Improved estimates of organic carbon using proximally sensed vis–NIR spectra corrected by piecewise direct standardization

ARTICLE *in* EUROPEAN JOURNAL OF SOIL SCIENCE · MAY 2015

Impact Factor: 2.65 · DOI: 10.1111/ejss.12271

3 AUTHORS:

Wenjun Ji
McGill University
17 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Raphael A Viscarra Rossel
The Commonwealth Scientific and Industri…
163 PUBLICATIONS **2,467** CITATIONS
SEE PROFILE

Zhou Shi
Zhejiang University
55 PUBLICATIONS **322** CITATIONS
SEE PROFILE

European Journal of Soil Science

European Journal of Soil Science, 2015 doi: 10.1111/ejss.12271

Special issue article

Improved estimates of organic carbon using proximally sensed vis–NIR spectra corrected by piecewise direct standardization

W. Ji[a,b,c], R. A. Viscarra Rossel[b] & Z. Shi[a]

[a]*Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, 866 Yuhangtang Road, Hangzhou, 310058, China,* [b]*Land and Water Flagship, CSIRO Bruce E. Butler Laboratory, GPO Box 1666, Canberra, Australian Capital Territory 2601, Australia, and* [c]*Department of Bioresource Engineering, McGill University, 21,111 Lakeshore Road, Montreal, H9X 3V9, Canada*

Summary

We investigated the use of piecewise direct standardization (PDS) to remove the effects of water and other environmental factors from proximally sensed (field) visible–near infrared (vis–NIR) spectra. Our hypothesis was that the PDS-standardized field spectra can be used to predict soil carbon effectively with calibrations derived from existing spectroscopic databases of spectra recorded in the laboratory on dried, ground and sieved samples. In our experiments we used field spectra recorded *in situ* with a portable spectrometer at 124 sites in 11 paddy fields in Zhejiang Province, China. We sampled the soil at these same sites, recorded their spectra in the laboratory and measured their soil organic carbon (SOC) contents with a conventional laboratory technique. Two-thirds of the samples were used to relate the laboratory spectra to SOC by partial least squares regression (PLSR), and the remaining one-third was used as an independent validation dataset. We selected a representative set of samples from corresponding field and laboratory spectra that we could use as the PDS transfer set. Piecewise direct standardization was used to relate each wavelength in the laboratory spectra to the corresponding wavelength and its neighbours in the field spectra. The field spectra of the validation samples were then corrected with PDS so that they acquired the characteristics of the spectra measured under laboratory conditions. The approach was evaluated by (i) quantifying the similarity between the PDS-standardized spectra and their corresponding laboratory spectra, (ii) measuring the accuracy of their SOC predictions on the independent validation dataset and (iii) comparing these results with those of direct standardization (DS). Both PDS and DS led to considerable improvements in the predictions of SOC ($R^2 = 0.71$, $R^2 = 0.60$, respectively), compared with those with original field spectra ($R^2 = 0.03$). However, fewer transfer samples were needed with PDS to obtain similar results.

Introduction

As the largest store of terrestrial organic carbon, soil has a profound effect on the global carbon cycle. Small changes in soil organic carbon (SOC) might lead to large fluctuations in atmospheric greenhouse gas concentrations (Lal, 2004). Accurate determination of the concentration and spatial variation in SOC is also important for precision agriculture, assessment of soil erosion and degradation, and understanding biogeochemical cycles. To obtain information about SOC, it is necessary to use methods of measurement that are rapid, accurate, inexpensive and field based (Viscarra Rossel *et al.*, 2011).

Proximal soil sensing (PSS) with visible and near infrared (vis–NIR) spectroscopy can quantify SOC or soil organic matter (SOM) efficiently in the field in stationary (Ben-Dor *et al.*, 2008; Guo *et al.*, 2013; Ji *et al.*, 2014) or mobile (Mouazen *et al.*, 2007; Christy, 2008; Bricklemyer & Brown, 2010) operations. Compared with laboratory-based soil spectroscopic analysis, PSS methods need little sample preparation, are non-destructive, rapid and more affordable. However, the spectra that are measured on soil samples under field conditions are often wet and affected by other environmental factors such as surface roughness, voids and

Correspondence: Z. Shi. E-mail: shizhou@zju.edu.cn

Received 13 April 2015; revised version accepted 13 April 2015

temperature, which might mask or alter the absorption features of SOC. Therefore, it is unlikely that vis–NIR spectra recorded in the field could be used to predict SOC with spectroscopic models that use spectra recorded in the laboratory on dried, ground and sieved samples.

The global need to develop large-scale (regional, national and continental) vis–NIR spectral databases has meant that considerable time, effort and expense have been spent on developing robust spectroscopic models that can predict soil properties locally. To exploit these databases fully, it is necessary to develop methods that can remove such environmental effects from proximally sensed field spectra.

In a previous study (Ji *et al.*, 2015), we investigated use of the direct standardization (DS) algorithm (Wang *et al.*, 1991) to remove the effects of environmental factors on proximally sensed spectra. Direct standardization effectively accounted for the effects of soil water and improved predictions of SOM with proximally sensed spectra using a calibration model derived from the Chinese soil spectral database. In Ji *et al.* (2015) we showed that the DS approach removed those effects more effectively than external parameter orthogonalization (EPO) (Roger *et al.*, 2003; Minasny *et al.*, 2011) and spiking (Guerrero *et al.*, 2010).

Piecewise direct standardization (PDS) (Wang *et al.*, 1991) is similar to DS in that PDS relates the absorbance of selected standard spectra recorded in the laboratory to their field spectra with the same model structure. However, in PDS the standardization is accomplished by using neighbouring wavelengths that are within a selected window size instead of an entire spectrum as in DS. Therefore, PDS will not only enable correction of differences in intensity but also wavelength shifts and peak broadening (Feudale *et al.*, 2002). Previous studies have used PDS successfully for the standardization of vis–NIR spectra as well as electrochemical and nuclear magnetic resonance data (Bouveresse & Massart, 1996; Herrero & Ortiz, 1997; Alam *et al.*, 2009).

