Geoderma 298 (2017) 46–53

Contents lists available at ScienceDirect

Geoderma

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

Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation

Radim Vašát *, Radka Kodešová, Aleš Klement, Luboš Borůvka

Czech University of Life Sciences Prague, Faculty of Agrobiology, Food and Natural Resources, Department of Soil Science and Soil Protection, Kamýcká 129, 165 00 Prague-Suchdol, Czech Republic

ARTICLE INFO

Article history:
Received 25 August 2016
Received in revised form 8 March 2017
Accepted 11 March 2017
Available online 24 March 2017

Keywords:
Soil organic matter
Visible– and near infrared spectroscopy
Signal pre-treatment
Predictive modeling
PLSR

ABSTRACT

As there is no single (or combination of) signal pre-processing method that works best with all data sets, choosing the most feasible one is a key aspect in soil diffuse reflectance spectroscopy in the visible– and near infrared region (400–2500 nm). The commonly used pre-processing methods include tools for spectra smoothing and/or noise reduction (e.g. Savitzky-Golay (SG) filtering or discrete wavelet transformation (DWT)), light scatter correction (multiplicative scatter correction (MSC), standard normal variate (SNV)), baseline normalization techniques to cope with vertical offset and/or slope effects (e.g. continuum removal (CR), first and second order derivative (FD and SD)), as well as other transformations (e.g. logarithmic-log(1/R)). All of these tools are aimed at eliminating or reducing unwanted side effects (artifacts) in the spectra and at enhancing the recognition of relevant information. For soil organic carbon content estimation using partial least square regression calibration technique, smoothing with SG filter and (or in combination with) CR usually ensures a reliable estimation. However, the common CR may suffer from a few shortcomings. An approximation is applied to connect the pivot points of the spectrum in order to derive a continuum, but more problematically, the CR procedure does not recognize the true essence of the vertical shift at the very beginning of the spectra (the CR value always equals one at that point). Therefore, we decided to modify the procedure in the way that the reflectance values at respective wavelengths were divided not by the continuum, but by the maximal reflectance value of the particular spectrum. This correction by the maximum reflectance (CMR) pre-processing was tested in comparison with eight other above mentioned methods at four different study sites that differ in the prevailing soil units. As a result, on site 1 (Haplic Chernozem), we achieved a significantly improved prediction accuracy using the CMR ($R^2_{cv} = 0.845$) compared to raw (but smoothed) soil spectra (0.815). On site 2 (Rendzic Leptosol), the most accurate prediction was achieved equally with CMR, MSC, SNV, log(1/R), DWT and raw spectra (R^2_{cv} from 0.560 to 0.592), and on site 3 (Haplic Cambisol) equally with MSC and CMR (both $R^2_{cv} = 0.767$), as only these two were significantly different from the raw spectra. On site 4 (Haplic Luvisol), the only one significantly more accurate prediction compared to raw spectra was achieved with FD ($R^2_{cv} = 0.611$), while for the rest of the methods, except SD, there was no difference if either raw spectra or other transformations were used (R^2_{cv} from 0.499 to 0.591). Finally, using the whole data set the differences between pre-processing methods were even less pronounced, when there was no significant difference between raw spectra and other methods (except SD which was significantly worse), although all the predictions were more accurate in general (R^2_{cv} from 0.811 to 0.831).

1. Introduction

Choosing the most feasible signal pre-processing method is a critical step in visible– and near infrared diffuse reflectance spectroscopy (Vis–NIR DRS, 400–2500 nm) (e.g. Engel et al., 2013). The right choice of the pre-processing strategy may improve the predictive model performance largely, and vice versa. However, there is no single (or combination of) pre-processing method that could be considered the best for all data sets. In soil spectroscopy, for different data sets (and so for different soil conditions) and by using different calibration techniques, usually a different pre-processing method works the best (e.g. Gholizadeh et al., 2013; Moron and Cozzolino, 2002; Mouazen et al., 2007; Stenberg et al., 2010; Udelhoven et al., 2003; Vašát et al., 2015a, b; Viscarra Rossel et al., 2006). At present, there are many techniques that are suitable for pre-processing the soil spectra, all of them are aimed at eliminating or reducing unwanted signal artifact of different nature and at highlighting the variation of interest. Most of the methods used in soil spectroscopy are actually adopted from other chemometrics disciplines and applications (Engel et al., 2013). These include tools for signal

