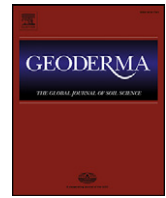




Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization



Nuwan K. Wijewardane^a, Yufeng Ge^{a,*}, Cristine L.S. Morgan^b

^a Department of Biological Systems Engineering, University of Nebraska–Lincoln, Lincoln, NE 68583, USA

^b Department of Soil and Crop Sciences, Texas A&M University, MS2474 TAMU, College Station, TX 77843-2474, USA

ARTICLE INFO

Article history:

Received 11 December 2015

Accepted 14 December 2015

Available online 8 January 2016

Keywords:

Chemometric modeling

External parameter orthogonalization

Soil carbon

Soil moisture

VNIR

ABSTRACT

Moisture is the single most important factor that affects soil reflectance spectra, particularly for field applications. Interest in using soil VNIR spectral libraries, which are commonly based on dry ground soils, to predict soils in the intact field-moist condition (in situ VNIR) is growing. External parameter orthogonalization (EPO) has been proposed as a useful method that links dry ground VNIR models to field moist scans. The goal of this study is to test EPO on a wider set of soil properties and four different modeling techniques, namely, Partial Least Squares Regression (PLS), Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machine (SVM). We selected and scanned 352 archived soil samples from Nebraska, USA, among which 185 samples were used to develop dry ground models and the remaining 167 samples were rewetted to eight different moisture levels for EPO development and testing. Two methods to determine optimum number of EPO components, model-coupled cross validation (Model-Coupled-CV) and Wilk's Λ were also compared. The results showed that EPO minimized the variability of soil spectra induced by moisture. Results suggest a preference for the Wilk's Λ method over Model-Coupled-CV for determining the number of EPO components g , as it produced smoother transformed spectra and more parsimonious models. Among the eight soil properties tested, EPO caused significant improvements for soil Organic Carbon (OC), Inorganic Carbon (IC), and Total Carbon (TC) prediction, marginal improvement for sand and clay, and no improvement for pH, Mehlich-3 Phosphorus, and Cation Exchange Capacity. The failed EPO for the latter three properties is attributable to the poor initial dry-ground models that EPO was built upon. For OC, IC, and TC, EPO coupled effectively with all four modeling methods, with ANN and SVM outperforming the other two slightly. This adds flexibility to the implementation of EPO in predicting field moist soils. As there are increasing demands of spatially-explicit soil data in many disciplines, EPO would be an important essential part for the future in situ VNIR based proximal soil sensing technology.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Visible and near infrared reflectance spectroscopy (VNIR) is widely used as a rapid and cost effective method to quantitatively infer soil properties (Chang et al., 2001; Stenberg, 2010; Viscarra Rossel et al., 2006). Numerous studies have shown VNIR soil spectra to successfully predict a wide array of soil properties including soil carbon (Brown et al., 2006; Minasny et al., 2011; Nocita et al., 2013; Sarkhot et al., 2011), texture (Brown et al., 2006; Sørensen and Dalsgaard, 2005; Waiser et al., 2007), moisture (Ben-Dor et al., 2008; Mouazen et al., 2005; Zhu et al., 2010), and plant macro- and micro-nutrients (Ge et al., 2007; Shepherd and Walsh, 2002).

For a long time, there has been interest in the soil community to develop large soil spectral libraries (Brown et al., 2006; Shepherd and Walsh, 2002). Legacy (or archive) soil samples with partial or full characterization are available all over the world. This makes the establishment of large spectral libraries a feasible task. One great potential of soil spectral libraries is that they can be used for rapid soil characterization in the field.

While conceptually attractive, a primary challenge for using soil spectral libraries in field applications is the wide range of soil moisture that will be encountered. Spectral libraries are most likely constructed from dry and ground soils; but field samples will be in quite different conditions in terms of moisture, small-scale heterogeneity, and temperature. Among these three factors, moisture has the most pronounced effects on VNIR spectra (Brickley and Brown, 2010; Kuang and Mouazen, 2013; Minasny et al., 2009; Sudduth and Hummel, 1993). The decrease of VNIR model performance, when dealing with field moist soil samples, is also documented in the literature (Brickley and Brown, 2010; Minasny et al., 2009). Several authors

* Corresponding author at: Department of Biological Systems Engineering, 209 Chase Hall, East Campus, University of Nebraska–Lincoln, Lincoln, NE 68583, USA.

E-mail addresses: yge2@unl.edu (Y. Ge), cmorgan@ag.tamu.edu (C.L.S. Morgan).

used VNIR spectra of intact soils and demonstrated successful prediction of intact soils (Gomez et al., 2008; Morgan et al., 2009; Waiser et al., 2007). But it is neither practical nor economical to develop both field moist spectral libraries and dry ground libraries due to the large number of field samples needed.

External parameter orthogonalization (EPO) has recently emerged as a promising method to remove (or minimize) the effect of soil moisture on VNIR spectra (Roger et al., 2003). EPO decomposes a spectrum into two components: a useful component that has a direct relationship with the response variable, and a parasitic component that is from an external influence. By removing the parasitic component through orthogonal transformation of spectra, the calibrated spectral model can be less sensitive to the external influence (soil moisture in our case) and more accurate when applied to field moist soils.

The usefulness of EPO in soil spectra was first demonstrated by Minasny et al. (2011). The authors used 391 soil samples collected from New South Wales, Australia and created five different moisture levels under the laboratory condition. Scans of 100 samples were used for EPO development, 271 for model calibration and the remaining 20 for testing. The results showed that EPO successfully removed the moisture effect from soil VNIR spectra and improved the prediction accuracy of soil Organic C (OC). In addition they suggested using more than 100 samples for model calibration and 60 samples for the EPO development.

Ge et al. (2014) expanded the study of EPO to clay content and soil OC. They used both rewetted samples in the lab and field moist scans for EPO development and testing. All samples were collected from Texas, USA; and model calibration was done using Texas Soil Spectral Library comprising more than 2000 samples. They showed that EPO from the field moist scans yielded large improvement in clay prediction. More recently, EPO was also tested for soils in China (Ji et al., 2015) and Brazil (Ackerson et al., 2015), both with positive results.

Previous research tested the validity of EPO in a narrow sense: they focused on the modeling of soil OC and clay, and PLS was the only modeling technique considered. In addition, all previous studies used the PLS-CV (Partial Least Squares—Cross Validation) method for determining the optimal number of EPO components. In Roger et al. (2003), an alternative method, known as Wilk's Λ , was also suggested for this purpose. As there are increasing demands for spatially explicit soil data at high spatial and temporal resolutions in many disciplines, in situ soil sensors based on VNIR will play a very important role to meet these demands at sufficient accuracy and throughput. EPO will become an essential part of the in situ VNIR soil sensing system.

Our long term goal is to use VNIR libraries for in field soil characterization. In this study we want to further test the usefulness of EPO with the following two objectives: (1) test EPO on a wider set of soil attributes beyond soil OC and clay, and (2) test the effectiveness of EPO when coupled with other nonlinear modeling techniques including Artificial Neural Network, Random Forest and Support Vector Machine. With the second objective, we also compare the model-coupled-CV and Wilk's Λ methods to determine the optimal number of EPO components.

