Geoderma 288 (2017) 175–183

Contents lists available at ScienceDirect

Geoderma

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

Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy

Federica Lucà [a,*], Massimo Conforti [a], Annamaria Castrignanò [a,b], Giorgio Matteucci [a], Gabriele Buttafuoco [a]

[a] *National Research Council of Italy - Institute for Agricultural and Forest Systems in the Mediterranean (ISAFOM), Rende, Cosenza, Italy*
[b] *CREA - Council for Agricultural Research and Economics, Bari, Italy*

ARTICLE INFO

Article history:
Received 7 March 2016
Received in revised form 26 October 2016
Accepted 9 November 2016
Available online 17 November 2016

Keywords:
Soil spectroscopy
Soil carbon
Calibration set size
Principal component regression
Partial least square regression
Support vector machine regression

ABSTRACT

Predicting soil properties through visible and near-infrared (Vis–NIR) spectroscopy by a limited number of calibration samples can reduce the cost and time for physic-chemical analyses. This study was aimed to assess the influence of calibration set size on the prediction of total carbon (TC) in the soil by Vis–NIR spectroscopy. In a forested area of 33 ha in southern Italy (Calabria), 216 soil samples were analyzed for TC concentration, and reflectance spectra were measured in the laboratory. The whole data set was randomly split into calibration and validation sets (70% and 30%, respectively). To study the effect of the number of samples on TC prediction, ten calibration subsets of samples between 14 and 144 were selected. Three techniques including principal components regression (PCR), partial least squares regression (PLSR) and support vector machine regression (SVMR) were used to develop 84 calibration models, validated through the same independent data. The models were compared through the coefficient of determination (R^2), the root mean square error of prediction (RMSEP) and the ratio of the interquartile distance (RPIQ). Validation results showed that to obtain not significant differences with models based on the full calibration set, 29, 72 and 115 samples were required for PCR, SVMR and PLSR respectively. Although PCR appeared less sensitive than PLSR and SVMR to calibration sample size, SVMR outperformed PLSR and PCR with higher R^2 and RPIQ values and lower RMSEP. To obtain RMSEP not significantly different from the best model achieved in this study, the required minimum number of samples was 72 for SVMR and 130 for PLSR. All PCR model were significantly poorest than the best model.

1. Introduction

Visible and near-infrared (350 to 2500 nm, Vis–NIR) spectroscopy is a powerful rapid and cost-effective method to predict chemical, physical and mineralogical attributes in the soil. It provides the opportunity to estimate many properties from one spectrum based on soil reflectance with minimal sample preparation and no chemical reagent (Shepherd and Walsh, 2002). Various chemical or physical properties of the soil, such as organic carbon (Ben-Dor et al., 2002; Stevens et al., 2010; Conforti et al., 2015a, Lucà et al., 2015), nitrogen (Selige et al., 2006, Lucà et al., 2015), Ca CO_3 (Ramirez-Lopez et al., 2014), texture (Selige et al., 2006; Conforti et al., 2015b) and moisture (Ben-Dor et al., 2002), have been estimated through the use of field, laboratory or air-borne sensing data. Among such properties, soil carbon has received much attention, since its dynamics is essential for sustainable land management, especially in the context of climate change (Lal, 2004).

The application of spectroscopy is based on the development of a calibration model, to build a mathematical relationship between the spectra and the soil property in question. Once a calibration model has been developed, it can be used to predict the chemical or physical property in unknown samples. However, different multivariate calibration methods can be used. The most commonly applied methods include multiple linear regression (MLR, Shibusawa et al., 2001), principal component regression (PCR, Chang et al., 2001), partial least squares regression (PLSR, Viscarra Rossel et al., 2006, Lucà et al., 2015), artificial neural networks (ANN, Fidêncio et al., 2002), support vector machine regression (SVMR, Stevens et al., 2010; Ramirez-Lopez et al., 2014), and regression tree (Vasques et al., 2008). There is no best method because each one has its advantages and drawbacks (Vasques et al., 2008; Mouazen et al., 2010; Stevens et al., 2010; Viscarra Rossel and Behrens, 2010 among many others). For example, compared to MLR, PCR and PLSR have the advantage of handling data multicollinearity but they only allow researchers to estimate linear relationships between the spectra and soil properties. On the contrary, more recent techniques like ANN and SVMR allow us to manage the non-linear behavior of soil reflectance (Viscarra Rossel and Behrens, 2010). In particular,

