Soil Biology & Biochemistry 68 (2014) 337–347

Contents lists available at ScienceDirect

Soil Biology & Biochemistry

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

Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach

Marco Nocita [a,b,*], Antoine Stevens [b], Gergely Toth [a], Panos Panagos [a], Bas van Wesemael [b], Luca Montanarella [a]

[a] European Commission, Joint Research Centre, Institute for Environment and Sustainability, Via E. Fermi 2749, I-21027 Ispra, VA, Italy
[b] Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute, Université Catholique de Louvain, 1348 Louvain La Neuve, Belgium

ARTICLE INFO

Article history:
Received 2 July 2013
Received in revised form
4 October 2013
Accepted 11 October 2013
Available online 23 October 2013

Keywords:
Soil organic carbon
Vis-NIR spectroscopy
Local approach
Spectral library
Sand content
Europe

ABSTRACT

Due to the large spatial variation of soil organic carbon (SOC) content, assessing the current state of SOC for large areas is costly and time consuming. Visible and Near Infrared Diffuse Reflectance Spectroscopy (Vis-NIR DRS) is a fast and cheap tool for measuring SOC based on empirical equations and spectral libraries. While the approach has been demonstrated to yield accurate predictions for databases containing samples belonging to soils with similar characteristics such as mineralogy, texture, iron, and $CaCO_3$ content, spectroscopic calibrations have been less successful when applied to large and diverse soil spectral libraries. The scope of this study was to predict SOC using a local partial least square regression approach. In total, 19,969 topsoil (0–20 cm) samples collected all over the European Union were analyzed for physical and chemical properties, and scanned with a Vis-NIR spectrometer in a single laboratory. The local regression method builds a different multivariate model for each sample to predict. Each local model is trained with neighbours' samples selected from a large spectral library, based on their spectral similarity with the sample to predict. We modified the local regression procedure by including other covariates (geographical and texture information) in the computation of the distance between samples. The results showed good prediction ability for mineral soils under cropland (RMSE = 3.6 g C kg^{-1}) and grassland (RMSE = 7.2 g C kg^{-1}). Predictions of mineral soils under woodland (RMSE = 11.9 g C kg^{-1}) and organic soils (RMSE = 51.1 g C kg^{-1}) were less accurate. The use of sand content in the computation of the sample similarities provided the most accurate SOC predictions due to its influence on light scattering properties of soils. In large datasets, using additional soil or environmental information allows to select neighbours that have overall the same soil composition as the samples to predict, resulting in more accurate models. This study shows that (i) it is possible to realize low-cost estimations of SOC at continental scale using large spectral libraries with a reasonable accuracy, and (ii) the local approach is a valuable tool to deal with large datasets, especially if existing soil property maps or soil legacy data could be used as covariates in the SOC prediction models.

1. Introduction

Recognizing the importance of soil organic carbon (SOC) for sustaining soil quality and food production, the European Union considers the decline of SOC in European soils as one of the main threats for soil degradation in its Thematic Strategy for Soil Protection (CEC, 2006). Moreover, there is an increasing demand for monitoring carbon levels in soils, as the depleted C levels particularly in croplands provide an opportunity for carbon sequestration through adequate management practices (Lal, 2004). Unfortunately, the cost of collecting soil information restricts the monitoring of soil properties at large scale (Conant et al., 2010). Visible (Vis, 400–700 nm) and Near-Infrared (NIR, 700–2500 nm) diffuse reflectance spectroscopy (DRS) has shown to be an efficient tool for the rapid and cheap prediction of soil properties (e.g. Islam et al., 2003). The absorptions occurring in the Vis and NIR regions of the electromagnetic spectrum are related to organic and inorganic phases of the soil that have molecular absorptions in the Mid Infra-Red (MIR: 2500–25,000 nm) wavelengths (Ben-Dor et al., 1999), but their overtones and combinations occur in the NIR region. Several absorptions occurring in the Vis are also caused by electronic transitions of atoms (Ben-Dor et al., 1999). Although MIR

* Corresponding author. Via Enrico Fermi 2749 – TP 280, I-21027 Ispra, VA, Italy. Tel.: +39 0332 78 3682; fax: +39 0332 78 6394.
E-mail addresses: marco.nocita@jrc.ec.europa.eu, marco.nocita@gmail.com (M. Nocita).

0038-0717/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.soilbio.2013.10.022

spectra are more specific, contain more information than Vis-NIR spectra, and typically provide more accurate soil properties predictions (Bellon-Maurel and McBratney, 2011), Vis-NIR spectroscopy has grown as an analytical method due to its cost-efficiency, ease of handling, rapidity, minimal sample preparation, and the development of chemometrics (Davies, 2005).

Vis-NIR DRS was first applied for soil analysis in the 1980's and showed that the SOC content, amongst other properties, can be accurately predicted (Dalal and Henry, 1986; McCarty et al., 2002; Cozzolino and Morón, 2003; Sørensen and Dalsgaard, 2005; Viscarra Rossel et al., 2006).

The prediction of soil properties requires the creation of a spectral library relating spectra with reference data. Such a library should be designed to represent the variation in soil properties of the soil types of interest. Multivariate regressions are then used to infer properties (Shepherd and Walsh, 2002). The majority of the existent spectral libraries are limited to small areas, and only few large scale soil spectral libraries have been developed so far. Shepherd and Walsh (2002) inferred SOC combining Vis-NIR DRS and Multivariate Adaptive Regression Splines (MARS) on more than 1000 topsoil samples from eastern and southern Africa. Brown et al. (2006) applied Vis-NIR DRS to characterize various soil properties of more than 4500 samples from the United States (3768), Africa (125), Asia (104), South and Central America (75), and Europe (112). Viscarra Rossel and Webster (2012) used 21,493 spectra of soil samples from approximately 4000 profiles to predict 24 soil properties.

