Title:

Penetrometer-mounted VisNIR spectroscopy: application of EPO-PLS to *in situ* VisNIR spectra

Authors:

Jason P. Ackerson^{a*}, C.L.S. Morgan^a, and Y. Ge^b

^aDepartment of Soil and Crop Science, Texas A&M University, College Station, United States of America

^bBiological Engineering Systems Department, University of Nebraska-Lincoln, United States of America

*Corresponding Author

Phone: +1 616 502 7642

Email: jackers2@tamu.edu

Address:

MS 2474 TAMU

College Station, TX, 77845, USA

Keywords:

Proximal Sensing; *In situ* VisNIR; External Parameter Orthogonalization; Partial least squares; Clay Content

Abbreviations:

EPO: External parameter orthogonalization

PLS: Partial least squares

VisNIR: Visible, near infrared spectroscopy

RMSE: Root mean squared error

1. Abstract

Visible near-infrared spectroscopy (VisNIR) has been used to measure many soil properties. Typically, VisNIR is used on air-dried and ground soils in the laboratory. Recent developments in VisNIR instrumentation have allowed for the collection of VisNIR spectra from *in situ* soils. In this study, we demonstrate the viability of an *in situ* VisNIR system. VisNIR spectra were collected using a penetrometer-mounted VisNIR probe. The penetrometer-mounted VisNIR system requires no field calibration or pre-drilling of access holes. This system has several advantages in that it: 1) allows for measurement of soil properties without sample collection, preparation, and laboratory analysis and 2) can provide rapid soil measurement at high-depth-resolutions (2cm). We applied an external parameter orthogonalization (EPO) to the *in situ* spectra to remove the effects of soil moisture and other *in situ* effects from the spectra. We calibrated partial least-squares (PLS) models using spectra from an existing library of air-dried and ground spectra. PLS models were then used to predict clay content of the EPO-transformed *in situ* spectra. Model results showed good predictive ability for *in situ* spectra with RMSE, bias, and R^2 of 88 g kg⁻¹, -15g kg⁻¹, and 0.76, respectively. A site-wise hold-out validation of EPO calibration demonstrated that EPO calibrations were robust to changes in soil characteristics and parent materials between study areas. These results show that by using the EPO-PLS method, *in situ* VisNIR is a viable tool for rapid, minimally invasive collection of soil data.

2. Introduction

Despite the continued demand for soil data in applications such as digital soil mapping and precision agriculture, these applications are still limited by the availability of reliable soil measurements. Data on soil properties are typically limited by the high cost of soil sample collection and laboratory analysis. Laboratory-based spectroscopy systems such as visible near-

infrared spectroscopy (VisNIR) can reduce the cost of laboratory analysis by replacing, or supplementing, traditional analytical approaches. Laboratory-based VisNIR has been used for prediction of many soil properties including clay content (e.g. Chang et al., 2001; Shepard and Walsh, 2002), organic and inorganic carbon content (e.g. Shepard and Walsh, 2002; McCarty et al., 2002), cation exchange capacity (e.g. Chang et al., 2001; Shepard and Walsh, 2002), and properties primarily related to clay content. Despite the success of laboratory-based VisNIR the method still requires collection and preparation of soil samples.

To reduce the need for sample collection, several researchers have been investigating the use of VisNIR for measurement on *in situ* soils. In the most elementary approaches, spectrometers are used to measure soils collected from sections of soil cores (Ge et al., 2014; Morgan et al., 2009; Waiser et al., 2007) or on soil profiles exposed during sampling pit-excavation (Viscarra Rossel et al., 2008). While these methods were successful, the methods still require collection or disturbance of the soil.

An alternative approach is to build an instrument that can be inserted into the soil where VisNIR spectra can be collected for the undisturbed *in situ* soils. Ben-Dor et al. (2008) developed such an instrument that could be inserted into bore-holes pre-drilled into the soil. Once inside the pre-excavated bore-hole, the instrument could collect VisNIR spectra. The basic idea behind this instrument was further refined by equipping soil penetrometers with optical instruments capable of measuring the VisNIR reflectance of soil *in situ* (Poggio et al., 2015; Chang et al., 2011). Penetrometer-mounted VisNIR probes can be inserted into the soil without excavation of a soil bore-hole. Penetrometer-mounted VisNIR probes can collect VisNIR spectra at high-depth-resolutions (i.e. 2 to 5 cm) with minimal soil disturbance. If successful,

VisNIR-equipped penetrometers could greatly reduce the need for expensive traditional soil sampling and laboratory approaches.

Typically, in VisNIR modeling, prediction models are calibrated using the collection of spectra measured from soils of known properties. These spectral collections, referred to as spectral libraries, can contain thousands to tens of thousands of soil spectra (Viscarra Rossel et al., 2016; Brown 2007) and represent a substantial financial investment. The vast majority of reference spectra in spectral libraries are collected from soils that have been air-dried and ground. A major challenge for *in situ* VisNIR is that *in situ* spectra are altered by the effects of soil moisture, ambient temperatures, and soil structure (Bricklemyer et al., 2010). These effects, henceforth referred to as *in situ* effects, alter the spectra enough that models calibrated with existing spectral libraries (i.e. calibrated with spectra from air-dried and ground soils) cannot be used successfully on *in situ* spectra (Ge et al., 2014).

