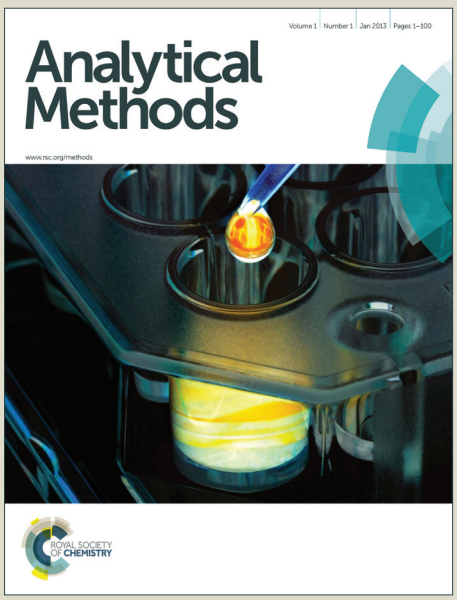
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Classification of soil samples based on Raman spectroscopy and X-ray fluorescence spectrometry combined with chemometric methods and variable selection

Aderval S. Luna*[a], Igor C. A. Lima[a], Werickson F.C. Rocha[b], Joyce R. Araújo[b], O. Kuznetsov[b], Erlon H. Martins Ferreira[b], Ricard Boqué[c] and J. Ferré[c]

[a]*Department of Analytical Chemistry, Rio de Janeiro State University, Rua São Francisco Xavier, 524 – Maracanã, Rio de Janeiro, RJ, 20550-013, Brazil. E-mail: asluna@uerj.br*

[b]*National Institute of Metrology, Quality and Technology-INMETRO, Av. Nossa Senhora das Graças, 50 – Xerém, Duque de Caxias, RJ, 25250-020, Brazil,*

[c]*Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, C/Marcel.lí Domingo, s/n – 43007, Tarragona, Spain*

Abstract

Soil classification is crucial for its cultivation preparation in countries that export several agricultural commodities. The soil classification system adopted in Brazil is based on chemical parameters and physical and morphological changes. This system possesses disadvantages because many analyses are time-consuming, especially during the sample preparation stage. Raman spectroscopy is a non-destructive technique that enables rapid soil sample characterisation. In this study, Raman spectroscopy was used to discriminate different soil samples. Although the Raman spectra of a substance can be used as a phase fingerprint due to its specificity, this technique is not adequate for sample discrimination and suffers from matrix interferences, especially during the analyses of soil samples. However, a synergic effect with satisfactory results regarding prediction and classification problems occurs when this method is coupled with chemometric tools. In this research, a robust classification method for analysing soils using Raman spectroscopy combined with a support vector machines (SVM-C) method and genetic algorithm (GA) for variable selection was developed. The results obtained from the combination of the proposed GA-SVM-C based on the figures of merit were sensitivity (1.000), specificity (1.000), and misclassification error (0.0%) in the validation step. This soil discrimination methodology was validated using X-ray fluorescence spectrometry. These tools can be used in routine analyses, reducing laboratory costs with good efficiency.

Keywords*:* soils, supervised classification, Raman spectroscopy, variable selection, support vector machines, genetic algorithm

1.Introduction

Soil classification enables adequate preparation for cultivation, which is important for countries that export various agricultural commodities. The Brazilian soil classification system is based on chemical parameters and physical and morphological changes. However, this system possesses disadvantages because some analyses are time-consuming, mainly due to the sample preparation stage. The chemical elemental composition of soil samples is routinely evaluated using X-ray fluorescence spectrometry; this technique is a primary method for inferring the origin and classification of a soil [1].

Raman spectroscopy is an invaluable technique for the chemical and structural characterisation of materials. Although the Raman signal of a substance can be used as a phase fingerprint due to its specificity, this method is not discriminatory and suffers from matrix interference, which can particularly complicate the analyses of complex materials. Recently, Ishikawa and Gulick showed that artificial neural networks (ANN) can be trained to accurately identify key minerals for characterising the composition of igneous rock using spectral data acquired via *in situ* Raman spectroscopy [2].

Soil contamination by energetic materials (EMs) is a serious concern in many countries. Calibration methods of the first order (Principal Component Regression and Partial Least Squares, at the percent level) have been used to obtain the Raman signatures of EMs from spectra obscured by extensive backgrounds caused by soil components. The authors demonstrated that Raman spectroscopy can be a useful screening technique at the percent level (by mass) if the soil fluorescence interference can be overcome [3]. The magnetic methods can be complemented with Raman spectroscopy to distinguish the different existing iron oxides in soil; however, some minerals cannot be easily distinguished from each other with these magnetic strategies. A magnetic method to differentiate lepidocrocite and ferrihydrite in soil samples was also validated. Raman spectroscopy is also a simple method for identifying magnetite and maghemite [4]. Laser Raman spectroscopy has been employed to examine four lunar soils. The Raman peak shift was used to calculate Mg/(Mg + Fe + Ca) and Ca/(Mg + Fe + Ca) for pyroxene and Mg/(Mg + Fe) for olivine, and thus, the compositional distributions of these two minerals in each of the four lunar soils were obtained. Classification of feldspar grains has been made based on their Raman patterns [5]. Raman spectroscopy has also been used to develop a method based on a tree-like classifier that differentiates between organic and inorganic particulate matter. This method employs a tree-like structure to classify Raman spectra as a decision tree. The optimal classifier is an ANN, linear discriminant analysis or SVM, in which different kernels are likely. SVM strategies are optimised using the simulated annealing method to achieve the optimal classifier. After training, an old-out experiment with two entirely different sets of Raman spectra were attempted to obtain the abilities of this method for real-world application [6]. Therefore, these studies indicate that Raman spectroscopy coupled to chemometric techniques can be used to classify Brazilian soils.

2.Chemometric methods

Two major areas, classification and calibration, exist in the chemometrics field [7]. Classification can be divided into two groups: supervised and unsupervised [8]. In supervised classification (or supervised pattern recognition), the class that produces each pattern in the modelling sample is known. The classifier is trained to replicate the correct decision for new samples. In unsupervised classification (or pattern recognition unsupervised) the training standards have not been classified, and therefore, algorithms must locate a data structure that allows them to be divided into groups. Once the information is less available, the classification is less accurate than that obtained with supervised methods. However, this method is the only possible solution to problems that do not have information regarding the groups that generated the data.

