Received: 31 March 2020 | Accepted: 12 July 2020

DOI: 10.1002/vzj2.20057

Vadose Zone Journal

ORIGINAL RESEARCH ARTICLES

Estimating coefficient of linear extensibility using Vis–NIR reflectance spectral data: Comparison of model validation approaches

Hafeez Ur Rehman[1] | **Emmanuel Arthur**[1] | **Andrej Tall**[2] | **Maria Knadel**[1]

[1] Dep. of Agroecology, Faculty of Technical Sciences, Aarhus Univ., Blichers Allé 20, PO Box 50, Tjele DK-8830, Denmark

[2] Institute of Hydrology, Slovak Academy of Sciences, Bratislava, Slovakia

Correspondence
Hafeez Ur Rehman, Dep. of Agroecology, Faculty of Technical Sciences, Aarhus Univ., Blichers Allé 20, PO Box 50, DK-8830 Tjele, Denmark.
Email: hafeez@agro.au.dk, hafeezrehman2011@hotmail.com

Funding information
Villum Fonden, Grant/Award Number: 13162; Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic, Grant/Award Number: VEGA: 2/0044/20

Abstract

The coefficient of linear extensibility (COLE) is used to classify soils according to their swell–shrink potential, and its estimation is crucial for engineering and agronomic applications. The aims of the study were (a) to develop a visible–near-infrared spectroscopy (Vis–NIRS, 400–2,500 nm) calibration model to estimate COLE, (b) to compare two model validation approaches (mixed data and country-wise), and (c) to test if a variable selection method improves the estimation accuracy of the calibration models. For this purpose, partial least square regression (PLSR) was used on the spectra of 53 soil samples from Slovakia and 24 samples from the United States. First, a calibration model based on 70% of the entire dataset (including samples from both locations) was developed and validated with the remaining 30% (mixed data approach). Second, a calibration model based on the Slovakian samples was validated with the U.S. samples (country-wise approach). Higher predictability for COLE with standardized root mean square error (SMRSE) of 0.099 was obtained for the mixed data approach than for the country-wise validation with SRMSE of 0.279. Furthermore, using interval PLSR (iPLSR) as a variable selection method did not improve the estimation accuracy of the mixed data approach (SRMSE of 0.099), and rather resulted in a twofold increase in SRMSE (0.560) for the country-wise validation approach. Overall, the good estimation of COLE from Vis–NIRS was attributed to the high correlation of COLE with clay content and spectrally active clay minerals.

1 | INTRODUCTION

The investigation of shrinkage and swelling of soils is crucial because it affects the physical conditions of the soil surface that sometimes create large and deep cracks during the dry season. The coefficient of linear extensibility (COLE) values are used to classify the soils according to their swelling tendencies (Kariuki, Woldai, & Van

Abbreviations: COLE, coefficient of linear extensibility; iPLSR, interval partial least squares regression; OC, organic carbon; OM, organic matter; PLSR, partial least squares regression; RMSECV, root mean square error of cross-validation; RMSEP, root mean square error of prediction; SRMSE, standardized root mean square error; Vis–NIRS, visible–near-infrared spectroscopy.

Der Meer, 2004). In civil engineering, accurate measurements of shrink–swell potential is crucial, as it determines the structural stability of buildings, roads, and bridges (Vaught, Brye, & Miller, 2006). In agricultural applications, the degree of shrink–swell potential is used to estimate structure development in compacted soils (Pillai & McGarry, 1999).

Existing traditional methods to quantify the shrink–swell potential of soils include the rod method, the Georgia volume-change test, and the swell–shrink test (Fityus, Cameron, & Walsh, 2005; Grossman, Brasher, Franzmeier, & Walker, 1968; Simon, Oosterhuis, & Reneau, 1987). There are some practical limitations to these methodologies—for instance, an intact soil core is needed to measure the volume change in wetting and drying conditions, which is expensive to obtain and time consuming for a large number of soils. Pedotransfer functions (PTFs) have also been used as an alternative to direct methods to estimate COLE (Tall, Gomboš, Kandra, & Pavelková, 2017), but existing PTFs are often site dependent, and thus a reliable rapid approach to estimate COLE will be beneficial.

One example of such an approach is the use of diffuse reflectance spectroscopy (e.g., visible [350–700 nm] and near-infrared [700–2,500 nm] spectroscopy, Vis–NIRS) combined with multivariate data analyses techniques. The Vis–NIRS approach has been widely used for the determination of soil properties having direct response in the Vis–NIR range (Stenberg, Viscarra Rossel, Mouazen, & Wetterlind, 2010; Viscarra Rossel, Walvoort, McBratney, Janik, & Skjemstad, 2006). For example, major absorption regions related to clay minerals are located around 1,400, 1,900, and 2,200–2,500 nm and are associated with spectrally active OH groups, which are present within the mineral lattice or as water adsorbed on the mineral surface (Ben-Dor, Patkin, Banin, & Karnieli, 2002). However, Vis–NIRS has been also used to estimate several soil properties with no direct responses in the electromagnetic spectrum through their correlation with other spectrally active soil components (e.g., cation exchange capacity, specific surface area, and Atterberg limits; Rehman et al., 2019; Soriano-Disla, Janik, Viscarra Rossel, Macdonald, & McLaughlin, 2014; Ulusoy, Tekin, Tümsavaş, & Mouazen, 2016). The possibility of estimating COLE using Vis—NIRS can be attributed to its covariation with clay content and mineralogy, moisture content, and organic C (OC). There are several advantages of Vis–NIRS application in soil science compared with existing methods. It is fast, environmentally friendly, nondestructive, highly reproducible, and repeatable. Moreover, one spectrum can be used to simultaneously estimate several properties. As the existing methods for the estimation of COLE values are expensive, slow, and of a low repeatability, using Vis–NIRS for their estimation would be advantageous, especially when a large number of samples is considered.

Core Ideas

- A Vis–NIRS technique in conjunction with PLSR and iPLSR was applied for estimation of COLE.
- A mixed data approach showed better accuracy than country-wise validation.
- Acceptable COLE predictions for the country-wise validation were due to similarities in clay mineralogy.
- The variable selection method did not improve the COLE estimation accuracy.