Here, we investigated the use of PDS to correct for the effects of water and other environmental factors from proximally sensed spectra to improve the prediction of SOC content. Our hypothesis was that PDS would outperform DS because the nonlinearities between field and laboratory spectra are modelled better by several local multivariate models than by a single global multivariate model as used in DS. Our aims were to: (i) describe PDS and its use to correct for the differences between spectra recorded under field conditions and in the laboratory, (ii) determine whether spectroscopic calibrations developed with laboratory spectra can be used to predict SOC using PDS-standardized field spectra and (iii) compare PDS and DS.

Materials and methods

Proximal vis–NIR sensing, soil sampling and laboratory measurements

A total of 124 sensing sites were selected at random from within 11 paddy fields (Table 1) around the China National Rice Research Institute in Zhejiang province, China. This is a major rice-producing region with more than 2000 years of history of continuous cultivation (Wissing *et al.*, 2011). The sites ranged in latitude from 29°03′N to 30°10′N and in longitude from 119°10′E to 122°48′E. The fields ranged in size from 0.25 to 1 ha and had been under rice cultivation for more than 10 years. The characterization of the 11 paddy fields is described in Table 1. We recorded the soil spectra by proximal *in situ* and stationary vis–NIR sensing, and at the same sites we also collected soil samples. These samples were transported to our laboratory for further measurements. The paddy fields were drained and the soil was allowed to dry for 10 days before sensing and sampling.

At each sensing site, the water content of the surface soil (0–20 cm) was measured twice by a FieldScout TDR 300 soil moisture meter (Spectrum Technologies Inc., Aurora, IL, USA) with a 20-cm guide, after which the measurements were averaged and recorded. A cubic soil sampler (10 cm × 10 cm × 20 cm) was used to take the soil sample. Then, 10 spectra were recorded at each of three randomly selected locations on one vertical side of the sample within the A-horizon. The average of the 30 spectra would represent the spectrum for the site. Before measurement, the soil sample was flattened, its surface made even and areas with stones, roots or voids were avoided. Subsequently, we refer to these spectra

Table 1 Characterization of the paddy fields studied

Field	N[a]	Soil genus	Subgroup	Texture classes	Soil parent materials
1	8	Red–purple mudstone soil	Submerged paddy soil	Clay loam	Purplish red sandstone slope deposit
2	10	Redeposited red soil	Hydrogric paddy soil	Clay loam	Red clay slope deposit
3	6	Fluvigenic loamy soil	Hydrogric paddy soil	Clay loam	Mainly streams alluvial deposit, also proluvial deposit
4	6	Fluvigenic loamy soil	Hydrogric paddy soil	Clay loam	Mainly alluvial deposit, also proluvial deposit
5	5	Fluvigenic loamy soil	Hydrogric paddy soil	Clay loam	Mainly alluvial deposit, also proluvial deposit
6	8	Redeposited red–purple sandstone soil	Hydrogric paddy soil	Clay loam	Purple sandy shale slope deposit
7	8	Fluvigenic loamy soil	Perco genic paddy soil	Clay loam	Recent river alluvial deposit
8	24	Calcareous purple mudstone soil	Perco genic paddy soil	Loam clay	Purple sandy shale slope deposit
9	29	Blue clayey soil	Degleyed paddy soil	Silty clay	Ancient lake and marine sediments
10	10	Blue clayey soil	Degleyed paddy soil	Silty clay	Ancient lake and marine sediments
11	10	Blue clayey soil	Degleyed paddy soil	Silty clay	Ancient lake and marine sediments

[a]N is the number of soil samples taken in individual fields.

as field spectra, or $\mathbf{X}_{\mathbf{F}}$. In this way, we recorded a total of 124 spectra measured under field conditions.

After the field spectroscopic measurement, samples of soil of 2 cm × 2 cm were taken at the three sensing locations within the larger 10 cm × 10 cm × 20 cm sample, and then bagged, labelled and taken to the laboratory. The 124 soil samples were air-dried, ground and sieved to a size fraction of less than 2 mm. The vis–NIR spectra of these samples were recorded again under laboratory conditions. Hereafter, we refer to these spectra as laboratory spectra, or $\mathbf{X}_{\mathbf{L}}$.

In the laboratory, we also measured the SOC content of the 124 samples by dry combustion at 1100°C with a multi N/C® 3100 (Analytik Jena AG, Jena, Germany).

For the spectroscopic measurements we used a Fieldspec Pro FR vis–NIR spectrometer and its high intensity contact probe (PANalytical, B.V, Boulder, Colorado, USA, formerly Analytical Spectral Devices) for both field and laboratory measurements. The spectroradiometer has a spectral range between 350 and 2500 nm, and a spectral resolution of 3 nm at 700 nm and 10 nm at 1400 and 2100 nm. The spectra have a sampling resolution of 1 nm. A Spectralon® panel was used to calibrate the spectroradiometer before field and laboratory measurements were made at each site.

Noisy spectra with wavelengths between 350–499 and 2451–2500 nm were removed. The reflectance spectra were transformed to the apparent absorbance ($\log_{10}$ 1/R), and the first derivatives were extracted to remove the additive baseline (Næs *et al.*, 2002).

Spectroscopic calibration of SOC

Two-thirds of the data (82 samples) were used to perform our experiments and the remaining third (42 samples) was used to test their outcomes. The selection was done with the Kennard-Stone (K-S) algorithm (Kennard & Stone, 1969). Partial least squares regression was used to predict SOC with a model derived from the laboratory spectra of the 82 samples. The predictions were tested on the 42 samples and their (i) laboratory spectra, (ii) field spectra, (iii) field spectra processed with PDS and (iv) field spectra processed with DS.

The piecewise direct standardization (PDS) algorithm

Piecewise direct standardization was developed to transfer spectra measured with one instrument so that it would seem they were measured by another (Wang *et al.*, 1991). We used the following steps to apply the PDS in our study.

