Contents lists available at ScienceDirect

Geoderma

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

Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region

Abhinav Gupta, Hitesh B. Vasava, Bhabani S. Das*, Aditya K. Choubey

Agricultural and Food Engineering Department, Indian Institute of Technology Kharagpur, WB 721302, India

ARTICLE INFO

Handling editor: A.B. McBratney

Keywords:
Diffuse reflectance spectroscopy
Locally-weighted partial-least-squares regression
Lasso
Ridge regression
Geometrical parameters

ABSTRACT

Robust calibration algorithms are needed for the accurate assessment of soil properties in the diffuse reflectance spectroscopy (DRS) approach. Despite several studies on different calibration algorithms, the prediction accuracy of soil properties using DRS need to be improved. Specifically, the utility of local modeling approaches for small spectral libraries is less examined compared with global modeling approaches. In this study, we compared global modeling approaches such as partial-least-squares regression (PLSR), lasso, ridge regression with several locally-weighted PLSR ($PLSR_{LW}$) approaches. We also examined seven different distance measures: Euclidean distance, covariance-based distance, correlation-based distance, surface difference spectrum, information-based distance, optimized principal component Mahalanobis, and locally-linear embeddings used in the $PLSR_{LW}$ approach for their effectiveness in modeling soil properties using DRS. A total of 954 soil samples were collected from three different states of India: West Bengal, Odisha, and Karnataka. Five soil properties such as sand content, clay content, soil organic carbon (SOC) content, extractable iron (Fe) content and extractable zinc (Zn) content were predicted using reflectance spectra over 350–2500 nm. Root-mean-squared error (RMSE) and the coefficient of determination (R^2) were used as performance statistics. Among the global modeling approaches, lasso performed better than the PLSR although it is computationally more intensive than the PLSR. In general, the correlation-based $PLSR_{LW}$ performed significantly better than the global approaches. Specifically, the test R^2 values increased from 0.66 to 0.72 for prediction of sand content, from 0.59 to 0.70 for prediction of SOC content, and from 0.70 to 0.74 for prediction of Fe content by using the $PLSR_{LW}$ over PLSR. We also used depth of absorption peak of spectra at approximately 1400, 1900 and 2200 nm for mineralogical characterization of soil samples. The results suggested that the improvement in prediction accuracy of soil properties using the $PLSR_{LW}$ was achieved because calibration samples which had same mineralogy as the test sample were given higher weights. These results suggest that the prediction accuracy of soil properties may also be improved in small spectral libraries if an appropriate local modeling scheme is selected.

1. Introduction

Diffuse reflectance spectroscopy (DRS) is emerging as a rapid and noninvasive method for soil analysis (Soriano-Disla et al., 2014). In this approach, several soil properties of a soil sample may be simultaneously estimated from a single set of reflectance values measured over the visible and near-infrared region (wavelength: 350–2500 nm) of the electromagnetic spectrum (Viscarra Rossel et al., 2006). Robust calibration algorithms and large spectral libraries are needed for implementing a DRS approach (Viscarra Rossel et al., 2016). Over the last three decades, several global calibration algorithms such as partial-least-squares regression (PLSR; Geladi and Kowalski, 1986), support vector regression (SVR; Smola and Schölkopf, 2004), multivariate adaptive regression splines (MARS; Friedman, 1991), random forests (RF; Breiman, 2002), etc. have been applied in soil DRS studies. While the PLSR approach captures the linear relationship between a soil property (response variable) and reflectance spectra (predictor variables), other algorithms such as SVR, MARS, and RF capture non-linear relationships between soil spectra and the response variable. Because of its simplicity and satisfactory performance, PLSR is one of the most frequently used algorithms in soil DRS studies (Viscarra Rossel and Behrens, 2009). Table 1 shows that despite using several calibration algorithms, the coefficients of determination (R^2) value achievable via these calibration algorithms are often low. Therefore, robust calibration algorithms are required to predict soil properties from spectra (Soriano-Disla et al., 2014).

Over the last decade, several large spectral libraries have been developed to improve the performance of DRS models (Viscarra Rossel

* Corresponding author.
E-mail address: bsdas@agfe.iitkgp.ernet.in (B.S. Das).

https://doi.org/10.1016/j.geoderma.2018.03.025
Received 8 August 2016; Received in revised form 10 May 2017; Accepted 22 March 2018

Table 1
Spectral range, calibration site and technique, coefficient of determination R^2 and reference to the recent article related to prediction of few soil properties.

Soil property	Spectral range, nm	Calibration site	Calibration technique	R^2	Reference
SOC, %	350–2500	Ethiopia	PLSR	0.62	Shiferaw and Hergarten (2014)
SOC, %	350–2500	Hawaii	PLSR	0.95	McDowell et al. (2012)
SOC, %	350–2500	Hawaii	RF	0.95	McDowell et al. (2012)
SOC, g kg^{-1}	400–2500	France	PLSR	0.02	Gomez et al. (2012)
SOC, g kg^{-1}	400–2500	Europe	GB	0.85	Liu et al. (2016)
SOC, g kg^{-1}	350–2500	France	PLSR	0.67	Clairotte et al. (2016)
SOC, g kg^{-1}	350–2500	France	PLSR	0.75	Cambou et al. (2016)
SOC, %	400–2447	India	PLSR	0.57	Sarathjith et al. (2016)
Sand, %	400–2447	India	PLSR	0.55	Sarathjith et al. (2016)
Sand, g kg^{-1}	400–2500	France	PLSR	0.20	Gomez et al. (2012)
Sand, %		Italy	PLSR	0.80	Curcio et al. (2013)
Sand, %		Czech Republic	PLSR	0.68	Gholizadeh et al. (2016)
Sand, %		Czech Republic	SVR	0.69	Gholizadeh et al. (2016)
Clay, %	400–2447	India	PLSR	0.47	Sarathjith et al. (2016)
Clay, %		Czech Republic	PLSR	0.79	Gholizadeh et al. (2016)
Clay, %		Czech Republic	SVR	0.82	Gholizadeh et al. (2016)
Clay, %		Italy	PLSR	0.87	Curcio et al. (2013)
Clay, g kg^{-1}	400–2500	France	PLSR	0.67	Gomez et al. (2012)
Fe, mg kg^{-1}	400–2447	India	PLSR	0.78	Sarathjith et al. (2016)
Fe, g/100 g	400–2500	France	PLSR	0.78	Gomez et al. (2012)
Zn, mg/L	400–2447	India	DWT-SVR	0.43	Sarathjith et al. (2016)
Zn, mg kg^{-1}	400–2400	China	PLSR-GA	0.70	Sun and Zhang (2017)

