	2–9	9–10	9/13/14	$2^{-15}/2^{-20}$
	S_{119}	(A)	*n*–*n*–0–0–0	1–9	6–10	11/12/13/15	$2^{-20}/2^{-25}$
		(D)	*n*/17–*n*/*y*–0/2–0/1–0/5/17/23	1, 3–5, 7–9	5–10	7/11/14/15	$2^{-15}/2^{-20}/2^{-25}$
	S_{71}	(A)	*n*–*n*–0–0–0	1–9	6–10	11/15	$2^{-20}/2^{-25}$
		(D)	*n*/17–*n*/*y*–0/2–0/1–0/11/23	2–5, 7–9	5–10	10/11/13/14	2^{-20}
pH	S_{263}	(A)	*n*–*n*–0–0–0	1–9	8–10	6/14/15	$2^{-15}/2^{-25}$
		(D)	17–*y*–2–1–23	4, 5, 7–9	8–10	7/8/11/13	$2^{-10}/2^{-15}$
	S_{144}	(A)	*n*–*n*–0–0–0	1–9	7–10	10/12/15	$2^{-20}/2^{-25}$
		(D)	*n*/17–*n*/*y*–0–0–0	1–9	7–9	12/13/14	2^{-20}
	S_{119}	(A)	*n*–*n*–0–0–0	1–9	10	13	2^{-20}
		(D)	*n*/17–*n*/*y*–0–0–0	7,8	9–10	14/15	2^{-15}
	S_{71}	(A)	*n*–*n*–0–0–0	1–9	8–10	11/12/13	2^{-20}
		(D)	17–*n*/*y*–0/2–0/1–0/23	2–4, 6–9	9–10	11/13/15	$2^{-15}/2^{-20}$
Clay	S_{263}	(A)	*n*–*n*–0–0–0	1–9	7–10	5/8	2^{-15}
		(D)	*n*/17–*n*/*y*–0/2–0/1–0/5/17	1, 3–8	5–10	3/4/5/9/12	$2^{-15}/2^{-20}$
	S_{144}	(A)	*n*–*n*–0–0–0	1–9	7–10	11/12/15	$2^{-20}/2^{-25}$
		(D)	*n*–*y*–0–0–0	1–8	6–10	7/8/12/13	$2^{-15}/2^{-20}$
	S_{119}	(A)	*n*–*n*–0–0–0	1–9	1–9	5/7/10/15	$2^{-15}/2^{-20}/2^{-25}$
		(D)	*n*/17–*n*/*y*–2–1/2–5/11/17/23	1–9	7–10	5/6/11/15	$2^{-15}/2^{-20}/2^{-25}$
	S_{71}	(A)	*n*–*n*–0–0–0	1–9	1–5	6/11/13/14	$2^{-15}/2^{-25}$
		(D)	*n*/17–*n*/*y*–0/2–0/1/2–0/11/17/23	1–9	4–10	4/5/9/14/15	$2^{-10}/2^{-15}/2^{-20}/2^{-25}$
Sand	S_{263}	(A)	*n*–*n*–0–0–0	1–9	5–8	4/5/6/9/10	$2^{-15}/2^{-20}$
		(D)	17–*n*/*y*–0/2–0/1–0/11/17	3–8	5–10	7/9/11/13	$2^{-15}/2^{-20}$
	S_{144}	(A)	*n*–*n*–0–0–0	1–9	4–6	4/6/14	$2^{-15}/2^{-25}$
		(D)	*n*/17–*n*/*y*–0/2–0/1–0/17	2–9	5–10	3/4/6/12/15	$2^{-15}/2^{-20}/2^{-25}$
	S_{119}	(A)	*n*–*n*–0–0–0	1–9	1–2	5/6/15	$2^{-15}/2^{-20}/2^{-25}$
		(D)	*n*/17–*n*/*y*–0/2–0/1/2–0/5/17	2–8	3–10	6/12/13/15	$2^{-15}/2^{-20}/2^{-25}$
	S_{71}	(A)	*n*–*n*–0–0–0	1–9	2–10	7/8/9/10/11	$2^{-20}/2^{-25}$
		(D)	*n*/17–*n*/*y*–0/2–0/1–0/5/11	1, 4–9	4–10	5/6/11/13	$2^{-15}/2^{-20}/2^{-25}$

[a]No use of moving averages (*n*) or averaging over 17 data points (17)–no resampling (*n*) or resampling (*y*) –polynomial degree–derivative–smoothing window size. Different optimal pretreatments for models produced from the five partitions of each sample are indicated by slashes.

[b]Wavenumber regions (defined in Section 2) were selected in at least one of the models produced from the five partitions of each sample.

TABLE 3 Performance measures (mean of the five partitions of the samples) for partial least squares regression (PLSR) and support vector machine regression (SVMR) models with a radial kernel for the variants A (original full spectrum absorption data) and D (optimal preprocessing approach and automatic selection of wavenumber ranges)