Partial least square regression (PLSR) is often used to correlate the spectra with measured soil properties. Another alternative to using the full Vis–NIRS spectral information is to apply a variable selection technique such as interval PLSR (iPLSR) to improve calibration models and reduce the wavelength range considered in the model (Hermansen et al., 2017; Knadel et al., 2018; Norgaard et al., 2000). The selection of the calibration and validation datasets is very important when developing predictive models. In order to obtain accurate models, the calibration set should not only cover the variability in soil properties to be determined but also the variation in the spectra, which reflects differences in other spectrally active soil properties. Moreover, the validation set should be independent to avoid the overestimation of the models' performance (Soriano-Disla et al., 2014; Stenberg et al., 2010). The commonly used approach for data partitioning is to proportionally divide the dataset into calibration and validation sets, with two-thirds in the calibration and one-third in the validation set (Stenberg et al., 2010). In this approach, first the entire dataset is mixed (if samples are from more than one location) and then partitioned into the calibration and validation sets, ensuring that the calibration set contains samples with similar soil properties, or spectral properties to the validation set. Another more stringent validation approach is one where samples from one geographic location are used to validate a calibration model based on samples from a different location. In the second approach, the calibration and validation sets can exhibit very different physical and chemical properties due to the differences in parent material or other soil properties. Even if

TABLE 1 General statistics for the calibration (mixed and country-wise samples, $n = 53$) and the independent validation (mixed and country-wise samples, $n = 24$) datasets

Statistics	Data splitting approach	COLE[a]	Clay	Silt	Sand	OC[b]
		cm cm^{-1}	%			
Mean	Mixed data validation	(0.09, 0.07)	(40, 35)	(42, 46)	(17, 19)	(0.84, 1.07)
Max.		(0.23, 0.21)	(76, 64)	(68, 64)	(60, 51)	(1.92, 3.19)
Min.		(0.02, 0.02)	(11, 14)	(21, 30)	(0, 3)	(0.10, 0.11)
SD		(0.06, 0.05)	(16, 14)	(11, 11)	(13, 13)	(0.49, 0.79)
Mean	Country-wise validation	(0.06, 0.13)	(33, 50)	(46, 37)	(20, 12)	(0.90, 0.13)
Max.		(0.23, 0.23)	(69, 76)	(64, 68)	(60, 35)	(3.1, 3)
Min.		(0.02, 0.04)	(11, 13)	(25, 21)	(4, 0)	(0.10, 0.11)
SD		(0.05, 0.04)	(13, 15)	(10, 10)	(14, 11)	(0.60, 0.62)

Note. The first value in parentheses is for the calibration dataset, and the second value in the parentheses is for the validation dataset.

[a] COLE, coefficient of linear extensibility.

[b] OC, organic C.

the soil properties exhibit the same range, their quality can vary substantially and affect soil spectra and model performance (Udelhoven, Emmerling, & Jarmer, 2003). Thus, it is challenging to develop calibration models from one geographic area and validate it with samples from a different geographic area with diverse soils in terms of texture, color, clay mineralogy, and differences in their interactions (Stenberg et al., 2010). It is nonetheless likely that if the basic soil properties (e.g., clay mineralogy) controlling the target property (e.g., COLE) are similar across the different geographic regions, acceptable country-wise validation results can be achieved.

The magnitude of COLE is determined by the cation exchange capacity, specific surface area, soil OC (Smith, Hadas, Dan, & Koyumdjisky, 1985), water content, clay mineralogy (Vaught et al., 2006), and soil texture (Gomboš & Tall, 2012). All these soil properties have been successfully determined by Vis–NIRS because they possess spectrally active components and clear responses in the visible and near-infrared (Vis–NIR) regions (Goetz, Chabrillat, & Lu, 2001; Knadel et al., 2018; Rehman et al., 2019; Viscarra Rossel et al., 2006). Thus, the Vis–NIRS technique has the potential to indirectly estimate COLE from spectral characterization of fine-grained soils, as reported by the very few studies currently available (Hallmark, Morgan, & Hutchison, 2011; Kariuki et al., 2004; Şenol & Akgul, 2012; Waruru, Shepherd, Ndegwa, Kamoni, & Sila, 2014). For instance, Hallmark et al. (2011) used Vis–NIRS (35–2,500 nm) and provided good accuracy for COLE (ranging from 0.001 to 0.240 cm cm^{-1}) with RMSD of 0.03 cm cm^{-1} ($n = 374$). Good prediction for COLE (ranging from 0.03 to 0.27 cm cm^{-1}) was also reported by Waruru et al. (2014) with a RMSE of prediction (RMSEP) of 0.05 cm cm^{-1} for 115 Kenyan samples. Şenol and Akgul (2012) reported good prediction accuracy for 68 Turkish soils with RMSEP of 0.039 cm cm^{-1} for COLE (0.01–0.20 cm cm^{-1}). However, in the previous studies, the Vis–NIRS calibration and validation models were based on samples from the same location. Moreover, none of the studies investigated the potential of combining Vis–NIRS with iPLSR for variable selection for model improvement. Thus, the aims of the study were (a) to develop an optimal Vis–NIRS calibration model to estimate COLE values, (b) to compare different validations approaches (i.e., mixed data partitioning and country-wise), and (c) to test if a variable selection method improves the estimation accuracy of the calibration models. Firstly, the Slovakian and U.S. soil samples were mixed and divided by the mixed data partitioning. Secondly, the calibration model developed on Slovakian samples was validated country wise with U.S. samples, which simulates the prediction of samples not represented in the calibration model.

2 | MATERIALS AND METHODS

2.1 | Soil sampling and reference measurements

The study was carried out using two datasets: 53 undisturbed soil samples from different localities of the East Slovak Lowland in Slovakia, and 24 undisturbed soil samples collected from different locations in Texas, USA. The samples from both geographic regions are dominated by calcium montmorillonite clay mineral. The samples were collected from different depths and are characterized by high textural diversity (Table 1, Supplemental Table S1).