In this paper several, supervised chemometric methods were used to treat unsupervised data with respect to soils; non-supervised methods are not commented because they have not produced satisfactory results.

The supervised classification methods used herein are described briefly below.

2.1 PLS-DA

PLS-DA is a supervised classification technique used in chemometrics. The concept of PLS-DA is identical to that of the PLS model used for calibration, however, using dummy variables of **Y** matrix. Briefly, the algorithm uses a set of spectra that is selected from each class for training (calibrating) the model, and the second set of spectra is reserved for testing (validating) or verifying the model. This input matrix (matrix **X**) is regressed against the second matrix (matrix **Y**), which contains information concerning each class [9]. Commonly in PLS-DA, the **Y** matrix contains 1 s and 2 s with 1 s, which indicate that the corresponding spectrum from the input matrix is a member of the class; the **Y** matrix contains 1 s to 5 s for the five classes in a data set: soil # 1, # 2 soil, soil # 3, # 4 soil and soil # 5.

2.1 SVM

SVM and chemometrics methods can be used for both calibration and classification in situations in which the data are not linear, i.e., when the linear models do not work well [10-13]. The SVMs are defined by means of the kernel function [14]. The most common ones are linear kernels for linear classification, polynomial kernels for polynomial decision boundaries (e.g., a parabola), and Gaussian kernels for general nonlinear, semiparametric decision boundaries. In this process, the data are mapped into a higher dimensional input space, and an optimal separating hyperplane is constructed in this space [11,15]. For details of applications and algorithms, see previous reports from Ivanciuc [16], Bishop [17] and Cherkassky [18].

2.2 Methods of variable selection

Variable selection is a technique that aims to identify a subset of variables that are, for a given problem, most useful for more precise and exact models. Furthermore, the selected subset may help make the model more easily interpretable chemically, which is crucial in different areas of chemistry because, in a set of hundreds or thousands of variables, some can be noise and/or contain irrelevant information. The objective of variable selection is to significantly reduce the number of variables to obtain simpler, more interpretable and robust models [19].

A large number of procedures for variable selection is available in the literature, most focused on the selection of wavelengths in spectroscopy. These procedures can be distinguished from each other using mathematical criterion to find the optimal variables.

In this work, two variable selection techniques were employed. The first method was based on intervals for the development of the iPLS-DA model, and the second one was based on Darwin's natural selection theory, in which a mathematic tool, genetic algorithm (GA), was used for the construction of GA-PLS-DA and GA-SVM models.

The iPLS-DA algorithm is a tool for the selection of spectral variables; the whole spectrum is divided into several equal subintervals in which PLS-DA models will be constructed for each subdivision [20]. Therefore, the optimal spectral regions are highlighted to construct the model based on a comparison of the root mean square error of cross validation (RMSECV) value for regions in comparison to the global model, which is used throughout the spectrum. Therefore, the regions that presented the lowest RMSECV value related to the global model are selected. Generally, the construction of PLS-DA models based on the selected variables have a number of latent variables different from the overall model to achieve a **Y** relevant variance mainly due to the width of the intervals, the number of substances that absorb/interfere and noise [21].

According to Leardi [22], GAs are quite a recent optimisation technique with the basic concept of mimicking the evolution of the species, according to the Darwinian theory of the "survival of the fittest." The application of GAs to complex problems generally produces much better results than those obtained with standard techniques.

GAs consider that the chromosome has "p" genes, where each gene represents one of the variables of the analytical signal (Raman spectrum in this case) being the number of genes equal to the number of variables contained in the signal. In the variable selection, it is used binary code to encode the problem. Each gene can assume a value of one or zero. When the position of a certain variable is equal to one, it is selected; if the position contains a zero value, the variable will not be selected. Techniques based on variable selection using a GA involve five steps: encoding of variables (as mentioned above), the generation of an initial population, evaluation of response, crossover and mutation. For details of each one of these steps, see Leardi [22] and Davis [23]. In this work, GA theory was used to select the variables for the development of GA-

PLS-DA and GA-SVM models. The PLS-DA and SVM model construction concept has already been described in items 2.1 and 2.1, respectively.

In this work, Raman spectroscopy coupled to chemometric data analysis was successfully used to classify Brazilian soils provided from different origins. This method of soil sample discrimination was validated using X-ray fluorescence in which similarities in chemical composition of soil samples were identified.

3. Experimental

3.1 Sample preparation

Five Brazilian soil samples were provided by Embrapa (Brazilian enterprise for research in agronomy) (Rio de Janeiro, Brazil). The soil samples were oven-dried for 48 hours, disaggregated with a rubber hammer and sieved to 2 mm before analyses. Figure 1 displays the five types (classes) of studied soil.

FIGURE 1

3.2 Sample identification

X-Ray fluorescence spectra were obtained using an S4-Pioneer spectrometer from Bruker AXS GmbH, Karlsruhe, Germany. X-Ray fluorescence analyses were realised using four different analyser crystal types because each analyser crystal is sensible to different energy ranges: lithium fluoride crystals (LiF 200), in which the lattice plane corresponds to (200) and an element range of > K Ka_1, pentaerythrite (PET), which includes an Al-Ti energy range, multilayer W/Si (OVO-55), which includes an O-Si energy range, and multilayer V/C, which is sensible to the energy of carbon.

Raman spectra were acquired with a Renishaw InVia spectrometer (Renishaw plc, Gloucestershire, UK) using a UV laser (325 nm), motorised stage and microscope to focus the sample. The laser spot was approximately 2 μm in diameter, and the laser intensity was approximately 10 mW at the objective exit.

Soil samples were homogenised and gently compacted in a tablet form before analyses. Raman spectrum analyses were then performed on 25 different points of each sample, forming a square matrix of 200 μm x 200 μm for representative sampling.

Figure 2 exhibits the data collection and transformation necessary to chemometric methods for classifying soils according to the following stages:

An image is obtained directly from the surface of the soil and consists of a spectral hypercube of raw data, where the x- and y-axes represent the spatial coordinate characteristic of the soil surface and the z-axis defines the wavelength (λ). Thus, for each pixel image, there is a full spectrum in the third dimension.