			PLSR				SVMR			
Property	**Sample**	**Variant**	**$RMSE_{CV}$**	**$RPIQ_{CV}$**	**$RMSE_V$**	**$RPIQ_V$**	**$RMSE_{CV}$**	**$RPIQ_{CV}$**	**$RMSE_V$**	**$RPIQ_V$**
SOC, g kg^{-1}	S_{263}	(A)	0.43	3.8	0.40	3.7	0.56	2.9	0.57	2.7
		(D)	0.38	4.4	0.37	3.9	0.34	4.6	0.37	4.6
	S_{144}	(A)	0.44	2.3	0.49	2.3	0.45	2.4	0.52	2.1
		(D)	0.38	2.7	0.44	2.6	0.39	2.7	0.41	2.7
	S_{119}	(A)	0.38	3.8	0.41	3.9	0.46	3.4	0.50	2.9
		(D)	0.22	6.6	0.28	5.7	0.25	6.2	0.28	5.3
	S_{71}	(A)	0.46	2.9	0.44	3.6	0.51	2.5	0.58	2.8
		(D)	0.20	7.0	0.27	5.9	0.32	4.1	0.31	5.3
N, g kg^{-1}	S_{263}	(A)	0.05	2.7	0.05	2.5	0.06	2.2	0.06	2.1
		(D)	0.05	2.8	0.06	2.3	0.05	2.6	0.06	2.3
	S_{144}	(A)	0.03	2.4	0.04	2.4	0.04	2.2	0.04	2.0
		(D)	0.03	2.5	0.04	2.4	0.03	2.4	0.04	2.3
	S_{119}	(A)	0.06	2.4	0.06	2.5	0.06	2.7	0.07	2.2
		(D)	0.06	2.7	0.07	2.2	0.05	3.0	0.06	2.4
	S_{71}	(A)	0.07	2.1	0.07	1.8	0.06	2.3	0.08	1.8
		(D)	0.05	2.9	0.07	1.8	0.04	2.9	0.07	2.0
pH	S_{263}	(A)	0.30	2.0	0.29	2.0	0.34	1.8	0.34	1.7
		(D)	0.22	2.7	0.23	2.5	0.18	3.3	0.22	2.4
	S_{144}	(A)	0.14	1.4	0.15	1.3	0.13	1.5	0.15	1.4
		(D)	0.13	1.5	0.15	1.3	0.12	1.5	0.13	1.6
	S_{119}	(A)	0.42	1.9	0.36	2.3	0.42	1.8	0.47	1.7
		(D)	0.22	3.7	0.22	3.9	0.26	2.8	0.27	3.2
	S_{71}	(A)	0.39	1.8	0.37	1.5	0.38	1.7	0.45	1.3
		(D)	0.22	3.1	0.30	1.9	0.24	2.8	0.47	1.3
Clay, %	S_{263}	(A)	1.6	3.4	1.6	3.2	1.6	3.4	1.6	3.3
		(D)	1.5	3.7	1.8	2.9	1.5	3.5	1.6	3.2
	S_{144}	(A)	1.2	2.7	1.3	2.6	1.2	3.1	1.3	2.2
		(D)	1.2	2.8	1.3	2.5	1.2	3.1	1.3	2.3
	S_{119}	(A)	1.6	3.0	1.9	2.8	1.6	3.3	1.8	2.8
		(D)	1.4	3.5	2.3	2.2	1.7	3.0	1.9	2.6
	S_{71}	(A)	1.5	3.3	1.5	2.6	1.4	3.0	1.5	3.0
		(D)	1.2	4.2	1.7	2.3	1.4	3.0	1.8	2.6
Sand, %	S_{263}	(A)	1.7	2.5	1.8	2.3	1.7	2.4	1.9	2.2
		(D)	1.5	2.8	1.7	2.4	1.6	2.6	1.6	2.8
	S_{144}	(A)	1.1	2.2	1.1	2.1	1.0	2.5	1.1	2.0
		(D)	1.0	2.5	1.2	2.1	1.0	2.6	1.1	2.0
	S_{119}	(A)	2.1	2.2	2.4	2.2	1.9	2.7	2.6	1.9
		(D)	1.9	2.4	2.7	1.9	1.9	2.5	2.7	1.8
	S_{71}	(A)	1.5	2.9	1.8	1.7	1.4	2.2	1.8	2.8
		(D)	1.3	3.5	2.0	1.6	1.4	2.2	2.0	2.5

SOC, soil organic C; $RMSE_{CV}$ and $RMSE_V$, RMSE of cross-validation and prediction; $RPIQ_{CV}$ and $RPIQ_V$, ratios of the interquartile range to $RMSE_{CV}$ and $RMSE_V$. Units are given in the first column.

parameter for the cross-validations ($RPIQ_{CV}$) and validations ($RPIQ_V$). Equations for the RMSE and RPIQ are

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}{N}} \quad (1)$$

$$\text{RPIQ} = \frac{\text{IQR}_y}{\text{RMSE}}, \quad (2)$$

where $\hat{y}$ is the estimated value, y is the measured value, N is the sample size, and IQR_y is the IQR for the measured y values.

The ranking system by Chang, Laird, Mausbach, and Hurburgh (2001), which was developed for the ratio of performance to deviation (RPD) values, is used for the ranking of RPIQ values by additionally considering that the IQR of a normal distribution equals 1.34896 × SD (Ludwig et al., 2019). Thus, the threshold for an unsuccessful estimation is RPIQ < 1.89 (RPD < 1.4) and good estimations have RPIQ ≥ 2.70 (RPD ≥ 2.0). However, it has to be noted that the thresholds proposed by Chang et al. (2001) were not based on any theory or experiment (for a normally distributed variable and large sample size, RPD < 1.4 corresponds to R^2 < .5), and the usefulness of a model is always defined in its specific context.

2.6 | Multiple partitioning into random calibration and validation samples

Since multiple partitioning of the data is essential in studies which focus on estimation accuracies (Cawley & Talbot, 2010), we created five different random partitions of the data into calibration (2/3 of the respective samples S_{71} to S_{263}) and validation (1/3 of the samples) using the set.seed() command in R with different seed numbers. The optimal mathematical pretreatment, wavenumber regions, and number of factors for the PLSR models as well as the optimal values for cost and sigma for the SVMR models thus depended on the random split. Parameterizations of the PLSR and SVMR models in variants A and D are shown in Table 2.