1 We selected a set of transfer samples from the experimental dataset (82 samples) with the K-S algorithm, and measured the spectra under laboratory and field conditions. We used K-S because the transfer samples should be selected to represent the range of variation in the experimental dataset and be able to characterize the differences between the field and laboratory spectra.
2 The PDS transfer parameters were determined by creating a linear relationship between the laboratory and field spectra of the transfer samples; at each wavelength, the apparent absorbance of transfer samples in the laboratory spectra was regressed against absorbance in the field spectra that corresponded to the particular wavelength and its neighbours within a predefined window. The partial least squares regression (PLSR) algorithm (Wold *et al.*, 2001) was used to formulate the regression models. The PDS transfer matrix was formed by recording the set of regression coefficients obtained from the PLSR models derived from within the moving window along a spectrum from beginning to end.
3 The field spectra from the validation dataset were then standardized using the PDS parameters so that they were directly comparable to those measured in the laboratory, which removed the environmental effects. An overview of the PDS algorithm is shown in Figure 1.

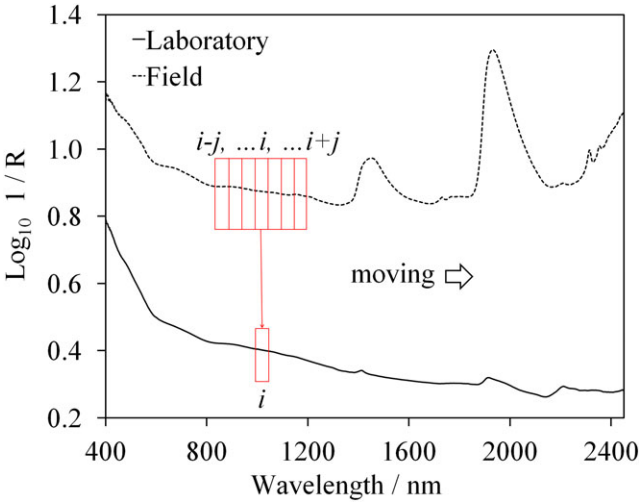

Figure 1 Overview of the piecewise direct standardization (PDS) algorithm.

To implement PDS one needs to know the number of transfer samples, the optimal number of PLSR factors and the size of the wavelength window to use. The PDS algorithm is described in the Appendix. For DS, which uses the entire spectra instead of the predefined window for the transfer, only the number of transfer samples is needed. We direct the reader to Ji *et al.* (2015) for details of the DS algorithm.

Implementation of PDS and comparison with DS

We investigated the effects on PDS of different numbers of transfer samples, the size of the wavelength window and the optimal number of PLSR factors:

1 With the K-S method we selected a series of transfer subsets, t, of corresponding field and laboratory spectra from the experimental dataset, where $t = 3, 4, 5, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80$ and 82.
2 We applied PDS when $j = 1, 2, 5, 10, 15, 20, 25$, and the corresponding window sizes were $ws = 3, 5, 11, 21, 31, 41$

and 51 nm, respectively. Because PDS works with a moving window of data, edge effects can occur at the beginning and end of the spectra. To make our results comparable, we removed wavelengths at 500–524 and 2426–2450 nm after each transfer procedure (to a maximum $j = 25$).

3 We used different numbers of PLSR factors ($NF = 1$, 2, 3, 5, 7 and 10) in the local PDS regressions. When 10 PLSR factors were used to compute the transfer coefficient, at least 12 transfer samples were needed in the PDS algorithm. When determining the optimal number of PLSR factors to make the results comparable, PDS transfers were started with 15 transfer samples.

For comparability with PDS, we did the DS transfer with the same number of transfer samples.

Spectral assessment of the standardized field spectra

In both PDS and DS, the field spectra were standardized for direct comparison with the laboratory spectra. The spectral correlation fitting (F) and spectral angles mapper (θ) (Kruse *et al.*, 1993) were used to measure their performance. The techniques are based on the degree of agreement between the standardized PDS and DS field spectra and their corresponding laboratory spectra.

The spectral correlation fitting (F) is defined by:

$$F = \frac{n\sum_{i=1}^{n} x_L x_F - \sum_{i=1}^{n} x_L \sum_{i=1}^{n} x_F}{\sqrt{\left[n\sum_{i=1}^{n} x_L^2 - \left(\sum_{i=1}^{n} x_L\right)^2\right]\left[n\sum_{i=1}^{n} x_F^2 - \left(\sum_{i=1}^{n} x_F\right)^2\right]}}, \quad (1)$$

and the spectral angles mapper (θ) by:

$$\theta = \arccos \frac{\sum_{i=1}^{n} x_L x_F}{\sqrt{\sum_{i=1}^{n} x_L^2}\sqrt{\sum_{i=1}^{n} x_F^2}}, \cdots \theta \in \left[0, \frac{\pi}{2}\right], \quad (2)$$

where x_L and x_F are the spectral responses of laboratory and field spectra at each wavelength, i, and n is the total number of wavelengths. Here, $n = 1901$ (525–2425 nm).

Values of F close to 1 and the smaller the angle θ indicate greater similarity between the standardized field spectra and laboratory spectra.

Predictions of SOC with the standardized field spectra

With the optimal parameters, the PDS- and DS-standardized field spectra were validated by the PLSR model formulated with the laboratory spectra of the experimental dataset (82 samples). The accuracy of predicted SOC values obtained with PDS and DS was determined by the root mean squared error (RMSE), the coefficient of determination (R^2) and the ratio of prediction to deviation (RPD). The predictability of the models was evaluated by Chang *et al.* (2001), where calibrations with RPD > 2.0 are considered good, those with RPD of 1.4–2.0 are considered moderate and calibrations with RPD of less than 1.4 are considered unacceptable.

The statistical and chemometric analyses were performed with the software R 2.15.0 (R Development Core Team, 2012) and ParLeS 3.1 (Viscarra Rossel, 2008).

Results

The volumetric water content of the 124 soil samples was between 40 and 50%, and the summary statistics of SOC (g kg^{-1}) values are given in Table 2. The SOC values were transformed to common logarithms, $\log_{10}$, to ensure that the data had a near-normal distribution before spectroscopic modelling.