In the next step, for each soil sample, the data are unfolded via spatial dimension elimination, resulting in matrix (xy)×λ. The pixel position information is maintained on the rows of this matrix.

The unfolded data, one for each soil sample, are used to obtain a data matrix. After this step, the spectra were divided into two sets, one for calibration and the other for data validation. The spectra selection was

performed using the Kennard-Stone algorithm for each class. This algorithm is a traditional method to extract a representative set of objects from a given data set [24-25]. Notably, each pixel obtained in this study is considered to be a sample. In addition, for the type of soil, # 2 pixel was discarded because it was very different from the others.

Finally, the following chemometric methods were used for soil classification: partial least squares discriminant analysis (PLS-DA), interval partial least squares discriminant analysis (iPLS-DA), GA with partial least squares discriminant analysis (GA-PLS-DA), and SVM for classification (SVM-C) with and without variables selection.

FIGURE 2

3.3 System used for data analyses

All calculations were carried out using MatLab software version 7.7 (MathWorks, USA) with PLS toolbox 6.5 (Eigenvector Research, USA).

4. Results and discussion

4.1 X-Ray fluorescence (XRF)

In this study, five different types of soil were classified and discriminated using X-ray fluorescence. Figure 3 shows the elemental compositions of the types of soil from different origins. All soil types contained Al (Figure 3a) and Si (Figure 3b), with the highest concentration of these elements in soil #1, #4 and #5 (Figure 3). Silicon is a compositional element in the most common soil minerals, such as silica (quartz) and aluminosilicates (clays) [26]. Raman spectroscopy corroborates the presence of quartz mineral in these soil samples. Aluminum and silicon oxide can also be available in the humic fraction of soils that include the sulphide fraction [27]. Aluminum occurs in the following soluble forms: Al^{3+}, Al-OH, and A1-SO_4 (especially in forest soils). A soluble form of Si can occur as orthosilicic acid, which exhibits increasing solubility with increasing pH; however, elements P, Al, Ca and Fe can replace this dependence.

FIGURE 3

Elements of Ca (Figure 4c), K (Figure 4d), Mg (Figure 4e) and Mn (Figure 4f) are usually present in soil in the formation of oxides or in association with the carbonate anion [28]. The presence of these metals is more pronounced in soil #2 and #5. Magnesium appears as magnesium oxide and is often associated with bioaccumulation by plants [29] and can also be found in association with other metals, such as Zn, Al and Fe in minerals [30]. Calcium cations are typically available in association with organic matter and with humic fractions [31].

Soil #3 demonstrated a high amount of Cu (Figure 4 g), Ni (Figure 4 h), P (Figure 4i) and S (Figure 4j). Sulphur constitutes sulphate anions, whereas phosphorous in soils appears in apatite and phosphate-

containing compounds [28]. Other transition metals, including Ni and Cu, are extremely vital in associations with organic compounds. These metals have greater mobility in alkaline soils [26].

The highest concentration of Fe was detected in soil #1 (Figure 4l). Fe can be present in carbonates and associated to organic matter [26-27]. Fractions of humic compounds serve a significant role in the distribution of Fe and in its solubility, which increases with decreasing pH. Fe-containing minerals and organic complexes are extremely vital in soil-forming processes. These compounds strongly influence the distribution of other elements. All Fe-containing minerals have a high sorption capacity for other metals. Fe mobility is increased in organic compounds, as is its bioavailability in plants [32]. The highest amount of titanium among the studied samples was also found in soil #1 (Figure 4m). Titanium in soils can exist as the TiO_2 rutile form, embedded in quartz minerals (SiO_2) [33].

Soil # 4 shows a variety of metals in its composition: Fe (Figure 4l), K (Figure 4d), Ca (Figure 4c), and Cr (Figure 4n) and high content of oxygen atoms (Figure 4o). These results suggest the presence of oxidised carbon groups, such as carboxylic acids, in this soil (see the carbon elemental spectra in Figure 4p). For the typical range of soil pH, carboxyl groups are deprotonated, generating carboxylate anions. Negative charge in the soil surface favours the metal adsorption phenomena and promotes high fertility [34-35].

Zirconium (Figure 4q) showed higher intensity in the FRX spectra in soil # 1. Zr and quartz minerals can generate high resistance to weathering in soils [36]. In the other soil samples, only traces of Zr were observed.

FIGURE 4

4.2 Raman spectra analyses

The Raman spectra of the analysed soils were remarkably similar in general, presenting sharp peaks in the low wavenumber region (200 cm^{-1} to 1000 cm^{-1}) and a broad band at approximately 1600 cm^{-1} with intense fluorescence. Fluorescence is commonly found in organic and amorphous compounds and can often prevent the detection of the weak Raman bands, particularly in the visible region of the spectrum. A way to circumvent this obstacle is changing the excitation energy to NIR or to low UV region [37], which was performed in this work.

The broad band found at approximately 1600 cm^{-1} is attributed to aromatic carbon stretching modes (C=C) (Figure 5) [38]. Although the carbon content in soils is not high, as observed in our XRF analyses, the C=C mode has a strong Raman scattering cross section, explaining the high intensity of this peak. In the low wavenumber region, Raman peaks that can be attributed to crystal lattice vibrations are observed [39]. Soils are formed primarily of silica (SiO_2), feldspar and other minerals. The XRF data showed that the most abundant elements are Si, Al, Fe, Ca, Mg, K, Na, and Mn, most likely associated with oxygen to form oxides. In such a complex matrix as soil, assigning mineral-specific bands is difficult, as most bands have remarkably similar spectra that can overlap with each other. The most predominant bands are observed at 264, 320, 354, 396, 465, 487, 637, and 800 cm^{-1}, and their relative intensities vary from soil to soil. One

particular case can be identified; however, soils # 4 and # 5 had distinctly different spectra from the other Raman spectra, providing an unusually strong band at 465 cm^{-1} and a small band at 354 cm^{-1}, which are the typical bands of quartz (α-SiO_2) [40].

FIGURE 5

4.3 Data analyses

After characterising the different soils through X-ray fluorescence spectroscopy, Raman spectral profiles were used to make chemometric models of classification.

The first step in the construction of these models was the data processing method selection. Several pre-processing methods and combinations were tested. The results indicated that the best estimate was mean-centring [41].