3 | RESULTS AND DISCUSSION

3.1 | Spectra and wavenumber regions

The spectra of the entire sample S_{263} (263 soils), which covers the MIR region (4,000–1,000 cm^{-1}) and also the longwave part of the NIR region (7,000–4,000 cm^{-1}), are shown in Figure 1. The spectra were separated into nine regions, which were used in the chemometric modeling. Regions 2–9 (6,336–1,030 cm^{-1}) contain—besides soil-matrix-related information—some organic matter-related spectral information (Soriano-Disla et al., 2014; Viscarra Rossel, Walvoort, McBratney, Janikc, & Skjemstad, 2006). Important spectral information for organic matter in the visNIR range include peaks around 6,250 and 5,882 cm^{-1} (Region 2), 5,556 cm^{-1} (Region 3), 5,000 cm^{-1}, and the area from 4,167 to 4,545 cm^{-1} (Regions 4 and 5) (Soriano-Disla et al., 2014). Absorbance in Region 6 (3,682–3,021 cm^{-1}) around 3,410–3,300 cm^{-1} is associated with OH in water (Johnston & Aochi, 1996), but also with phenols, alcohols, acids, hydrochinons, and inorganic hydroxides. Overlapping with that region, O–H, N–H, and C–H stretching results in various peaks in the 3,500- to 3,000-cm^{-1} region (Baes & Bloom, 1989). Regions 7 (3,021–2,359 cm^{-1}), 8 (2,359–1,694 cm^{-1}), and 9 (1,694–1,030 cm^{-1}) are important because of aliphatic CH stretching (2,930–2,850 cm^{-1}), vibrations related to carboxylic groups (1,720 cm^{-1}), protein amide (1,670 and 1,530 cm^{-1}), associated water (1,630 cm^{-1}), carboxylate groups (1,600 and 1,400 cm^{-1}), and aromatic groups (1,600–1,570 cm^{-1}) (Baes & Bloom, 1989; Senesi, D'Orazio, & Ricca, 2003; Skoog and Leary,1992; Soriano-Disla et al., 2014). The same importance of the Regions 2 to 9 can be expected for the N contents, not only because of the existence of a large number of C-N and N-H vibrations in these regions (Soriano-Disla et al., 2014), but also due to the high Spearman rank correlation coefficients (r_s) between SOC and N ranging from .90 to .94 for the samples S_{71} to S_{263}.

For the prediction of clay contents, Regions 4–6 (5,009–3,021 cm^{-1}) and 9 (1,694–1,030 cm^{-1}) may be especially important because of the vibrations associated with kaolinite (4,545 and 3,690–3,620 cm^{-1}), and smectite and illite (4,545, 4,274, 4,090, 3,630–3,620, 3,400–3,300, and 1,630 cm^{-1}; Kuligiewicz, Derkowski, Szczerba, Gionis, & Chryssikos, 2015; Soriano-Disla et al., 2014). For the prediction of sand contents, Regions 7–9 (3,021–1,030 cm^{-1}) may be important because of the vibrations associated with quartz (2,500–1,500 and 1,100–1,000 cm^{-1}, Soriano-Disla et al., 2014; Wood, 1960). Moreover, Region 6 (3,682–3,021 cm^{-1}) provides information on sand, since it contains the signature for kaolinite which often forms sand-sized grains or particles and is not associated with quartz.

In the absence of carbonates, pH prediction may be indirectly estimated based on the presence of proton-rich clays, Al oxyhydroxide minerals and sulfides, oxidizable ammonium and organic N as amides, and carboxylic acids and phenols (Leenen, Welp, Gebbers, & Pätzold, 2019; Soriano-Disla et al., 2014); therefore, many of the abovementioned regions could be useful for pH prediction.

All regions (1–9) were used in at least one of the optimal models with wavenumber region selection (variants B and D) for each soil parameter, indicating the broad spread of useful information for direct or indirect estimation of soil parameters across the spectral region from 7,000 to 1,030 cm^{-1}.

