Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon

B. Minasny, A.B. Mcbratney, Véronique Bellon Maurel, J.M. Roger, Alexia Gobrecht, L. Ferrand, S. Joalland

Author-produced version of the article published in Geoderma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon

Budiman Minasny[1*], Alex. B. McBratney[1], Veronique Bellon-Maurel[2], Jean-Michel Roger[2], Alexia Gobrecht[2], Laure Ferrand[1,2], Samuel Joalland[1,2]

[1] Australian Centre for Precision Agriculture, The University of Sydney, Sydney NSW 2006, Australia.

[2] Montpellier Supagro-Cemagref, UMR ITAP, BP 5095, 34033 MONTPELLIER Cedex 1, France .

* Corresponding author:

Budiman.minansy@sydney.edu.au

Tel. +61 2 9036 9043

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

1. Introduction

Near Infrared Diffuse Reflectance Spectroscopy (NIR-DRS) is a promising technology for high resolution digital soil mapping and precision agriculture (Stenberg et al., 2010). NIR reflectance is particularly sensitive to organic and mineral soil composition, thus it permits the prediction of many soil properties from a single measurement. The technology has been used for several decades in many industries including crop production, food processing, pharmaceutical and petrochemicals. The NIR-DRS instruments are diverse, including portable units that can be easily transported, allowing both in situ and laboratory soil analysis.

Field measurement using Near Infrared Diffuse Reflectance Spectroscopy (NIR-DRS) spectroscopy has become popular for in situ prediction of various soil properties (Mouazen et al., 2007; Viscarra Rossel et al., 2009; Waiser et al., 2007). In particular, the use of NIR spectroscopy for prediction of soil organic carbon (SOC) content in the field is highly desirable for soil quality assessment and carbon accounting purposes (Bricklemyer and Brown, 2010; Christy, 2008). SOC represents a key parameter in evaluating the quality of soils, and is one of the most commonly and successfully predicted parameters using NIR-DRS. Once the spectra have been calibrated for SOC, the method can provide rapid and inexpensive estimation of SOC in the field.

However NIR reflectance is quite sensitive to external environmental conditions, such as temperature, soil moisture, and soil structural conditions. The influence of soil moisture on the NIR reflectance spectra has been reported by many authors (Bowers and Hanks, 1965; Bricklemyer and Brown, 2010; Lee and Boregki, 2006; Lobell and Asner, 2001; Minasny et al., 2009; Sudduth and Hummel, 1993). Figure 1 shows an example of the absorbance spectra of a soil with 1% SOC content with varying moisture content. The absorbance spectra clearly increase with increasing moisture content.

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

Figure 1. The effect of moisture on the absorbance spectra for a loamy soil with 17% clay and 1.1% SOC content. The spectra from the lower to higher absorbance are for the soil at gravimetric moisture content of 4, 5, 6, 12, 14, 16, and 20%.

While in the laboratory, soil can be scanned under standard air-dried conditions; in the field it is very difficult to control the water content. Most studies have demonstrated that NIR spectra calibrated using field soil samples can be used for field prediction of soil properties (Bricklemyer and Brown, 2010; Christy, 2008; Kusumo et al., 2008). Morgan et al. (2009) showed that when using NIR on field samples, the variability of soil moisture in the field reduced the prediction accuracies of SOC content. Sudduth and Hummel (1993) in a laboratory study showed NIR measurements on a wide range of soil moisture tensions. Although they found best SOC prediction with dry samples, reasonable estimates were obtained across the full range of tensions. These authors suggested that including a wide range of water contents in the calibration set could take care of the issue of moisture variation.

Although, many studies have successfully used spectra collected in field conditions to calibrate against measured SOC, the variation of soil moisture content can really have a huge impact on the prediction of SOC. Wu et al. (2009) identified a range of wavelengths in the NIR region where the first derivative of the reflectance spectra seems independent of the moisture content of the soil samples. They suggested to only use these selected wavelength intervals, to obtain moisture-independent estimates of SOC under field conditions.

This paper investigates the influence of soil moisture on the NIR-DRS signal and its effects on prediction of SOC. The ultimate objective is to remove the effect of moisture on the spectra for prediction of SOC content. Specifically, we shall investigate and evaluate the external parameter orthogonalisation (EPO) method to remove the moisture effect from the spectral calibration. EPO algorithm projects all the soil spectra orthogonal to the space of unwanted variation (i.e. moisture), and thus the variations of soil moisture can be effectively removed.

