Geoderma 268 (2016) 147–155

Contents lists available at ScienceDirect

Geoderma

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

Characterizing typical farmland soils in China using Raman spectroscopy

Zhe Xing [a,b], Changwen Du [a,*], Yin Zeng [a,b], Fei Ma [a], Jianmin Zhou [a]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] University of Chinese Academy of Sciences, Beijing 100039, China

ARTICLE INFO

Article history:
Received 15 June 2015
Received in revised form 18 January 2016
Accepted 19 January 2016
Available online 5 February 2016

Keywords:
Raman spectroscopy
Soil identification
Principal component analysis
Probabilistic neural network
Partial least square regression

ABSTRACT

Raman spectroscopy has been rarely applied in soil characterization due to the interference of fluorescence resulting from soil organic matter (SOM). However, valuable structural information is likely stored in the soil Raman spectra. This study was undertaken to investigate the potential of Raman spectroscopy in soil identification as an alternative tool to traditional methods, and the partial least squares (PLS) model was developed from Raman spectra to make a quantitative prediction of SOM. Diverse soil samples ($n = 200$) representing four typical farmlands in China were scanned with a portable Raman spectrometer (i-Raman® Plus, USA) from the spectra range of 180 to 3200 cm^{-1}. The Raman shift range of 1000–2000 cm^{-1} was selected to establish an identification model. Probabilistic neural network (PNN) combined with principal component analysis (PCA) of the spectra data was employed to identify 200 soils and an acceptable result was obtained with an accuracy of 96% in validation. The PLS model with cross-validation was constructed to predict the content of SOM, and the best prediction model was calculated using spectra with the linear exclusion in selected range ($RPD_v = 1.92$, $R_v^2 = 0.74$, and RMSEP = 8.16 g kg^{-1}). This study illustrates that soil Raman spectra contain information of soil constituents even in the presence of fluorescence interference; moreover, it demonstrates that this contained information can play a vital role in the characterization of symmetric structures of SOM, which can provide complementary knowledge to understand molecular structure of SOM that infrared spectroscopy cannot offer.

1. Introduction

As a composition of mineral particles, humus, liquids, gas, and living organisms, agriculture soil plays a vital part in food production and environment health protection and agro-ecosystem balance. Soil organic matter (SOM) is a heterogeneous mixture of organic residue from plants and animals in the soil (Bardy et al., 2008; Deshmukh et al., 2007). Over 90% of the N and S, and 75% of P are contained in the organic matter of the topsoil (Olk, 2006). As the major reservoir of terrestrial carbon (Ferreira et al., 2014; Wang et al., 2015), certain amounts of organic matter are stored in agriculture soils for both the preservation of soil fertility and the improvement of the soil structure (Tobiasova, 2011).

The demand for the efficient identification and organic matter prediction of arable soil has been increasing with the development of precision agriculture. However, the routine physicochemical analysis of soil is time and labor consuming, and faces challenges in seeking structure information. Furthermore, the massive chemical reagents used can increase risks for both the operators and ecological environment. These deficiencies are mainly responsible for the development of rapid determination methods for soil analysis.

Raman spectroscopy is a technology stemming from the inelastic scattering of light, which reflects molecular vibrational and rotational modes. The advantages include not only being nondestructive, cost-efficient, easy for sample preparation, and less exposure to hazardous chemicals compared to conventional laboratory chemical determination, but also has unique independence from water interference compared to infrared spectroscopy (Barancikova, 2008; Martins et al., 2011; St Luce et al., 2014; Zimmermann et al., 2007). However, the widespread application of Raman spectroscopy is limited due to the fluorescence effect of soil, which often weakens or even masks Raman signals and further complicates the extraction of spectral information. Lowering the excitation energy, pretreating soils with an oxidative method, and absorbing samples to metal matrix are some common means to reduce fluorescence influence and enhance the Raman signal proposed in previous researches (Edwards et al., 2012; Francioso et al., 2000; Luna et al., 2014). With these methods, researchers have suggested that Raman spectroscopy could be particularly available in reflecting skeletal vibrations of both aliphatic and aromatic fractions of soil components, which may be helpful to further reveal the structure of soil (Carrero et al., 2012; Leyton et al., 2008; Ribeiro-Soares et al., 2013). The progress on the application of Raman spectroscopy in soil science was elaborated in a review paper by Parikh et al. (2014).

Nonetheless, to date, there have been few studies focused on original Raman spectra under a fluorescence background. The relationship between soil Raman spectra and chemical properties have been rarely reported in the open published literature. In this study, we consider, for the first time, that the information contained in the original soil

* Corresponding author.
E-mail address: chwdu@issas.ac.cn (C. Du).

http://dx.doi.org/10.1016/j.geoderma.2016.01.029
0016-7061/

Raman spectra can be used in characterizing arable soil, and the specific objectives of this study are: (i) to summarize the Raman spectra features of different soils, (ii) to perform soil identifications using spectral principal component analysis (PCA) and probabilistic Neural Network (PNN) model, and (iii) to build the prediction model for SOM based on Raman spectra.

2. Materials and methods

2.1. Soil samples

200 soil samples were collected from the top layer (0–20 cm depth) of typical farmlands in China. Field sampling was performed before crop establishment and consisted of four different soil types: black soil from Heilongjiang Province, paddy soil from Jiangsu Province, red soil from Jiangxi Province, and Fluvo-aquic soil from Shandong Province. All sampling plots were geo-referenced using GPS.

2.2. Laboratory analyses

2.2.1. Chemical methods

All bulk soils were submitted to the laboratory, air dried at room temperature (25 ± 2 °C), homogenized, and ground through a 2 mm mesh sieve prior to analysis. SOM was chemically determined using a potassium dichromatic oxidation titration method (Walkley and Black, 1934).