Figure 6 below shows the step-by-step of the construction of the calibration and validation arrays. The spectra of all the samples were arranged in matrix $\mathbf{X}_{calibration}$ (75 x 1562) and in matrix $\mathbf{X}_{validation}$ (49 x 1562), where each row represented a sample (pixel), and the columns contained the equivalent spectroscopic responses. Thus, the calibration and validation sets were constructed with 75 spectra (15 of each soil) and 49 spectra (soils # 1, # 3, # 4 and # 5 with ten spectra, and nine spectra in soil # 2), respectively, in total.

FIGURE 6

4.4 Figures of merit

In routine analytical chemistry, a large body of literature has been devoted to verifying the accuracy and precision of most techniques; however, validation, in the context of pattern recognition, has not been strongly emphasised. Understanding how classification methods can be validated is necessary for establishing a criterion for success [41]. In this work, the comparison of the models was performed using the following statistical parameters: sensitivity (sens), specificity (spec) and misclassification error (ME%). The misclassification error was calculated based on the following equation (1):

$$ME\ \% = \frac{y_i - y_{ref}}{y_{ref}} \times 100\ \%, \quad (1)$$

where y_i represents the class observed, and y_{ref} denotes the reference class. For calculation details, see previous work [25].

The other figures of merit can be defined as the following: N^a represents the number of spectra in each soil; N^b defines the number of misclassified spectra soils; ME (%) is the Misclassification Error; TP denotes the proportion of positive cases that were correctly identified; FP is the proportion of negatives cases that were incorrectly classified as positive; TN represents the proportion of negatives cases that were classified correctly; and FN defines the proportion of positive cases that were incorrectly classified as negative.

Therefore, the sensitivity (sens) of the method is defined by TP/(TP + FN), and the specificity (spec) of the method is defined by TN/(TN + FP) [25].

Thus, four different approaches were used to classify the five distinct classes of soils using various chemometric methods for classification. The first strategy, PLS-DA and SVM-C were used without variable selection methods, but the other three strategies employed different chemometric methods with variable selection. Described below are the four strategies used.

4.5 Application of the classification methods for Raman spectra data

4.5.1 Strategy 1: Methods of classification using PLS-DA and SVM-C without variable selection

For the construction of the PLS-DA model, the first step is to determine the number of latent variables. The PLS-DA method was constructed with two latent variables based on Haaland and Thomas criteria [42].

To create the SVM model, choosing an appropriate kernel function and determining its optimal parameters for generating the optimal SVM model is essential. That is, the model should possess a lower RMSECV value. In this work, we optimised the cost parameters (C = 10), number of support vectors (75), and gamma (γ = 1e-006). As reported by Liu, the regularisation parameter, γ, determines the tradeoff between minimising the training error and minimising the model complexity [43].

The results obtained were inadequate for both models, as shown in the misclassification errors of classification (% ME) for both calibration and/or validation sets. For the PLS-DA method, soil # 4 exhibited % Mes of 53.3% and 70.0% in the calibration and validation sets, respectively. When using the SVM-C model, soil # 3 showed % MEs equal to 100% and 30% in the calibration and validation sets, while soil # 5 showed 100% in both sets. These results showed high % MEs. An alternative method with which we could test our hypothesis was the variable selection strategy; we proposed that this strategy would enhance the discriminant analysis.

Three analysis strategies were applied for variable selection. The first method uses interval PLS (iPLS), and the others employ GAs coupled to PLS-DA or SVM-C, respectively. Thus, the spectral regions were selected for each algorithm used for building the chemometric models.

4.5.2 Strategy 2: Methods of classification using iPLS-DA

In constructing the model iPLS forward-DA, we used the following parameters; number of intervals, 23; range size, 25 and maximum, 10 latent variables. Thus, using the iPLS-DA algorithm selected the following spectral regions: 1494-1421, 1265-1190, 874-797, 714-556 and 389-147 cm^{-1}, as shown in Figure 7. These selected intervals showed a lower RMSECV.

FIGURE 7

The results obtained using the iPLS-DA method for the classification of five soils are shown in Table 1, which lists the parameters used to evaluate them in the calibration and validation sets. From the data in Table 1, in the calibration model, one spectrum was incorrectly classified in soil # 2 (% ME = 6.7), and three spectra were incorrectly classified in soil # 5 (% ME% = 20). The validation step showed no incorrectly classified samples with a sensitivity equal to 1 (i.e., all of the samples were designated for that class assigned to the appropriate class) for all of the classes (soil). The specificities were equal to 1 (i.e., no sample was incorrectly classified), from soil # 1 to # 3. However, soils # 4 and # 5 showed specificities of 0.744 and 0.462, respectively.

The theoretical model has high sensitivity and specificity. These results indicate that the probability that the sample was not properly being classified as not belonging to that class is high. Due to the low specificity of soil # 5 in the validation set, another tool for variable selection was used.

4.5.3 Strategy 3: Methods of classification using GA-PLS-DA

Because the obtained results were now inadequate, the GA was chosen for variable selection. The parameters used for optimisation were the following: population size, 76; window width, 5%; initial terms, 30; mutation rate, 0.05; and crossover, single. The selected wavelengths are shown in Figure 8, in which the frequency with which genes were crossed are displayed (crossing-over).

FIGURE 8

The calibration and validation sets used to conduct GA-PLS-DA models were the same as previously described. The results obtained with the GA-PLS-DA strategy for the classification of five soils are shown in Table 2, which lists the parameters used to evaluate it in both calibration and validation sets. In the calibration step, the misclassification errors were 6.7 and 20.0% for soils # 2 and # 5, respectively. All soils had high sensitivity; however, soil # 5 showed small specificity in the calibration and validation sets. In the validation stage, none of the five soils were classified incorrectly, with sensitivities equal to 1 in all classes and specificities equal to 1 (soils # 1 to # 3), 0.744 (soil # 4) and 0.462 (soil # 5), as observed in Table 2. Even with the variable selection procedure change, a misclassification error of 20% for soil # 5, which is indicative of the inadequate model, as well as unsatisfactory results for the specificities of soils # 4 and # 5. Thus, in predictions, the probability that the sample is not classified correctly as not belonging to that class is high. Therefore, a non-linear method was used to overcome this problem.