Determination of parameters used in the PDS

If the standardized spectra matched the laboratory spectra after spectral transfer by PDS and DS, the effects of water and other environmental factors were deemed to have been removed successfully. In PDS, first we assumed that most of the information was obtained by the first PLSR factor. Graphs of the spectral correlation fitting, F, and the spectral angles mapper, θ, for different transfer sets and window sizes are shown in Figure 2(a,b, respectively). The similarity between the standardized field spectra and the laboratory spectra increased when the window size factor, j, was smaller than 15 nm ($ws < 31$ nm), whereas it stabilized when j was larger than 15 nm ($ws > 31$ nm). Therefore, in this study we used 31 nm as the window size parameter for PDS. The two spectral matching factors, F and θ, became more stable when we selected 20 or more samples

Table 2 Summary statistics of soil organic carbon (SOC) of the samples used in this study

	n	Mean	SD	Median	Min.	Max.	Skewness coefficient
SOC / g kg^{-1}	124	16.40	7.98	14.49	4.12	36.29	422.81
	82	17.31	8.65	14.66	4.12	36.29	481.93
	42	14.63	6.19	14.42	4.13	32.99	117.45
$\log_{10}$ (SOC g kg^{-1})	124	1.16	0.23	1.16	0.61	1.56	−0.01
	82	1.18	0.23	1.17	0.61	1.56	−0.01
	42	1.12	0.21	1.16	0.62	1.52	−0.01

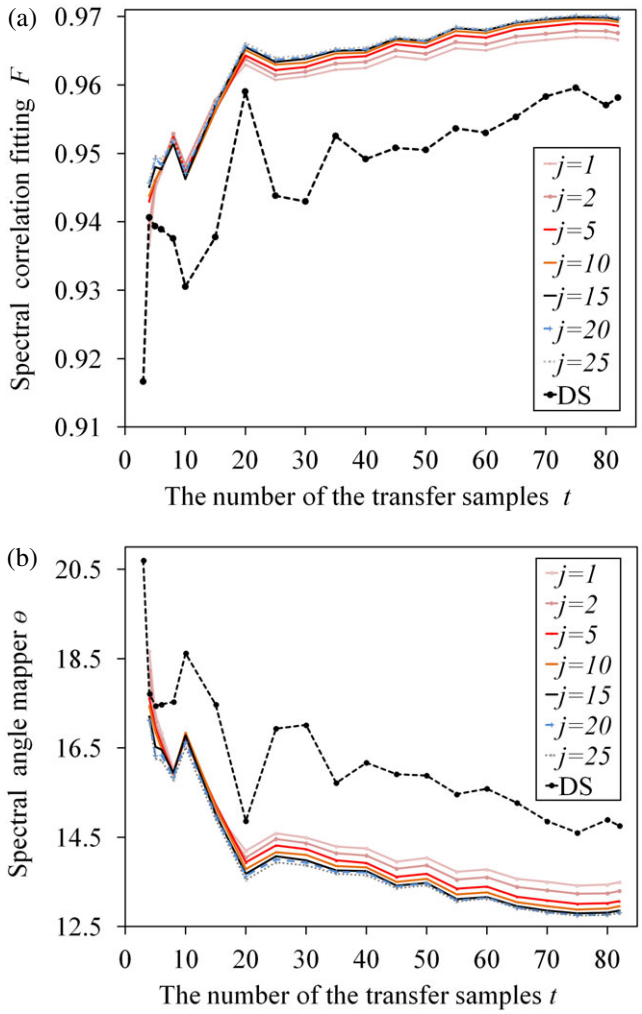

Figure 2 (a) The spectral correlation fitting (F) and (b) spectral angles mapper (θ) used to measure the standardization performance of the piecewise direct standardization (PDS) algorithm with different numbers of transfer samples and different window sizes.

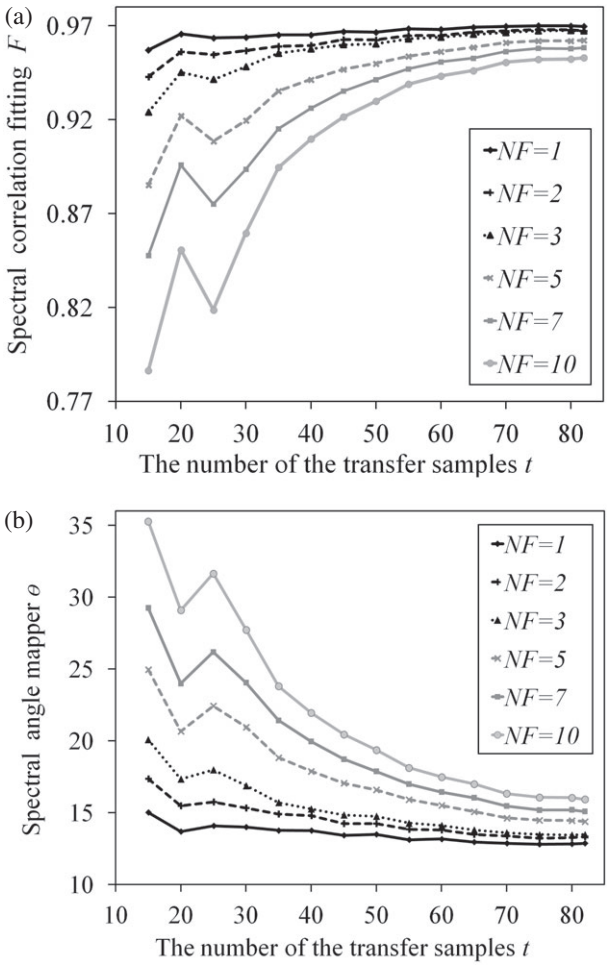

Figure 3 (a) The spectral correlation fitting (F) and (b) spectral angles mapper (θ) used to measure the standardization performance of the piecewise direct standardization (PDS) algorithm with different numbers of transfer samples and partial least squares regression (PLSR) factors with a window size of 31 nm.

from the transfer set, whereas they fluctuated when fewer transfer samples were used.