[*] Corresponding author.
E-mail address: federica.luca@isafom.cnr.it (F. Lucà).

http://dx.doi.org/10.1016/j.geoderma.2016.11.015
0016-7061/© 2016 Elsevier B.V. All rights reserved.

SVMR is based on the statistical learning theory (Vapnik, 1995) and performed well also when few samples are used for training the calibration model. Comparative studies were inconclusive on which method can be considered as the most efficient and accurate, as the findings on this issue are.

Nevertheless, regardless of the calibration method used, effectiveness and accuracy are heavily dependent on the calibration set. In fact, to get robustness and reliability in prediction, a representative data set should be selected on the basis of spectral features (Guerrero et al., 2010), analytical properties (Aïchi et al., 2009), or both (Williams, 2001). Although several works used different algorithms for selecting representative subsets from a large data pool (Minasny and McBratney, 2006; Ramirez-Lopez et al., 2014), the influence of the number of calibration samples on model prediction has received less attention (Brown et al., 2005; Guerrero et al., 2010, 2016; Kuang and Mouazen, 2012; Debaene et al., 2014; Ramirez-Lopez et al., 2014). The optimal calibration set size might vary depending on the geographical scale and on the pedodiversity of the study area; splitting the dataset based on auxiliary information (soil type, topography, vegetation) and running local calibrations separately would result in an improvement of the model's performance (Stevens et al., 2010). Moreover, Ramirez-Lopez et al. (2014) found that when the number of calibration samples was small, the accuracy of the Vis-NIR models was also influenced by the sampling algorithm. On the one hand, analyzing a large amount of calibration samples is not computationally effective and it would cause an increase in the noise, resulting in inadequate models (Sáiz-Abajo et al., 2005). On the other hand, calibrations based on few samples sometimes lead to unsatisfactory and unreliable results (Shepherd and Walsh, 2002; Russell, 2003; Siebielec et al., 2004). In any case, spiking global libraries with local samples could improve the predictive accuracy of the target site (Brown, 2007; Guerrero et al., 2016).

The aim of this work was to investigate the influence of calibration sample number on the predictive performance of Vis–NIR models for total carbon (TC) prediction at local scale for soil samples collected in a forested area. Furthermore, to evaluate the effects of the statistical approach on the results, three multivariate methods were selected among those most widely applied: PCR, PLSR and SVMR. The purpose was to compare, by varying the number of calibration samples, the performance of traditional regression methods (PCR and PLSR) with a more recent technique like SVMR, which is able to handle nonlinear data. For each sample size, the same independent validation dataset was used to test the models. For each method, the optimal calibration set size, that is the smallest number of soil samples that can be used without significantly compromising the prediction, was identified. In addition, the comparison with the best model resulting from this study was also performed to identify the optimal combination between sample size and statistical approach.

2. Materials and methods

2.1. Study area

The study area (about 33 ha) is located within the “Marchesale” Biogenetic Natural Reserve in the Serre Massif (Calabria, southern Italy) at elevations ranging from 1136 to 1212 m above sea level (Fig. 1). It is covered by a high forest beech dominated by *Fagus sylvatica*. The climate is typical upland Mediterranean (Csb, sensu Köppen, 1936), whereas the pedoclimate is characterized by mesic regime for soil temperature associated with udic regime for soil moisture (ARSSA, 2003).