The soil texture analysis of the Slovakian samples was performed by the hydrometer method, whereas the pipette method was used for the U.S. samples (Gee & Or, 2002). The OC was determined on milled subsamples by oxidizing C at 950 °C with an elemental analyzer (Thermo Fisher Scientific).

To estimate the COLE values, the volume change of the undisturbed soils was measured. Kopecky cylinders (100 cm^3) were used to collect undisturbed soil samples. The samples were fully saturated with water, followed by air drying, weighing, and oven drying at 105 °C (Bronswijk & Eversvermeer, 1990). Afterwards, the average values of heights and diameters of the cylinders from clipper (in perpendicular directions) were noted to estimate the volume change. The equation by Grossman et al. (1968) was used to estimate COLE values:

$$\text{COLE} = \sqrt[r_\text{s}]{\frac{V_\text{wet}}{V_\text{dry}}} - 1 \tag{1}$$

where V_wet is soil volume in a saturated state (cm), V_dry is soil volume in oven-dried state (cm), and r_s is the shrinkage geometric function. In this study, we assumed shrinkage was distributed equally in all directions in the soil; hence, $r_\text{s} = 3$ was used.

2.2 | Spectral measurements

For spectral analysis, the air-dried soil samples were ground and sieved to <2 mm. Subsequently, the sieved sample was thoroughly mixed with a spoon, and 50 g was taken from different parts of the sample mixture and placed in a 7-cm sample cup with a 6-cm quartz window at the bottom. We used a Vis–NIR spectrometer, NIRSTM DS2500 (FOSS), which is equipped with silicon (400–1,100 nm) and lead sulfide (1,100–2,500 nm) detectors. Prior to scanning the samples, the spectrometer was calibrated using a white reference (Spectralon). All the samples were scanned on the same day under the same laboratory conditions (average temperature of 21 °C and relative humidity of 59.1%). The spectral reflectance (with a spectral resolution of 8.75-nm full width at half height and a sampling interval of 0.5 nm) was averaged from seven spectra collected at different spots of the sample and was used for further analysis. Thereafter, the recorded reflectance data were converted to absorbance by $A = [\log (1/R)]$, where R is the reflectance.

2.3 | Multivariate regression model

Partial least squares regression using Unscrambler X 10.5.1 software (Camo ASA) was used to correlate the Vis–NIR spectra with the COLE reference values, whereas iPLSR was tested using Matlab software and the PLS toolbox 8.2 (Eigenvector Research). The PLSR is an extension of multiple linear regression, which correlates the dependent variable (e.g., soil property) with the independent variable (spectral measurements). Compared with traditional multiple regression models, the PLSR can handle missing and noisy variables and provides maximum correlation by compressing the spectra to few representative factors between dependent and independent variables (Wold, Sjöstrom, & Eriksson, 2001). Before developing the PLSR calibration models, the dataset was partitioned into calibration and validation subsets using mixed data and country-wise validation approaches described below.

First, the entire dataset was partitioned into calibration (70%) and validation (30%) sets using the Kennard–Stone algorithm (Kennard & Stone, 1969) applied to the first three principle components. The Kennard–Stone algorithm is a sequential function that selects uniform distribution over the objective space (in our case, this was the spectral space). The algorithm selects two extreme samples on Euclidean distance, and the third sample selected is farthest from the two already selected samples. The process continues until the required samples have been selected for the calibration set and the algorithm assures an enhanced coverage of the spectral variability.

For the second approach, the Slovakian samples were assigned to the calibration dataset, and the samples from the United States (Texas) were used as the validation dataset.

2.3.1 | Spectral data preprocessing

Prior to calibration model development, the most widely used preprocessing techniques were tested. Preprocessing techniques such as Savitzky–Golay smoothing, scatter correction, and derivatives were used to remove the noise from the spectra (Rinnan, Berg, & Engelsen, 2009). As the preprocessing techniques had no impact on the performance of the calibration models, we used the raw spectra (absorbance) to develop the calibration models followed by leave-one-out full cross-validation.

2.3.2 | Interval partial least square regression

To evaluate the effect of variable selection compared with the full PLSR, iPLSR regression was tested as a variable selection method. The iPLSR method divides the entire reflectance spectral data into few intervals and then applies the PLSR model on each interval separately (Norgaard et al., 2000). The forward mode iPLSR method was used such that the additional intervals of spectra were consecutively incorporated in the analysis in search of the best combination of variables. The iPLSR identifies important intervals that when used in PLSR could result in better model calibration accuracy compared with using the full spectral data. Since the interval size can affect the accuracy of the calibration model (Hermansen et al., 2017), the following spectral interval sizes (20, 40, 50, 70, and 100) were tested for both the mixed data and country-wise validation approaches. Similar to the PLSR analyses, the robustness of the selected iPLSR calibration model was evaluated by leave-one-out full cross-validation.

2.4 | Evaluation of model performance

The RMSE of cross-validation (RMSECV) and Pearson correlation coefficient (R) were used to evaluate the predictive accuracy of the Vis–NIRS calibration model for COLE, and independent validation of the models was evaluated with the RMSE of prediction (RMSEP), standardized RMSE (SRMSE), and R^2. The RMSE was calculated as follows:

$$\mathrm{RMSE} = \sqrt{\sum \frac{(\mathrm{COLE_p} - \mathrm{COLE_r})^2}{N}} \tag{2}$$

where $\mathrm{COLE_p}$ and $\mathrm{COLE_r}$ are the predicted and reference values of COLE, respectively, and N is the number of samples.

Further, the SRMSE (Equation 3) was used to make a comparison with previous studies, since the RMSE is strongly influenced by the range of the measured data (Rehman et al., 2019):

$$\mathrm{SRMSE} = \frac{\mathrm{RMSE}}{\mathrm{Range}} \tag{3}$$

where “Range” represents the largest to the smallest values of the COLE for the respective dataset or from published studies. Smaller SRMSE values denote more accurate estimations of COLE, and vice versa.