2. Material and methods

2.1. Soil samples and dataset description

Soil samples used in this study were selected from the soil archive maintained by Kellogg Soil Survey Laboratory of USDA-NRCS. Samples in the archive are stored in air dried and ground (<2.0 mm) condition. The following criteria were used for sample selection from the archive. First, the samples are originated from Nebraska, USA. Second, the samples had the full characterization of eight chemical and physical properties: OC, Inorganic Carbon (IC), Total Carbon (TC), sand, clay, 1:1 water pH, Mehlich-3 Phosphorus (P), and Cation Exchange Capacity (CEC). Since OC is an emphasis property of this study, the third criterion involved the selection of samples representing the full range of OC in the archive through a 20-stratum stratified random sampling. As a result, a total of

352 samples were selected and used in this study. These samples were further divided into three non-overlapping subsets. They were described as follows, and similar notations as in Roger et al. (2003) are used.

Dry ground set (S^0) – This set consisted of 185 samples to develop dry ground multivariate models for different soil properties. Samples in this set were scanned once in dry ground condition.

EPO development set (S^1) – This set consisted of 100 samples for EPO development. Each sample in this set was scanned nine times: one scan in dry ground condition and eight scans at eight moisture levels (900 scans in total). A detailed description of soil rewetting procedure is in the next section.

Testing set (S^2) – This set consisted of 67 samples for independent EPO validation. The samples in this set were also scanned nine times and in the same moisture condition as S^1 (603 scans in total).

The summary statistics of soil properties in the three different sets are given in Table 1. Median and Interquartile Range (IQR) are presented instead of mean and standard deviation, because all properties exhibited non-normal distribution.

2.2. Rewetting procedure and VNIR scanning

Sample rewetting ($n = 167$) was carried out in 4 batches, where batch 1 had 35 soil samples; batch 2 had 46 soil samples; batch 3 had 46 soil samples; and batch 4 had 40 soil samples for easy handling during the experiment. Approximately 10 g of each sample in batch 1 were placed in petri dishes and scanned in the air dried condition first. A known amount of deionized water was then added to each sample to achieve a moisture content of ~33% (dry basis, gravimetric). Petri dishes were covered with glass lids to avoid moisture loss to the outside and

Table 1

Summary statistics of the soil properties in three different subsets (S^0 , S^1 , and S^2) in this study.

Soil property		S^0	S^1	S^2	
		No. of samples (n)	185	100	67
		No. of scans (N)	185	900	603
Organic C (%)	Min	0.0	0.1	0.0	
	Median	0.8	1.4	1.1	
	Max	6.2	6.5	5.7	
	IQR*	1.2	1.3	1.5	
Inorganic C (%)	Min	0.0	0.0	0.0	
	Median	0.1	0.1	0.1	
	Max	3.8	1.9	3.9	
	IQR	0.3	0.2	0.8	
Total C (%)	Min	0.1	0.1	0.2	
	Median	1.2	1.5	1.5	
	Max	6.4	6.5	6.4	
	IQR	1.3	1.4	1.4	
Sand (%)	Min	0.2	0.5	0.9	
	Median	8.9	5.4	8.2	
	Max	99.2	98.4	96.5	
	IQR	34.8	23.3	28.6	
Clay (%)	Min	0.3	1.3	1.3	
	Median	30.0	28.5	29.3	
	Max	74.0	67.6	48.2	
	IQR	18.9	10.8	11.2	
pH	Min	4.4	4.7	5.2	
	Median	7.8	6.4	7.1	
	Max	10.1	8.4	8.6	
	IQR	0.8	1.5	1.6	
P (mg kg ⁻¹)	Min	0.2	0.0	0.0	
	Median	30.7	24.9	14.7	
	Max	244.2	162.4	195.7	
	IQR	52.7	40.2	42.9	
CEC (cmol + kg ⁻¹)	Min	1.7	1.4	1.9	
	Median	21.8	21.5	21.4	
	Max	51.8	47.1	43.5	
	IQR	12.4	7.9	6.9	

* IQR is Interquartile Range; CEC is Cation Exchange Capacity; and P is Mehlich-3 Phosphorous.

facilitate uniform moisture distribution within samples. After ~24 h, the petri dishes were weighed and then scanned to obtain the first set of moist scans. The lids were then kept open to enhance evaporation of water with continuous monitoring of moisture content. When samples reached the next desired moisture level, petri dishes were again covered with lids for 24 h (for moisture homogenization within samples) followed by scanning and weighing for the second moisture level. This step was repeated to obtain scans at eight different moisture levels. This procedure was applied to other batches with the exception that 15 g of soil were used in subsequent batches to better control moisture evaporation. Care was given to maintain the intended moisture levels to resemble the first batch. The average moisture levels for all the batches were 33, 29, 25, 21, 17, 13, 10 and 8% by weight; and all moisture contents were calculated as dry basis adjusted with the ADOD (Air Dried to Oven Dried weight) values for each sample. For the two rewetting sets (S^1 and S^2), ADOD values ranged from 1.002 to 1.082 with an average of 1.027.

An ASD LabSpec® spectrometer with a mug light (Analytical Spectral Devices, Boulder, Colorado, USA) was used to acquire VNIR reflectance spectra from 350 to 2500 nm (spectral sampling interval of 1 nm). A Spectralon panel (Labsphere Inc., North Sutton, NH, USA) was set up in a petri dish and used as white reference to convert radiometric digital numbers to reflectance. Each spectrum was an average of 100 instantaneous internal scans. The spectral range from 350 to 499 nm was excluded from data analysis because of the low signal to noise ratio of these bands. Spectra were averaged with 5 nm interval to reduce the dimensionality of data for statistical modeling.

2.3. Model calibration/validation and EPO transformation with model-coupled-CV and Wilk's Λ

Fig. 1 shows the flow chart of model calibration and validation in this study. First, the dry ground set (S^0) was used to calibrate a dry ground model (Model A). This model was then tested on the validation set (S^2) at different moisture levels (Prediction A which we expected poor performance). The EPO development set (S^1) was used to develop EPO transformation matrix P . P matrix was then applied to S^0 and moisture-insensitive EPO model B was calibrated. This model was tested on S^{2*} (resulted from the EPO transformation of S^2 and the results (Prediction B) were compared to those of model A.

Data analysis was performed in R statistical environment (R Core Team, 2015) using pls (Mevik et al., 2013), caret (Kuhn et al., 2015), randomForest (Liaw and Wiener, 2002), kernlab (Karatzoglou et al., 2004), nnet (Venables and Ripley, 2002), psych (Revelle, 2015) and ggplot2 (Wickham, 2009) packages. R^2 (Coefficient of Determination), RPIQ (Ratio of Performance to Interquartile Range, Bellon-Maurel et al., 2010) and RMSE_p (Root Mean Squared Error of Prediction) were

used to compare the performance of EPO across different soil properties and modeling techniques. RPIQ was used instead of RPD (Ratio of Prediction to Deviation) because none of the soil properties were normally distributed.

The principle and mathematical implementation of EPO transformation were covered in detail in Roger et al. (2003). For the completeness of the article, a summary is given below. A flow diagram is given in Fig. 2, with the important matrices and parameters being annotated.

EPO assumes that spectra matrix X can be decomposed into two systematic components: a useful component XP and a parasitic component XQ , as indicated in Eq. (1). R is the noise component originated from lack of fitting.

$$X = XP + XQ + R \quad (1)$$

The procedure to find XP is through spectra matrix D , which is the difference between the spectra matrix with and without external influence. Q is estimated through singular value decomposition of D , and XP is then calculated as $X(I - Q)$; I is the identity matrix.

