applied sciences

Article

Potential Model Overfitting in Predicting Soil Carbon Content by Visible and Near-Infrared Spectroscopy

Lizardo Reyna [1,2,†], Francis Dube [3], Juan A. Barrera [1] and Erick Zagal [1,*]

[1] Department of Soils and Natural Resources, Faculty of Agronomy, Universidad de Concepción, Vicente Méndez 595, Casilla 537, Chillán 3812120, Chile; lreyna@udec.cl or lreyna@utm.edu.ec (L.R.); jbarrera@udec.cl (J.A.B.)

[2] Doctoral Program in Agronomic Sciences, Faculty of Agronomy, Universidad de Concepción, Vicente Méndez 595, Casilla 537, Chillán 3812120, Chile

[3] Department of Silviculture, Faculty of Forest Sciences, Universidad de Concepción, Victoria 631, Casilla 160-C, Concepción 4030000, Chile; fdube@udec.cl

* Correspondence: ezagal@udec.cl; Tel.: +56-42-2208853

† Current address: Facultad de Ingeniería Agrícola, Universidad Técnica de Manabí, Casilla 82, Lodana, Manabí, Ecuador.

Academic Editor: Johannes Kiefer
Received: 13 June 2017; Accepted: 5 July 2017; Published: 8 July 2017

Abstract: Soil spectroscopy is known as a rapid and cost-effective method for predicting soil properties from spectral data. The objective of this work was to build a statistical model to predict soil carbon content from spectral data by partial least squares regression using a limited number of soil samples. Soil samples were collected from two soil orders (Andisol and Ultisol), where the dominant land cover is native *Nothofagus* forest. Total carbon was analyzed in the laboratory and samples were scanned using a spectroradiometer. We found evidence that the reflectance was influenced by soil carbon content, which is consistent with the literature. However, the reflectance was not useful for building an appropriate regression model. Thus, we report here intriguing results obtained in the calibration process that can be confusing and misinterpreted. For instance, using the Savitzky–Golay filter for pre-processing spectral data, we obtained $R^2 = 0.82$ and root-mean-squared error $(RMSE) = 0.61\%$ in model calibration. However, despite these values being comparable with those of other similar studies, in the cross-validation procedure, the data showed an unusual behavior that leads to the conclusion that the model overfits the data. This indicates that the model should not be used on unobserved data.

Keywords: chemometrics; SOC; spectral diffuse reflectance; partial least squares regression; cross-validation

1. Introduction

Soil total carbon (TC) is composed of organic (all organic components mainly derived from the decomposition of plants and animals; and including living organisms) and inorganic (non-living C, typically as carbonates) carbon forms. Due to the short-term cycle of soil organic carbon (SOC) and its key role for soil functions, the quantitative evaluation of SOC is essential for determining a suitable management practice to conserve or increase soil carbon stock [1–4]. Monitoring SOC over large areas or long periods of time requires analysis of substantial numbers of samples which can be labor-intensive and expensive. Under those circumstances, the soil spectroscopy technique is an effective method to predict SOC rapidly at minimal cost [5,6]. Soil spectroscopy uses the visible and near-infrared (VIS-NIR, 400–2500 nm) and mid-infrared (2500–25,000 nm) spectral reflectance to infer soil properties from a scanned sample [7]. This technique has been used mainly under laboratory conditions, but it can also be applied in the field for a specific site or in an instrument setup for ongoing scanning [6].