* Corresponding author.
E-mail address: vasat@af.czu.cz (R. Vašát).

http://dx.doi.org/10.1016/j.geoderma.2017.03.012
0016-7061/

smoothing and/or noise reduction such as Savitzky-Golay (SG) algorithm or discrete wavelet transformation (DWT), methods aimed at reducing the effect of light scattering such as multiplicative scatter correction (MSC) or standard normal variate (SNV), baseline normalization techniques aimed at eliminating vertical offset and/or slope effects such as continuum removal (CR), first and second order derivative (FD and SD), as well as other transformations (e.g. logarithmic- log(1/R)), and finally the combinations thereof (e.g. smoothing is frequently carried out prior to further signal processing) depending on the nature of the artifacts.

When using partial least squares regression (PLSR), one of the most common calibration techniques (as documented by Stenberg et al., 2010), the raw (or SG smoothed) reflectance spectra or CR spectra often ensures a reliable SOC content estimation (e.g. Gholizadeh et al., 2013; Vašát et al., 2015a). With the CR method (Clark and Roush, 1984) one achieves a set of distinct absorption features (AF), whose parameters (e.g. width, depth or area) can be directly related to the soil variable (e.g. content or quality indicator of any of its constituents) of interest (e.g. Bayer et al., 2012; Gomez et al., 2008; Vašát et al., 2014; Vašát et al., 2015b). When using PLSR, however, the AF parameters are mostly neglected as only the actual CR values at respective wavelengths are used for the calibration. Furthermore, the resulting CR spectra may suffer from approximation that is usually applied to connect the pivot points (local maxima) of the spectrum in order to derive a continuum. On occasions, it may happen that some distinct spectral responses may be reduced or eliminated within this process, which consequently leads to worsening of the prediction accuracy. But more importantly, the CR method does not recognize whether the vertical shift at the very beginning of the signal (which part is especially important for SOC delineation; e.g. Stenberg et al., 2010; Viscarra Rossel and Hicks, 2015) is due to different absorption characteristics of the material, or whether it is due to an overall vertical offset caused by light scattering. The CR value equals one either way at that point. Therefore, since there is no need for AF parameters (when using PLSR), we decided to modify the commonly used CR method in order to eliminate the negative effects, but still preserve the overall vertical offset correction as with the common CR. The modified procedure relies upon division of the reflectance values at each wavelength by the maximum reflectance value of the respective spectrum. We believe that by omitting the approximation from the normalization process and by preserving the nature of the vertical shift at the beginning of the signal, more spectral signatures can be preserved and consequently a more accurate prediction of SOC content can be achieved. This correction by the maximum reflectance (CMR) method was compared to eight other well known methods (SG, FD, SD, MSC, SNV, log(1/R), DWT and CR), as well as to raw soil spectra.

2. Materials and methods

2.1. Study sites, soil sampling, laboratory and soil spectra measurements

The four study sites represent intensively farmed arable land of acreage 100, 3, 3 and 8 (site 1, 2, 3 and 4, respectively), with different dominating soil units and significant slope across the area. Soil samples (in total 106, 53, 63 and 76; site 1, 2, 3 and 4, respectively) were taken similarly for all four sites in a rectangular grid from the topsoil layer (at a depth of 25 cm) in 2010 (site 1), 2012 (site 4) and 2013 (site 2 and 3). On site 1, the prevailing soil units were classified as Haplic Chernozem, Regosol, Colluvial Chernozem and Colluvial soil (according to World Reference Base for Soil Resources (IUSS Working Group WRB, 2014)) developed on loess substrate, where the latter three units were developed more recently as a result of strong water erosion impact (Jakšík et al., 2015). Site 2 was characterized by the occurrence of Rendzic Leptosol and Colluvial soil (the most bottom parts) developed on spongolite substrate. Cambisol (on slate substrate) was the sole soil unit identified on site 3. On site 4, the soil units were identified as Haplic Luvisol, Regosol and Colluvial soil (all developed on loess substrate), in that order from the top to bottom parts of the area (Zádorová et al., 2014).