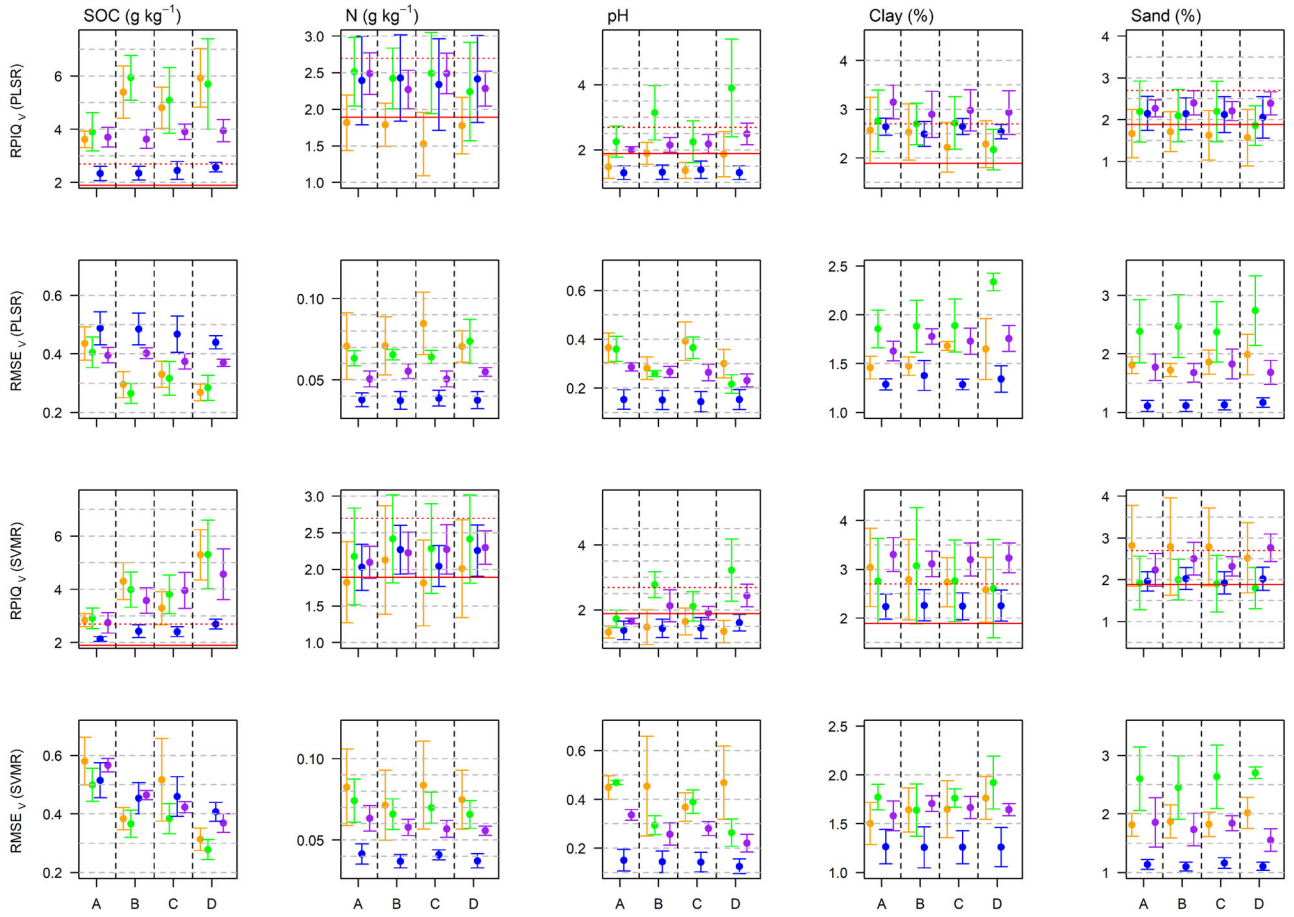

FIGURE 2 Ratio of the interquartile range to the RMSE of validation ($RPIQ_V$) and RMSE of validation ($RMSE_V$) (units are given at the top) values for soil organic C (SOC), N, pH, clay and sand for different sample sizes (orange: 71; green: 119; blue: 144; purple: 263) and mathematical treatments and wavenumber regions (A: no pretreatment, all regions; B: no pretreatment, optimum regions; C: optimum pretreatment, all regions; D: optimum pretreatment and regions). Means and standard deviations for five different partitions are shown. Unsuccessful estimation accuracies ($RPIQ_V < 1.89$) are below the solid red line. Good estimation accuracies ($RPIQ_V \geq 2.70$) are on and above the dotted red line. PLSR, partial least squares regression; SVMR, support vector machine regression

3.2 | Effects of mathematical pretreatments and selection of wavenumber regions for the multiple partitions

The accuracies of validations (as indicated by $RPIQ_V$) slightly decreased compared with those of the cross-validations for the soil properties of all samples S_{71} to S_{263} in all variants A to D. Mean decreases of RPIQ ($RPIQ_{CV} - RPIQ_V$) were 0.4 for PLSR and 0.3 for SVMR in variants A and D (Table 3).

Mathematical pretreatments and wavenumber region selections had pronounced effects on the estimation accuracy of SOC contents using PLSR for small sample sizes and using SVMR for all sample sizes. The $RPIQ_V$ values (means and standard deviations from the five partitions) of the PLSR models increased for small sample sizes from 3.6 ± 0.3 (S_{71}) and 3.9 ± 0.7 (S_{119}) in variant A (no pretreatment, full spectrum) to 5.4 ± 1.0 (S_{71}) and 5.9 ± 0.8 (S_{119}) in variant B (no-pretreatment, optimum region) (Figure 2). Variant B and C (optimum pretreatment, full region) were similarly successful, but there was no combined benefit for variant D (optimum pretreatment and region). For higher sample sizes, the multiple partitions indicated no benefit of the variants B to D over A for the PLSR models. For SVMR, the gain in SOC estimation performance as indicated by increasing $RPIQ_V$ values and decreasing $RMSE_V$ values was not restricted to small sample sizes, and combined benefits of optimizing the mathematical pretreatment and the region selection are evident. However, performance variability between partitions also increased for the combined optimization of pretreatment and wavenumber regions (Figure 2).

For pH, which had either no or only weak Spearman correlations with SOC (data not shown), the benefits of pretreatments and wavenumber region selections were less pronounced: PLSR estimation accuracy improved with both mathematical pretreatments and wavenumber region selection for sample S_{71} (variant A vs. D mean and standard deviation of $RPIQ_V$ of 1.5 ± 0.4 vs. 1.9 ± 0.7), S_{119} (2.3 ± 0.5 vs. 3.9 ± 1.5), and S_{263} (2.0 ± 0.1 vs. 2.5 ± 0.3), whereas SVMR

only notably improved for S_{119} (1.7 ± 0.3 vs. 3.2 ± 1.0) and S_{263} (1.7 ± 0.1 vs. 2.4 ± 0.3) from variant A to D. Combined mathematical pretreatments and wavenumber region selection improved sand estimation accuracy only by SVMR for S_{263} (from 2.2 ± 0.4 to 2.8 ± 0.3 for variant A vs. D), had little effect on N estimation accuracy with either algorithm (despite close correlations to SOC), and even a negative effect in some cases for clay estimation accuracy (decrease in $RPIQ_V$ from variant A and D by PLSR for S_{119} from 2.8 ± 0.6 to 2.2 ± 0.4 and SVMR for S_{71} from 3.0 ± 0.8 to 2.6 ± 0.7).