3 | RESULTS

3.1 | Descriptive statistics of soil properties

A summary of laboratory-measured COLE and basic soil properties is provided in Table 1. The average of the clay content and COLE values for the entire dataset were 38% and 0.08 cm cm^{-1}, respectively. The average OC for the Slovakian and U.S. samples was 0.89 and 0.93%, respectively. The mixed data calibration set covered a wider range of clay content that better represented the validation set compared with the country-wise calibration set. Other soil properties in mixed dataset also covered the ranges slightly better than the country-wise approach. Supplemental information on the soil properties for each individual sample is provided in Supplemental Table S1.

3.2 | Cross-validation of partial least square and interval partial squares regression

The performance of the PLSR and iPLSR calibration models for mixed data and country-wise approaches is shown in Figures 1a, 1b, 2a, and 2b. When the full spectrum was used for the analyses (PLSR), the COLE was very well estimated for the calibration dataset of the mixed data approach (R^2 = .84, RMSECV = 0.022 cm cm^{-1}, and SRMSE = 0.108; Figure 1a) and for the country-wise approach (R^2 = .88, RMSECV = 0.016 cm cm^{-1}, and SRMSE = 0.076; Figure 2a). For the iPLSR variable selection method, the five interval sizes tested showed slight differences in estimation accuracy for the country-wise approach, whereas similar results were obtained for the mixed data approach (Table 2). We discarded the models with interval sizes of 20, 40, and 100 for the mixed data approach and 40 and 100 for the country-wise approach due to a significantly higher RMSECV and lower R^2 compared with the other models. The remaining iPLSR models showed only slight differences in model performance. Similar model accuracy was obtained for spectral intervals of 50 and 70 nm equaling 150 and 280 variables for the mixed data approach, using three factors each, and 20- and 50-nm intervals with 120 and 100 variables, using eight and four factors, for the country-wise approach, respectively (Table 2; Figures 1b and 2b). Overall, calibration models for the country-wise approach showed slightly higher accuracy in estimating COLE values than the mixed data approach (Table 2).

Based on the interval size, number of variables, and the calibration performance, we selected models with 50 and 70 intervals for the mixed data approach and country-wise approach, respectively. The selected models had cross-validation performance values of R^2 = .84 and RMSECV = 0.022 cm cm^{-1} for mixed data, and R^2 = .91 and RMSECV = 0.014 cm cm^{-1} for country-wise validation.

3.3 | Regression coefficients of PLSR and iPLSR calibration models

The regression coefficients for the wavelength ranges considered for the two models (PLSR and iPLSR) built on the mixed calibration dataset and country-wise calibration dataset are presented in Figures 1c, 1d, 2c,

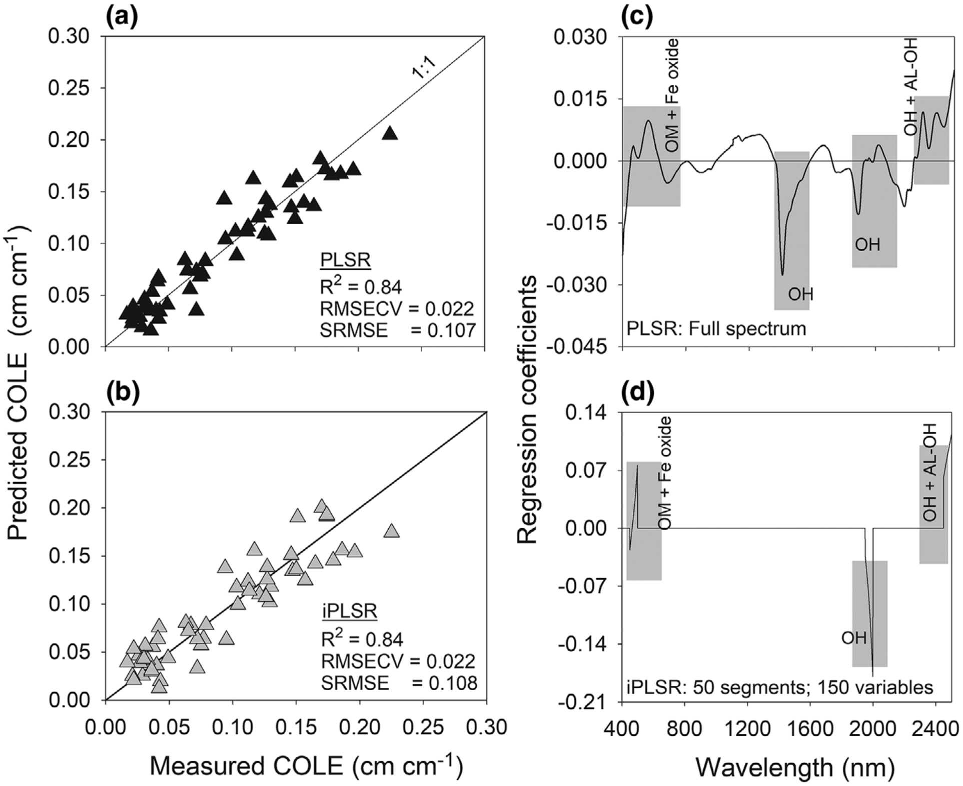

FIGURE 1 Scatterplot of cross-validation for measured vs. visible–near-infrared spectroscopy (Vis–NIRS)-predicted coefficient of linear extensibility (COLE), and regression coefficients of the Vis–NIRS cross-validation model using (a and c, respectively) partial least square regression (PLSR) and (b and d, respectively) interval partial least square regression (iPLSR) for mixed database approach. RMSECV, RMSE of cross-validation; SRMSE, standardized RMSE; RMSEP, RMSE of prediction