PLSR: partial-least-squares regression; RF: random forest; GB = gradient boosting; SVR: support vector regression; DWT: discrete wavelet transform; PLSR-GA: genetic algorithm based PLSR.

et al., 2016). In a large spectral library, soil samples are often collected from a large geographical area introducing large heterogeneity in the spectral library. Relationship between a soil property and spectra in such a library may be complex and non-linear (Savvides et al., 2010) and therefore, the prediction accuracy achieved by a model calibrated from a large spectral library may deteriorate because the underlying assumption (e.g., linearity in PLSR) of the model may not be valid (Hastie et al., 2001). To overcome such a challenge, the local modeling approach is often proposed (Christy and Dyer, 2006; Nocita et al., 2014). For example, Nocita et al. (2014) selected a subset of K nearest samples from an available pool of calibration samples for a sample to be predicted (target sample or test sample) based on the Euclidean distance as a similarity measure. A linear model was calibrated by PLSR using these K samples and resulting regression coefficients were then used to predict the soil organic carbon of the target sample. This procedure was repeated for all the samples to be predicted and is referred hereinafter as the K-nearest neighbor PLSR ($PLSR_{KNN}$). Though $PLSR_{KNN}$ is a good modeling approach, it may often eliminate useful information from calibration data because it uses only a few calibration samples. Thus, the information content of all the soil samples is not utilized in $PLSR_{KNN}$. Another drawback of $PLSR_{KNN}$ is that all the calibration samples selected for a target sample are considered equally relevant which may not always be true. These drawbacks may be avoided in locally weighted-PLSR ($PLSR_{LW}$; Kim et al., 2011). In $PLSR_{LW}$, an independent model is calibrated for each target sample. Subsequently, each sample in the calibration set is assigned a weight based on its similarity with the target sample and weighted-PLSR is performed to obtain regression coefficients. The weights may be determined as exponential (Kim et al., 2011) or tricubic (Hastie et al., 2001) function of the distance between calibration and test spectra. The regression coefficients are then used to estimate the response of the target sample. Thus, all relevant information for a target sample may be retained and, also, the redundant information may be eliminated. Therefore, $PLSR_{LW}$ is expected to perform better than $PLSR_{KNN}$ and PLSR. Over the last few decades, several other approaches have been proposed in the different chemometric literature (Hazama and Kano, 2015), which should also be tested for soil applications involving small spectral libraries. Generally, local regression approaches have shown better performance than global approaches in large spectral libraries (Genot et al., 2011; Gogé et al., 2012; Ramirez-Lopez et al., 2013a; Christy and Dyer, 2006; Nocita et al., 2014).

Selection of an appropriate distance metric used for calculating the similarity between a test sample and the calibration samples is a key step in local modeling approaches (Hazama and Kano, 2015). Ramirez-Lopez et al. (2013b) compared different distance measures in the VNIR and projected spectral space for their efficacy in identifying similar soil samples in terms of composition. In the VNIR space, the surface difference spectrum (SDS) was observed to be the best performing distance measure compared to the Euclidean distance (ED), Mahalanobis distance (MD), spectral angle mapper (SAM), and spectral information divergence (SID) distance. In projected spaces, principal component distance (PC-M), optimized principal component distance (oPC-M) and local linear embeddings (LLE-M) and σ-local linear embedding (σLLE) were compared. In general, distance measures in projected spaces performed better than those in the VNIR space. For example, LLE-M and σLLE-M performed better than the other distance measures. In addition to the aforesaid distance measures, Hazama and Kano (2015) proposed a covariance-based distance measure in conjunction with PLSR to predict unreacted NaOH in a petrochemical process and the concentration of residual drug substance in a pharmaceutical process. This distance measure considers the covariance between response and predictor variables; hence, it may be expected to perform better than other distance measures. Another distance-based weighing scheme may be based on the correlation between response and predictor variables. Because different soil properties influence different regions of the VNIR spectrum differently, a correlation-based distance measure may capture this importance while building a weighted PLSR model.

$PLSR_{LW}$ is yet to be tested in soil DRS literature. Particularly, the utility of local modeling approaches for small spectral libraries in is less examined. A comprehensive analysis of the suitability of different distance measures is also not available. We hypothesized that the $PLSR_{LW}$ would perform better than global PLSR even in a small spectral library. Because the $PLSR_{LW}$ allows a different model for each target sample and it has the capability to use all the relevant information as well as eliminate redundant information. Therefore, the objective of this study is (a) to test the suitability of different local-modeling approaches for the prediction of soil properties from diffuse reflectance spectra for small spectral libraries and (b) to compare different distance measures that may be used in local modeling approaches.