Application of spectral pretreatments and wavenumber region selection are known to be especially helpful in improving variable, noisy spectra, enhancing weak spectral signals, and removing nonlinearities (Stenberg et al., 2010). Thus, the lack of a consistent benefit of these measures in the present study may be a result of the already good quality of laboratory (dried, ground) IR spectra and prior removal of the noisy region below 1,000 cm^{-1}. Furthermore, hand selection of regions or wavelengths may result in more robust validations than automatic wavenumber region selection by retaining information of importance for direct spectral estimation and removing spectral information used for indirect estimation.

3.3 | Effects of sample size and algorithms

Relating $RMSE_V$ to median values of each soil parameter in each sample and calculating average values across all optimal models, relative $RMSE_V$ was lowest for the homogeneous S_{144} (relative $RMSE_V$ = 3.5%), intermediate for the combined sample S_{263} (relative $RMSE_V$ = 4.9%), and higher for the smaller and more heterogeneous S_{119} ($RMSE_V$ = 5.9%) and S_{71} ($RMSE_V$ = 6.0%). In contrast with this ranking, average $RPIQ_V$ values across all optimal models show very good average performance for S_{119} ($RPIQ_V$ = 2.86) and S_{263} ($RPIQ_V$ = 2.73), and satisfactory performance for S_{71} ($RPIQ_V$ = 2.52) and S_{144} ($RPIQ_V$ = 2.11). Therefore, we see in this study that differences in performance were related to sample variability, with opposite effects on $RMSE_V$ and $RPIQ_V$ (i.e., higher error but also higher $RPIQ_V$ with increasing IQR of the validation sample [IQR_V]), whereas larger sample size (S_{263}) generally improved both performance measures (Figure 2).

The PLSR and SVMR always resulted in successful validation estimates for clay when the sample size was 144 or greater (Figure 2). However, for sand, the sample size needed to be 263 to obtain successful validations. For this sample size, only one partitioning resulted in an unsuccessful validation in the SVMR (which can be seen on the standard deviation being below the threshold of 1.89 in variant A, Figure 2). Median validation estimation accuracies of the five partitions indicated higher success of SVMR for SOC estimation ($RPIQ_V$ = 4.81) over PLSR in variant D for S_{263} (Figure 3). However, for the other soil properties, differences in $RPIQ_V$ were less between the algorithms (Figure 3). These effects of sample size on $RPIQ_V$ can be attributed to the combined effect of spectral information in the cross-validation and spatial autocorrelation (Guo, Chen, et al., 2017; Guo, Zhao, et al., 2017), since both increase with increasing sample size in this field-scale study.

3.4 | Synopsis: Importance of the studied factors

A plot of $RMSE_V$ against IQR_V shows the importance of the various factors studied, where the colors indicate increasing sample sizes from S_{71} (orange) to S_{263} (purple) and the letters A (no pretreatment, full spectrum) to D (optimized pretreatments and region selections) refer to the variants of data pretreatment and wavenumber region selection. In each plot, the red solid and dotted lines have an intercept of 0 and slopes of 1/1.89 and 1/2.70, respectively, and thus all symbols above the red solid lines refer to unsuccessful estimations ($RPIQ_V$ < 1.89, based on Chang et al. [2001] for RPD with the additional accounting for the relationship between SD and IQR [Ludwig et al., 2019]), and all symbols below the red dotted lines refer to good estimation accuracies ($RPIQ_V \geq 2.70$). The black solid lines refer to the average regression lines of the $RMSE_V$ against the IQR_V without an intercept, and the $RPIQ_V$ values given in subfigures are the reciprocals of the slopes (Figure 4).

Our hypothesis of a pronounced improved accuracy of SVMR over PLSR with increasing sample size holds only to some extent for SOC (for variant D, average $RPIQ_V$ for S_{263} was 3.9 for PLSR vs. 4.6 for SVMR, whereas SVMR was comparable or worse than PLSR for smaller samples) and to some extent for clay (for variant D, average $RPIQ_V$ for S_{263} was 2.9 for PLSR vs. 3.2 for SVMR, whereas for S_{144}, average $RPIQ_V$ was 2.5 for PLSR vs. 2.3 for SVMR), but the findings are not robust (i.e., they vary across the different partitions; Figures 2 and 4).

In contrast with our hypothesis, the key variable affecting the accuracy of the validation estimations for all soil properties in this field-scale study is the IQR_V, which is dependent on the sample size and specific partitioning. For almost all subfigures in Figure 4, there is an IQR_V threshold (black vertical dotted lines) for which all $RPIQ_V$ values are ≥ 2.70. For SOC, all partitions, pretreatments and sample sizes with an $IQR_V \geq 1.3$ g kg^{-1} (PLSR) or 1.8 g kg^{-1} (SVMR) had good estimation accuracies. The IQR_V threshold for SOC in SVMR only in variant D is much smaller than 1.8 g kg^{-1} (Figure 4), which points to the importance of the combined mathematical pretreatment and region selection for SVMR in the SOC modeling.