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

2. Materials and methods

2.1 Soil samples and the experiment

We used a soil library (391 samples) collected from agricultural areas in southern New South Wales, Australia (Minasny et al., 2009). The samples covered a range of soil types and were taken from different soil horizons up to 1 m depth. The soil samples have been ground and passed through a 2 mm sieve. SOC was determined using the dry combustion method, and ranges between 0.06 to 12.7 g $100g^{-1}$. The data were split into three independent datasets:

- Dataset A is the EPO development data, which is used to study the influence of soil moisture on the spectra. We systematically selected 100 samples from the whole data to represent the full range of SOC contents. The data were ranked from the lowest to highest values in SOC, and were stratified into 100 strata. A sample was then systematically taken from each stratum. These 100 samples were used for deriving a pre-processing method to remove the effect of moisture using the external parameter orthogonalisation (EPO) algorithm.
- Dataset B is the validation data. We selected 20 samples randomly from the rest of the library (291 samples) and used these as a validation dataset.
- Dataset C is the model calibration data. The rest of the data (271 samples) were used for model calibration, relating the spectra to SOC content using partial least squares (PLS) regression.

The statistics of the data are given in Table 1.

We used the AgriSpec™ instrument with a contact probe (Analytical Spectral Devices, Boulder, Colorado, USA) for collection of the Vis-NIR soil reflectance spectra (350 – 2500 nm). The samples were illuminated by a halogen lamp and the reflected light was transmitted to the spectrometer through a fibre optic bundle. A Spectralon (Labsphere Inc., North Sutton, N.H., USA) was used as a reflectance standard and employed to convert raw spectral data to reflectance. Each soil spectrum was obtained as the mean of 40 scans.

To investigate the effect of the moisture on the accuracy of the prediction formula, samples from dataset A (n = 100) were wetted evenly to approximately the sticky limit. This is the moisture

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

content at which a sieved soil will not stick to a metal spatula, which was required so that the soil did not stick to the contact probe. The samples, which were approximately 5 mm thick, were covered and left to equilibrate for 2 days. After equilibration, each of the soil samples were scanned and weighed. The samples were then left to dry under laboratory conditions (25^{o} C). Although there was a slight variation in moisture in the samples due to differential drying (dryer at the surface), this variation did not cause much difference in the NIR spectra. Subsequently each day (for 5 to 6 days) the samples were weighed and then scanned. Thus, for each sample at each day, we have the moisture content and a spectrum. We also conducted the same experiment for the 20 validation samples.

We only used the reflectance data that have a high signal to noise ratio (in the spectral range 500-2450 nm) for analysis. The spectra were transformed to absorbance (log 1/Reflectance) and sampled to a resolution of 2 nm, and smoothed using the Savitzky-Golay algorithm with a window size of 11 and polynomial of order 2 (Savitzky and Golay, 1964). To remove the baseline effect, we normalised each spectra by taking its mean and divided by its standard deviation, the so-called standard normal variate (snv) transformation (Barnes et al., 1989).

We used the Partial Least Squares (PLS) regression (Wold et al., 2001) to calibrate the spectra against measured SOC content. Since the distribution of SOC content is positively skewed, we used a logarithmic transformation of SOC for calibration. The value is back transformed for accuracy assessment. The accuracy of prediction was assessed using R^2 (coefficient of determination), ME (mean error), and RMSE (root mean squared error). Mean error (ME) is the mean difference between predicted and observed SOC content for the samples, positive value indicates overestimation, while negative value indicates underestimation. Root mean squared error indicates the accuracy of the prediction.

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

Table 1. Statistics of soil organic carbon (SOC) and clay content for the three datasets used in this study.