4.5.4 Strategy 4: Methods of classification using GA-SVM-C

The calibration and validation sets used to the conduct GA-SVM-C models were the same aforementioned sets. In the calibration step, the GA-SVM-C was constructed with the following adopted parameters: radial basis function, cost (C = 31.6), gamma (γ = 1e-006) and the number of support vectors (SV = 56). The

results obtained using the GA-DA-C for classifying the five soils are shown in Table 3, which lists the parameters used to evaluate it in both calibration and validation sets.

With these parameters, only soil sample # 2 was incorrectly classified (% ME = 6.7) in the calibration set, and all samples were classified correctly in the validation set. Notably, in this discriminant analysis, all calibration (except soil # 2 with sens = 0.933) and validation samples demonstrated sensitivities equal to 1 and specificities equal to 1 (except for soil # 1 in the calibration set with spec = 0.983).

The GA-SVM-C model was the best classification model found as presented for almost all soils a sensitivity equal to one (except for soil # 2 in the rankings) and specificity equal to one (except for the soil # 1 in the calibration set). The sensitivity values for soil # 2 indicate that the probability of a false positive is < 7%, and the probability of a false negative for soil # 1 is < 2%. This model is the most appropriate for forecasting the classification of samples.

5. Conclusions

Based on this research, an XRF method that allows immediate elemental analysis of soil samples from different regions was established. Raman data analyses revealed the presence of aromatic rings, as evidenced by the high-frequency region of different spectra. The low-frequency region of the Raman spectra indicates the presence of metal oxides and minerals in the samples. Although Raman spectroscopy does not directly provide the phase composition of complex systems such as soil, the presented approach to data analyses offers insight into this characteristic of the studied system. In the examined soil samples, the detected elements constituted the oxides, organic matter and humic compounds in metal complexes (Fe, Ca).

The use of Raman spectroscopy using classification models and variable selection provides a simple and powerful tool for differentiating different soil samples. In particular, for this study, our optimal result was obtained from the combination of GA-SVM-C based on the figures of merit, sensitivity, specificity and misclassification error, in which distinguishing all five types of soil was possible. Our results demonstrate the potential use of these tools in routine analyses, reducing laboratory costs.

Acknowledgements

S. Luna greatly acknowledges CNPq, Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), for project financial support and Programa Prociência (UERJ/FAPERJ) for grants. I. C. A. Lima thanks FAPERJ for a scholarship.

W. F. C. Rocha thanks FAPERJ for financial support of E-26/111.051/2013 and E-26/110.637/2013 research projects.

The authors thank EMBRAPA SOLOS researchers Maurício Rizzato Coelho and André Marcelo de Souza for providing the soil samples.

References

1 F. L. Melquiades, L.F.S. Andreoni and E.L. Thomaz, *Appl. Radiat. Isotopes,* 2013; **77**, 27-31.

2 S.T. Ishikawa and V.C. Gulick, *Computers & Geosciences,* 2013; **54**, 259-268.

3 D. S. Moore, *Fresen. J. Anal. Chem.,* 2001; **369**, 393-396.

4 M. Hanesch, *Geophys. J. Int.,* 2009; **177**, 941-948.

5 Z.C. Ling, A.Wang and B.L. Jolliff, *Icarus,* 2011; **211**, 101-113.

6 W. Schumacher, M. Kühnert, P. Rösch and J. J. Popp, *Raman Spectrosc.,* 2011; **42**, 383-392.

7 M. Otto, Chemometrics: statistics and computer application in analytical chemistry. Wiley-VCH, New York, 2007.

8 M. R. de Almeida, D. N. Correa, W. F.C. Rocha, F. J.O. Scafi, R. J. Poppi, *Microchem. J.,* 2013; **109**, 170–177.

9 T. M. C. Pereira, J. A. Q. Júnior, R. S. Ortiz, W. F. C. Rocha, D. C. Endringer, P. R. Filgueiras, R. J. Poppi, W. Romão, *Anal. Methods*, 2014; **6**, 2722-2728.

10 B. G. Vaz, P. V. Abdelnur, W. F. C. Rocha, A. O. Gomes, R. C. L. Pereira, *Energ. Fuels*, 2013; **27**(4) 1873 – 1880.

11 W. F. C. Rocha, B. G. Vaz, G. F. Sarmanho, L. H. C. Leal, R. Nogueira, V. F. Silva, C. N. Borges, Anal. Lett., 2012; **45**, 2398 – 2411.

12 C. J. C. Burges, Data Min. Knowl. Discov., 1998; **2,** 121–167.

13 N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, Cambridge, 2000.

14 S. D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics - Chemical and Biochemical Data Analysis, Elsevier Science B. V., Amsterdam, 2009.

15 H. Li, Y. Liang, Q. Xu, Support vector machines and its applications in chemistry, *Chemometr. Intell. Lab. Syst.,* 2009; **95**(2), 188 – 198.

16 O. Ivanciuc, Applications of Support Vector Machines in Chemistry. In: Reviews in Computational Chemistry, volume 23, Eds.: K. B. Lipkowitz, T. R. Cundari, Wiley-VCH, Weinheim, 2007, pp.291-400.

17 C. M. Bishop, Neural networks for pattern recognition, Oxford University Press, Oxford, UK, 1995.

18 V. S. Cherkassky, F. M. Mulier, Learning from data: concepts, theory, and methods. Wiley & Sons, Chichester, UK, 2007.

19 D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetry, Part A, Elsevier Science B. V., Amsterdam, 1997.

20 J. Wagner, Guideline for Interval PLS, p.2-10, 2000. http://www.models.kvl.dk/source/ipls. Accessed on June 08th, 2006.

21 S. D. Osborne, R. B. Jordan, R. Kunnemeyer, R. Analyst, 1997; **122**, 1531–1537.

22 R. Leardi, *J. Chromatogr. A,* 2007; **1158** (1–2) 226–233.

23 L. Davis, (Ed.) Handbook of genetic algorithms, Van Nostrand Reinhold, New York, 1991

24 R. Wehrens, Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences, Springer, Berlin, 2011.

25 M. J. C. Pontes, R. K. H. Galvão, M. C. U. Araújo, P. N. T. Moreira, O. D. P. Neto, G. E. José and T. C. B. Saldanha, *Chemometr. Intell. Lab. Syst.,* 2005; **78**, 11-18.