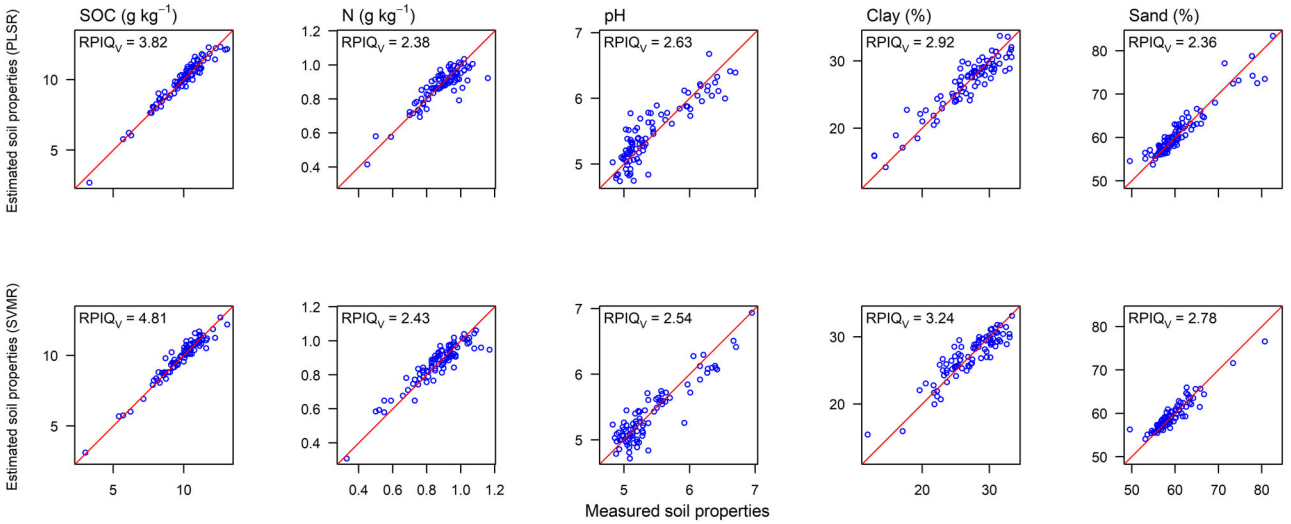

FIGURE 3 Estimated against measured soil properties for sample S_{263} in the variant D (optimum pretreatment and region). Results refer to the respective partition with median accuracy in the partial least squares regression (PLSR) and support vector machine regression (SVMR). The ratio of the interquartile range to RMSE of validation ($RPIQ_V$) values and 1:1 lines are also given

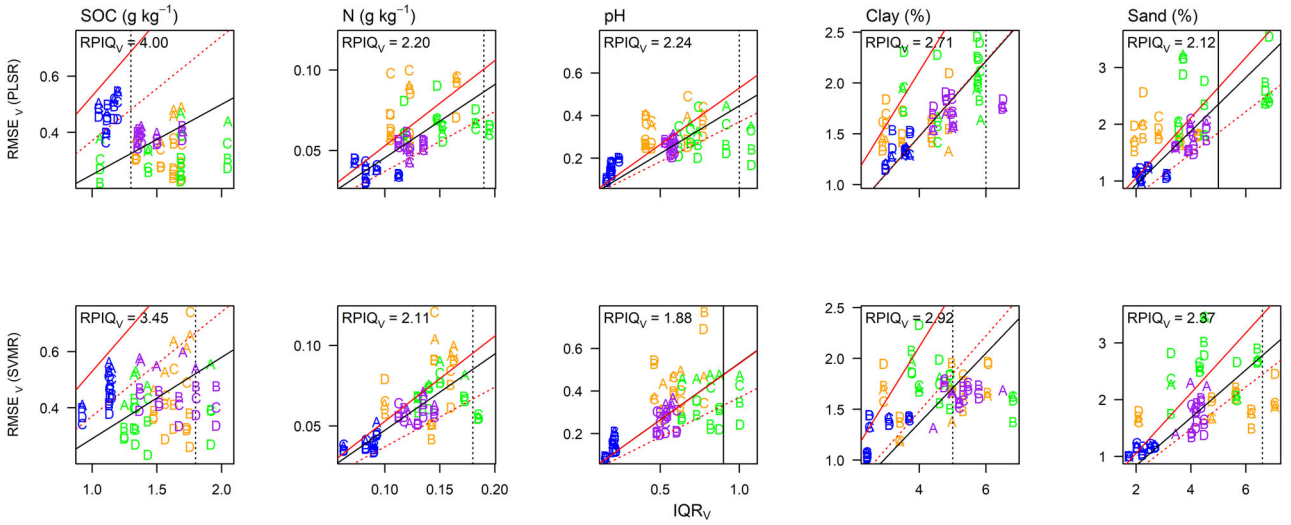

FIGURE 4 The RMSE of validation ($RMSE_V$) values for soil organic C (SOC), N, pH, clay, and sand for different sample sizes (orange: 71; green: 119; blue: 144; purple: 263) and mathematical treatments and wavenumber regions (A: no pretreatment, all regions; B: no pretreatment, optimum regions; C: optimum pretreatment, all regions; D: optimum pretreatment and region) against the interquartile ranges of the validation samples (IQR_V) for partial least squares regression (PLSR) and support vector machine regression (SVMR) models. Values above the red solid red lines refer to unsuccessful validation estimates (the ratio of the interquartile range to RMSE of validation [$RPIQ_V$] < 1.89) and values on and below the red dotted lines refer to good validation estimation accuracies ($RPIQ_V \geq 2.70$). Black solid lines are the average regression lines of the $RMSE_V$ against the interquartile range without an intercept, and the given $RPIQ_V$ values are the reciprocals of the slopes. Vertical lines give interquartile range thresholds above which all $RPIQ_V$ values are satisfactory (>1.89, solid lines) or good (>2.70, dotted lines)

For N, the IQR_V thresholds are 0.19 g kg^{-1} (PLSR) and 0.18 g kg^{-1} (SVMR) for good estimation accuracies. For pH, the IQR_V threshold for good estimation accuracies is 1 for PLSR, whereas for SVMR, only an IQR_V threshold of 0.9 for satisfactory estimation accuracies ($RPIQ_V \geq 1.89$) exists. Support vector machine regression performed better for clay estimations than PLSR as indicated by the smaller IQR_V threshold for good estimation accuracies of 5% in SVMR compared with 6% in PLSR (Figure 4). The same was noted for the sand estimations: for PLSR only a satisfactory IQR_V threshold exists at 5% sand ($RPIQ_V \geq 1.89$), whereas for SVMR, an IQR_V threshold for good estimation accuracies is 6.6% (Figure 4).