One approach for VisNIR modeling with *in situ* spectra is to generate new spectral libraries specifically for *in situ* spectra. While this approach has been implemented (Roudier et al., 2015; Morgan et al., 2009; Waiser et al., 2007), it is costly and can still result in errors due to the effects of differential soil moisture (Waiser et al., 2007). An alternative approach is to remove the *in situ* effects from the spectra using a spectral projection (Ji et al., 2015; Ackerson et al., 2015; Ge et al., 2014; Minasny et al., 2011). In this study we used a projection called an external parameter orthogonalization (EPO). The EPO rotates spectra in such a way the *in situ* effects are removed from the spectra (Roger et al., 2003). EPO has been used to remove soil-moisture effects from ground soil samples (Minasny et al., 2011) and spectra from intact soil cores (Ackerson et al., 2015; Ge et al., 2014). By removing *in situ* effects from spectra using the

EPO, models calibrated with an existing library of spectra from air-dried and ground soils can be used on spectra collected *in situ*.

EPO is not the only projection that has been used to remove water content effects from VisNIR spectra. Direct Standardization (DS) has also been used (Ji et al., 2015). The goal of this study was primarily to evaluate the penetrometer-equipped VisNIR system and a thorough comparison of the performance differences between DS and EPO is beyond the scope of this paper. For a more thorough discussion of DS and EPO readers are directed to Wijewardane et al., (2016) and Ji et al., (2015).

In this study we demonstrate the utility of a penetrometer-mounted VisNIR probe as a viable tool for measuring soil properties *in situ*. To do this, we collected two sets of VisNIR spectra. One set of spectra was collected from *in situ* soils using a penetrometer-mounted VisNIR probe. The second set of spectra was collected from the same soils in the air-dried and ground condition. Using this data we test: 1) if the EPO can remove the *in situ* effects from *in situ* spectra and 2) that after application of EPO, clay content of *in situ* spectra can be estimated with models calibrated using a spectral library of air-dried and ground spectra. We will compare the performance of predictions made from *in situ* spectra with the performance made using spectra from the same soils in the air-dried and ground condition.

3. Materials and Methods

3.1 Instrumentation for collection of *in situ* VisNIR spectra

A penetrometer-mounted VisNIR probe was used to collect *in situ* VisNIR spectra. The probe is similar to that used by Poggio et al. (2015). The probe consists of a stainless steel outer case, 32 mm in diameter (Fig. 1). The probe is attached to a hollow 25-mm diameter steel tube 1.2 meters in length. This tube is then attached to a hydraulic soil probe (Giddings Machine,

Fort Collins, CO, USA) which is used to insert the probe into the soil. The tube is hollow to allow power supply cables and optical fibers to be routed inside the tube itself, thus connecting the probe to instruments and a power supply located at the soil surface.

Inside the stainless steel case is a lamp which generates the initial light source for the probe. This light is reflected via a mirror across a sapphire window mounted on the side-wall of the spectrometer (Fig. 1). The light then interacts with soil and is reflected back into the probe where it is intercepted by an optical fiber. The optical fiber, housed inside the hollow steel tube, connects the probe to the spectrometer located on the soil surface. The optical fiber transmits light reflected from the soil to the spectrometer. An ASD AgriSpec spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA) was used for collection of all spectra in this study.

The optical fiber used in the penetrometer is five meters long, which is substantially longer than most fibers used for VisNIR. The longer fiber does result in minor signal loss; however, laboratory tests have indicated that the the signal to noise ratio of the longer fiber is only increased by roughly 1.5%. Additionally, the signal processing used in this study, including the use of EPO, greatly minimizes any negative effects due to increased optical fiber length.

After initial setup, collection of VisNIR spectra using the penetrometer-mounted VisNIR probe is straight-forward. The probe is first calibrated by placing a spectralon panel against the sapphire window and by following typical VisNIR spectroradiometer calibration procedures (i.e. instrument optimization and standardization). Next, using the hydraulic soil probe, the instrument is inserted 5 cm into the soil and a spectra is collected. The probe is then inserted an additional 5 cm into the soil and a second spectra is collected. This procedure is repeated on 5-cm intervals until the probe has traveled the maximum distance of 120 cm.

3.2 Soil sampling

The penetrometer mounted VisNIR probe was tested at four sampling areas in Burleson and Brazos counties in the state of Texas, USA (Fig. 2). The sampling areas consist of a floodplain, a stream terrace, and two upland locations. These sampling locations were chosen because they offer a diverse range of parent materials and geologic ages. Soils on the floodplain and stream terrace were developed in alluvial materials dating from the Holocene and Pleistocene epochs, respectively. Soils from the upland areas were developed in coastal plain sediments of the Eocene-aged Yegua formation (Soil Survey Staff, 2002).

The soil moisture and temperature regimes of the sampling area are Ustic and Thermic, respectively. The clay mineralogy of the soils in the stream terrace and uplands is often smectitic, although soils located in the floodplain generally have mixed mineralogies (e.g mica, smectite, and kaolinite). In the floodplain, soils are mapped predominantly as Vertisols and Inceptisols. On the stream terrace and upland, soils are mapped as Alfisols, Vertisols, and their respective intergrades (Soil Survey Staff, 2002).

Within each sampling area, we chose several sampling locations, in a fashion that maximized the diversity of soil properties observed within each study site. At the stream terrace, where soils exhibited the largest range in physical and morphological properties, 20 locations, or soil profiles, were sampled. At the floodplain and uplands where soils were less morphologically and physically diverse, fewer locations were sampled. Seven locations were sampled at the floodplain and six locations were sampled at each of the upland sites.