One of the most important parameters to be determined during EPO development is the number of EPO components g (the same notation as in Roger et al. (2003) and Minasny et al. (2011); Ge et al. (2014) used c for the same meaning). Roger et al. (2003) suggested two methods to determine g : (1) Cross validation of PLS calibration on transformed spectra S^{1*} (PLS-CV); and (2) calculating Wilk's Λ of the transformed spectra S^{1*} as:

$$\text{Wilk's } \Lambda = \frac{\text{Trace}(\mathbf{B})}{\text{Trace}(\mathbf{T})} \quad (2)$$

where T is the variance–covariance matrix of the EPO transformed spectra S^{1*} , and B is the variance–covariance matrix of S^{1*} aggregated by sample (i.e., averaging across all moisture levels for each sample).

In PLS-CV, there is optimal coupling between g and PLS latent variable n_{LV} (Roger et al., 2003; Minasny et al., 2011; Ge et al., 2014). This optimal coupling is found by a certain combination of g and n_{LV} that gives minimum RMSE in cross validation. We hypothesized that this optimal coupling might be more important for the nonlinear modeling techniques of RF, ANN, and SVM. Therefore we extend PLS-CV to these three nonlinear modeling techniques by coupling one important tuning parameter for each modeling technique with EPO through cross validation. Instead of PLS-CV, we refer to it as Model-Coupled-CV. The paragraph below gives the detail on the procedure.

All the nonlinear modeling techniques have tuning parameters analogous to n_{LV} in PLS. The major tuning parameters are m_{try} (the number of variables randomly sampled as candidates at each tree node split) for RF, s (the number of nodes in the hidden layer) for ANN, and C (the

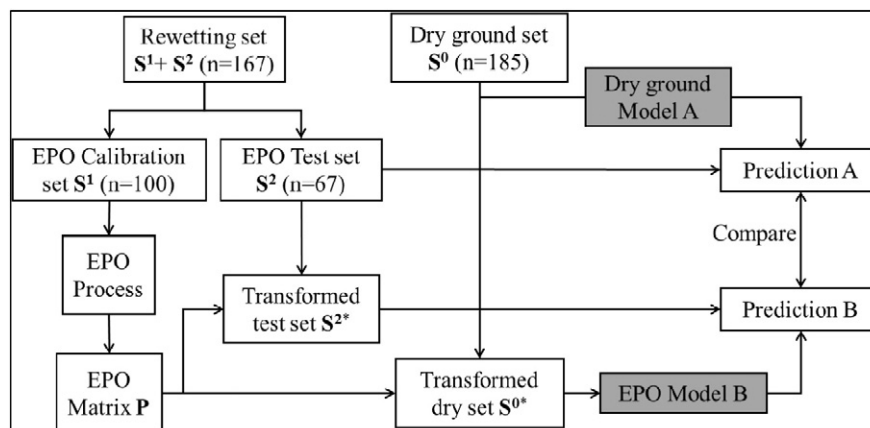


Fig. 1. The flow chart of external parameter orthogonalization (EPO) development and validation scheme in this study.

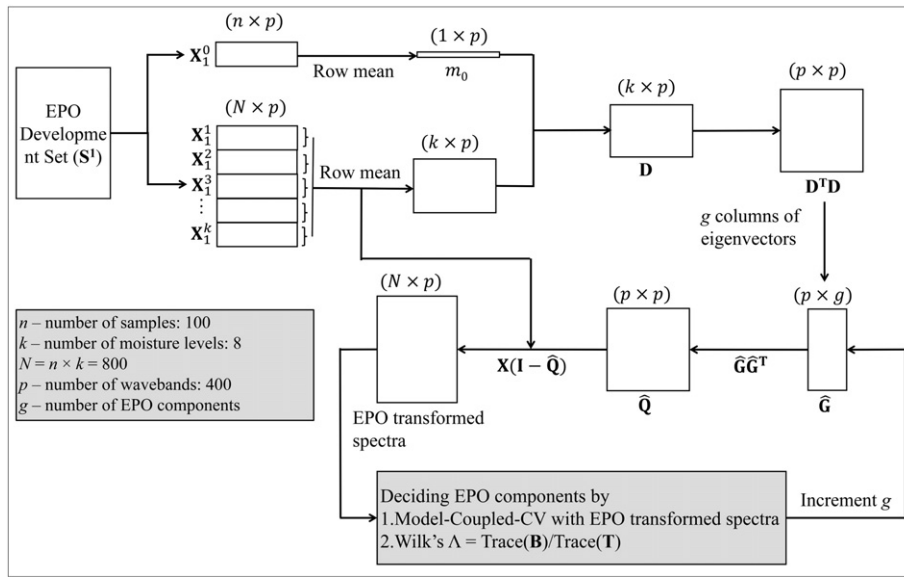


Fig. 2. Implementation of external parameter orthogonalization (EPO) transformation with the model-coupled-Cross Validation (model-coupled-CV) and Wilk's Λ method. The matrix symbols drawn in the figure are for the understanding of matrix operations in the EPO procedure.

number of violations to the margin) for SVM (Hastie et al., 2009; James et al., 2013). In RF, m_{try} was varied from 5 to 125 (increment by 15 at each step) to find the best coupling with EPO. In ANN, s was allowed to vary 1 to 9 (step 2) for EPO coupling. In SVM, C was varied from 8 to 64 (increment by 8 at each step). In all three nonlinear modeling techniques, g was allowed to vary from 1 to 10, and the optimal coupling was found by searching for the best combination of g and the respective coupling parameters (m_{try} , s , and C) that gives the lowest $RMSE_{CV}$ (Root Mean Squared Error of cross validation).

3. Results and discussion

The effect of moisture on soil VNIR reflectance spectra has been documented in several previous publications such as Lobell and Asner (2002); Zhu et al. (2010), and Minasny et al. (2011). Our findings are consistent to these studies. In general, there is a systematic decrease in soil reflectance with increasing moisture content. The shift is, however, not uniform along the wavelengths. The decrease between two neighboring moisture levels is more pronounced and well separated at the longer wavelengths than shorter wavelengths. This can be attributed to the general lower reflectance of dry ground soil in the visible region than the NIR region. Based on this phenomenon, Lobell and Asner (2002) suggested that longer wavelengths are more suited to observe moisture effect on spectra and estimate moisture content.

3.1. Model-coupled-CV versus Wilk's Λ to determine the optimal number of EPO components g

The characteristic of the EPO transformation matrix \mathbf{P} is dependent on the number of EPO components g . In this study EPO components were determined by two methods: Model-Coupled-CV and Wilk's Λ . A key difference between the two methods is that Model-Coupled-CV considers the cross validation of model calibration on the transformed spectra \mathbf{S}^{1*} (Fig. 2). This means g is dependent on two factors: (1) the coupling between model calibration and EPO, and (2) the response variable Y . Conversely, Wilk's Λ only results in one value of g regardless of these two factors. As Roger et al. (2003) pointed out, Wilk's Λ is a cluster separation measurement where it measures the potential classification in a group of samples. Before EPO transformation, models are usually poor because the spectra of the same sample at different moisture levels (intra-sample variation) could differ more than the spectra of two different samples (inter-sample variation). If the transformation is

successful in removing or minimizing the moisture effect, different samples should be well separated from the viewpoint of classification in the spectral space (Roger et al., 2003). Larger Wilk's Λ implies better separation of samples. Fig. 3 shows Wilk's Λ as a function of the number of EPO components, which suggests that g is 2 for our dataset.