Figure 3 shows the performance of the PDS-standardized field spectra with different numbers of transfer samples and PLSR factors for a window size of 31 nm. Values of θ increased and those of F decreased as the number of PLSR factors increased. The performance of PDS decreased as the number of PLSR factors increased in the PDS regressions. The use of a single PLSR factor produced the best transfer.

Figure 4 shows the first-derivative absorbance spectra of three randomly selected samples from the 42 validation samples measured under both laboratory and field conditions and their PDS-standardized ($NF = 1$, $j = 15$ nm and number of transfer samples = 20) field spectra. The field and laboratory spectra differed mainly at the wavelengths of water absorption, around 1400 and 1900 nm. With PDS, the standardized field spectra were similar to the corresponding laboratory spectra (Figure 4) and almost coincided. The field spectra standardized by PDS were similar to the original laboratory spectra, especially at wavelengths that characterize absorption related to soil water.

The PDS algorithms for prediction of SOC

The PDS-standardized field spectra were similar to their corresponding laboratory spectra (Figure 4); however, the predictability of spectroscopic models for SOC derived from them needs to be tested further.

A PLSR model with 10 factors, determined by leave-one-out cross-validation, was formulated with the laboratory spectra and SOC content of 82 samples (Table 3). Predictions of SOC for laboratory spectra of the 42 validation samples had an R^2 of 0.93, RMSE of 0.06 $\log_{10}$ (SOC g kg^{-1}) and RPD of 3.35, whereas the predictions of SOC from the field spectra were poor, with R^2 of 0.03, RMSE of 0.56 $\log_{10}$ (SOC g kg^{-1}) and RPD of 0.38.

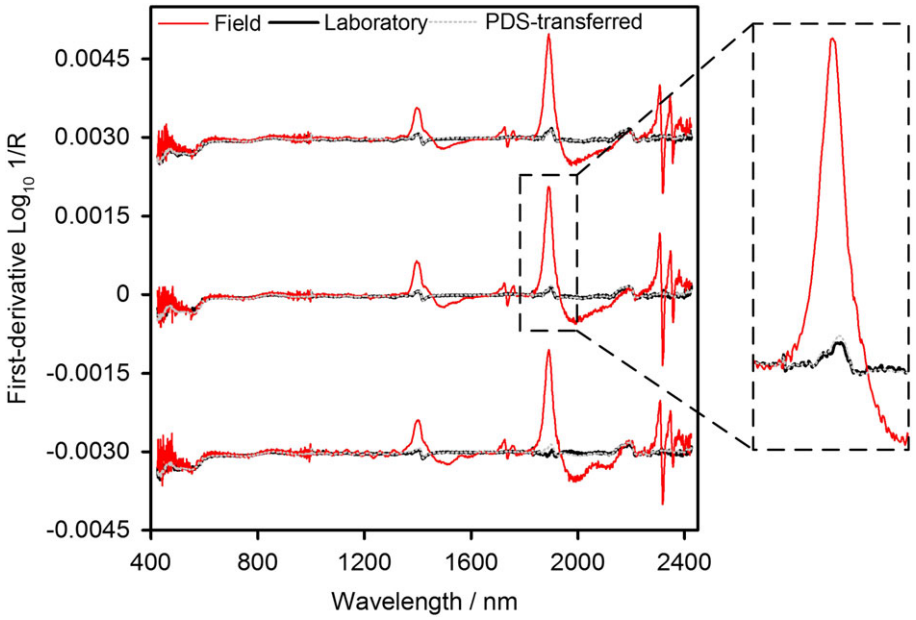

Figure 4 First-derivative absorbance spectra of three samples selected at random from the 42 validation samples: field spectra, laboratory spectra and the field spectra standardized with the piecewise direct standardization (PDS) algorithm ($NF = 1$, $j = 15$ nm and number of transfer samples = 20). An offset has been added to each spectrum for clarity.

Table 3 Prediction of soil organic carbon (SOC) for samples with laboratory spectra, field spectra and standardized field spectra with piecewise direct standardization (PDS) and direct standardization (DS) using partial least squares regression (PLSR) calibrated on air-dried and ground training samples

		RMSE			
Validation dataset	R^2	Log_{10} (SOC g kg^{-1})	SOC_{min} / g kg^{-1}	SOC_{max} / g kg^{-1}	RPD
Laboratory 42	0.93	0.06	0.61	5.38	3.35
Field 42	0.03	0.56	10.84	95.47	0.38
Field PDS 42 ($t = 20$)	0.71	0.12	1.31	11.55	1.74
Field DS 42 ($t = 20$)	0.52	0.16	1.84	16.17	1.36

PDS = piecewise direct standardization; DS = direct standardization; t = number of transfer samples used in the transfer procedure.
RMSE values at SOC_{min} and SOC_{max} are back-transformed to original scale.

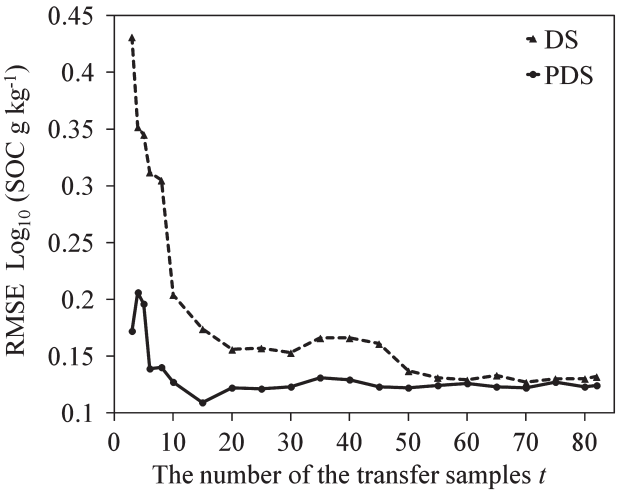

Figure 5 Accuracy of predictions when different numbers of transfer samples were used in piecewise direct standardization (PDS) and direct standardization (DS).