		Dataset A (EPO development data)	Dataset B (Validation data)	Dataset C (Calibration data)
	n	100	20	271
SOC content (g $100g^{-1}$)	Median	0.93	1.27	0.80
	Min	0.09	0.27	0.06
	Max	12.74	5.90	8.65
Clay content (g $100g^{-1}$)	Median	22.5	17.5	20.0
	Min	5.0	8.0	5.0
	Max	74.0	30.0	73.0

2.2 External parameter orthogonalisation (EPO)

We used a parameter orthogonalisation algorithm (Roger et al., 2003) to remove the effect of soil moisture from the spectra. The algorithm finds the areas in the spectra which are affected by moisture and projects the spectra orthogonal to this variation, and the unwanted variations of soil moisture can be effectively removed. This analysis is related to principal component analysis (PCA), but PCA only performs a transformation to take into account all of the variability in the spectra. Meanwhile, the orthogonolisation takes into account the variability which is due to an external factor (i.e. moisture). The algorithm was initially developed by Roger et al. (2003) to remove the effect of temperature from the spectra for the prediction of measurement of sugar content of fruit.

Let S be the m-dimensional space of the n measured spectra, the spectra can be written as

$$S = C + G + R$$

C is useful chemical spectral responses;

G is the part of the spectra that is caused by the external parameter and independent from C; and

R is the independent residual.

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

In matrix form, the spectra **X** (size $n \times m$) can be written as:

$\mathbf{X} = \mathbf{XP} + \mathbf{XQ} + \mathbf{R}$

P is the projection matrix (size $m \times m$) of the useful part of the spectra: $\mathbf{X^*} = \mathbf{XP}$;

Q is the projection matrix (size $m \times m$) of the not useful part (influenced by moisture) of the spectra: $\mathbf{X}^{\#} = \mathbf{XQ}$; and

R is the residual matrix (size $n \times m$).

The aim of EPO is to obtain the useful spectra $\mathbf{X^*} = \mathbf{X}(\mathbf{I} - \mathbf{Q})$, while matrix **Q** can be written as $\mathbf{Q} = \mathbf{G}\,\mathbf{G}^T$.

To estimate **G**, the uninformative part of the spectra that is orthogonal to the useful part of the spectra, Roger et al. (2003) suggested using the principal component of the difference spectra **D**. In this work, we define **D** as the difference between the moist and dry spectra. Figure 2 shows the difference spectra for the 100 soil samples of the EPO dataset. The spectra have been normalised using snv, and the plot shows the between the spectra of air-dried samples and samples at the wettest moisture condition (day 1). We can see the water peak at 1450 nm and 1940 nm, which are exaggerated with increasing moisture content. We then performed a Principal Component Analysis (PCA) on **D** to extract the variation subspace.

In summary, the EPO algorithm works as follows:

- Calculate difference spectra **D** (size $n \times m$): $\mathbf{D} = \mathbf{X}_{moist} - \mathbf{X}_{dry}$.
- Perform a principal component analysis (PCA) on $\mathbf{D}^T\mathbf{D}$. The PCA can be obtained using a singular value decomposition (SVD) of $\mathbf{D}^T\mathbf{D}$ to obtain $\mathbf{USV}^T$, where **U** is an $n \times n$ matrix, **S** is an $n \times m$ diagonal matrix, and **V** is an $m \times m$ matrix. See Press et al. (1992) for details on singular value decomposition, and Wall et al. (2003) for the relationship between PCA and SVD.
- Define c number of factors, and obtain $\mathbf{V}_s$ a subset of m by c of **V**.
- Estimate **Q** from $\mathbf{V}_s\mathbf{V}_s^T$.
- Calculate the projection matrix as $\mathbf{P} = \mathbf{I} - \mathbf{Q}$ (P is an m by m symmetric matrix)
- The transformed spectra is calculated as $\mathbf{X^*} = \mathbf{X}\,\mathbf{P}$.

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

The only parameter that needs to be chosen in EPO is c, the number of dimension of the principal component of $\mathbf{D}^T\mathbf{D}$, which represents the un-useful space that needs to be removed.

The data analysis taken in this paper is summarised as follows:

1. Perform EPO on dataset A, to obtain projection matrix **P**.
2. Using **P**, transform spectra of dataset C to transformed spectra.
3. Calibrate a partial least squares (PLS) regression using transformed spectra of dataset C to predict SOC.
4. For validation, transform spectra of dataset B using **P**. Using the transformed spectra, predict SOC using established PLS (from step 3).