Table 2 gives g determined by Model-Coupled-CV and Wilk's Λ methods for the four modeling techniques with their respective coupling parameter. For PLS (the only linear modeling technique), g ranges from 8 to 10, larger than 2 as determined by Wilk's Λ . At the same time, the coupling parameter n_{try} is also smaller for all soil properties (except for Sand). This indicates that, for PLS, Wilk's Λ leads to more parsimonious models for EPO correction and prediction.

An examination of other three nonlinear modeling techniques in Table 2 reveals more interesting pattern. The CV method coupled with RF and SVM yielded g values of either 1 or 2 for all eight soil properties. This is in agreement with Fig. 3 where Wilk's Λ increases from 0.47 ($g = 0$) to 0.92 ($g = 1$), peaks at 0.95 ($g = 2$), then decreases and fluctuates around 0.90 ($g = 3$ to 10). This indicates when $g = 1$ or 2, the separation of spectra between different moisture levels are best; and modeling of EPO transformed spectra with RF and SVM favors lower g with cross validation. ANN modeling gives higher g than RF and SVM with cross validation (but still lower than PLS for OC, IC, and TC).

Fig. 4 visualizes EPO transformation for different g , by Model-Coupled-CV and Wilk's Λ methods. The top and bottom rows are the original and EPO-transformed spectra at selected four levels of moisture. The mid row shows the \mathbf{P} matrices that resulted from different g . As previously shown, g determined by Wilk's Λ is 2, and those determined by Model-Coupled-CV range from 1 to 10. Here only the transformations with g equal to 2, 6 and 9 are shown.

Fig. 4 shows that EPO effectively removes the variability in soil reflectance spectra caused by moisture, yielding much similar spectra of the same sample after EPO transformation. One striking pattern revealed in Fig. 4 is that the transformation with $g = 2$ (Fig. 4a) gave a much smoother transformed spectra, comparing to the noisy spectra by higher EPO components (Fig. 4b and c). The primary reason, as we speculate, is that higher EPO components by including a large number of eigenvectors from the decomposition of the $\mathbf{D}^T \mathbf{D}$ matrix could potentially introduce extra spectral noise into parasitic matrix \mathbf{Q} (a similar effect as in principal component analysis where higher principal components are usually associated with noise). This in turn makes the EPO projection matrix \mathbf{P} noisier ($\mathbf{P} = \mathbf{I} - \mathbf{Q}$). In addition, it is quite clear in Fig. 4 that the two soil samples are better separated when

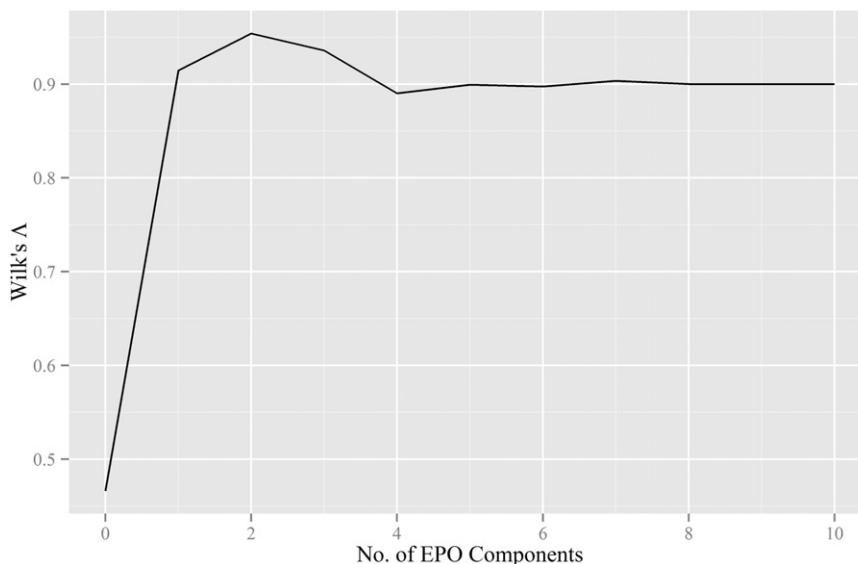


Fig. 3. Wilk's Λ as a function of the number of external parameter orthogonalization (EPO) components g (zero component means no EPO transformation). Higher Wilk's Λ indicates a higher degree of separation of different samples with respect to the same sample of different moisture levels in the spectral space.

$g = 2$ compared to $g = 6$ and 9 . We reason that the smoother spectra, together with the better between-sample spectral separation (both achieved when $g = 2$), lend the transformed spectra amenable to develop moisture insensitive soil property models.

3.2. Comparison of EPO modeling for different soils properties and modeling techniques

Table 3 gives the validation result of dry ground models (Model A from S^0) for predicting zero moisture spectra of test set (S^2) for the eight soil properties with the four modeling techniques. It shows that, for our dataset, OC, IC, and TC can be predicted reasonably accurately with R^2 ranging from 0.56 to 0.88 and RPIQ from 1.79 to 2.64. The dry ground models for clay and sand vary significantly among the modeling techniques. They are best predicted with PLS; but their modeling with

RF, ANN and SVM is somewhat poor. The models for pH, P, and CEC are quite poor, with R^2 lower than 0.35. This is consistent with the literature: VNIR models for soil carbon contents and textures are usually superior to other properties such as pH, P and CEC (Stenberg et al., 2010).

Table 4 gives the results of model performance with EPO (prediction B in Fig. 1) and without EPO (prediction A in Fig. 1) for the eight soil properties and four modeling techniques studied. Again, EPO transformation was implemented with two methods: Model-Coupled-CV and Wilk's Λ . Note that R^2 , RPIQ and $RMSE_p$ in Table 4 were calculated using all eight moisture levels in S^2 . For prediction A, it can be seen that all predictions fail as indicated by low R^2 , RPIQ and high RMSE. For OC, IC, and TC, EPO effectively removes the moisture effect and yielded significant improvement in prediction for all four modeling techniques. Note that TC, OC, and IC also have the highest accuracy for the initial dry ground models in Table 3. Consistent improvement by EPO is observed for sand with the decreasing of $RMSE_p$ values, although the improvement is quite marginal. Clay shows improvement only with PLS, but not with other three techniques. For pH, P, and CEC, no improvement was achieved. When comparing these findings with Table 3, it appears that a good initial dry ground model is quite important for EPO to work. This is not surprising. EPO in principle is an information removal process that removes spectral components associated with moisture. If the initial dry ground model is poor, meaning a weak correlation between certain soil property and spectra, EPO would not make the correlation stronger (because no spectral information is added).

Since EPO works best for OC, IC and TC, our discussions in the following focus on these three properties. A cross comparison between Tables 2 and 4 provides us some insight on how differently EPO couples the linear (PLS) and nonlinear (RF, ANN, and SVM) techniques. For PLS, g as determined by PLS-CV is much larger than Wilk's Λ . But the prediction result in Table 4 shows that the performance is superior with Wilk's Λ for OC and IC, and only slightly inferior for TC. Larger g seems not beneficial to improvement the EPO modeling performance. This is in agreement with Fig. 4 where EPO-transformed spectra from lower g are much smoother. For ANN, the same pattern can be seen between ANN-CV and Wilk's Λ (smaller g and s but comparable model performance). For RF and SVM, it is notable that their CV gives the almost same g (1 or 2) as Wilk's Λ . From these analyses it seems that the nonlinear modeling techniques can be better coupled with EPO through cross validation, as they yielded g comparable to Wilk's Λ (which we know is better as seen in Fig. 3) and better model performance than PLS.