The accuracy of SOC predictions (RMSE) from validation of the 42 field spectra standardized by PDS is shown in Figure 5. Predictions of SOC were relatively stable when 20 or more transfer samples were used in the PDS transfer, and were similar to the results obtained with spectral matching (Figure 2). The prediction of SOC with 20 transfer samples and the PDS algorithm was $R^2 = 0.71$, RMSE = 0.12 $\log_{10}$ (SOC g kg^{-1}) and RPD = 1.74. This approach predicted SOC with an error of 1.31 g kg^{-1} for soil with small SOC contents (4.12 g kg^{-1}) and 11.57 g kg^{-1} for soil rich in SOC (36.29 g kg^{-1}).

Comparison of PDS and DS

We selected 50 transfer samples to use in the DS (Figures 2, 5) because the two spectral matching factors became more or less stable with this number. The prediction of SOC with DS-standardized spectra and 50 transfer samples resulted in an R^2 value of 0.60, a RMSE of 0.14 $\log_{10}$ (SOC g kg^{-1}) and RPD of 1.55. There were considerable improvements in the predictions made with the DS-standardized spectra compared with those that used the original field spectra.

When comparing the spectral matching parameters of the field spectra standardized by PDS and DS with laboratory spectra (Figure 2) with the same number of transfer samples regardless of the size of the wavelength window, field spectra standardized by PDS produced better results than those standardized by DS (Figure 2).

Comparisons of the prediction of SOC with PDS- and DS-standardized field spectra are shown in Figure 5. When

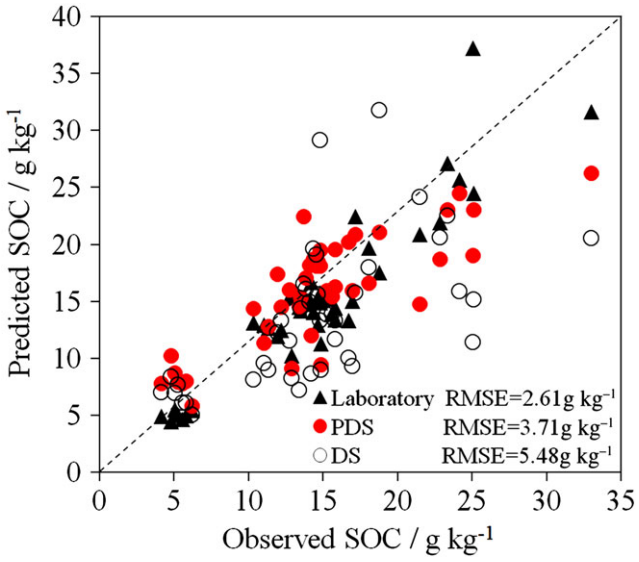

Figure 6 Scatter plot of the predicted and observed values of soil organic carbon (SOC).

fewer than 50 samples were used in the transfer procedures, the predictions obtained from PDS-transferred field spectra were better than those from DS-standardized field spectra. The RMSE values with PDS-standardized spectra were smaller than those obtained with DS-standardized spectra. However, the accuracy of predictions of the two methods was comparable when 50 or more samples were used. Similar predictions were obtained with 60 transfer samples between DS ($R^2 = 0.62$, RMSE = 0.13 $\log_{10}$ (SOC g kg^{-1}) and RPD = 1.65) and PDS ($R^2 = 0.70$, RMSE = 0.13 $\log_{10}$ (SOC g kg^{-1}) and RPD = 1.68), which was also the case when more transfer samples were used.

Predictions of SOC by both PDS and DS methods were not as good as those from laboratory spectra only. In terms of RMSE, R^2 and RPD, the predictions made with PDS- and DS-standardized field spectra were considerably better than those made with the original field spectra. The prediction of SOC with 20 transfer samples and the PDS algorithm with the other two optimized parameters was more accurate than that with 50 transfer samples and the DS method ($R^2 = 0.60$, RMSE = 0.14 $\log_{10}$ (SOC g kg^{-1}) and RPD = 1.55), and was better than that obtained with the same number of transfer samples ($t = 20$) with the DS method ($R^2 = 0.52$, RMSE = 0.16 $\log_{10}$ (SOC g kg^{-1}) and RPD = 1.36). Figure 6 shows the predicted values of SOC plotted against the observed values. To help with interpretation, we back-transformed the $\log_{10}$ predictions to the original scale.

Discussion

Artefacts occur in the transfer spectra if the transfer parameters are not well determined (de Noord, 1994). The main causes of artefacts are the size of the local wavelength region, the number of transfer samples and the number of PLSR factors used in the local regressions. By comparing the similarity between the PDS-standardized field spectra and laboratory spectra, we determined that, for our experiment, a 31-nm window size was adequate. As many soil materials have diagnostic absorption features that are 20–40 nm wide at half the band depth (Hunt, 1977), the field spectra within a window size of 31 nm have stronger correlations with laboratory spectra than those within the entire wavelength region. The results of the present study indicated that the first PLSR factor only should be used in the local regressions in PDS transfer. This is because of collinearity of the spectra in the small wavelength domain with fewer transfer samples.

As for the DS algorithm, a subset of samples that represented the entire experimental dataset well was needed to measure the difference in the response of spectra measured under field and laboratory conditions. Too few or too many samples in the transfer set can lead to under- or over-fitting, which reduces the performance of the standardization. Here, we determined the optimal number of transfer samples by using both the spectral similarities and the ensuing prediction accuracies of SOC. The optimal number of transfer samples used for the PDS algorithm was 20, whereas 50 transfer samples were needed for the DS algorithm. Better transfer performance was achieved by PDS with fewer transfer samples because the local rank of each window in PDS was smaller than the rank of the total data matrix used in DS. Air-drying, grinding and measuring only 16% (20 out of 124) of the samples in the laboratory will take less time than for the whole set. However, the ratio of

Table 4 Comparison of the performances of external parameter orthogonalization (EPO), direct standardization (DS), piecewise direct standardization (PDS) and spiking

Requirements	EPO	DS	PDS	Spiking
If transfer samples are needed	Y	Y	Y	Y
Number of transfer samples needed				
In this study	–	50 (82, 42)	20 (82, 42)	–
Minasny *et al.* (2011)	60 (271, 20)	–	–	–
Ge *et al.* (2014)	177 (2017, 58)	–	–	–
Ji *et al.* (2015)	50 (1581, 54)	50 (1581, 54)	–	50 (1581, 54)
Spectral pre-treatment	N	N	Y or N	N
Recalibration	Y	N	Y or N	Y

The numbers in parentheses represent the number of samples from the calibration and validation sets, respectively.

transfer samples to the whole dataset will depend on the variation in the soil and environment in the study region.