At each sampling location, we collected *in situ* VisNIR spectra using the penetrometer-mounted VisNIR probe. Spectra were collected on 5-cm intervals between 5 and 120 cm deep, and two profiles of spectra were collected at each sampling location. These adjacent profiles

were separated by a horizontal distance of no more than 10 cm. Spectra at each depth were averaged across both profiles. In addition to collection of *in situ* VisNIR spectra, a soil core was also collected.

Soil cores were divided into 5-cm segments corresponding to the depths at which *in situ* VisNIR spectra were collected. Segments were air-dried and ground to pass through a 2-mm sieve. From every distinct horizon from each core, a representative 5-cm segment was selected. Clay content and particle size class of each selected segment was measured using the pipette method (Gee and Orr, 2002). A summary of the results for particle size analysis for each sample area is shown in Table 1.

For all the sampled profiles combined, every USDA soil textural class is represented (Fig. 3). In general, the stream terrace and upland sites tended to be sandy/clayey, while the floodplain sites tended to be silty. Each area represented a large range in clay contents with all areas having samples with clay contents greater than 40 percent and less than 10 percent.

For each soil sample where clay content was measured, VisNIR spectra were collected from air-dried and ground soils in the laboratory. Spectra were collected using the same spectrometer that was used for collection of *in situ* spectra (i.e. ASD AgriSpec). These spectra, collected from air-dry and ground soil, represent the VisNIR reflectance from soils collected under laboratory conditions. Spectra collected from air-dried and ground soil in the laboratory will henceforth be referred to as laboratory spectra. For each laboratory spectra, there is a corresponding spectra collected from the same soil collected under *in situ* conditions. The resulting dataset consisted of a series of 155 pairs of VisNIR spectra from *in situ* and air-dried and ground soils.

3.3 Spectral datasets

As outlined in the previous section, we collected VisNIR spectra from soil under *in situ* and laboratory conditions. This collection of paired spectra were used to test the effectiveness of the penetrometer-mounted VisNIR for predictions of clay content. The data were divided into four separate datasets by sampling area. Clay content of soils from each dataset are summarized in Table 1.

Partial least squares modeling (PLS) was used to translate spectral data in to predictions of clay content. Partial least squares models were calibrated using the PLS package in the statistical software R (R Core Team, 2015). To calibrate PLS models, we used the Texas Soil Spectral Library (TSSL), which consists of spectra from over 2,000 soils collected across the state of Texas. Spectra in the TSSL were collected from soil under laboratory conditions (i.e. air-dried and ground soils). For further details on the TSSL readers are directed to Ge et al. (2014).

Prior to analysis, we filtered all spectra using the Savitzky-Golay transformation with a second order filter and a window size of 11 nm (Savitzky and Golay, 1964). To remove the regions of poor signal to noise ratio, reflectance from wavelengths below 500 nm and above 2450 nm were removed. To reduce the size of the data, spectra were resampled on 10-nm intervals. Finally, the filtered reflectance spectra were transformed to absorbance spectra ($\log 1/\text{reflectance}$).

3.4 External Parameter Orthogonalization (EPO)

As mentioned in previous sections, the major limitation to the use of *in situ* VisNIR is that *in situ* spectra are altered by the effects of soil moisture, ambient temperatures, and soil structure. By correcting for these *in situ* effects, we can use spectral models calibrated on existing laboratory spectral libraries to make predictions on *in situ* spectra. To remove *in situ*

effects from the spectra collected using the penetrometer-mounted VisNIR probe, an External Parameter Orthogonalization (EPO) was used.

In situ effects include the effects of soil water content as well as soil aggregation and roughness. Ge et al. (2014) tested separate EPO projections for wet-intact soil cores and dry-intact soil cores. Their results showed that soil water contributed to the majority, but not all, of error in VisNIR predictions. They found that an EPO calibrated to remove water content effects and intactness effects (e.g. roughness and aggregation) performed better than an EPO calibrated to remove only water effects. The EPO may also account for possible effects on the VisNIR spectra due to smearing of the soil by the probe; however, the effects of smearing on VisNIR spectra have been shown to be negligible (Waiser et al., 2007). To achieve the highest accuracy in *in situ* VisNIR predictions, we will calibrate the EPO to account for all *in situ* effects including soil water content, aggregation, roughness, and ambient temperature

The EPO, developed by Roger et al. (2003), can be used to remove unwanted effects from VisNIR spectra. While Roger et al. (2003) used the EPO to remove temperature effects from VisNIR spectra of fruit juices, EPO has been used to standardize between VisNIR instruments (Amat-Tosell et al., 2009) and remove soil moisture effects from soil spectra (Minasny et al., 2011). External parameter orthogonalization has been used to correct for *in situ* effects for soil spectra in soils from the United States (Ge et al., 2014) and tropical soils (Ackerson et al., 2015). The EPO has yet to be tested on spectra collected using a penetrometer-mounted VisNIR probe.

EPO removes unwanted spectral effects by rotating or projecting VisNIR spectra in a way that orthogonalizes the spectra to the effects of interest. Essentially, the EPO removes *in situ* effects by repositioning spectra in such a way that *in situ* effects are no longer apparent in

the spectra. The orthogonalization is achieved by multiplying the spectra by a projection matrix and estimation of the projection matrix is the critical step in the EPO procedure. For details on the derivation and testing of the EPO procedure, readers are directed to Roger et al. (2003) and Minasny et al. (2011).