The bedrock consists of Palaeozoic granitoid rocks, deeply fractured, weathered and frequently covered by thick regolith and/or colluvial deposits (Borsi et al., 1976; Calcaterra et al., 1996); the latter mostly found in concave areas and at the foot of the slopes. The average slope gradient is around 10° and landscape morphology is characterized by paleosurfaces, representing the remnants of flat or gently-sloping highlands, often sharply separated by steep slopes and V-shaped valleys (Sorriso-Valvo, 1993; Robustelli et al., 2009; Lucà et al., 2011).

The soils are from shallow to moderately deep (0.20 to 1 m) and coarse- to medium- textured, ranging from sandy loam to loam (sensu USDA, 2010). They have acidic pH (3.7–5.8) and bulk density ranging from 0.2 g cm^{-3} in the upper horizons to 1.7 g cm^{-3} for the deepest ones (Conforti et al., 2016). Soils are young and poorly developed with profiles generally characterized by O-A-Bw-Cr or O-A-Cr horizons (ARSSA, 2003). The upper A-horizon commonly shows a very dark brown color due to the accumulation of organic matter (umbric epipedon). Following USDA (2010), the soils of the study area can be classified as Inceptisols and Entisols (ARSSA, 2003).

2.2. Soil sampling and analyses

After removing surface litter, 216 soil samples were randomly collected within the study area up to 0.20-m depth by means of a metal cylinder, and each sampling location (Fig. 1) was recorded with the Real Time Kinematic (RTK) technique, using a differential global positioning system (Leica Geosystems GPS1200).

In the laboratory, the soil samples were oven dried at 40 °C for 48 h, gently crushed and passed through a 2 mm mesh sieve to collect fine earth fraction, removing coarse roots and rock fragments. About 100 mg from each soil sample were used to determine TC with Shimadzu TOC-analyzer, using a SSM-5000A solid sample module operated at 900 °C (Shimadzu Corporation, Kyoto, Japan).

For the sieved soils, Vis–NIR reflectance was obtained under controlled laboratory conditions with an ASD FieldSpec IV spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA), at 1 nm intervals, over the spectral range from 350 to 2500 nm. A 50-Watt halogen lamp with a zenith angle of 30°, located at a distance of approximately 25 cm from the soil sample, was used as artificial illumination. The spectroradiometer was placed in a nadir position at a distance of 10 cm from the sample, which was placed in a 9-cm diameter petri dish. For each soil sample, to reduce instrumental noise, remove any possible spectral anomalies due to geometry of measurement and minimize the possible errors associated with stray light, four scans, each rotated by an angle of 90°, were performed and then averaged to obtain a spectrum for each sample. A Spectralon® panel (20 cm × 20 cm, Labsphere Inc., North Sutton, USA) was used as white reference to compute reflectance values, measured under the same illumination conditions. A reference spectrum was acquired immediately before the first scan and after every set of five samples. For more details on spectral acquisition, see Conforti et al. (2015b).

The average spectrum for each sample was later used for transformation and chemometric modeling. To reduce the amount of data and computation time, the spectra were resampled every 10 nm. Spectral reflectance (R) was transformed to apparent absorbance (A) by $A = \text{Log}(1/R)$ and the light scattering effects were removed by applying a standard normal variate (SNV) pre-processing (Barnes et al., 1989).

2.3. Sample selection

The whole dataset ($n = 216$) was randomly split into a validation set (including 1/3 of total samples, $n = 72$) and a full calibration set (the remaining 2/3 of soil samples, $n = 144$). To analyze the effect of the number of calibration samples on the predictive performance of TC, 10 calibration subsets corresponding to different percentages of samples (from 10%, to 100% of the full calibration set) were created, with the number of samples varying from 14 to 144.

To ensure that the selected samples were representative of the variation of the soil property in the study area, a stratified random partition by 10 strata was performed as follows:

1) the 144 calibration samples were sorted from the lowest to the highest value of TC;