and 2d. In Figure 1c for mixed samples, one pronounced peak around 561 nm and a less prominent peak at 479 nm can be observed. Similarly, a pronounced peak around 590 nm and a less prominent peak at 490 nm for country-wise samples can be seen in Figure 2c. For mixed data samples, in the NIR range, prominent peaks around 1,400, 1,890, 2,180, 2,300, and 2,340 nm and weak peaks around 1,130, 1,670, 2,020, and 2,390 nm were observed (Figure 1c). On the other hand, prominent peaks were observed around 830, 1,140, 1,720, 1,900, 2,210, 2,350, and 2,400 nm, and less pronounced peaks were observed at 1,390 and 1,930 nm for the country-wise dataset (Figure 2c). For the iPLSR models, the 150 variables that were considered ranged from 450 to 499, 1,950 to 1,999, and 2,450 to 2,4,99 nm for the mixed calibration dataset, and the 210 variables that were considered ranged from 1,730 to 1,799 and 2,220 to 2,359 nm for the country-wise calibration dataset (Figures 1d and 2d). The mixed calibration dataset showed different peaks around 500, 2,000, and 2,440 nm, and the country-wise calibration dataset showed distinct peaks around 1,800, 2,220, 2,260, 2,290, and 2,360 nm, with smaller peaks around 1,730 and 2,320 nm (Figures 1d and 2d).

3.4 | Independent validation

The Vis–NIRS calibration models developed on mixed and country-wise datasets for COLE were independently validated with the samples from the mixed and country-wise validation approaches (U.S. samples) (Figures 3a and 3b). Independent validation showed very good accuracy for PLSR with R^2 = .88 and SRMSE = 0.099 for the mixed dataset compared with the set from the country-wise validation approach, with R^2 = .64 and SRMSE = 0.279. The estimation errors stems from overestimation of some of the samples with COLE values >0.10 cm cm^{-1} for the country-wise dataset.

On the other hand, iPLSR showed similar accuracy (R^2 = .87 and SRMSE = 0.100) compared with PLSR for mixed samples. The iPLSR had lower estimation accuracy (R^2 = .56 and SRMSE = 0.560) compared with PLSR for country-wise validation (Figures 3a and 3b). The errors for iPLSR in the country-wise approach were from two negative estimates for samples with COLE < 0.10 cm cm^{-1} and overestimations for four samples with higher COLE values (Figure 3b).

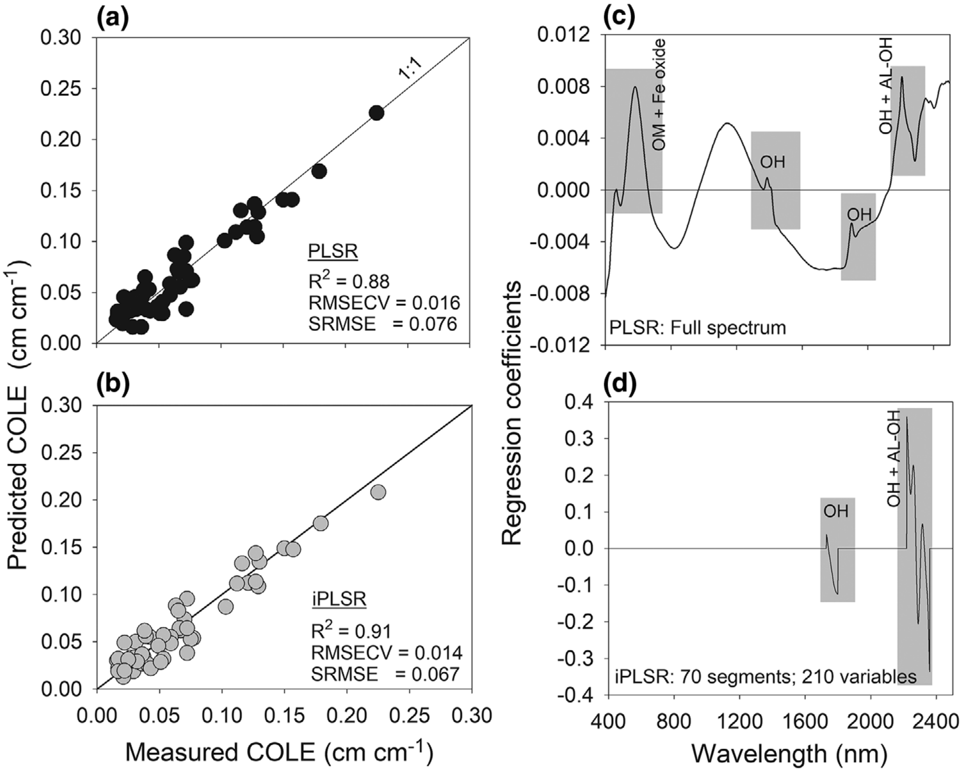

FIGURE 2 Scatterplot of cross-validation for measured vs. visible–near-infrared spectroscopy (Vis–NIRS)-predicted coefficient of linear extensibility (COLE), and regression coefficients of the Vis–NIRS cross-validation model using (a and c, respectively) partial least square regression (PLSR) and (b and d, respectively) interval partial least square regression (iPLSR) for country-wise validation approach. RMSECV, RMSE of cross-validation; SRMSE, standardized RMSE; RMSEP, RMSE of prediction

TABLE 2 Partial least square regression (PLSR) and interval partial least square regression (iPLSR) calibration results for the coefficient of linear extensibility (COLE)

Method	Data splitting approach	Interval size (segments)	No. of variables	Leave-one-out cross-validation		
				R^2	RMSECV[a]	Factors
PLSR	Mixed data validation		2,100	.84	0.022	8
iPLSR		20	60	.79	0.025	3
		40	120	.73	0.028	3
		50	150	.84	0.022	3
		70	280	.84	0.022	3
		100	300	.74	0.028	3
PLSR	Country-wise validation		2,100	.88	0.016	6
iPLSR		20	120	.89	0.015	8
		40	240	.82	0.020	4
		50	100	.90	0.015	4
		70	210	.91	0.014	3
		100	200	.84	0.020	3

[a] RMSECV, RMSE of cross-validation.