To test the effectiveness of the EPO on *in situ* spectra collected using the penetrometer-mounted VisNIR probe, a site-wise holdout validation process was used. For site-wise holdout, the EPO-PLS algorithm is calibrated using spectra from all but one of the sampling areas. The EPO-PLS algorithm is then tested using the remaining sampling area. For example, all spectra from the floodplain area would be reserved as a validation dataset and the spectra from the remaining sites (i.e. stream terrace, and upland 1 and upland 2) would be reserved as a calibration dataset and used for calibration of the EPO-PLS algorithm. For each EPO-PLS calibration, an EPO projection is estimated using the air-dry and *in situ* spectra from the calibration dataset. This EPO projection is then used to transform the *in situ* spectra from the validation dataset and the air-dry spectra from the TSSL. A PLS model calibrated using the EPO-transformed TSSL and is then tested on the EPO-transformed *in situ* spectra from the validation dataset.. The TSSL contains only data from air-dry soils and therefore cannot be used for EPO calibration and is used only for PLS model calibration.

Site-wise holdout validation is a rigorous validation procedure. By calibrating the EPO-PLS algorithm on soils that are dramatically different from the soils used to validate the algorithm, there is little change of EPO-PLS performance being linked to site-specific interactions. In other words, site-wise holdout gives us the most confidence that the performance of the EPO-PLS algorithm is the result of a true orthogonalization rather than a site-specific correlation. All the EPO-PLS results shown in this paper were the result of site-wise validation.

3.5 Model evaluation

The effectiveness of the EPO-PLS algorithm was assessed in terms of root-mean squared error (RMSE), bias, and the coefficient of determination (R^2). RMSE was calculated via:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

where x_i and y_i are the i th paired observations from populations X and Y of measured and predicted values, respectively and n is the number of observation pairs. A RMSE value represents the accuracy of the EPO-PLS algorithm.

Bias, which represents the systematic over or under prediction of the EPO-PLS algorithm, was calculated via:

$$Bias = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i).$$

And R^2 was calculated via:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where $\hat{y}_i$ is the value of the least squares regression line between X and Y . The value of R^2 used in this discussion does not represent the R^2 of the EPO-PLS predictions per se, but rather represents the R^2 of a least-squares regression line between EPO-PLS predicted clay content and the measured clay content. In this way, the values of R^2 used in this analysis are not effected by any bias in EPO-PLS predictions and therefore can represent the precision of the EPO-PLS algorithm.

4. Results and discussion

4.1 Principle component analysis

We performed principle component analysis (PCA) on the laboratory VisNIR spectra from all five datasets (Fig. 4). The first two principle components (PCs) summarized the

majority of the variability in the data, accounting for 94% of the total variance. The dataset covering the largest extent in PC space was the TSSL. This is not surprising as the TSSL contains spectra from a large number of soils covering a large geographic area. The centroid of the TSSL lies outside the convex hull of each of the four remaining datasets of laboratory spectra indicating that majority of spectra in the TSSL are unlike spectra from the other datasets. PLS models calibrated with the TSSL will be optimized for spectra represented in the TSSL and therefore PLS models calibrated with the TSSL may show less than optimal performance for other datasets. This is a common problem with using large regional-scale spectral libraries such as the TSSL. Regional models are spectrally diverse leading to increased generality at the expense of local specificity which may result in impaired model performance (Viscarra-Rossel et al., 2016).

The four datasets used for EPO testing, the stream terrace, floodplain, and upland 1 and 2, covered a much smaller region of PC space than the TSSL. These datasets were constrained to a relatively confined region of PC space with all four datasets overlapping somewhat. The stream terrace, and upland datasets overlapped the most in PC space with the convex hull of each dataset containing the centroids of each other dataset.

Of the four datasets for EPO testing, the floodplain dataset is the most spectrally unique. The centroid of the floodplain dataset lies outside the convex hull of the other datasets in the PC. Additionally, the floodplain dataset covers a much smaller region of PC space than the remaining three EPO test datasets. These factors indicate that the soils in floodplain dataset are slightly different than the other datasets. Observed differences are likely due to the fact the floodplain soils are much younger than the other datasets and are therefore less weathered. Floodplain soils contain relatively large proportion of silt and very little sand (Fig. 3) and are still calcareous

throughout from lack of weathering. Typically, soils in the Brazos river floodplain are classified as having mixed minerologies while on the surrounding terraces and uplands, minerologies are typically classified as smectitic (Soil survey staff, 2002). When evaluating the EPO and PLS models, differences in the spectral properties of soils from the floodplain site may cause the EPO-PLS algorithm to behave differently than for spectra from the remaining study areas.

A second PCA on the *in situ* and laboratory spectra from the four study areas prior to application of the EPO was also performed (Fig. 5). For each study area, the *in situ* spectra are separated from laboratory spectra in PC space. In only the stream terrace (Fig. 5a) and upland 1 areas (Fig. 5b) do the convex hulls of the *in situ* and laboratory spectra overlap. This overlap however, is quite small consisting of one or two spectra. Without the EPO, the differences between *in situ* and laboratory spectra will likely result in differing PLS model performance for the *in situ* and laboratory spectra.