Compared with other samples, S_{144} had low $RMSE_V$ and tightly clustered performance across model variants and partitions for all parameters besides SOC (Figure 4), suggesting that a relatively homogeneous sample can produce well-calibrated, consistent models, but performance may be satisfactory to poor according to $RPIQ_V$ due to the narrow IQR_V. The largest sample, S_{263}, also resulted in models with relatively tightly clustered performance across variants and partitions, supporting our hypothesis that larger samples have higher robustness.

It is also notable that of all the factors studied in this experiment (i.e., multiple partitions, sample, algorithm, and pretreatment and wavenumber region selection), the sample caused the largest difference between minimum and maximum $RPIQ_V$ of the factors levels (S_{71}, S_{119}, S_{144}, and S_{263}) for estimation of all soil parameters, whereas the algorithm (PLSR or SVMR) caused the smallest differences in $RPIQ_V$ for all soil parameters except sand (Figure 2). Thus, the random partitions caused more variability in model performance than the algorithm applied for SOC, N, pH, and clay estimation. This supports the findings of Clingensmith et al. (2019) and Somarathna et al. (2017) that the characteristics and division of the sample into training and test sets affect model performance to a greater extent than the algorithm applied.

4 | CONCLUSIONS

Multiple partitioning of the data is essential in IR studies, which focus on estimation accuracies. Considerable variations in estimation accuracies were found between the different partitions and indicated an expected range of accuracies for future sampling campaigns. However, despite the variations, some advantages of SVMR over PLSR could be detected as indicated by smaller IQR_V thresholds for good estimation accuracies for clay and sand, whereas the benefits of SVMR for SOC were dependent on spectral preprocessing and sufficiently large sample size. In this field-scale study, the effects of sample size can be attributed to the combined effect of spectral information in the cross-validation and spatial autocorrelation.

Our field-scale results showed for all five soil properties studied that the IQR_V—which is, among other factors, affected by the sample size and specific partitioning—is a key parameter affecting the estimation accuracy as indicated by $RPIQ_V$. If these results also hold for other field- and larger-scale studies, then $RPIQ_V$ and $RMSE_V$ values for spectrally active soil properties in IR studies should not be interpreted independently, but in the context of the respective IQR_V values.

ACKNOWLEDGMENTS

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.

AUTHOR CONTRIBUTIONS

Bernard Ludwig: Conceptualization; Data curation; Funding acquisition; Investigation; Methodology; Writing-original draft; Writing-review & editing. Isabel Greenberg: Investigation; Validation; Writing-review & editing. Anja Sawallisch: Data curation; Formal analysis. Michael Vohland: Funding acquisition; Methodology; Validation; Writing-review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

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

REFERENCES

Baes, A. U., & Bloom, P. R. (1989). Diffuse reflectance and transmission Fourier transform infrared (DRIFT) spectroscopy of humic and fulvic acids. *Soil Science Society of America Journal*, *53*, 695–700. https://doi.org/10.2136/sssaj1989.03615995005300030008x

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 Research*, *51*, 577–595. https://doi.org/10.1071/SR13077

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends in Analytical Chemistry*, *29*, 1073–1081. https://doi.org/10.1016/j.trac.2010.05.006

Bellon-Maurel, V., & McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils: Critical review and research perspectives. *Soil Biology & Biochemistry*, *43*, 1398–1410. https://doi.org/10.1016/j.soilbio.2011.02.019

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.

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

Clingensmith, C. M., Grunwald, S., & Wani, S. P. (2019). Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment. *European Journal of Soil Science*, *70*, 107–126. https://doi.org/10.1111/ejss.12753

Deiss, L., Margenot, A. J., Culman, S. W., & Demyan, M. S. (2020). Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma*, *365*, 114227. https://doi.org/10.1016/j.geoderma.2020.114227

DIN ISO 10390. (2005). *Bodenbeschaffenheit: Bestimmung des pH-Wertes* (ISO 10390). Beuth.

DIN ISO 11277. (2002). *Bodenbeschaffenheit: Bestimmung der Partikelgrößenverteilung in Mineralböden: Verfahren mittels Siebung und Sedimentation* (ISO 11277: 1998/Cor.1:2002). Beuth.

Dotto, A. C., Dalmolin, R. S. D., Grunwald, S., ten Caten, A., & Filho, W. P. (2017). Two preprocessing techniques to reduce model covariables in soil property predictions by Vis-NIR spectroscopy. *Soil & Tillage Research*, *172*, 59–68. https://doi.org/10.1016/j.still.2017.05.008

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

Gholizadeh, A., Boruvka, L., Vašát, R., Saberioon, M. M., Klement, A., Kratina, J., … Drábek, O. (2015). Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: A case study. *PLOS ONE*, *10*, e0117457. https://doi.org/10.1371/journal.pone.0117457

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

Grunwald, S., Yu, C., & Xiong, X. (2018). Transferability and scalability of soil total carbon prediction models in Florida, USA. *Pedosphere*, *28*, 856–872. https://doi.org/10.1016/S1002-0160(18)60048-7

Guo, L., Chen, Y., Shi, T., Zhao, C., Liu, Y., Wang, S., & Zhang, H. (2017). Exploring the role of the spatial characteristics of visible and near-infrared reflectance in predicting soil organic carbon density. *International Journal of Geo-Information*, *6*, 308. https://doi.org/10.3390/ijgi6100308