26 R. Baranowski, A. Rybak and I. Baranowska, *Pol. J. Environ. Stud.,* 2002; **11**, 473-482.

27 J. C. Joo, M.-S. Song and J.-K. Kim, *J. Environ. Sci. Heal. A,* 2012; **47**, 909-918.

28 F. J. Stevenson and M.A. Cole, in Cycles of Soils: Carbon, Nitrogen, Phosphorus, Sulfur, Micronutrients, John Wiley & Sons, New York, 1999, pp. 70-72.

29 J. Kaiser, M. Galiová, K. Novotný, R. Červenka, L. Reale, J. Novotný, M. Liška, O. Samek, V. Kanický, A. Hrdlička, K. Stejskal, V. Adam and R. Kizek, *Spectrochim. Acta Part B,* 2009; **64**, 67-73.

30 M.S. Torn, S.E. Trumbore, O.A. Chadwick, P.M. Vitousek and D.M. Hendricks, *Nature,* 1997; **389**, 170-173.

31 J. Lehmann, D. Solomon, J. Kinyangi, L. Dathe, S. Wirick and C. Jacobsen, *Nature Geosciences,* 2008; **1**, 238-242.

32 J. Six, R.T. Conant, E.A. Paul and K. Paustian, *Plant Soil,* 2002; **241**, 155-176.

33 G. Meinhold, *Earth-Sci. Rev.,* 2010; **102**, 1-28.

34 J. Ni, J.J. Pignatello and B. Xing, *Environ. Sci. Technol.,* 2011; **45**, 9240-9248.

35 D. Xu, S. Zhu, H. Chen and F. Li, *Colloid Surface A,* 2006; **276**, 1-7.

36 R. Parahyba, M.C. Santos and F.C.R. Neto, *Rev. Bras. Ciênc. Solo,* 2009; **33**, 1991-2000.

37 S. A. Asher, in Handbook of Vibrational Spectroscopy, (Eds: J. M. Chalmers, P. R. Griffiths), John Wiley & Sons Ltd., Chichester, 2002, pp. 557 – 571.

38 A.C. Ferrari and J. Robertson, *Phys. Rev. B,* 2000; **61**, 14095-14107.

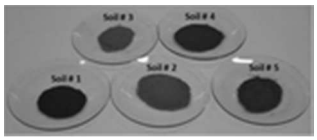
39 http://www.ijvs.com/volume3/edition4/section1.html#Feature. R.L. Frost, J. Kloprogge, J. Schmidt, *The Internet Journal of Vibrational Spectroscopy* 3(4).

40 C. V. Raman and T. M. K. Nedungadi, *Nature,* 1940; **145**, 147-147.

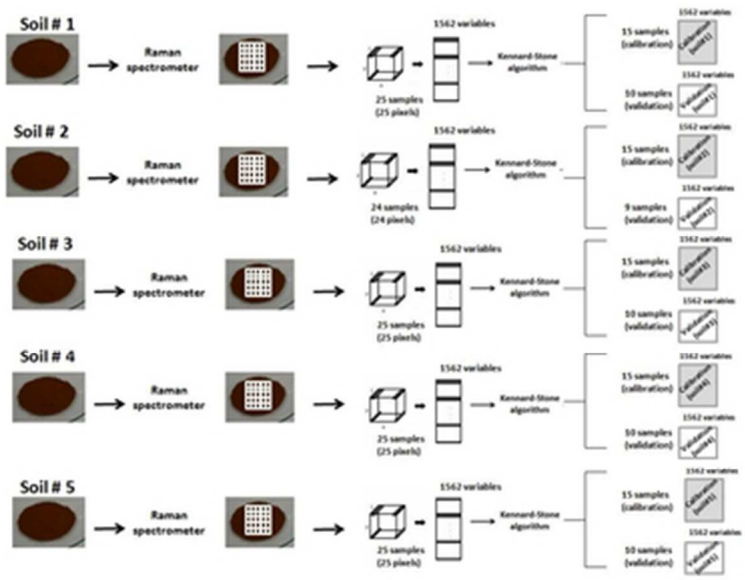
41 R.G. Brereton, Chemometrics for Pattern Recognition, John Wiley & Sons Ltd., Chichester, 2009.

42 D. M. Haaland and E.V. Thomas, *Anal. Chem.,* 1988; **60**, 1193-1202.

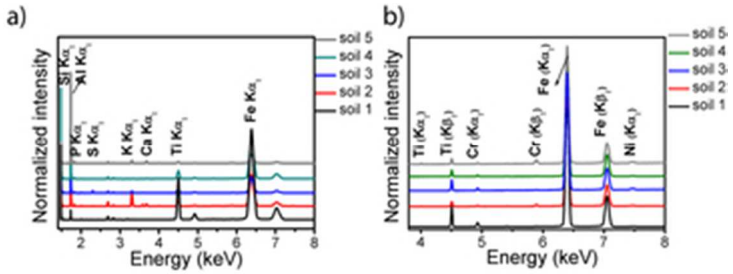
43 F. Liu, Y. Jiang and Y. He, *Anal. Chim. Acta,* 2009; **635**, 45-52.