The first and second PCs of the *in situ* and laboratory spectra prior to transformation with the EPO represent 84 and 15% of the variance in the spectra, respectively. The *in situ* and laboratory spectra span the same range in the first component, with component scores ranging from -20 to 40. However the *in situ* and laboratory spectra occupy different ranges in the second PC with scores ranging from -15 to 5 and 5 to 10 for the laboratory and *in situ* spectra, respectively. This difference in the range of component score is why the second PC appears to be a superior discriminator of *in situ* and laboratory spectra. The second PC contains mostly information on the *in situ* effects on the spectra and this component comprises 15% of the variability in the spectra.

After application of the EPO, we performed a final PCA on the EPO-transformed laboratory and *in situ* spectra from all four study areas (Fig. 6). As opposed to the PCA prior to

application of the EPO (Fig. 5), there is little discernable difference between EPO-transformed laboratory and *in situ* spectra in PC space. For each study area, the convex hulls of the *in situ* and laboratory spectra overlap. The floodplain is the only study area where the centroid of the laboratory data is not contained within the convex hull of the *in situ* data. This discrepancy may be due to the fact that the laboratory spectra from the floodplain site are inherently different from the laboratory spectra of the remaining sites (Fig. 4) and the EPO is not fully accounting for *in situ* effect on spectra from the floodplain area.

The first and second PC of the EPO-transformed spectra account for 99 and 1 % of the variability of the spectra, respectively. The fact that the first PC accounts for the majority of the variability is due to the orthogonalization performed by the EPO. One way to conceptualize the EPO is that the EPO removes or subtracts the portion of the spectra effected by *in situ* effects. In the case of the spectra prior to EPO-transformation, this portion of the spectra would be analogous to the second PC which accounted for 15% of the variability in the spectra. If we were to remove this 15% of spectral variability, the remaining components would account for larger proportions of the total variability. For example, the first PC prior to EPO covered 84% of the variability, after EPO (i.e. removal of the second PC) the same component covered 99% of the variability of the spectra (i.e. $84/(100-15) = 99$).

4.2 Partial least-squares (PLS) performance on laboratory and *in situ* spectra without EPO

We tested the performance of PLS for predicting clay content on laboratory and *in situ* spectra prior to application of the EPO (Fig. 7, Table 2). The PLS model, henceforth referred to as the non-EPO model, was calibrated using the laboratory spectra from TSSL that had not been transformed using the EPO. The non-EPO model consisted of 15 PLS latent variables. We selected this number of latent variables by minimization of clay content prediction RMSE from a

five-fold cross-validation using only the TSSL data. During cross-validation, the non-EPO model had an RMSE and bias of 92 and 1 g kg⁻¹, respectively and an R² of 0.87 (Table 2). While this number of latent variables is high, models of similar size have been used previously (e.g. Ge et al., 2015). The high number of latent variables needed for adequate prediction likely reflects the spectral diversity in the TSSL.

We tested the effectiveness of the non-EPO model on both the laboratory and *in situ* spectra from each sample area. In general predictions made using the laboratory spectra were good (Fig. 7, Table 3). For all sample areas except the floodplain, RMSE of laboratory spectra was less than that of the TSSL cross-validation. The study area with the poorest model performance on laboratory spectra was the floodplain. Despite having the highest R² (0.93) of all study locations, the floodplain area had the highest RMSE (116 g kg⁻¹). This elevated RMSE is the result of the large bias in model predictions; -104 g kg⁻¹. We were unable to determine any concrete source of this systematic error,; however these results are not wholly unsurprising given that floodplain spectra were slightly separated from spectra at the other sites in PC space (Fig. 4). This separation in PC space suggests that some underlying spectral variation at the floodplain sites is unaccounted for in the non-EPO model thus, generating a systematic error in clay content predictions. The underlying source of spectral variation deviation from the TSSL may be due to the relative paucity of sand in the floodplain soils (Fig. 3).

While performance of the non-EPO model on laboratory spectra was generally good, the model performed more poorly on untransformed *in situ* spectra. RMSE for *in situ* spectra ranged from 252 to 418 g kg⁻¹ and bias ranged from 185 to 382 g kg⁻¹. This result is consistent with other results where untransformed models were applied to *in situ* spectra (Ji et al., 2016; Ge et al.,

2015, Minasny et al., 2009). Without the EPO or other transformation techniques, successfully predicting the clay content from *in situ* spectra is unlikely.

4.3 PLS performance for *in situ* spectra with EPO

We tested the performance of the EPO by evaluating the accuracy and precision of clay content predictions of PLS on EPO-transformed spectra (Fig. 8, Table 3). As mentioned in section 4.5, a whole-site holdout was used for EPO evaluation. Because a separate cross-validation was used for each study area, the EPO projection and accompanying PLS models were calibrated for each cross validation. The EPO and PLS calibrations required estimation of the appropriate number of EPO Eigenvectors, and PLS latent variables; parameters c and k , respectively. For all cross-validations, parameter optimization resulted in the same values for c and k : 1 and 5, respectively.

When the EPO-PLS algorithm was applied to *in situ* spectra, the prediction accuracy and precision improved, compared to predictions for untransformed *in situ* spectra (Fig. 8, Table 3). For each study area, the RMSE, bias, and R^2 of EPO-PLS predictions were improved compared to predictions for the same *in situ* spectra without application of the EPO. RMSE for EPO-transformed *in situ* spectra ranged from 98 to 70 g kg⁻¹ and R^2 ranged from 0.60 to 0.82. In general, the performance of EPO-PLS algorithm was similar the performance of the cross-validation of the laboratory spectral model. For two out of four sites, RMSE of EPO-PLS predictions for *in situ* spectra was less than that of the cross-validated TSSL predictions.