Soil samples were air-dried, ground, and mixed thoroughly using a mortar and pestle, and sieved to particle fraction ≤0.25 mm. The SOC measurements were carried out in two sub-steps following the dichromate redox titration method (Skjemstad and Baldock, 2008). First, the samples were oxidized with $K_2Cr_2O_7$, and subsequently the solution was potentiometrically titrated with ferrous ammonium sulphate. The soil pH was measured using a 1:5 (w/v) ratio of soil and water suspension with inoLab Level 1 pH-meter. The descriptive statistics of SOC contents (%) and soil pH is shown in Table 1. The correlation (Pearson correlation coefficient) analysis showed a vague, or no relationship at all, between SOC and soil pH (it was −0.46, 0.01, 0.23, 0.13 for site 1, 2, 3 and 4, respectively, and it was −0.46 for the whole data set). The reason for including soil pH was to assess its possible effects on SOC estimation accuracy.

The spectral measurements were carried out ex-situ (under laboratory conditions) using FieldSpec® 3 (PANanalytical Inc., Boulder, Colorado, USA) spectroradiometer device combined with high-intensity contact probe. The spectral resolution was 3 nm from 350 to 1000 nm, and 10 nm from 1000 to 2500 nm. The bandwidth was 1.4 nm from 350 to 1000 nm and 2 nm from 1000 to 2500 nm. The sensor was periodically re-calibrated after each ten samples using Spectralon® (Labsphere, North Sutton, NH, USA) standard white reference. The raw signal was transformed into spectral reflectance. The sampling resolution was 1 nm, and hence each spectrum comprised of reflectance at 2151 wavelengths. Due to extensive noise at the beginning of the signal, the part 350–399 nm was omitted from further calculations.

2.2. Soil spectra pre-processing

All the considered signal pre-processing methods: i.e. SG filtering, FD, SD, MSC, SNV, log(1/R), DWT, CR, and the CMR, were calculated with R software (R Development Core Team, 2015). For the SG filtering we employed *sgolayfilt* function (adjusted for second-order polynomial fit with 31 smoothing points) from *signal* R package (Signal developers, 2013). The MSC was calculated using *pls* R package (Mevik and Wehrens, 2007), particularly the *msc* function. The SNV was calculated by subtracting every reflectance value from the mean reflectance value of the particular spectrum, and by dividing this value by the standard deviation of the whole spectrum. DWT was calculated with *dwt* function from *wavelets* R package (Aldrich, 2013). CR was calculated by the division of the original spectrum by the continuum, that was calculated as convex hull fit over the spectrum using Delaunay triangulation contained in *tripack* R package (Renka, 1996). Finally, the CMR was calculated following a simple manner, that each reflectance value of the spectrum was divided by the maximal reflectance value of the respective spectrum. The CMR was applied on every spectrum individually, and hence it is set-independent. The rest of the pre-processing methods were calculated using standard R functions. To visualize differences between different pre-processing methods, all of the signal transforms were plotted in Fig. 1. In addition, all the methods (except SG) were calculated in two ways, i.e. if either raw reflectance spectra or SG smoothed spectra were provided as input data. In Fig. 1, to reduce the space, only transforms computed from SG spectra are shown.

2.3. Predictive modeling and validation

The predictive models were calibrated using PLSR technique, which is well known for its strong predictive capability, and hence widely used across different scientific disciplines. The method can be advantageously used for regression problems where the response variable has to be related to a large number of predictor variables. It can be employed even in situations when the number of predictor variables is larger than the number of observations. Moreover, the method is robust to