Table 2
Optimum number of external parameter orthogonalization (EPO) components (g) and the coupling parameters for the four modeling techniques determined by the Model-Coupled-CV and Wilk's Λ method.

Modeling Technique	EPO method	Tuning parameter	Soil property								
			OC	IC	TC	Sand	Clay	pH	P	CEC	
PLS	PLS-CV	n_{LV}	15	20	10	20	20	20	20	20	
		g	9	10	9	9	9	8	8	10	
	Wilk's Λ	n_{LV}	11	14	5	22	12	20	5	15	
		g	2	2	2	2	2	2	2	2	
RF	RF-CV	m_{try}	110	5	125	35	125	20	125	125	
		g	1	1	1	2	2	1	1	2	
	Wilk's Λ	m_{try}	50	125	110	20	125	80	5	65	
		g	2	2	2	2	2	2	2	2	
ANN	ANN-CV	s	5	9	5	5	3	5	5	9	
		g	6	7	4	10	10	6	9	9	
	Wilk's Λ	s	1	1	3	3	1	1	1	7	
		g	2	2	2	2	2	2	2	2	
SVM	SVM-CV	C	64	64	64	64	64	64	64	64	
		g	1	1	1	1	1	1	1	1	
	Wilk's Λ	C	64	64	64	64	64	64	64	64	
		g	2	2	2	2	2	2	2	2	

PLS is Partial Least Squares Regression; RF is Random Forest; ANN is Artificial Neural Networks; SVM is Support Vector Machine.
 P is Mehlich-3 Phosphorus; CEC is Cation Exchange Capacity.
 n_{LV} is the number of PLS latent variables; g is number of EPO components; m_{try} is the number of variables randomly sampled as candidates at each split in RF modeling; s is the number of nodes in the hidden layer in ANN modeling; and C is the severity of the violations to the margin in SVM modeling.

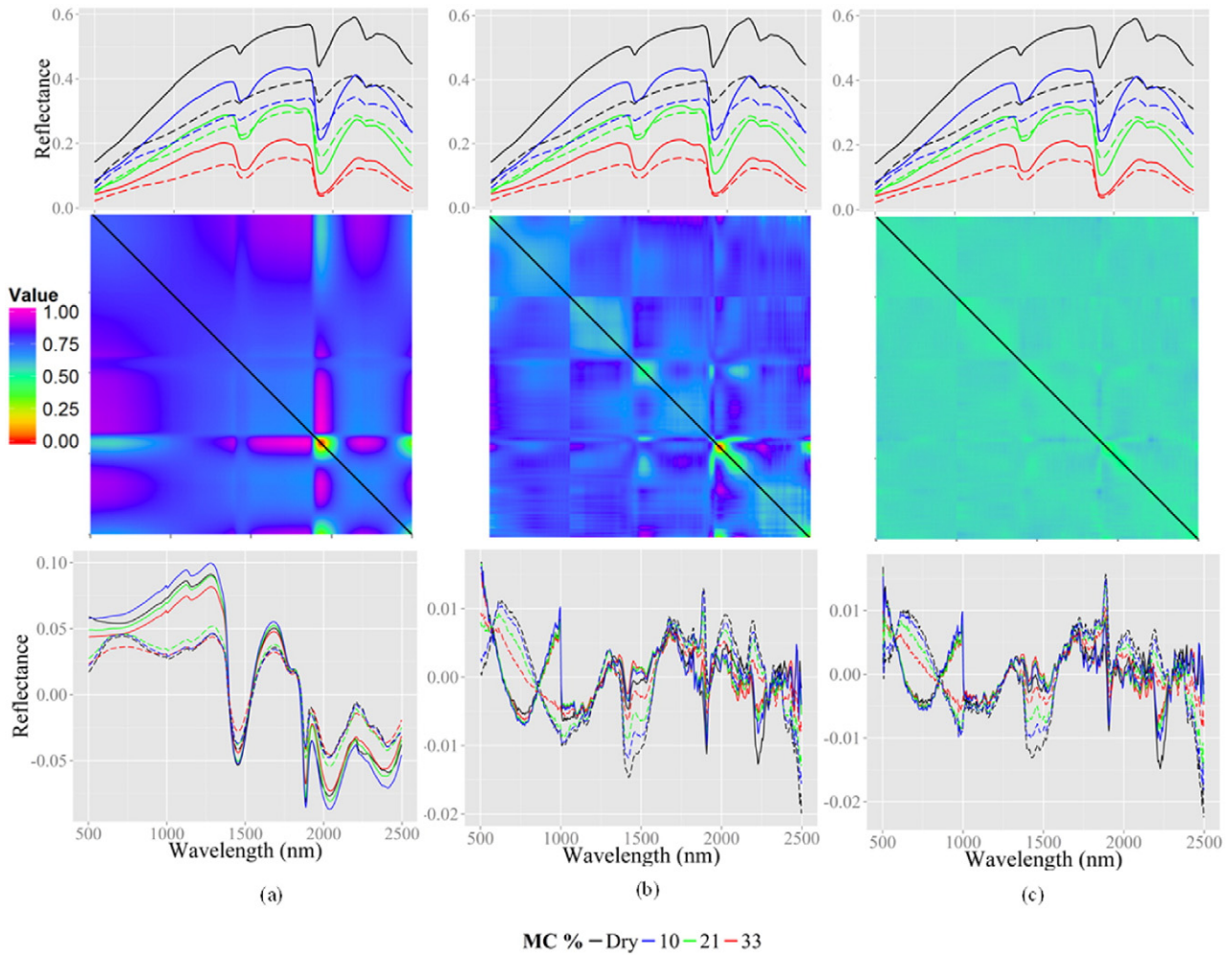


Fig. 4. Visualization of external parameter orthogonalization (EPO) transformation with two (a), six (b) and nine (c) EPO components. Solid and dash lines represents two different soil samples while black, red, green and blue colors represent different soil moisture levels.

Fig. 5 provides a pictorial view of how EPO improves modeling for samples at different moisture levels. Fig. 5a to c show OC prediction with ANN. It can be seen that, without EPO (Model A applied directly

to S^2), as moisture level increases, the prediction becomes poorer as suggested by increasing deviation from 1:1 line. The lack of linear response and a systematic over prediction indicate that the variation in spectra is dominated by moisture, thus failing OC prediction, particularly at higher moisture levels. Fig. 5b and c are predictions with ANN-CV and Wilk’s Λ , respectively. It is obvious that the prediction with EPO greatly improve the linear response of predictions across all moisture levels, resulting in significant improvement in model statistics (R^2 , RPIQ, and $RMSE_p$). Same patterns are observed for IC prediction with PLS (Fig. 5d–f); and TC with SVM (Fig. 5g–i) modeling. The variation induced by moisture was greatly suppressed, resulting in strong linear responses to the respective soil properties and minimal differences across different moisture levels.