In general, EPO-PLS predictions for *in situ* spectra (Fig. 7, Table 3) were not as accurate or precise as PLS predictions for non-EPO-transformed laboratory spectra of the same soils (Fig. 6, Table 2). This is not wholly surprising as the laboratory spectra were unadulterated by *in situ* effects and were not subjected to any extraneous transformations (i.e. EPO). However, for

spectra collected from the floodplain area, the bias of EPO-PLS predictions for *in situ* spectra is significantly lower than that of laboratory spectra. This result suggests, that under some circumstances, spectral projection can improve model performance.

If the EPO is an effective tool it must remove enough of the *in situ* effects from the spectra to make accurate and precise predictions possible. However, whenever transforming or rotating a spectra, there is a chance that useful information may be lost. This loss of useful information would result in loss of predictive power for spectral models. One way to assess the possible loss of such information is to apply the EPO transformation to laboratory spectra. Because the laboratory spectra are unaffected by *in situ* effects, any decrease in PLS performance on EPO-transformed laboratory spectra relative to non-EPO transformed laboratory spectra can be attributed to over-correction by the EPO.

To test whether the EPO is over-correcting spectra we applied the EPO-PLS algorithm to laboratory spectra for each of the study areas (Fig. 8, Table 3). With the exception of the stream terrace site, where performance did not change, there is a general pattern of slightly decreasing model performance for prediction on laboratory spectra after application of the EPO. It should be noted that while accuracy and precision of predictions for laboratory spectra did decrease after application of the EPO, the performance at these sites still exceeded the performance of the spectra library during internal cross-validation. This suggests that, even if the EPO is removing some spectral information critical to clay content prediction, the resulting decrease PLS performance does not significantly degrade model performance. EPO-transformed models are performing as well as we could expect given the performance of TSSL model cross-validation.

4.4 High-depth-resolution profiles of clay content

One advantage of the penetrometer-mounted VisNIR probe is that the instrument allows collection of VisNIR spectra at high-depth-resolutions. Spectra can be collected at depth resolutions of 2 to 5 cm. These high-depth-resolution spectral datasets can be used to construct near-continuous depth-profiles of soil properties (Fig. 9). High-depth-resolution profiles of soil clay content can be used to identify morphological features in a soil profile. For example, the soil profile shown in Fig. 9a, was collected from a location that had been mapped as Silawa fine sandy loam; a soil classified as an Ultic Paleustalf. The VisNIR profile exhibits two distinct horizons; an upper horizon with low clay content, and a lower, high-clay argillic horizon.

Typically in soil sampling, soils are sampled on fixed intervals or by horizon. Soil from within a horizon or depth interval is homogenized and analyzed as a discrete unit. One disadvantage of this approach is that analysis of a homogenized layer provides no data on the variability of soil properties within a layer. High-depth-resolution soil data from the penetrometer-mounted VisNIR probe can provide data on the variability of soil within a defined depth-interval. Coupling high-depth-resolution VisNIR with traditional soil sampling could provide a more comprehensive measurement of soil profile properties; supplying heretofore unavailable data on within layer variability of soil properties.

5. Conclusions

In this study we assessed the efficacy of a penetrometer-mounted VisNIR probe for collection of VisNIR spectra from soils *in situ*. Samples were collected from 38 sampling locations dispersed across four study areas in central Texas. For each sampling location, VisNIR spectra were collected *in situ* using the penetrometer-mounted probe and from air-dried and ground soil cores. An EPO was applied to *in situ* VisNIR spectra to remove the effects of soil moisture and intactness from the spectra. After application of the EPO, PLS models calibrated

using a spectral library containing only spectra from air-dried and ground soils were used to predict clay content from *in situ* spectra.

The PLS models performed well for air-dried and ground spectra with average RMSE and R^2 across all sampling areas of 70 g kg^{-1} and 0.86, respectively. Prior to application of the EPO, PLS was unable to accurately predict clay content from *in situ* spectra. After application of the EPO, PLS performance on *in situ* spectra was quite good with average RMSE and R^2 across all sampling areas of 88 g kg^{-1} and 0.76, respectively. While PLS performance for *in situ* spectra was slightly poorer than that of air-dried and ground spectra, the performance of *in situ* predictions is comparable to that of the performance of the spectral library under cross-validation.

These results demonstrate that, with application of the EPO, *in situ* spectra collected using a penetrometer-mounted VisNIR probe can be used for prediction of clay content. The EPO has two distinct advantages. Firstly, EPO correction does not require information on the water content of the soil and the EPO can therefore be applied without collecting additional measurements from the soil (i.e. soil water content). Secondly, because the EPO removes the effect of soil water and other *in situ* effects from the spectra, PLS models calibrated using spectra from air-dried and ground soils can be applied to *in situ* data. This allows for the utilization of existing spectral libraries for PLS model calibration, thus negating the need for collection of additional calibration data specifically for *in situ* spectra.