In this study, the local PDS regressions derived in small windows enabled better and more detailed standardization of the laboratory and field spectra than DS, and PDS was good at correcting the more subtle differences between them. More accurate predictions of SOC were obtained with PDS than with DS when a smaller number of transfer samples were used. However, for PDS we needed to take the first derivatives of the spectra, whereas this was not necessary for DS. The use of PDS with the original spectra produced artefacts in the wavelength regions that are commonly attributed to water.

In a previous study (Ji *et al.*, 2015) we reported that DS performed better than EPO and spiking for removing the effects of water and environmental factors from proximally sensed spectra. Table 4 summarizes the comparison between PDS, DS, EPO and spiking for removing the effects of water and environmental factors from proximally sensed (field) spectra. Readers are directed to Ji *et al.* (2015) for details on DS, EPO and spiking.

Conclusions

We concluded that both PDS and DS can remove the effects of water and other environmental factors from proximally sensed field spectra. The standardized field spectra can be used to predict SOC effectively from spectroscopic calibrations formulated with laboratory spectra. Fewer transfer samples were needed to implement PDS adequately than to implement DS. In this study, the ratio of transfer samples to the whole dataset was 16% for PDS and between 40 and 50% for DS. Compared with DS, which broadly removes the effects of water and environmental factors on field spectra, PDS works more locally. Therefore, it can be used to remove these effects, but it can also account for the more subtle differences between laboratory and field spectra. Adequate spectral pre-treatment might be required before the use of PDS, whereas it is not necessary for DS.

Our results indicate the potential of *in situ* proximal vis–NIR sensing to estimate rapidly the soil organic carbon content of local soil with calibrations developed with large laboratory-derived spectroscopic databases.

Acknowledgements

Our work was supported by the Research Fund of the State Key Laboratory of Soil and Sustainable Agriculture, Nanjing Institute of Soil Science, Chinese Academy of Science (Y412201430), National Natural Science Foundation of China (No.41271234) and Zhejiang University Project 985 for the PhD scholarship of Wenjun Ji. Wenjun Ji thanks the Commonwealth Scientific and Industrial Research Organisation's (CSIRO) Land & Water Flagship for hosting her visit there. Raphael Viscarra Rossel also thanks the Land & Water Flagship for their support.

References

Alam, T.M., Alam, M.K., McIntyre, S.K., Volk, D.E., Neerathilingam, M. & Luxon, B.A. 2009. Investigation of chemometric instrumental transfer methods for high-resolution NMR. *Analytical Chemistry*, **81**, 4433–4443.

Ben-Dor, E., Heller, D. & Chudnovsky, A. 2008. A novel method of classifying soil profiles in the field using optical means. *Soil Science Society of America Journal*, **72**, 1113–1123.

Bouveresse, E. & Massart, D.L. 1996. Standardization of near-infrared spectrometric instruments: a review. *Vibrational Spectroscopy*, **11**, 3–15.

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

Chang, C.W., Laird, D.A., Mausbach, M.J. & Hurburgh, C.R. 2001. Near-infrared reflectance spectroscopy – principal components regression analyses of soil properties. *Soil Science Society of America Journal*, **65**, 480–490.

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

Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D. & Ferre, J. 2002. Transfer of multivariate calibration models: a review. *Chemometrics & Intelligent Laboratory Systems*, **64**, 181–192.

Ge, Y., Morgan, C.L. & Ackerson, J.P. 2014. VisNIR spectra of dried ground soils predict properties of soils scanned moist and intact. *Geoderma*, **221**, 61–69.

Guerrero, C., Zornoza, R., Gomez, I. & Mataix-Beneyto, J. 2010. Spiking of NIR regional models using samples from target sites: effect of model size on prediction accuracy. *Geoderma*, **158**, 66–77.

Guo, Y., Ji, W., Wu, H. & Shi, Z. 2013. Estimation and mapping of soil organic matter based on Vis–NIR reflectance spectroscopy. *Spectroscopy & Spectral Analysis*, **33**, 1135–1140 (in Chinese).

Herrero, A. & Ortiz, M.C. 1997. Multivariate calibration transfer applied to the routine polarographic determination of copper, lead, cadmium and zinc. *Analytica Chimica Acta*, **348**, 51–59.

Hunt, G.R. 1977. Spectral signatures of particulate minerals in the visible and near infrared. *Geophysics*, **42**, 501–513.

Ji, W., Shi, Z., Huang, J. & Li, S. 2014. In situ measurement of some soil properties in paddy soil using visible and near-infrared spectroscopy. *PLoS ONE*, **9**, 1–11.

Ji, W., Viscarra Rossel, R.A. & Shi, Z. 2015. Accounting for the effects of water and the environment on proximally sensed vis–NIR spectra and their calibrations. *European Journal of Soil Science*, doi: 10.1111/ejss.12239.

Kennard, R.W. & Stone, L.A. 1969. Computer aided design of experiments. *Technometrics*, **11**, 137–148.

Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J. *et al.* 1993. The spectral image processing system (SIPS) – interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, **44**, 145–163.

Lal, R. 2004. Soil carbon sequestration impacts on global climate change and food security. *Science*, **304**, 1623–1627.

Minasny, B., McBratney, A.B., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L. *et al.* 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma*, **167–168**, 118–124.

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

Næs, T., Isaksson, T., Fern, T. & Davies, T. 2002. *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester.

de Noord, O.E. 1994. Multivariate calibration standardization. *Chemometrics & Intelligent Laboratory Systems*, **25**, 85–97.

R Development Core Team 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0 [WWW document]. URL http://www.R–project.org [accessed on 1 March 2012].

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 & Intelligent Laboratory Systems*, **66**, 191–204.