Table 3

The validation results of dry ground models (Model A from S^0 in Fig. 1) to predict for the dry ground spectra of test set (zero moisture in S^2) for the eight soil properties with four modeling techniques.

Modeling technique	Parameter	Soil property							
		OC	IC	TC	Sand	Clay	pH	P	CEC
PLS	R^2	0.71	0.81	0.67	0.55	0.44	0.08	0.12	0.24
	RPIQ	2.03	2.59	1.79	1.41	0.89	1.24	1.10	1.04
	$RMSE_p$	0.74	0.31	0.78	20.22	12.54	1.29	38.97	6.61
RF	R^2	0.76	0.73	0.64	0.00	0.16	0.35	0.11	0.12
	RPIQ	2.05	2.09	1.82	0.99	1.32	1.93	1.11	1.06
	$RMSE_p$	0.73	0.38	0.77	29.00	8.51	0.83	38.78	6.54
ANN	R^2	0.76	0.79	0.88	0.40	0.27	0.07	0.05	0.23
	RPIQ	2.46	2.58	2.64	1.58	0.91	0.67	0.99	1.01
	$RMSE_p$	0.61	0.31	0.53	18.09	12.25	2.39	43.22	6.85
SVM	R^2	0.77	0.56	0.69	0.26	0.15	0.07	0.03	0.08
	RPIQ	2.50	1.81	2.03	1.36	1.11	1.38	1.02	0.99
	$RMSE_p$	0.60	0.44	0.69	21.03	10.07	1.16	41.94	6.97

PLS is Partial Least Squares Regression; RF is Random Forest; ANN is Artificial Neural Network; and SVM is Support Vector Machine.

P is Mehlich-3 Phosphorus; CEC is Cation Exchange Capacity.

$RMSE_p$ units are % for OC, TC, IC, Sand and Clay, $mg\ kg^{-1}$ for P, and $cmol\ +\ kg^{-1}$ for CEC.

3.3. The performance of EPO across different moisture levels

In field applications, soils will be at different field moisture levels. Therefore it is an important question to ask if EPO performs equally well across different moisture levels. Fig. 6 gives the summary statistics of R^2 (on the left Y axis) and RPIQ (on the right Y axis) as a function of moisture level for OC prediction with ANN (Fig. 6a) and TC prediction with SVM (Fig. 6b). Without EPO, both R^2 and RPIQ drop very quickly from 0 (dry ground) to 8% moisture (the first level) and remain low for higher moisture levels. With the EPO transformation, models still predict best for 0 moisture, but only drop slowly from 0 to the next three moisture levels (8, 10, 13%) and then level off at higher moisture

Table 4
Results of model performance using moist soil VNIR spectra without (A) and with (B) external parameter orthogonalization (EPO) for predicting the eight soil properties of S^2 set by using four modeling techniques.

Modeling technique	Prediction	Parameter	Soil property							
			OC	IC	TC	Sand	Clay	pH	P	CEC
PLS	A	R ²	0.11	0.19	0.03	0.13	0.00	0.04	0.00	0.04
		RPIQ	0.31	0.21	0.20	0.32	0.12	0.30	0.57	0.38
		RMSE _p	4.80	3.78	6.87	88.83	94.78	5.36	74.63	17.97
	B/PLS-CV	R ²	0.08	0.31	0.51	0.27	0.01	0.04	0.08	0.08
		RPIQ	1.10	1.44	1.49	1.33	0.72	1.49	1.02	0.67
		RMSE _p	1.36	0.55	0.94	21.58	15.60	1.08	42.00	10.34
	B/Wilk's Λ	R ²	0.56	0.61	0.29	0.20	0.00	0.00	0.11	0.08
		RPIQ	1.77	1.87	1.31	1.25	0.48	0.87	1.09	0.64
		RMSE _p	0.85	0.43	1.07	22.83	23.19	1.84	39.39	10.80
RF	A	R ²	0.19	0.05	0.07	0.00	0.09	0.12	0.04	0.01
		RPIQ	0.56	0.99	0.62	0.73	1.01	1.53	1.21	0.59
		RMSE _p	2.70	0.81	2.26	39.23	11.09	1.05	35.49	11.71
	B/RF-CV	R ²	0.63	0.70	0.55	0.19	0.01	0.21	0.02	0.08
		RPIQ	1.89	1.96	1.64	1.51	0.78	1.73	0.90	0.75
		RMSE _p	0.79	0.41	0.85	18.98	14.45	0.92	47.89	9.19
	B/Wilk's Λ	R ²	0.61	0.71	0.55	0.18	0.01	0.12	0.02	0.08
		RPIQ	1.86	1.82	1.65	1.52	0.77	1.65	0.94	0.76
		RMSE _p	0.81	0.44	0.85	18.76	14.46	0.97	45.72	9.10
ANN	A	R ²	0.01	0.40	0.24	0.08	0.01	0.00	0.01	0.00
		RPIQ	0.49	1.51	0.44	0.54	0.64	0.65	0.83	0.44
		RMSE _p	3.05	0.53	3.19	52.95	17.39	2.48	51.57	15.74
	B/ANN-CV	R ²	0.64	0.68	0.71	0.11	0.02	0.12	0.02	0.09
		RPIQ	1.86	2.08	2.11	1.21	0.78	1.52	0.75	0.69
		RMSE _p	0.81	0.38	0.66	23.58	14.45	1.05	56.95	10.00
	B/Wilk's Λ	R ²	0.77	0.32	0.64	0.16	0.00	0.00	0.01	0.06
		RPIQ	2.47	1.45	1.86	1.22	0.69	0.88	1.12	0.72
		RMSE _p	0.61	0.55	0.75	23.50	16.22	1.81	38.22	9.63
SVM	A	R ²	0.12	0.08	0.06	0.04	0.03	0.06	0.02	0.00
		RPIQ	1.19	1.05	1.04	0.60	1.06	1.31	1.15	0.91
		RMSE _p	1.27	0.76	1.34	47.55	10.58	1.22	37.30	7.62
	B/SVM-CV	R ²	0.71	0.55	0.67	0.11	0.00	0.12	0.00	0.08
		RPIQ	1.99	1.61	1.86	1.26	0.52	1.53	0.95	0.56
		RMSE _p	0.75	0.50	0.75	22.70	21.42	1.05	45.34	12.43
	B/Wilk's Λ	R ²	0.75	0.70	0.70	0.18	0.00	0.03	0.01	0.07
		RPIQ	2.40	1.72	2.02	1.37	0.68	1.47	0.94	0.68
		RMSE _p	0.63	0.46	0.69	20.89	16.54	1.09	45.42	10.17

PLS is Partial Least Squares Regression; RF is Random Forest; ANN is Artificial Neural Network; SVM is Support Vector Machine.

P is Mehlich-3 Phosphorus; CEC is Cation Exchange Capacity.

Prediction method: A means the models are based on the dry ground spectra without EPO correction; B/Model-CV means the EPO models developed from the model-coupled-CV method; B/Wilk's Λ means the EPO models developed from the Wilk's Λ method.

RMSE_p units are % for OC, TC, IC, Sand and Clay, mg kg⁻¹ for P, and cmol + kg⁻¹ for CEC.

levels. This shows that EPO can substantially improve prediction for all moisture level. In field applications, it is expected that EPO has the potential to be used for a wide range of field moist samples, and there would be a slight decrease in prediction performance for higher moisture soils than lower moisture soils.