2.2.2. Raman spectra acquisition and spectra pretreatments

Raman spectra of the soils were recorded for each soil at 1 cm^{-1} intervals over the range 180–3200 cm^{-1} using a portable Raman Spectrometer (i-Raman® Plus, USA) with the thermoelectric cooled charge-coupled device detector (TEC-CCD). The excitation wavelength was 785 nm and the resolution was 4.5 cm^{-1}. The same integration time of 60,000 ms was adopted for measuring the spectra of all the soils. In the process of sample measurement, each soil sample was thoroughly mixed and placed in a plastic zip-lock bag with a length of 60 cm and width of 10 cm for laser irradiation (the laser was verified to pass through the zip-lock bag). Dark current was measured before each test of the soil samples to enhance the signal-to-noise ratio. For each sample, the spectra was automatically measured three times in situ and then averaged for further analysis.

The measured spectra were smoothed using Fourier transform filtering (FFT) before data transformation to minimize the systematic noise, and the number of spectra variables was reduced from 3022 to 1000 from this operation. Prior to PCA analysis, linear baseline exclusion was performed using the software PeakFit v4.12 in the spectra range of 1,000 to 2000 cm^{-1} to mitigate disturbances from environments and experimental operations. The function 'map std' in Matlab 2013a was performed, and by which the input and target data can be transformed to a uniform scale by mapping its mean and standard deviations to 0 and 1, respectively (Du et al., 2010).

2.3. Statistical analysis

2.3.1. Probabilistic neural network (PNN)

The probabilistic neural network (PNN) model was built in this study for soil identification. The PNN algorithm was first introduced by Donald Specht (Kusy and Zajdel, 2014)., it is a feed forward network for pattern recognition with a faster training rate and higher prediction accuracy than traditional neural networks (Bhatt and Helle, 2002). Based on a Bayesian classification rule, and the approach realizes pattern prediction by constructing best correlations from the input data. A typical PNN model sequentially consists of input layers, pattern layers (model layers), sum layers, and output layers (contest layers) (Fig. 1). x_k (k = 1, 2…n) denotes the feature vectors to be input and y_k (k = 1, 2…n) denotes the output vectors (Hampson et al., 2001; Jeong et al., 2014). The neurons of the input layers, which are responsible for receiving and linearly transferring the variables to pattern layers, equal in amount to the dimensions of the feature vectors of the learning samples. The numbers of nodes in the pattern layers are determined by the product of the input samples and categories to be matched with. The pattern layer is designed to calculate the matching relation between the feature vector and the samples of the training data. And then the sum layers add the input data from the same type of pattern layers to obtain the maximum probability that the input samples are within this category. By receiving various probability density functions from the sum layers, the output layers present the final judging results (Li et al., 2010).

150 soil samples (75% of total soil samples) were randomly selected to train the network, and the remaining 48 samples (approximately 25%) were used for model validation. The PNN model was configured in Matlab R2013a using the code "net = Newman (p, t, Spread)", where p represented the input vectors, which referred to the first ten PC scores, and t represented the target vectors, the four soil groups.

2.3.2. Partial least squares regression (PLSR)

In this study PLSR was implemented in Matlab 2013a software to relate the soil Raman intensity to the measured SOM. PLSR is one of the most commonly used multivariate algorithms for regressing in the analysis of spectral data (Shi et al., 2014). The PLSR algorithm can produce reliable prediction results by means of integrating independent spectral variables with the properties studied. Information regarding the PLS algorithm was detailed by Wold et al. (2001).

Soil samples were randomly split into 150 samples (75%) for the calibration (cross-validation) set and 50 samples (25%) for the validation set. The statistics of the calibration and validation set for PLS models are provided in Table 1. The optimal number of latent variables were determined by leave-one-out cross validation for better model construction while avoiding over-fitting (Awiti et al., 2008).

The performance of the constructed models are usually evaluated using the coefficient of determination (R^2) and root mean squared error (RMSE) of the test sets. RMSE was calculated as Eq. (1):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (x_i - y_i)^2}, \quad (1)$$

where (x_i-y_i) refers to the differences between the measured and predicted values using PLSR and n is the number of samples. Another commonly used parameter to assess the robustness of the model is the ratio of performance to standard deviation (*RPD* value) (Wu et al., 2010), which can be calculated by dividing the standard deviation of the measured values by the standard deviation of the predicted values. In soil science analysis, if *RPD* < 1.0, the model is unacceptable for prediction

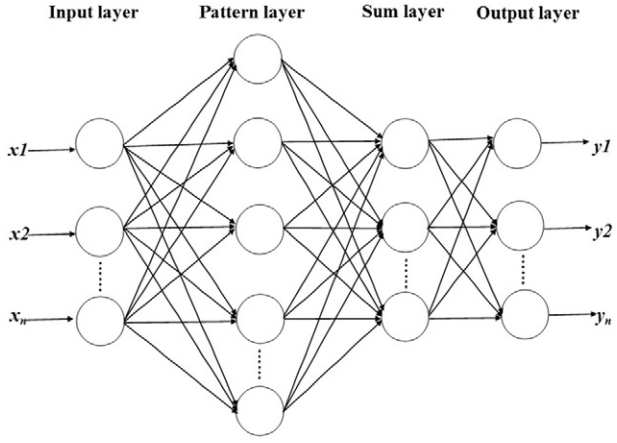

Fig. 1. Illustration of a probabilistic neural network (PNN), including input, pattern, sum and output layers. x_k (k = 1, 2…n) denotes the feature vectors to be input and y_k (k = 1, 2…n) denotes the output vectors.