In large and complex datasets, the relationship between soil properties and spectral data can be highly non-linear and spatial-dependent (Savvides et al., 2010; Stenberg et al., 2010). However, spectral variation associated to soil properties can be locally stable (Ramirez-Lopez et al., 2013). Hence, one possible approach for predicting soil properties in large scale databases is the use of local regressions. In local regressions, each sample is predicted with a different calibration equation (Næs et al., 2002). The samples used for training each local model, called nearest neighbours, are selected from a spectral library based on their similarity with the sample to predict, expressed as a measure of distance between the spectra (Fernandez Pierna and Dardenne, 2008). Such approach is able to effectively remove uninformative/unrelated samples from calibrations, thereby allowing combining the advantages of having a spectral library covering a large domain with the accuracy obtained by specific/local calibration models (Pérez-Marín et al., 2007). Overall, local regressions demonstrated good performance for soil properties inference (Genot et al., 2011; Gogé et al., 2012; Ramirez-Lopez et al., 2013).

Several local regression approaches have been developed for the calibration of spectroscopic data, the most commonly used being the LOCAL (Shenk et al., 1997) and Locally Weighted Regression (LWR; Næs and Isaksson, 1992) algorithms, implemented respectively in the WinISI (Infrasoft International LLC, PA, USA) and Unscrambler (Camo, Norway) softwares. Ramirez-Lopez et al. (2013) developed a novel algorithm, specifically designed for spectroscopic predictions of soil properties, called the Spectrum-Based Learner (SBL). One of the improvements over other local algorithms consists in using optimized principal component distance for retrieving the set of nearest neighbours. This procedure selects the optimal number of principal components by taking into account the soil composition of the samples.

Within the framework of the Land Use/Cover Area frame Statistical Survey (LUCAS), about 20,000 topsoil samples have been collected to assess the state of the soil across Europe. Soil samples were analyzed for several physical and chemical parameters, and scanned with a Vis-NIR spectrometer in a single laboratory. The aim of our research was to predict the SOC content of the LUCAS soil samples using the spectral library coupled with a local partial least square regression approach. We modified the local regression procedure by including other covariates (geographical and texture information) in the computation of the distance between samples. To produce accurate predictions, a critical step is indeed the selection of the nearest neighbours for building the local models. Using additional information in the estimation of the similarity between samples allows selecting meaningful neighbours i.e. having roughly the same composition of the sample to predict. This modification might favour more accurate predictions of SOC through a better account of the complexity of the soil spectral response.

2. Methodology

2.1. LUCAS sampling campaign

The LUCAS survey is based on the visual assessment of land use and land cover parameters that are deemed relevant for agricultural policy (Land Use/Land Cover Area Frame Survey; Eurostat, 2012). This visual survey was complemented with a soil sampling in 2009, with the aim to produce the first coherent pan-European physical and chemical topsoil database, which can serve as baseline information for an EU wide soil monitoring. A total of 19,969 topsoil samples were collected in twenty-five EU Member States (EU-28 except Bulgaria, Romania, and Croatia) (Toth et al., 2013). This survey is the first effort to build a consistent spatial database of the topsoil across Europe, based on a single sampling protocol and analysis carried out in a single laboratory. A stratified sampling design was realized to select sampling locations that are representative of the major landforms and land cover types of the participating countries (Montanarella et al., 2011).

A single soil sampling protocol was followed by all surveyor teams in order to guarantee comparability of the data. The sampling scheme consisted in the collection of a composite of 5 subsamples. The first subsample was collected in the selected location, while the other 4 subsamples were collected at a distance of 2 m following the cardinal directions. All the soil samples were air dried in drying room at temperature of 40 °C for an average of 3 days. The soil samples were crushed, sieved and the fraction smaller than 2 mm was kept for further analysis. The following properties were then analyzed: particle size distribution, pH in H_2O, pH in $CaCl_2$, organic carbon, carbonates, Nitrogen, Phosphorus, and Potassium (NPK), cationic exchange capacity (CEC), and Vis-NIR diffuse reflectance. The database is available for download for non-commercial purposes at http://eusoils.jrc.ec.europa.eu/projects/Lucas/data.html.

2.2. Chemical and spectral analysis

Total carbon was measured with a VarioMax CN Analyzer (Elementar Analysis, Germany) after heating the soil to 900 °C. The SOC content was then obtained subtracting the carbonate content (measured according to ISO 10693:1995) to total carbon. The particle size distribution was determined by a combination of sieving and sedimentation, starting from air-dried soil sample (ISO 11277:1998). Soil particles were divided in sand (particle size: 0.063–2 mm), silt (0.002–0.063 mm), and clay (particle size $<$ 0.002 mm). The Vis-NIR absorbance of soil was measured using a FOSS XDS Rapid Content Analyzer (FOSS NIRSystems Inc., Denmark), operating in the 400–2500 nm wavelength range, with 2 nm spectral resolution and 0.5 nm spectral data interval. Each sample was put in a sample cell with dimensions of 140 $\times$ 40 $\times$ 50 mm, and scanned twice in both directions. The mean spectrum of the two repetitions was calculated and samples with a