Viscarra Rossel, R.A. 2008. ParLeS: software for chemometric analysis of spectroscopic data. *Chemometrics & Intelligent Laboratory Systems*, **90**, 72–83.

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J. & Lobsey, C. 2011. Proximal soil sensing: an effective approach for soil measurements in space and time, Chapter 5. *Advances in Agronomy*, **113**, 237–283.

Wang, Y., Veltkamp, D.J. & Kowalski, R. 1991. Multivariate instrument standardization. *Analytical Chemistry*, **63**, 2750–2756.

Wang, Z., Dean, T. & Kowalski, B.R. 1995. Additive background correction in multivariate instrument standardization. *Analytical Chemistry*, **67**, 2379–2385.

Wissing, L., Kölbl, A., Vogelsang, V., Fu, J.R., Cao, Z.H. & Kögel–Knabner, I. 2011. Organic carbon accumulation in a 2000–year chronosequence of paddy soil evolution. *Geoderma*, **87**, 376–385.

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

Appendix

The piecewise direct standardization (PDS) algorithm used in this study

We measured the transfer samples in both environments, field and laboratory, and used the spectral responses to create a linear relationship between the two measurements as in the DS algorithm (Wang *et al.*, 1991, 1995; Ji *et al.*, 2015).

$$\mathbf{X}_{\mathrm{L}} = \mathbf{X}_{\mathrm{F}}\mathbf{B} + \mathbf{E}, \tag{A1}$$

Both $\mathbf{X}_{\mathrm{L}}$ and $\mathbf{X}_{\mathrm{F}}$ matrices have a size of $t \times p$, where t represents the number of transfer spectra and p the number of wavelengths. We tested different numbers of transfer spectra, and we describe how we did this next. The transfer matrix, $\mathbf{B}$, of size $p \times p$ of unknown parameters accounts for the variation in both $\mathbf{X}_{\mathrm{L}}$ and $\mathbf{X}_{\mathrm{F}}$, and $\mathbf{E}$ is the residual matrix. A detailed description of the calculation of $\mathbf{E}$ is given in Wang *et al.* (1995) and Ji *et al.* (2015).

To calculate the transfer matrix $\mathbf{B}$, first we removed the additive noise denoted by $\mathbf{E}$ (Equation (1)), by mean centring the laboratory spectra $\mathbf{X}_{\mathrm{L}}$ and field spectra $\mathbf{X}_{\mathrm{F}}$ of the transfer samples. Therefore, Equation (A1) may be rewritten as:

$$\overline{\mathbf{X}}_{\mathrm{L}} = \overline{\mathbf{X}}_{\mathrm{F}}\mathbf{B}, \tag{A2}$$

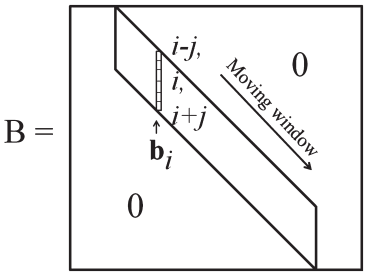

Figure A1 Structure of the diagonal transfer matrix **B** for the piecewise direct standardization (PDS) algorithm showing the transfer coefficients.

where $\overline{\mathbf{X}}_{\mathrm{L}}$ and $\overline{\mathbf{X}}_{\mathrm{F}}$ represent the mean-centred spectra of $\mathbf{X}_{\mathrm{L}}$ and $\mathbf{X}_{\mathrm{F}}$, respectively.

As an improvement over the DS algorithm, PDS relates the spectral response of the transfer samples recorded at wavelength i in the laboratory (denoted by $\mathbf{X}_{\mathrm{L}i}$) to the wavelengths in a window around i (denoted by $\mathbf{X}_{\mathrm{F}i}$) recorded in the field (Wang *et al.*, 1991). The wavelength windows are made up of a set of spectral responses recorded at each wavelength (denoted by $\mathbf{x}_{\mathrm{F}i}$), and symmetrical around every wavelength in this study as follows:

$$\mathbf{X}_{\mathrm{F}i} = \left[\mathbf{x}_{\mathrm{F}(i-j)}, \mathbf{x}_{\mathrm{F}(i-j+1)}, \ldots\ldots, \mathbf{x}_{\mathrm{F}(i-1)}, \mathbf{x}_{\mathrm{F}i}, \mathbf{x}_{\mathrm{F}(i+1)}, \ldots\ldots\mathbf{x}_{\mathrm{F}(i+j-1)}, \mathbf{x}_{\mathrm{F}(i+j)}\right]. \tag{A3}$$

The laboratory spectra at wavelength i were then regressed on to the field spectra with data from within the selected window as follows:

$$\mathbf{x}_{\mathrm{L}i} = \mathbf{X}_{\mathrm{F}i}\mathbf{b}_i, \tag{A4}$$

where $\mathbf{x}_{\mathrm{L}i}$ is a vector of the spectral value at wavelength i with a size of $t \times 1$, $\mathbf{X}_{\mathrm{F}i}$ has a size of $t \times ws$, where ws is the window size, $ws = 2j + 1$ and $\mathbf{b}_i$ is a vector with the transfer coefficient for the *ith* wavelength with a size of $ws \times 1$.

Partial least squares regression was used to calculate the transfer coefficient $\mathbf{b}_i$. During movement of the *ith* wavelength from the beginning to the end of the entire wavelength range, a set of the transfer coefficients was obtained for each wavelength. The transfer matrix was formed diagonally as follows:

$$\mathbf{B} = \mathrm{diag}\left(\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2^{\mathrm{T}}, \ldots, \mathbf{b}_i^{\mathrm{T}}, \ldots, \mathbf{b}_m^{\mathrm{T}}\right), \tag{A5}$$

where m is the number of the wavelengths used in the PDS algorithm and the other elements are zero. The transfer coefficients and the diagonal transfer matrix are shown in Figure A1.

Standardization of the field spectra was done with the validation set and the same model structure as in Equation (A1):

$$\mathbf{X}'_{\mathrm{F}42} = \mathbf{X}_{\mathrm{F}42}\mathbf{B} + \mathbf{E}. \tag{A6}$$