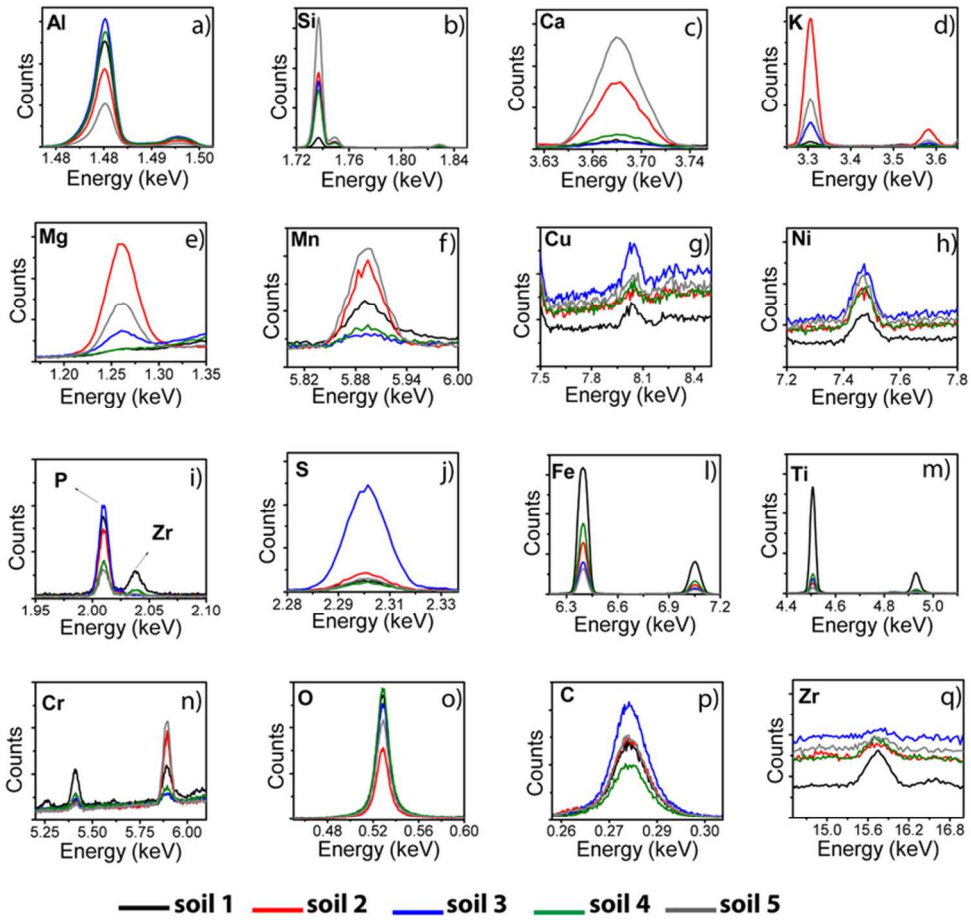
6x2mm (600 x 600 DPI)



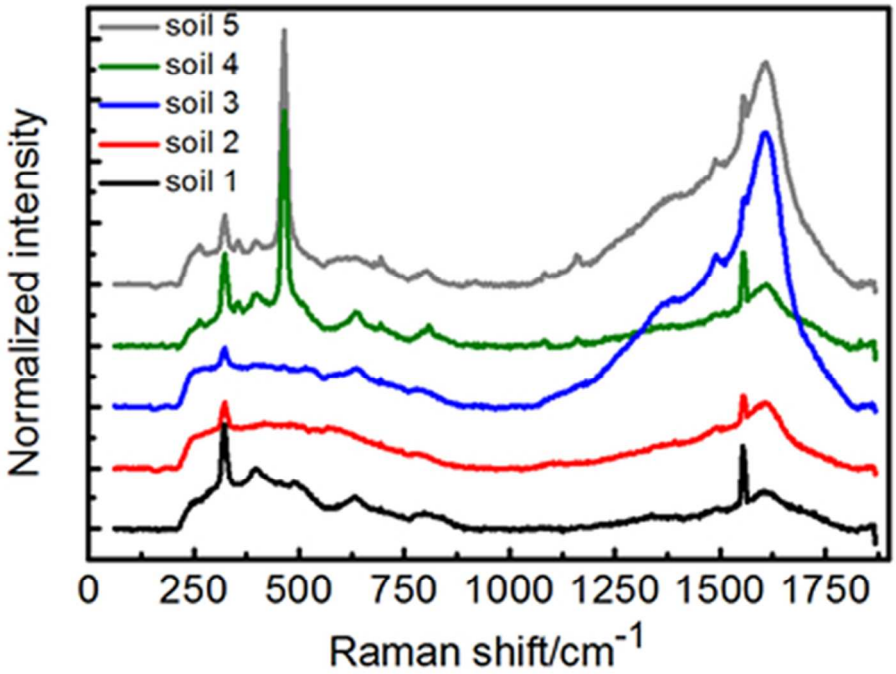
15x12mm (600 x 600 DPI)



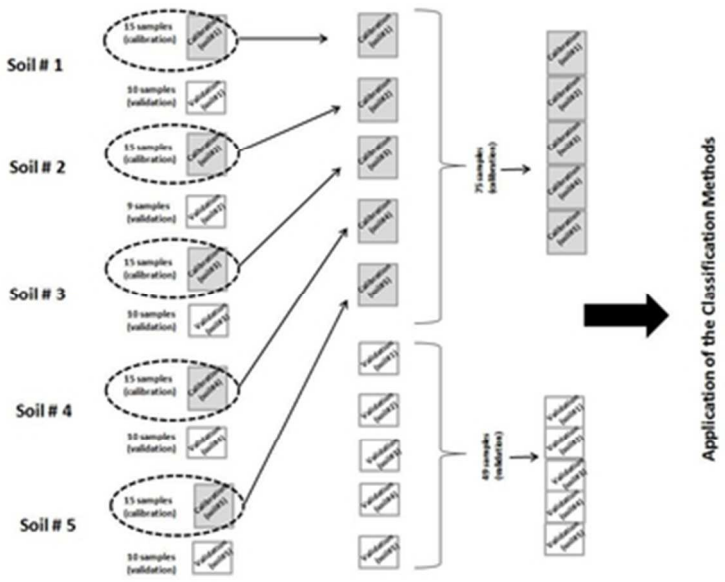
30x11mm (300 x 300 DPI)



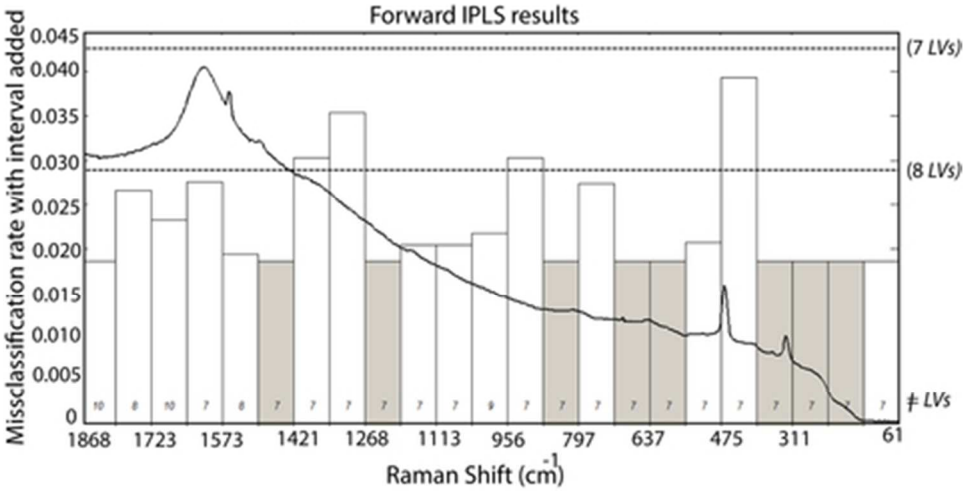
73x67mm (300 x 300 DPI)