4. Conclusions

We conducted an in-depth study of EPO for removing moisture effect from soil VNIR spectra and improving model performance. We expanded the investigation to eight soil properties and four different modeling techniques. We also compared the two methods for determining the optimum number of EPO components g : model-coupled-CV and Wilk's Λ . The major conclusions drawn from this study are as follows.

1. EPO is effective in removing moisture effect from soil VNIR spectra at different moisture levels. Its effectiveness can be readily seen by comparing the spectra before and after EPO transformation. Before transformation, the variation induced by moisture is so large that it masks the variation between samples. After EPO transformation, moisture-induced variation was largely suppressed and between-sample variation becomes dominant again.

2. Wilk's Λ is a viable method for determining g . Compared to model-coupled-CV, Wilk's Λ only relies on reflectance spectra for g determination and yields smoother spectra after transformation. With PLS, Wilk's Λ generally gives rise to lower g and n_{LV} than PLS-CV for all soil properties, indicating its advantage in model parsimony. For the nonlinear modeling techniques, their respective CV methods tend to favor smaller g comparable to the Wilk's Λ method (especially for RF and SVM). The coupling between EPO and these nonlinear modeling techniques is superior to PLS.
3. Among the eight soil properties, EPO improved the prediction of OC, IC, and TC significantly. The prediction of sand and clay, were improved only marginally, and the improvement is not consistent among different modeling techniques. For the predictions of pH, Mehlich-3 P, and CEC, EPO did not show improvement. Having good initial dry ground models is important for EPO to work.
4. For OC, IC, and TC, EPO substantially improves prediction at all different moisture levels. However, there is a slight decrease in the prediction accuracy when samples have higher moisture contents.

Acknowledgments

This work was funded by the U.S. Department of Agriculture – Natural Resources Conservation Services. The authors would like

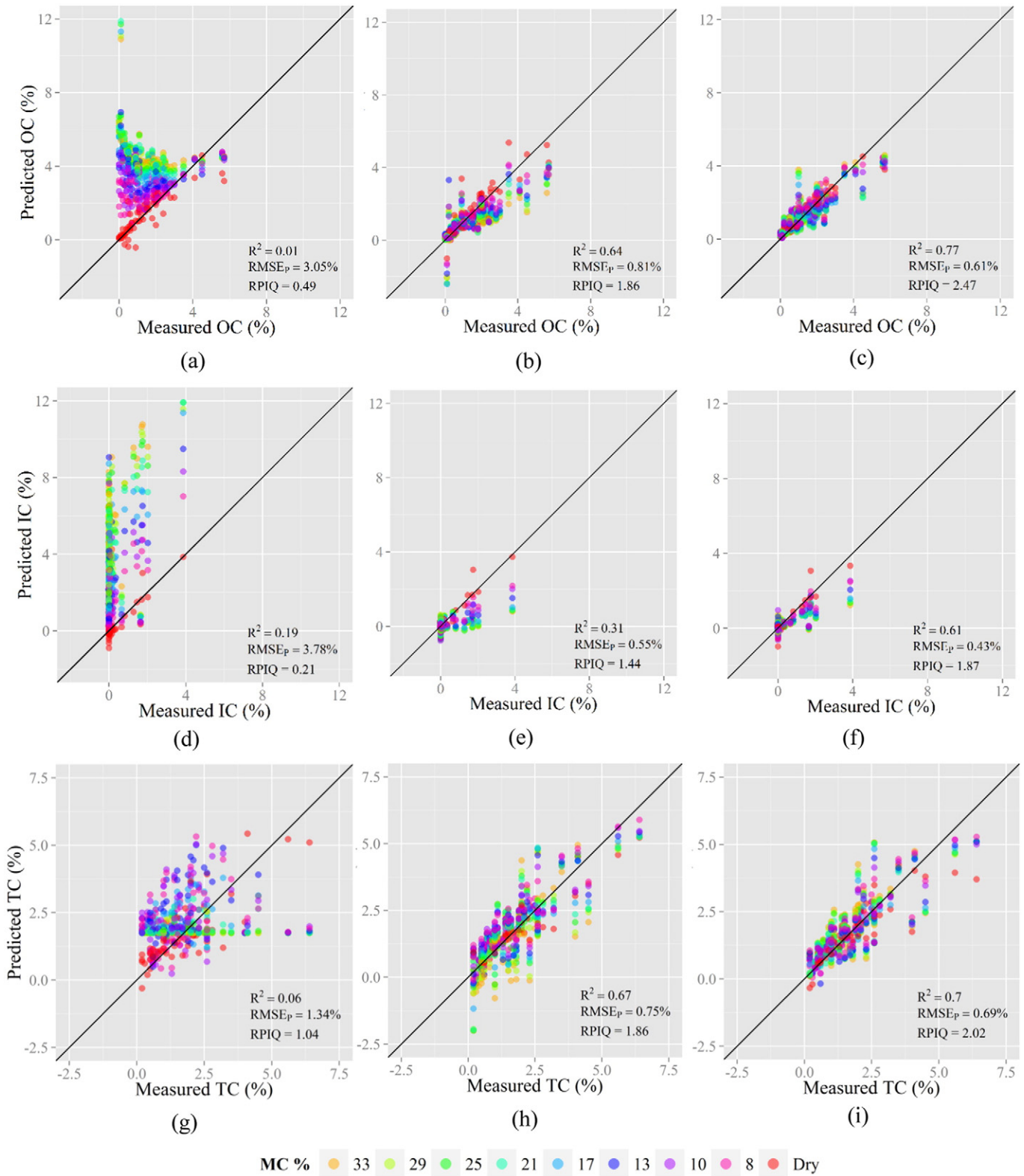


Fig. 5. Prediction of different soil properties with and without external parameter orthogonalization (EPO) transformation. First, second and third rows correspond to Organic C with ANN, Inorganic C with PLS, and Total C with SVM modeling respectively. First, second and third columns correspond to prediction plots without EPO transformation, EPO transformation with Model-coupled-CV method, and EPO transformation with Wilk's Λ method.

to thank the staff at Kellogg Soil Survey Lab (Dr. Richard R. Ferguson, Scarlett Bailey and Michael J. Pearson) for their assistance in sample retrieval from the archive and VNIR scanning.

References

Ackerson, J.P., Demattê, J.A.M., Morgan, C.L.S., 2015. Predicting clay content on field-moist intact tropical soils using a dried, ground VisNIR library with external parameter orthogonalization. *Geoderma* 259–260, 196–204.