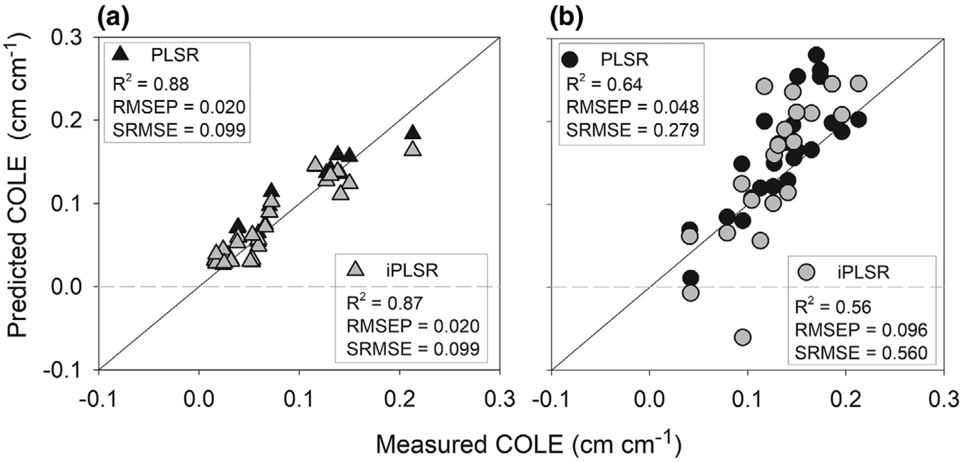

FIGURE 3 Scatterplot of independent validation for measured vs. visible–near-infrared spectroscopy (Vis–NIRS) predicted coefficient of linear extensibility (COLE), using partial least square regression (PLSR) and interval partial least square regression (iPLSR). (a) Mixed data approach and (b) country-wise validation approach. RMSECV, RMSE of cross-validation; SRMSE, standardized RMSE; RMSEP, RMSE of prediction

4 | DISCUSSION

4.1 | Calibration models and regression coefficients

The PLSR and iPLSR models developed by Vis–NIRS for COLE showed good correlations with measured COLE values for mixed data calibration and country-wise calibration datasets. The calibration model developed on mixed data showed slightly lower accuracy than the model developed on the country-wise samples. The slightly higher accuracy for the calibration model of the country-wise dataset could be due to the more homogenous parent material at each location. By using iPLSR, the number of variables decreased from 2,100 to 150 nm for the mixed samples calibration dataset and to 210 nm for country-wise calibration dataset (Table 2). Together with a reduction in the variable number for some segments, the number of factors was also reduced. After the variable reduction, the calibration model showed similar accuracy for mixed samples calibration dataset and slightly better accuracy than PLSR for country-wise samples (Figures 1a, 1b, 2a, and 2b; Table 2). The results showed that the variable selection method did not significantly improve the estimation accuracy; hence, the full spectrum would be recommended in the case of samples from different geographic areas.

The COLE is a soil property that does not have a direct spectral response in the Vis–NIRS range. Therefore, the estimation of COLE is possible only through its covariation with primary soil properties such as clay content and mineralogy, moisture content, and organic matter (OM) that have direct responses in the Vis–NIRs range (Stenberg et al., 2010; Viscarra Rossel et al., 2006). Consequently, the observed peaks in the plot of the regression coefficients and wavelengths are indicative of the primary soil properties that are strongly related to the COLE (Goetz et al., 2001; Kariuki et al., 2004; Viscarra Rossel et al., 2006). For example, the pronounced peaks in PLSR observed around 561 nm for the mixed data samples and around 590 nm for the country-wise samples can be assigned to the chromophore Fe-OOH found in goethite. The peaks in PLSR regression coefficients around 2,180 nm for the mixed data samples and the peak at 2,210 nm for the country-wise samples can be related to OH + Al-OH, overtones of O-H and H-O-H stretch vibrations of free water, carbonates, and clay minerals (Mortimore, Marshall, Almond, Hollins, & Matthews, 2004; Viscarra Rossel & Behrens, 2010). The presence of aromatic C-H bonds may be the reason for the significant absorption around 1,112 nm for mixed sample and 1,140 nm for the country-wise sample regression coefficients (Viscarra Rossel & Behrens, 2010). In both approaches (mixed data and country-wise), the peaks around 1,400, 1,890, 1,900, 2,300, and 2,350 nm in the PLSR regression coefficients and around 1,551–1,600, 2,051–2,100, 1,730–1,799, and 2,220–2,359 nm for iPLSR can be associated with overtones of O-H and H-O-H stretch vibrations of free water, OM, and clay minerals (Knadel, Viscarra Rossel, Deng, Thomsen, & Greve, 2013; Viscarra Rossel & Behrens, 2010). The two negative peaks around 1,410 and 1,890 nm for mixed data samples and the two negative peaks around 490 and 1,900 nm for country-wise samples, observed in the PLSR regression coefficients, can be assigned to aromatic C-H bonds. Additionally, the two peaks at 2,290 and 2,360 nm for iPLSR for the Slovakian sample-based regression coefficients can be related to OM (Stenberg et al., 2010; Viscarra Rossel & Behrens, 2010). The abovementioned wavelengths confirm the importance of spectrally active clay mineralogy, as well as OC, in COLE estimation. To further confirm the covariation of the COLE with the primary properties, we investigated how the measured COLE was related to the clay and OC. A multiple regression analyses comparing COLE with clay and OC as regressors (Equation 4), for all the samples, showed that ~91% of the variation in COLE values could be explained by clay and OC contents:

$$\text{COLE} = -0.0482 + 0.0033\,(\text{Clay}) + 0.0083\,(\text{OC}) \tag{4}$$

where clay and OC are in percentage weight, adjusted $R^2 = .91$, standard error of the estimate (SEE) = 0.017, the number of samples (n) = 77, and $p < .001$. The SEE is the a measure of the accuracy of predictions.

The PLSR calibration models presented here showed better accuracy (based on mixed samples [$R^2 = .84$; RMSECV = 0.022] and country-wise samples [$R^2 = .88$; RMSECV = 0.016]) after cross-validation when compared with earlier studies on COLE. For example, Waruru et al. (2014) used NIRS to develop a calibration model on 136 samples from the Lake Victoria basin in Kenya and reported an R^2 of .70. Similarly, Şenol and Akgul (2012) reported lower accuracy (R^2 of .55 and RMSECV of 0.024) of the calibration model they presented for 68 samples from different regions in Turkey.