Guo, L., Zhao, C., Zhang, H., Chen, Y., Linderman, M., Zhang, Q., & Liu, Y. (2017). Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology. *Geoderma*, *285*, 280–292. https://doi.org/10.1016/j.geoderma.2016.10.010

Hutengs, C., Ludwig, B., Jung, A., Eisele, A., & Vohland, M. (2018). Comparison of portable and bench-top spectrometers for mid-infrared diffuse reflectance measurements of soils. *Sensors*, *18*, 993. https://doi.org/10.3390/s18040993

Hutengs, C., Ludwig, B., Oertel, F., Seidel, M., & Vohland, M. (2019). In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of soil organic carbon. *Geoderma*, *355*, 113900. https://doi.org/10.1016/j.geoderma.2019.113900

Johnston, C. T., & Aochi, Y. O. (1996). Fourier transform infrared and Raman spectroscopy, In D. L. Sparks (Ed.), *Methods of soil analysis, Part 3: Chemical methods* (pp. 269–322). SSSA.

Karatzoglou, A., Smola, A., & Hornik, K. (2018). *Kernlab: Kernel-based machine learning lab*. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/kernlab/kernlab.pdf

Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Hentent, E. J. (2012). Sensing soil properties in the laboratory, in situ, and on-line: A review. *Advances in Agronomy*, *114*, 155–223. https://doi.org/10.1016/B978-0-12-394275-3.00003-1

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A. et al. (2019). caret: Classification and regression training (R package version 6.0-84). Comprehensive R Archive Network. https://CRAN.R-project.org/package=caret

Kuligiewicz, A., Derkowski, A., Szczerba, M., Gionis, V., & Chrysikos, G. D. (2015). Revisiting the infrared spectrum of the water-smectite interface. *Clays and Clay Minerals*, *63*, 15–29. https://doi.org/10.1346/CCMN.2015.0630102

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

Lucà, F., Conforti, M., Castrignanó, A., Matteucci, G., & Buttafuoco, G. (2017). Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. *Geoderma*, *288*, 175–183. https://doi.org/10.1016/j.geoderma.2016.11.015

Ludwig, B., Murugan, R., Parama, V. R. R., & Vohland, M. (2018). Use of different chemometric approaches for an estimation of contents of soil properties in an Indian arable field with near infrared spectroscopy. *Journal of Plant Nutrition and Soil Science*, *181*, 704–713. https://doi.org/10.1002/jpln.201800130

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

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2019). pls: Partial least squares and principal component regression (R package version 2.7-1). Comprehensive R Archive Network. https://CRAN.R-project.org/package=pls

Moeckel, T., Dayananda, S., Nidamanuri, R. R., Nautiyal, S., Hanumaiah, N., Buerkert, A., & Wachendorf, M. (2018). *Remote Sensing*, *10*, 805. https://doi.org/10.3390/rs10050805

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., & Mouazen, A. M. (2016). Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil and Tillage Research*, *155*, 510–522. https://doi.org/10.1016/j.still.2015.07.021

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

O'Rourke, S. M., Minasny, B., Holden, N. M., & McBratney, A. B. (2016). The synergistic use of Vis-NIR, MIR, and XRF spectroscopy for the determination of soil geochemistry. *Soil Science Society of America Journal*, *80*, 888–899. https://doi.org/10.2136/sssaj2015.10.0361

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

Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Demattê, J. A. M., & Scholten, T. (2014). Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma*, *226–227*, 140–150. https://doi.org/10.1016/j.geoderma.2014.02.002

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reeves, J. B. III (2010). Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma*, *158*, 3–14. https://doi.org/10.1016/j.geoderma.2009.04.005

Senesi, N., D'Orazio, V., & Ricca, G. (2003). Humic acids in the first generation of EUROSOILS. *Geoderma*, *116*, 325–344. https://doi.org/10.1016/S0016-7061(03)00107-1

Skoog, D. A., & Leary, J. J. (1992). *Principles of instrumental analysis* (4th ed.). Saunders College Publishing.

Somarathna, P. D. S.N., Minasny, B., & Malone, B. P. (2017). More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. *Soil Science Society of America Journal*, *81*, 1413–1426. https://doi.org/10.2136/sssaj2016.11.0376

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

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, *107*, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7

Stevens, A., & Ramirez-Lopez, L. (2020). *An introduction to the prospectr package. R package Vignette R package version 0.2.0*. Comprehensive R Archive Network. https://cran.rproject.org/web/packages/prospectr/index.html

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janikc, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, *131*, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

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

Wehrens, R. (2011). *Chemometrics with R*. Springer.

Wood, D. L. (1960). Infrared absorption of defects in quartz. *Journal of Physics and Chemistry of Solids*, *13*, 326–336. https://doi.org/10.1016/0022-3697(60)90017-2

Xu, S., Lu, B., Baldea, M., Edgar, T. F., & Nixon, M. (2018). An improved variable selection method for support vector regression in NIR spectral modeling. *Journal of Process Control*, *67*, 83–93. https://doi.org/10.1016/j.jprocont.2017.06.001

Zou, X., Zhao, J., Povey, M. J. W., Holmes, M., & Mao, H. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, *667*, 14–32. https://doi.org/10.1016/j.aca.2010.03.048

How to cite this article: Ludwig B, Greenberg I, Sawallisch A, Vohland M. Diffuse reflectance infrared spectroscopy estimates for soil properties using multiple partitions: Effects of the range of contents, sample size, and algorithms. *Soil Sci Soc Am J*. 2021;*85:*546–559. https://doi.org/10.1002/saj2.20205