By combining the penetrometer-mounted VisNIR probe with the EPO, we have established a system capable of *in situ* measurements of soil clay content along a profile at high-depth-resolutions. While further research is needed to assess the ability of the system to measure other soil properties (e.g. organic carbon content, cation exchange capacity, etc.), the current

system represents a strong forward step in spatially explicit, depth-wise proximal sensing of soil profiles. *In situ* VisNIR diminishes the need for traditional soil sampling and laboratory analysis, and therefore greatly reduces the operational costs of soil survey applications. By combining an *in situ* VisNIR system with existing proximal and remote sensors, high-resolution soil maps for precision agriculture will be easier and cheaper to develop.

6. Acknowledgements

This work was funded by the U.S. Department of Agriculture Natural Resource Conservation Service (USDA-NRCS)

464

465 7. References

- 466 Ackerson, J.P., Demattê, J. a. M., Morgan, C.L.S., 2015. Predicting clay content on field-moist
467 intact tropical soils using a dried, ground VisNIR library with external parameter
468 orthogonalization. *Geoderma* 259-260, 196–204.
- 469 Amat-Tosello, S., Dupuy, N., Kister, J., 2009. Contribution of external parameter
470 orthogonalisation for calibration transfer in short waves-Near infrared spectroscopy
471
- 472 Ben-Dor, E., Heller, D., Chudnovsky, A., 2008. A Novel Method of Classifying Soil profiles in
473 the Field using Optical Means. *Soil Sci. Soc. Am. J.* 72, 1113-1123.
474
- 475 Bricklemeyer, R.S., Brown, D.J., 2010. On-the-go VisNIR: Potential and limitations for mapping
476 soil clay and organic carbon. *Comput. Electron. Agric.* 70, 209–216.
477
- 478 Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and
479 landscape modeling in a 2nd-order Uganda watershed. *Geoderma* 140, 444–453.
480
- 481 Chang, C.C., Garcia-Urbe, A., Zou, J., Morgan, C.L.S., 2011. Micro side-viewing optical probe
482 for VNIR-DRS soil measurement. *IEEE Sens. J.* 11, 2527–2532.
483
- 484 Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh Jr., C.R., 2001. Near-infrared reflectance
485 spectroscopy - principal components regression analyses of soil properties. *Soil Sci. Soc.*
486 *Am. J.* 65, 480–490.
- 487 Gee, G.W., and D. Or. 2002. Particle-size analysis. p. 255–293 *In* J.H. Dane and G.C. Topp (ed.)
488 *Methods of soil analysis. Part 4. SSSA Book Ser. 5. SSSA, Madison, WI.*
- 489 Ge, Y., Morgan, C.L.S., Ackerson, J.P., 2014. VisNIR spectra of dried ground soils predict
490 properties of soils scanned moist and intact. *Geoderma* 221-222, 61–69.
491
- 492 Ji, W., Viscarra Rossel, R.A., Shi, Z., 2015. Accounting for the effects of water and the
493 environment on proximally sensed vis-NIR soil spectra and their calibrations. *Eur. J. Soil*
494 *Sci.* 66, 555–565.
495
- 496 McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-infrared and
497 near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Sci. Soc.*
498 *Am. J.* 66, 640-646.
499
- 500 Minasny, B., McBratney, A.B., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L.,
501 Joalland, S., 2011. Removing the effect of soil moisture from NIR diffuse reflectance
502 spectra for the prediction of soil organic carbon. *Geoderma* 167-168, 118–124.
503
- 504 Morgan, C.L.S., Waiser, T.H., Brown, D.J., Hallmark, C.T., 2009. Simulated in situ

505 characterization of soil organic and inorganic carbon with visible near-infrared diffuse
506 reflectance spectroscopy. *Geoderma* 151, 249–256.

507

508 Poggio, M., Brown, D.J., Bricklemyer, R.S., 2015. Laboratory-based evaluation of optical
509 performance for a new soil penetrometer visible and near-infrared (VisNIR) foreoptic.
510 *Comput. Electron. Agric.* 115, 12–20.

511 R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for
512 Statistical Computing, Vienna, Austria.

513 Roudier, P., Hedley, C.B., Ross, W., 2015. Prediction of volumetric soil organic carbon from
514 field-moist intact soil cores. *Eur. J. Soil Sci.* 66, 651–660.

515 Roger, J.M., Chauchard, F., Bellon-Maurel, V., 2003. EPO-PLS external parameter
516 orthogonalisation of PLS application to temperature-independent measurement of sugar
517 content of intact fruits. *Chemom. Intell. Lab. Syst.* 66, 191–204.

518 Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least
519 squares procedures. *Analytical Chem.* 36, 1627-1639.

520 Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for
521 characterization of soil properties. *Soil Sci. Soc. Am. J.* 66, 988.

522

523 Soil Survey Staff, 2002. Soil survey of Brazos country. USDA, Washington, DC.

524

525 Viscarra Rossel, R. A., Cattle, S.R., Ortega, a., Fouad, Y., 2009. In situ measurements of
526 soil colour, mineral composition and clay content by vis-NIR spectroscopy.
527 *Geoderma* 150, 253–266.

528

529 Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M.,
530 Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H.,
531 Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L.,
532 Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A.,
533 Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L.,
534 Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G.,
535 Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to
536 characterize the world's soil. *Earth-Science Rev.*

537

538 Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. In situ characterization of soil
539 clay content with visible near-infrared diffuse reflectance spectroscopy. *Soil Sci. Soc. Am.*
540 *J.* 71, 389.