4.2 | Independent validation

For the majority of studies involving estimation of soil properties from Vis–NIRS, the samples partitioned into the calibration and validation datasets are often from the same geographic origin or location. Application of such calibration models to estimate the target property from different geographic areas that have soils with variations in texture, color, and clay mineralogy can be challenging (Stenberg et al., 2010).

For the present study, independent validation of the calibration model developed on mixed samples showed very good accuracy (PLSR: SRMSE of 0.099) compared with country-wise validation (PLSR: SRMSE of 0.279, samples from a different geographic region [Texas, USA]) (Figure 3). The iPLSR showed similar estimation accuracy to PLSR for mixed samples, whereas PLSR showed better accuracy than iPLSR for country-wise validation. After mixing the sample, accuracy was improved as expected, as both calibration and validation datasets of the mixed approach may contain similar parent material and clay content range and can provide better accuracy than the country-wise validation dataset (Xu & Goodacre, 2018). The acceptable accuracy from the country-wise validation may be due to the similarity in the range of the measured COLE and the similar clay mineralogy and OC contents of the two countries. The independent validation showed higher accuracy for mixed samples validation and lower estimation accuracy (according to the R^2 values) for country-wise validation than that of Hallmark et al. (2011), who found $R^2 = .67$ for a larger database of samples ($n = 374$) from Texas. However, the range of the COLE was not available to calculate the SRMSE from their study. Lower accuracy than observed in the present study was obtained by Waruru et al. (2014) for Kenyan sample (R^2 of .46 and SRMSE of 0.23), whereas slightly higher accuracy than in our study was obtained by Şenol and Akgul (2012) (R^2 of .50 and SRMSE of 0.20) for Turkish samples. Furthermore, the accuracy obtained from the country-wise validation dataset in this study is comparable with the results obtained in previous studies based on mixed data validation approaches. However, to have better comparison between the mixed data approach and country-wise validation, more data should be included with respect to the number of samples and a wider range in the clay mineralogy and OM contents. Our study shows that potential estimation errors may not necessarily be related to the differences in geographic location of the samples in the calibration and validation datasets. Rather, they could be due to differences in the range and quality of spectrally active properties that affect the COLE values; in our case, clay content range was different in mixed and country-wise calibration datasets.

Despite the slight improvement of calibration model by using iPLSR for the country-wise model, the prediction accuracy of Vis–NIRS for COLE compared with PLSR was reduced. On the other hand, no improvement was noticed for the mixed data calibration model. However, previously, Hermansen et al. (2017) reported improved accuracy using iPLSR for 431 Danish soils to estimate soil texture, and Knadel et al. (2018) reported slightly improved accuracy for 550 soils from different geographic origins for specific surface area. The possible reasons for the acceptable accuracy of the country-wise validation in this study could be (a) the similar range of COLE for the two datasets, and (b) the identical clay mineralogical composition (montmorillonite) of the samples from both locations.

It is important to note that a limitation of the study is the small number of samples (<100) that were used for calibration and validation. However, the wide range of the COLE values in both datasets reduces the uncertainties associated with the small sample number. Applying the developed calibration model to a different geographical areas or samples will be constrained by differences in clay mineralogy between the target area and the sites used for our study.

Thus, the calibration models are applicable to samples rich in montmorillonite.

5 | CONCLUSIONS

In this study, mixed data and country-wise validation approaches were compared to estimate the COLE using Vis–NIRS. Additionally, two modeling methods (PLSR and iPLSR, which represents a variable selection method) were tested. Independent validation of the calibration model with the mixed data approach showed higher accuracy than country-wise validation, which reveals that mixed data calibration and validation sets may contain not only a wide range in clay content and but probably similarities in parent materials. Furthermore, no significant improvement was observed after using the iPLSR for the mixed sample-based calibration model, whereas country-wise validation indicated that PLSR showed slightly better estimation accuracy than iPLSR. Results of iPLSR reveal that in the case of samples from different geographic origins, it is probably better to use the entire spectrum instead of a variable selection method. Validating Vis–NIRS calibration models with a geographically different set of samples than used in the calibration model usually presents a challenge; however, this study shows that similarity in the range of the property being investigated and clay mineralogy in the calibration and validation sets can enable acceptable estimations for samples from different geographical locations. A more systematic study with a larger dataset with a wide range in clay mineralogy and OM is recommended for better comparison between mixed data and country-wise validation approaches.

DATA AVAILABILITY STATEMENT

The authors confirm that the data supporting the findings of this study are available within the article and its supplemental materials.

ACKNOWLEDGMENTS

This research was funded by a research grant (13162) from VILLUM FONDEN and a grant from the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic (Project VEGA 2/0044/20). We also thank Dr. Cristine Morgan, formerly of Texas A&M University, for providing the samples from Texas and the associated data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Hafeez Ur Rehman https://orcid.org/0000-0002-7400-4026
Emmanuel Arthur https://orcid.org/0000-0002-0788-0712
Andrej Tall https://orcid.org/0000-0002-1934-1353
Maria Knadel https://orcid.org/0000-0001-7539-6191

REFERENCES

Ben-Dor, E., Patkin, K., Banin, A., & Karnieli, A. (2002). Mapping of several soil properties using DAIS-7915 hyperspectral scanner data: A case study over clayey soils in Israel. *International Journal of Remote Sensing*, *23*, 1043–1062. https://doi.org/10.1080/01431160010006962

Bronswijk, J. J. B., & Eversvermeer, J. J. (1990). Shrinkage of Dutch clay soil aggregates. *Wageningen Journal of Life Sciences*, *38*, 175–194.