3. Results and discussion

3.1 The effect of moisture content

Since the SOC content has a skewed distribution, we normalised it using a logarithmic transformation. For accuracy assessment, the value is transformed back. We first calibrated a PLS regression (Wold et al., 2001) using the soil samples from dataset A (calibration data) to predict SOC (using 6 factors, determined by cross validation). The resulting PLS has an R^2 value of 0.72. The PLS regression formula was then used to predict SOC content from the spectra of the EPO development data (100 samples) which have different moisture contents. The gravimetric moisture content (w) of the samples varied from an average of 12% to 7%. Figure 3 shows the prediction of the SOC using spectra at three moisture conditions. We can see that the prediction deteriorates significantly and also has a bias, a tendency to over-predict the SOC content with increasing moisture content (Table 2). The R^2 value declines to 0.28 for the highest moisture content with large bias and RMSE values. As the moisture content decreases towards air-dried conditions, the prediction accuracy improves. The difference in SOC content visually can be shown at wavelengths around 1400, 1900, and 2200 nm. Thus moisture can mask these peaks due to other OH bands present; it can also cause spectral differences due to interactions

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

between water and other components (Reeves, 2010). Meanwhile the moisture bands also show up at wavelengths 1400 and 1900 nm. The variation in soil moisture has been reported to cause poorer calibrations for other soil properties (e.g. SOC, clay content, CEC) (Bogrekci and Lee, 2005; Lobell and Asner, 2002; Minasny et al., 2009).

Figure 3. Correlation between predicted SOC and measured SOC for samples at three moisture conditions using PLS calibrated on air-dried samples. w is gravimetric moisture content (g 100g^{-1}).

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

Table 2. Prediction of SOC content using NIR spectroscopy. PLS is prediction using partial least squares regression calibrated on spectra from dataset A. EPO-PLS refers to the spectra after EPO pro-processing, and the PLS is calibrated using transformed spectra of dataset A. *w* is gravimetric moisture content (g $100g^{-1}$). ME is mean error, and RMSE is root mean square error. Note that the prediction is based on logarithmic transformation of C content.

		PLS			EPO-PLS		
	Average *w* (%)	R^2	ME (g $100g^{-1}$)	RMSE (g $100g^{-1}$)	R^2	ME (g $100g^{-1}$)	RMSE (g $100g^{-1}$)
Dataset A: EPO development data (n = 100)							
Air-dry	6	0.53	-0.14	0.73	0.54	-0.05	1.06
Day 1	12	0.28	2.13	4.01	0.48	-0.13	0.96
Day 2	9	0.52	0.24	0.92	0.52	0.07	0.73
Day 3	8	0.34	0.39	1.79	0.53	-0.01	0.62
Day 4	8	0.55	0.01	0.67	0.53	-0.04	0.75
Day 5	7	0.56	0.19	1.11	0.48	0.18	1.34
Dataset B: Validation data (n=20)							
Air-dry	4	0.80	-0.01	0.85	0.78	0.18	1.07
Day 1	18	0.68	2.55	3.22	0.89	0.09	0.49
Day 2	16	0.72	2.11	2.95	0.83	0.08	0.59
Day 3	14	0.74	1.57	2.30	0.86	0.08	0.55
Day 4	11	0.73	1.04	1.74	0.81	0.12	0.64
Day 5	7	0.67	0.62	1.44	0.71	0.22	1.07
Day 6	5	0.70	0.70	1.91	0.72	0.44	1.45
Dataset C: Calibration data (n=271)	7	0.72	-0.11	0.56	0.79	-0.08	0.46

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

3.2 EPO-PLS for prediction of SOC

We used the External Parameter Orthogonalisation (EPO) algorithm which transforms the spectra into a space which is not affected by water content, i.e. orthogonal to the moisture effect. The formulation is described as follows:

$$\mathbf{X^*} = \mathbf{X}\,\mathbf{P}$$

where $\mathbf{P}$ is the transformation matrix. The new spectra $\mathbf{X}^*$ then can be used to build a calibration function which is not affected by moisture. Using the dataset A (EPO development data), we first construct a set of difference spectra $\mathbf{D}$ between spectra in the moist conditions (at day 1 with average w = 12% and at day 3 with average w = 8%) and spectra at air-dried condition (average w = 6%). In matrix terms, this is written as:

$$\mathbf{D} = [\mathbf{X}_{d1} \mid \mathbf{X}_{d3}] - [\mathbf{X}_{ad} \mid \mathbf{X}_{ad}],$$

where $\mathbf{X}_{ad}$ is the spectra at air-dried condition, and $\mathbf{X}_{d1}$ and $\mathbf{X}_{d3}$ are spectra at the first and third day after wetting. We then perform a principal component analysis on $\mathbf{D}^T\mathbf{D}$ which defines the projection matrix $\mathbf{P}$, which is described in section 2.2 and summarised in Figure 4.