541

542 Wijewardane, N.K., Ge, Y., Morgan, C.L.S., 2016. Prediction of soil organic and inorganic
543 carbon at different moisture contents with dry ground VNIR : a comparative. *Eur. J. Soil*
544 *Sci.* 67, 605–615

8. Tables

Table 1. Clay content summary statistics for the partial least squares (PLS) calibration spectral library (Texas Soil Spectral Library, TSSL) and each study area.

Dataset	n	Minimum	Median	Maximum	Standard Deviation	IQR†
-----gkg ⁻¹ -----						
TSSL‡	2022	0	276	882	200	315
Terrace	83	27	245	531	152	311
Floodplain	25	93	322	636	144	221
Upland1	24	67	285	537	130	166
Upland2	23	34	294	681	172	193

†IQR is the interquartile range

‡TSSL is the Texas soil spectral library.

554

555

Table 2. Partial least squares model performance for clay content predictions prior to application

556

of the external parameter orthogonalization.

Sample Area	Laboratory spectra			<i>In situ</i> spectra		
	RMSE†	Bias	R ²	RMSE†	Bias	R ²
-----	-----g kg ⁻¹ -----	-----	-----	-----g kg ⁻¹ -----	-----	-----
TSSL‡	92	1	0.87	-	-	-
Terrace	81	-36	0.77	374	305	0.29
Floodplain	116	-104	0.97	418	382	0.55
Upland1	41	-18	0.93	252	185	0.37
Upland2	42	-12	0.94	317	282	0.59

557

†RMSE is the root mean squared error

558

‡TSSL is the Texas soil spectral library. Results for the TSSL were generated using a five-fold

559

cross-validation.

560

561

561

562

Table 3. Partial least squares model performance for clay content predictions after application of

563

the external parameter orthogonalization.

Sample Area	Laboratory spectra			<i>In situ</i> spectra		
	RMSE†	Bias	R ²	RMSE†	Bias	R ²
	-----g kg ⁻¹ -----		-----	-----g kg ⁻¹ -----		-----
Terrace	72	-9	0.78	97	4	0.60
Floodplain	196	-176	0.92	86	-53	0.82
Upland1	85	69	0.85	70	-14	0.74
Upland2	63	30	0.92	98	3	0.72

564

† RMSE is the root mean squared error.

565

566

9. Figure Captions

Figure 1. Schematic of the penetrometer-mounted VisNIR probe. The upper photograph shows the probe exterior and the lower diagram shows the internal structure of the probe. White arrows represent the path of light inside the probe.

Figure 2. Map of the sampling areas where soil samples and *in situ* spectra were collected. The inset map shows the location of the sample areas within the state of Texas. The map on the right shows the location of each sampling location using white circles. The color of the map background represents the elevation of the sampling area in m.

Figure 3. Soil texture of each study area plotted on a USDA textural triangle. Solid lines represent the boundaries of USDA textural classes.

Figure 4. Principle component biplots for laboratory spectra for each study area and the Texas Soil Spectral library (TSSL). Lines represent the convex hull of each dataset and circles represent the centroids of each dataset.

Figure 5. Principle component biplots of *in situ* and laboratory spectra for each study area prior to application of External Parameter Orthogonalization (EPO). Solid and dashed lines represent the convex hull of laboratory and *in situ* spectra, respectively. The centroids of laboratory and *in situ* spectra are represented by the “X”, and “+” signs, respectively. Data from the stream terrace, floodplain, upland 1, and upland 2 areas are represented in Figs. 5a, 5b, 5c, and 5d, respectively.

590

591 Figure 6. Principle component biplots of *in situ* and laboratory spectra for each study area after
592 application of External Parameter Orthogonalization (EPO). Solid and dashed lines represent the
593 convex hull of laboratory and *in situ* spectra, respectively. The centroids of laboratory and *in*
594 *situ* spectra are represented by the “X”, and “+” signs, respectively. Data from the stream
595 terrace, floodplain, upland 1, and upland 2 areas are represented in Figs. 6a, 6b, 6c, and 6d,
596 respectively.

597

598 Figure 7. Partial least squares predictions of clay content of laboratory and *in situ* spectra prior to
599 application of the External Parameter Orthogonalization (EPO). Prediction for laboratory and *in*
600 *situ* spectra are represented by the circles and X’s, respectively. The solid line represents the 1:1
601 correspondence line. Data from the stream terrace, floodplain, upland 1, and upland 2 are
602 represented in Figs. 7a, 7b, 7c, and 7d; respectively.

603

604 Figure 8. Partial least squares predictions of clay content of laboratory and *in situ* spectra prior to
605 application of the External Parameter Orthogonalization (EPO). Prediction for laboratory and *in*
606 *situ* spectra are represented by the circles and X’s, respectively. The solid line represents the 1:1
607 correspondence line. Data from the stream terrace, floodplain, upland 1, and upland 2 are
608 represented in Figs. 8a, 8b, 8c, and 8d; respectively.

609

610 Figure 9. Example high-resolution-depth profiles of soil clay content. Open circles represent
611 clay content predictions made using the laboratory spectra without external parameter

612 orthogonalization (EPO) and solid circles represent clay content predictions made using *in situ*
613 spectra with the EPO. Measured values for clay content are represented by the X's.

614

615

Figure1

Figure2

Elevation [m]

56 72 88 104

2 km

Figure3

Figure4

Figure5

Figure6

Figure7

Figure8

Figure9