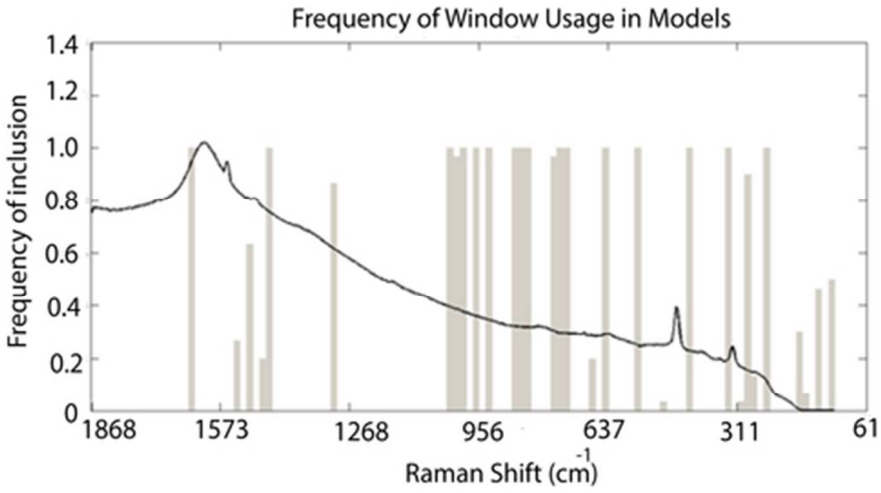
56x39mm (300 x 300 DPI)



15x12mm (600 x 600 DPI)



40x20mm (300 x 300 DPI)



46x27mm (300 x 300 DPI)

Captions of figures

Figure 1-Photo of the five types (classes) of the studied soil.

Figure 2- Scheme of collect of data hypercube following the unfolded data for construction of the data matrix.

Figure 3 X-ray fluorescence spectra of five soil samples using two different analyzer crystals: a) pentaerythite (PET) and b) lithium fluoride (LiF 200). 30x11mm (300 x 300 DPI)

Figure 4 X-ray fluorescence spectra by chemical element of the five soil samples showing relative proportion of a) Aluminum, b) Silicon, c) Calcium, d) Potassium, e) Magnesium, f) Manganese, g) Copper, h) Nickel, i) Phosphorous, j) Sulfur l) Iron, m) Titanium, n) Chromium, o) Oxygen, p) Carbon and q) Zirconium.

Figure 5 Raman spectra of the five soil samples. Part of fluorescence background has been removed using a double linear baseline subtraction in the intervals 0 to 1000 cm^{-1} and 1000 to 1800 cm^{-1}.

Figure 6 Strategy for the separation of samples to be used in the chemometric model for classification

Figure 7 Selected regions of the Raman spectrum (gray) by forward iPLS in the calibration model.

Figure 8 Selected regions of the Raman spectrum (gray) by genetic algorithm in the calibration model.

Table 1 Results obtained using iPLS-DA (2 latent variables for each class)

Subset	Calibration					Validation				
Group	Soil 1	Soil 2	Soil 3	Soil 4	Soil 5	Soil 1	Soil 2	Soil 3	Soil 4	Soil 5
N^a	15	15	15	15	15	10	9	10	10	10
N^b	0	1	0	0	3	0	0	0	0	0
ME (%)	0.0	6.7	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0
TP	1.00	0.93	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00
FP	0.02	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TN	0.98	0.97	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FN	0.00	0.07	0.00	0.00	0.20	0.00	1.00	0.00	0.00	0.00
Sens	1.00	0.93	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00
Spec	0.98	0.93	0.98	0.73	0.45	1.00	1.00	1.00	0.74	0.46

N^a: number of spectra in each soil;

N^b: number of misclassified spectra soils;

ME (%): Misclassification Error;

TP: True Positive;

FP: False Positive;

TN: True Negative;

FN: False Negative;

Sens: Sensitivity;

Spec: Specificity.

Table 2 Results obtained using GA-PLS-DA (2 latent variables for each class)

Subset	Calibration					Validation				
Group	Soil 1	Soil 2	Soil 3	Soil 4	Soil 5	Soil 1	Soil 2	Soil 3	Soil 4	Soil 5
N^a	15	15	15	15	15	10	9	10	10	10
N^b	0	1	0	0	3	0	0	0	0	0
ME(%)	0.0	6.7	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0
TP	1.00	0.93	1.00	1.00	0.86	1.00	1.00	1.00	1.00	1.00
FP	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TN	0.98	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FN	0.00	0.07	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00
Sens	1.00	0.93	1.00	1.00	0.86	1.00	1.00	1.00	1.00	1.00
Spec	0.98	0.93	1.00	0.73	0.43	1.00	1.00	1.00	0.74	0.46

N^a: number of spectra in each soil;

N^b: number of misclassified spectra soils;

ME (%): Misclassification Error;

TP: True Positive;

FP: False Positive;

TN: True Negative;

FN: False Negative;

Sens: Sensitivity;

Spec: Specificity.

Table 3 Results obtained using GA-SVM-C

Subset	Calibration					Validation				
Group	Soil 1	Soil 2	Soil 3	Soil 4	Soil 5	Soil 1	Soil 2	Soil 3	Soil 4	Soil 5
N^a	15	15	15	15	15	10	9	10	10	10
N^b	0	1	0	0	0	0	0	0	0	0
ME(%)[c]	0.0	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TP	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FP	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TN	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FN	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sens	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Spec	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

N^a: number of spectra in each soil;

N^b: number of misclassified spectra soils;

ME (%): Misclassification Error;

TP: True Positive;

FP: False Positive;

TN: True Negative;

FN: False Negative;

Sens: Sensitivity;

Spec: Specificity.