*Figure 4. External Parameter Orthogonolisation (EPO) pre-processing algorithm of the spectra. The original spectra **X** is multiplied by transformation matrix **P** to obtain transformed spectra **X*** which is free from soil moisture effect.*

The parameters that need to be optimised for EPO-PLS are c, the number of EPO dimensions and k, the number of PLS factors. We optimised this by using dataset A (EPO development data, n = 100). Using c values from 1 to 10, for each c, we calculated the projection matrix $\mathbf{P}$. Then with a range of PLS factors k (from 1 to 10), we derived a PLS model to predict SOC content using the transformed air-dried spectra. The PLS was then used to predict SOC of the transformed spectra at air-dried and day 1 after wetting (from dataset A). We then calculated the RMSE of the SOC prediction for different numbers of c and k, as shown in Figure 5. The results show that RMSE values do not decrease significantly after 6 PLS factors, and the RMSE is lowest an EPO dimension of 4. Therefore who chose c = 4 and k = 6.

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

Figure 5. Root mean square error (RMSE) of the prediction of SOC content using transformed spectra at different number of EPO dimension and different number PLS factors (k).

After establishing the optimum EPO dimensionality, we calculated the transformation matrix **P** from dataset A. Spectra for different water content can be transformed using **P**, resulting in transformed spectra $\mathbf{X}^*$. Figure 6 shows the spectra and EPO-transformed spectra of the soil sample shown in Fig. 1 for different moisture contents. Here we can see that the spectra of a soil with different moisture contents now look similar. The moisture variation at the peaks around 1400 and 1900 nm is now not visible.

Figure 6. (a) Original spectra of a soil with various moisture content and (b) the spectra after EPO transformation.

Using the projection matrix **P**, we transformed the spectra from dataset C (calibration data, n = 271) to become $\mathbf{X}^*_{\mathbf{ad}}$. We then calibrated the transformed spectra for SOC content prediction using PLS regression (6 factors). The PLS prediction on dataset C has an $R^2 = 0.79$, and RMSE = 0.46 g $100g^{-1}$. The PLS model was used to predict soil samples that have various moisture contents (dataset A and B). The resulting prediction of SOC for the 100 soil samples of dataset A at three different moisture conditions is shown in Figure 7. We can see that the prediction is now less affected by moisture content. Table 2 lists the accuracy of the prediction using the EPO preprocessing and then PLS prediction (EPO-PLS). For dataset A, the R^2 values are around 0.50, with RMSE between 0.62 and 1.34 g $100g^{-1}$. Overall the bias is reduced (ME values close to zero) when compared to the results obtained without EPO pre-processing. The accuracy of prediction on the wettest samples (day 1 of the experiment) is comparable to the prediction using air-dried samples. This shows that the EPO has effectively removed the effect of moisture from the spectra.

Figure 8 shows the prediction for the 20 independent validation samples (dataset B) which were collected at 7 different moisture contents. The prediction for all the samples looks similar despite of the variation in moisture content. Table 2 shows the improved accuracy for prediction of the

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

samples at the highest moisture content (day 1), in terms of R^2 values (0.68 to 0.89), less bias (ME near zero) and lower RMSE values (3.2 to 0.5 g $100g^{-1}$). Overall, the prediction accuracy of the soil with varying moisture content is comparable to the prediction of air-dried soil samples.

Figure 6. Prediction of soil organic carbon content on samples from dataset A at three different moisture conditions using EPO-PLS.

Figure 7. Prediction of soil organic carbon content on the validation dataset (dataset B) at different moisture contents using EPO-PLS.