Fityus, S., Cameron, D., & Walsh, P. (2005). The shrink swell test. *Geotechnical Testing Journal*, *28*, 92–101. https://doi.org/10.1520/GTJ12327

Gee, G. W., & Or, D. (2002). Particle-size analysis. In J. H. Dane & G. C. Topp (Eds.), *Methods of soil analysis*: *Part 4. Physical methods* (pp. 255–293). Madison, WI: SSSA. https://doi.org/10.2136/sssabookser5.4.c12

Goetz, A. F. H., Chabrillat, S., & Lu, Z. (2001). Field reflectance spectrometry for detection of swelling clays at construction sites. *Field Analytical Chemistry & Technology*, *5*, 143–155. https://doi.org/10.1002/fact.1015

Gomboš, M., & Tall, A. (2012). Soil clay fraction impact on coefficient of linear extensibility. *Ovidius University Annals*, *2012*(14).

Grossman, R. B., Brasher, B. R., Franzmeier, D. P., & Walker, J. L. (1968). Linear extensibility as calculated from natural-clod bulk density measurements. *Soil Science Society of America Journal*, *32*, 570–573. https://doi.org/10.2136/sssaj1968.03615995003200040041x

Hallmark, C. T., Morgan, C. L., & Hutchison, K. M. (2011). *Characterization of soil shrink-swell potential using the Texas VNIR diffuse reflectance spectroscopy library*. College Station: Texas A&M University.

Hermansen, C., Knadel, M., Moldrup, P., Greve, M. H., Karup, D., & de Jonge, L. W. (2017). Complete soil texture is accurately predicted by visible near-infrared spectroscopy. *Soil Science Society of America Journal*, *81*, 758–769. https://doi.org/10.2136/sssaj2017.02.0066

Kariuki, P. C., Woldai, T., & Van Der Meer, F. (2004). Effectiveness of spectroscopy in identification of swelling indicator clay minerals. *International Journal of Remote Sensing*, *25*, 455–469. https://doi.org/10.1080/0143116031000084314

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*, 137–148. https://doi.org/10.1080/00401706.1969.10490666

Knadel, M., Arthur, E., Weber, P., Moldrup, P., Greve, M. H., Chrysodonta, Z. P., & de Jonge, L. W. (2018). Soil specific surface area determination by visible near-infrared spectroscopy. *Soil Science Society of America Journal*, *82*, 1046–1056. https://doi.org/10.2136/sssaj2018.03.0093

Knadel, M., Viscarra Rossel, R. A., Deng, F., Thomsen, A., & Greve, M. H. (2013). Visible–near infrared spectra as a proxy for topsoil texture and glacial boundaries. *Soil Science Society of America Journal*, *77*, 568–579. https://doi.org/10.2136/sssaj2012.0093

Mortimore, J. L., Marshall, L.-J. R., Almond, M. J., Hollins, P., & Matthews, W. (2004). Analysis of red and yellow ochre samples from Clearwell Caves and Çatalhöyük by vibrational spectroscopy and other techniques. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *60*, 1179–1188. https://doi.org/10.1016/j.saa.2003.08.002

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, *54*, 413–419. https://doi.org/10.1366/0003702001949500

Pillai, U., & McGarry, D. (1999). Structure repair of a compacted vertisol with wet-dry cycles and crops. *Soil Science Society of America Journal*, *63*, 201–210. https://doi.org/10.2136/sssaj1999.03615995006300010029x

Rehman, H. U., Knadel, M., Jonge, L. W.d., Moldrup, P., Greve, M. H., & Arthur, E. (2019). Comparison of cation exchange capacity estimated from Vis–NIR spectral reflectance data and a pedotransfer function. *Vadose Zone Journal*, *18*. https://doi.org/10.2136/vzj2018.10.0192

Rinnan, Å., Berg, F.v.d., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, *28*, 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Şenol, H., & Akgul, M. (2012). Determination of some soil properties with near-infrared reflectance spectroscopy. *Tarim Bilimleri Dergisi*, *18*, 197–213.

Simon, J., Oosterhuis, L., & Reneau, R. B. J. (1987). Comparison of shrink-swell potential of seven Ultisols and one Alfisol using two different COLE techniques. *Soil Science*, *143*, 50–55.

Smith, C. W., Hadas, A., Dan, J., & Koyumdjisky, H. (1985). Shrinkage and Atterberg limits in relation to other properties of principal soil types in Israel. *Geoderma*, *35*, 47–65. https://doi.org/10.1016/0016-7061(85)90055-2

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

Tall, A., Gomboš, M., Kandra, B., & Pavelková, D. (2017). Pedotransfer function for calculating the potential of volume changes in soils. *IOP Conference Series: Earth and Environmental Science*, *92*. https://doi.org/10.1088/1755-1315/92/1/012067

Udelhoven, T., Emmerling, C., & Jarmer, T. (2003). Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study. *Plant and Soil*, *251*, 319–329. https://doi.org/10.1023/A:1023008322682

Ulusoy, Y., Tekin, Y., Tümsavaş, Z., & Mouazen, A. M. (2016). Prediction of soil cation exchange capacity using visible and near infrared spectroscopy. *Biosystems Engineering*, *152*, 79–93. https://doi.org/10.1016/j.biosystemseng.2016.03.005

Vaught, R., Brye, K. R., & Miller, D. (2006). Relationships among coefficient of linear extensibility and clay fractions in expansive, stoney soils. *Soil Science Society of America Journal*, *70*, 1983–1990. https://doi.org/10.2136/sssaj2006.0054

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

Waruru, B. K., Shepherd, K. D., Ndegwa, G. M., Kamoni, P. T., & Sila, A. M. (2014). Rapid estimation of soil engineering properties using diffuse reflectance near infrared spectroscopy. *Biosystems Engineering*, *121*, 177–185. https://doi.org/10.1016/j.biosystemseng.2014.03.003

Wold, S., Sjöstrom, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*, 249–262. https://doi.org/10.1007/s41664-018-0068-2

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Rehman HU, Arthur, E, Tall A, Knadel M. Estimating coefficient of linear extensibility using Vis–NIR reflectance spectral data: Comparison of model validation approaches. *Vadose Zone J*. 2020;19:e20057. https://doi.org/10.1002/vzj2.20057