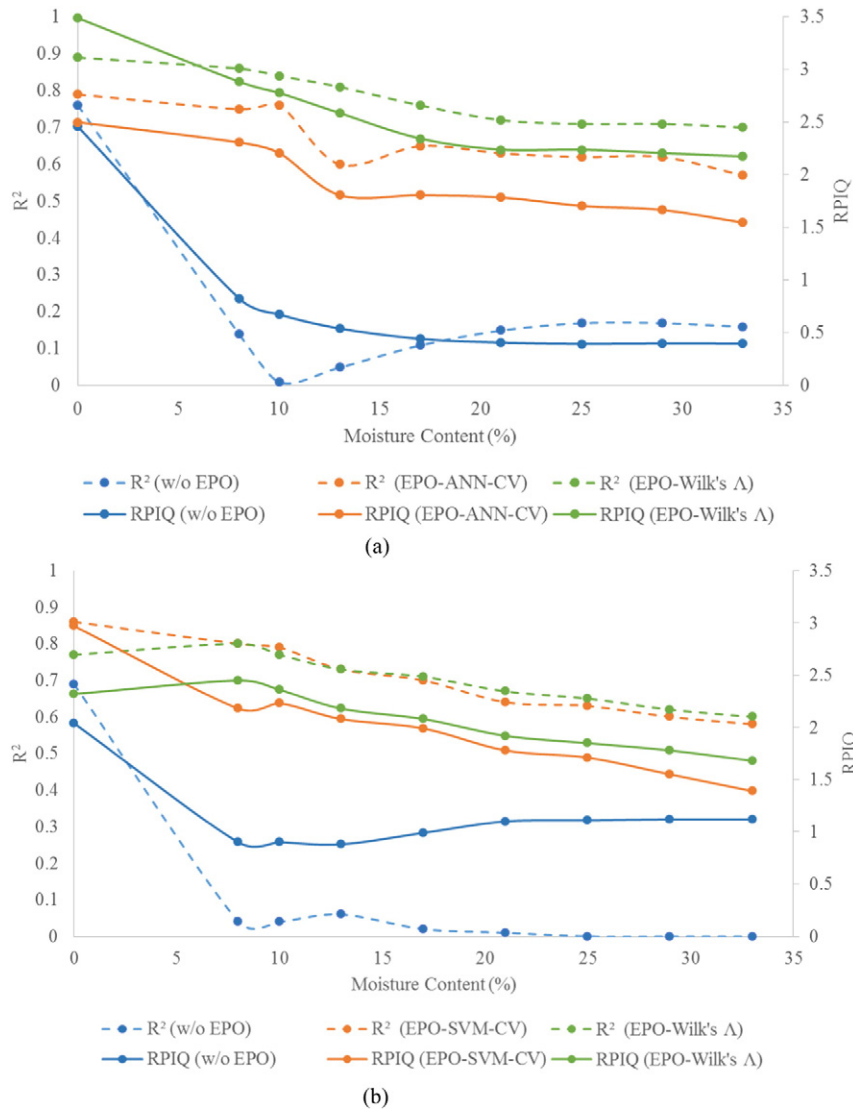


Fig. 6. The performance of external parameter orthogonalization (EPO) correction on (a) soil Organic C with ANN modeling, and (b) Total C with SVM modeling, at different soil moisture levels.

- 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. *TrAC Trends in Analytical Chemistry* 29 (9), 1073–1081.
- Ben-Dor, E., Heller, D., Chudnovsky, A., 2008. A novel method of classifying soil profiles in the field using optical means. *Soil Sci. Soc. Am. J.* 72 (4), 1113–1123.
- Brickley, R.S., Brown, D.J., 2010. On-the-go VisNIR: potential and limitations for mapping soil clay and organic carbon. *Comput. Electron. Agric.* 70 (1), 209–216.
- Brown, D.J., Shepherd, K.D., Walsh, M.G., Dwayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* 132 (3–4), 273–290.
- Chang, C.-W., Laird, D.A., Maudsby, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy—principal components regression analyses of soil properties. *Soil Sci. Soc. Am. J.* 65 (2), 480–490.
- Core Team, R., 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ge, Y., Thomasson, J.A., Morgan, C.L., Searcy, S.W., 2007. VNIR diffuse reflectance spectroscopy for agricultural soil property determination based on regression-kriging. *Trans. ASABE* 50 (3), 1081–1092.
- Ge, Y., Morgan, C.L.S., Ackerson, J.P., 2014. VisNIR spectra of dried ground soils predict properties of soils scanned moist and intact. *Geoderma* 213, 61–69.
- Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: an Australian case study. *Geoderma* 146 (3–4), 403–411.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning*, 2. Springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An Introduction to Statistical Learning*. Springer.
- Ji, W., Viscarra Rossel, R.A., Shi, Z., 2015. Accounting for the effects of water and the environment on proximally sensed vis-NIR soil spectra and their calibrations. *Eur. J. Soil Sci.* 66 (3), 555–565.
- Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kernlab — an S4 package for kernel methods in R. *J. Stat. Softw.* 11 (9), 1–20.
- Kuang, B., Mouazen, A.M., 2013. Non-biased prediction of soil organic carbon and total nitrogen with vis-NIR spectroscopy, as affected by soil moisture content and texture. *Biosyst. Eng.* 114 (3), 249–258.
- Max, K., Jed, W., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., 2015. *Caret: classification and regression training*.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R News* 2 (3), 18–22.
- Lobell, D.B., Asner, G.P., 2002. Moisture effects on soil reflectance. *Soil Sci. Soc. Am. J.* 66 (3), 722–727.
- Mevik, B.-H., Wehrens, R., Liland, K.H., 2013. pls: partial least squares and principal component regression.
- Minasny, B., McBratney, A.B., Pichon, L., Sun, W., Short, M.G., 2009. Evaluating near infrared spectroscopy for field prediction of soil properties. *Soil Res.* 47 (7), 664–673.
- Minasny, B., McBratney, A.B., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L., Joalland, S., 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma* 167–168, 118–124.
- Morgan, C.L.S., Waizer, T.H., Brown, D.J., Hallmark, C.T., 2009. Simulated in situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma* 151 (3–4), 249–256.
- Mouazen, A.M., De Baerdemaeker, J., Ramon, H., 2005. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil Tillage Res.* 80 (1), 171–183.

- Nocita, M., Stevens, A., Noon, C., van Wesemael, B., 2013. Prediction of soil organic carbon for different levels of soil moisture using vis-NIR spectroscopy. *Geoderma* 199, 37–42.
- Revelle, W., 2015. *Psych: procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois.
- Roger, J.-M., Chauchard, F., Bellon-Maurel, V., 2003. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemom. Intell. Lab. Syst.* 66 (2), 191–204.
- Sarkhot, D.V., Grunwald, S., Ge, Y., Morgan, C.L.S., 2011. Comparison and detection of total and available soil carbon fractions using visible/near infrared diffuse reflectance spectroscopy. *Geoderma* 164 (1–2), 22–32.
- Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. *Soil Sci. Soc. Am. J.* 66 (3), 988–998.
- Sørensen, L., Dalsgaard, S., 2005. Determination of clay and other soil properties by near infrared spectroscopy. *Soil Sci. Soc. Am. J.* 69 (1), 159–167.
- Stenberg, B., 2010. Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on vis-NIR predictions of clay and soil organic carbon. *Geoderma* 158 (1–2), 15–22.
- Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* 107, 163–215.
- Sudduth, K., Hummel, J., 1993. Soil organic matter, CEC, and moisture sensing with a portable NIR spectrophotometer. *Trans. ASAE (USA)*.
- Venables, W.N., Ripley, B.D., 2002. *Modern Applied Statistics with S*. fourth ed. Springer, New York.
- 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 (1–2), 59–75.
- Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. *Soil Sci. Soc. Am. J.* 71 (2), 389.
- Wickham, H., 2009. *ggplot2: elegant graphics for data analysis*.
- Zhu, Y., Weindorf, D.C., Chakraborty, S., Haggard, B., Johnson, S., Bakr, N., 2010. Characterizing surface soil water with field portable diffuse reflectance spectroscopy. *J. Hydrol.* 391 (1–2), 133–140.