3.3 The effect of the number of samples used for EPO development

As explained in the Methods section, we used three independent datasets for this approach, and the success of the EPO processing will depend on the number and representative samples used in dataset A (EPO development data). Representative samples here meaning that they should represent a range of soil types and SOC contents. Here we perform a numerical experiment to examine the effect of the number of samples required to develop an effective EPO spectral pre-processing method. From the 100 samples (of dataset A), we sampled for a range number of samples: 5, 10, 20, 30, 40 …, 100 to create EPO test datasets. The sampling is based on a systematic sampling approach: first the data is sorted based on their SOC content, to obtain *s* number of samples, the data was divided into *s* section, and a sample was taken in the middle of the section. We now have a range test datasets, each having different number of samples (5, 10, 20, 30, 40 …, 100). As previously, for each of the test dataset, we developed an EPO transformation matrix **P**. We then applied the transformation matrix to dataset C (calibration data) to create transformed spectra. The transformed spectra were calibrated for SOC using PLS. The PLS function was used to predict SOC on dataset A and B.

Figure 8 shows the RMSE values for dataset A and B as a function of number of samples used in developing the EPO transformation matrix. The boxplot represents the RMSE values at various

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

moisture contents. When the number of samples was between 5 and 50, the RMSE values in both datasets A and B showed variable results. For example, using 10 samples showed a low RMSE, but the RMSE increased when using 20 samples. After 60 samples, the RMSE values appeared to be more stable and below the general mean value of 0.35 (log[g 100^{-1}]). Therefore in this example, we can conclude that we need at least 60 samples to develop a stable EPO transformation matrix. And of course, the choice of the right samples is essential (covering the full range of soil type, SOC and represented in the spectra).

Figure 8. The effect of the number of samples used in the EPO development on the root mean square error (RMSE) values for the prediction of soil organic carbon content on datasets A and B. The boxplot represents RMSE values at different moisture contents.The line in the middle of the plot represents the grand mean.

4. Conclusions

Near infrared diffuse reflectance spectroscopy could constitute a revolution in monitoring SOC content in the field if robustness of calibration models can be improved. Moisture content undoubtedly affects the model prediction and is an environmental factor that is not easy to control while analysing soils in situ. This first laboratory experiment shows a feasible and robust method to predict SOC independent of (field) moisture content. Pre-processing the data with the EPO method allows the removal the moisture effect, to a large degree, which improves the quality of the prediction model. This method will facilitate a field method for rapid measurement of SOC content.

We are yet to test the transferability of a dried ground spectra library to a field intact soil spectra prediction using the EPO-PLS approach. In this case, there are two variables that need to be addressed: the soil structure effect and soil moisture. Since spectroscopy is a volumetric measurement, we need to deal with the soil structure issue, as the measurement also implies the change in mass by volume.

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

We propose to use 3 independent datasets, as demonstrated in this paper, for the field NIR calibration procedure:

- The calibration dataset contains soil samples with measured spectra and SOC content under standard (or laboratory) condition (air-dried). The number of samples is preferably greater than 100.
- The EPO development dataset contains spectra under laboratory condition (air-dried samples) and spectra collected under field conditions (varying soil moisture content). In this paper, we found that at least 60 samples to be optimum.
- The validation dataset contains spectra collected under field condition and measured SOC content.

The EPO development data were used to develop the EPO pre-processing of the spectra. The spectra from the calibration dataset were then transformed using EPO, and a PLS regression model (or other multivariate prediction techniques) was developed to predict SOC content from the transformed spectra. The validation dataset can be used to check the accuracy of the prediction model.

We need to further investigate the effect of the EPO spectral transformation on the prediction of other soil properties. Further work is ongoing to validate this method for field NIR measurements.

Acknowledgments

Our work is supported by the Australian Research Council through its Linkage program, and by the **French Environment and Energy Management Agency (ADEME) through its GESSOL program, project INCA. The authors thank two anonymous reviewers for their helpful comments.**

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

References

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra. Applied Spectroscopy 43, 772–777.

Bogrekci, I., Lee, W.S., 2006. Effects of soil moisture content on absorbance spectra of sandy soils in sensing phosphorus concentrations using uv-vis-nir spectroscopy. Transactions of the ASABE 49, 1175-1180.

Bowers, S.A., Hanks, R.J., 1965. Reflection of radiant energy from soils. Soil Science 100, 130-138.

Bricklemyer, R.S., Brown, D.J., 2010. On-the-go VisNIR: Potential and limitations for mapping soil clay and organic carbon. Computers and Electronics in Agriculture 70, 209-216.

Christy, C.D., 2008. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. Computers and Electronics in Agriculture 61, 10-19.

Kusumo, B.H., Hedley, C.B., Hedley, M.J., Hueni, A., Tuohy, M.P., Arnold, G.C., 2008. The use of diffuse reflectance spectroscopy for in situ carbon and nitrogen analysis of pastoral soils. Australian Journal of Soil Research 46, 623–635.

Lobell, D.B., Asner, G.P., 2001. Moisture effects on soil reflectance. Soil Science Society of America Journal 66, 722-727.

Minasny, B., McBratney, A.B., Pichon, L., Sun, W., 2009. Evaluating near infrared spectroscopy for field prediction of soil properties. Australian Journal of Soil Research 47, 664–673.

Morgan, C.L.S., Waiser, 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, 249-256.

Mouazen, A.M., Maleki, M.R., De Baerdemaeker, J., Ramon, H., 2007. On-line measurement of selected soil properties using a VIS-NIR sensor. Soil & Tillage Research 93, 13–27.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterlin, W.T., 1992. Numerical Recipes in Fortran 77. The Art of Scientific Computing, 2nd Edition. Cambridge University Press, Cambridge.

Reeves, J.B., 2010. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

to be done? Geoderma 158, 3-14.

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. Chemometrics and Intelligent Laboratory Systems 66, 191-204.

Savitzky, A. Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36, 1627–1639.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Advances in Agronomy 107, 163-215.

Sudduth, K.A., Hummel, J.W., 1993. Soil organic matter, CEC, and moisture sensing with a prototype NIR spectrometer. Transactions of the American Society of Agricultural Engineers 36, 1571–1582.

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

Waiser, T.H., Morgan, C.LS., Brown, D.J., Hallmark, C.T., 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Science Society of America Journal 71, 389-396.

Wall, M.E., Rechtsteiner, A., Rocha, L.M., 2003. Singular value decomposition and principal component analysis. In: A Practical Approach to Microarray Data Analysis. (Berrar, D.P., Dubitzky, W., Granzow, M., eds.) pp. 91-109. Kluwer, Norwell, MA.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58, 109–130.

Wu, C-Y., Jacobson, A.R., Laba, M., Baveye, P.C., 2009. Alleviating moisture content effects on the visible near-infrared diffuse-reflectance sensing of soils. Soil Science 174, 456-465

Author-produced version of the article published in Gederma, 2011, 167-168,118-124.
The original publication is available at http://www.sciencedirect.com
doi:10.1016/j.geoderma.2011.09.008

Figure captions

Figure 1. The effect of moisture on the absorbance spectra for a loamy soil with 17% clay and 1.1% SOC content. The spectra from the lower to higher absorbance are for the soil at gravimetric moisture content of 4, 5, 6, 12, 14, 16, and 20%.

Figure 2. The difference spectra of 100 soil samples (from dataset A), showing the difference in absorbance between the soil at moist condition (first day after wetting) and at air-dried.

Figure 3. Correlation between predicted SOC and measured SOC for samples at three moisture conditions using PLS calibrated on air-dried samples (dataset A).

Figure 4. External Parameter Orthogonolisation (EPO) pre-processing algorithm of the spectra. The original spectra **X** is multiplied by transformation matrix **P** to obtain transformed spectra $\mathbf{X}^*$ which is free from soil moisture effect. The transformed spectra are then used in a partial least squares (PLS) regression calibration.

Figure 5. Root mean square error (RMSE) of the prediction of SOC content on dataset A using transformed spectra at different number of EPO dimension and different number PLS factors (k).

Figure 6. (a) Original standard normal variate (snv) spectra of a soil at different moisture contents and (b) the spectra after EPO transformation.

Figure 7. Prediction of soil organic carbon content at 3 different moisture conditions on dataset A using the EPO- PLS regression.

Figure 8. Prediction of soil organic carbon content on dataset B (validation data) at different moisture contents using the EPO-PLS regression.

Figure 9. The effect of the number of samples used in EPO development on the root mean square error (RMSE) values for the prediction of SOC content on dataset A and B. The boxplot represents RMSE values at different moisture contents. The line in the middle of the boxplot represents the grand mean.