Geoderma 354 (2019) 113793

Contents lists available at ScienceDirect

Geoderma

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

The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges

José A.M. Demattê[a,*], André Carnieletto Dotto[a], Ariane F.S. Paiva[a], Marcus V. Sato[a], Ricardo S.D. Dalmolin[b], Maria do Socorro B. de Araújo[c], Elisângela B. da Silva[d], Marcos R. Nanni[e], Alexandre ten Caten[f], Norberto C. Noronha[g], Marilusa P.C. Lacerda[h], José Coelho de Araújo Filho[i], Rodnei Rizzo[j], Henrique Bellinaso[k], Márcio R. Francelino[l], Carlos E.G.R. Schaefer[l], Luiz E. Vicente[m], Uemeson J. dos Santos[n], Everardo V. de Sá Barretto Sampaio[n], Rômulo S.C. Menezes[n], José João L.L. de Souza[o], Walter A.P. Abrahão[l], Ricardo M. Coelho[p], Célia R. Grego[m], João L. Lani[l], Antonio R. Fernandes[q], Deyvison A.M. Gonçalves[q], Sérgio H.G. Silva[r], Michele D. de Menezes[r], Nilton Curi[r], Eduardo G. Couto[s], Lúcia H.C. dos Anjos[t], Marcos B. Ceddia[t], Érika F.M. Pinheiro[t], Sabine Grunwald[u], Gustavo M. Vasques[v], José Marques Júnior[w], Airon J. da Silva[x], Marcos C. de Vasconcelos Barreto[x], Gabriel N. Nóbrega[y], Marcelo Z. da Silva[z], Sara F. de Souza[aa], Gustavo S. Valladares[ab], João Herbert M. Viana[ac], Fabricio da Silva Terra[ad], Ingrid Horák-Terra[ad], Peterson R. Fiorio[ae], Rafael C. da Silva[a], Elizio F. Frade Júnior[af], Raimundo H.C. Lima[ag], José M. Filippini Alba[ah], Valdomiro S. de Souza Junior[ai], Maria De Lourdes Mendonça Santos Brefin[aj], Maria De Lourdes P. Ruivo[ak], Tiago O. Ferreira[a], Marny A. Brait[al], Norton R. Caetano[am], Idone Bringhenti[am], Wanderson de Sousa Mendes[a], José L. Safanelli[a], Clécia C.B. Guimarães[a], Raul R. Poppiel[h], Arnaldo Barros e Souza[a], Carlos A. Quesada[an], Hilton T. Zarate do Couto[ao]

[a] Department of Soil Science, Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP), Ave. Pádua Dias 11, Cx. Postal 9, 13418-900, Piracicaba, São Paulo, Brazil
[b] Department of Soil, Federal University of Santa Maria, Av. Roraima 1000, 97105-900 Santa Maria, Rio Grande do Sul, Brazil
[c] Geographical Sciences Department, Federal University of Pernambuco, Av. Ac. Hélio Ramos, s/n, 50740-530, Recife, Pernambuco, Brazil
[d] Agricultural Research and Rural Extension Corporation of Santa Catarina, Rodovia Admar Gonzaga 1347, 88034-901, Florianópolis, Santa Catarina, Brazil
[e] Department of Agronomy, State University of Maringá, Av. Colombo 5790, 87020-900, Maringá, Paraná, Brazil
[f] Department of Agriculture, Biodiversity and Forestry, Federal University of Santa Catarina, Rodovia Ulysses Gaboardi 3000 - Km 3, 89520-000 Curitibanos, Santa Catarina, Brazil
[g] Federal Rural University of Amazon, Ave. Presidente Tancredo Neves 2501, 66077-530 Belém, Pará, Brazil
[h] Faculty of Agronomy and Veterinary Medicine, University of Brasília, 70910-900 Brasília, Distrito Federal, Brazil
[i] EMBRAPA - Solos, R. Antônio Falcão, 402, Boa Viagem, 51020-240 Recife, Pernambuco, Brazil
[j] Center of Nuclear Energy in Agriculture (CENA), USP, Av. Centenário 303, 13416-000 Piracicaba, São Paulo, Brazil
[k] CDRS/Secretary of Agriculture of São Paulo State, R. Campos Salles 507, 13400-200 Piracicaba, São Paulo, Brazil
[l] Department of Soils, Federal University of Viçosa, Ave. Peter Henry Rolfs s/n, 36570-900 Viçosa, Minas Gerais, Brazil
[m] EMBRAPA – Informática Agropecuária, Ave. André Tosello, 209, 13083-886 Campinas, São Paulo, Brazil
[n] Department of Nuclear Energy, Federal University of Pernambuco, Av. Prof. Luis Freire 1000, 50740-540 Recife, PE, Brazil
[o] Department of Geography, Federal University of Rio Grande do Norte, R. Joaquim Gregório s/n, 59300-000 Caicó, Rio Grande do Norte, Brazil
[p] Agronomic Institute of Campinas (IAC), Ave. Barão de Itapura 1481, 13020-902, Campinas, São Paulo, Brazil
[q] Institute of Agricultural Sciences, Federal Rural University of Amazônia, Ave. Presidente Tancredo Neves 2501, 66.077-830, Belém, Pará, Brazil
[r] Department of Soil Science, Federal University of Lavras, 37200-000 Lavras, Minas Gerais, Brazil
[s] Federal University of Mato Grosso, Cuiabá, Av. Fernando Corrêa da Costa 2367, 78060-900 Mato Grosso, Brazil
[t] Department of Soils, Federal Rural University of Rio de Janeiro, Rodovia BR 465, Km 07 s/n, 23890-000, Seropédica, Rio de Janeiro, Brazil
[u] Soil and Water Sciences Department, University of Florida, 2181 McCarty Hallr, PO Box 110290, 32611, Gainesville, Fl, USA

* Corresponding author at: Department of Soil Science, Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP), Ave. Pádua Dias 11, Cx. Postal 9, 13418-900 Piracicaba, São Paulo, Brazil
E-mail address: jamdemat@usp.br (J.A.M. Demattê).
URL: http://www.bv.fapesp.br/en/pesquisador/2291/jose-alexandre-melo-dematte (J.A.M. Demattê).

https://doi.org/10.1016/j.geoderma.2019.05.043

Available online 05 August 2019

[v] EMBRAPA - Solos, R. Jardim Botânico, 1024, 22460-000 Rio de Janeiro, RJ, Brazil
[w] Department of Soils and Fertilizers, School of Agricultural and Veterinary Studies, São Paulo State University (FCAV-UNESP), Via de Acesso Prof. Paulo Donato Castellane s/n, 14884-900 Jaboticabal, SP, Brazil
[x] Federal University of Sergipe, São Cristóvão, Av. Marechal Rondon s/n, 49100-000, Sergipe, Brazil
[y] Graduate Program in Earth Sciences (Geochemistry), Department of Geochemistry, Federal Fluminense University, Outeiro São João Batista, s/n, 24020-141 Niterói, RJ, Brazil
[z] Federal Institute of the Southeast of Minas Gerais, R. Monsenhor José Augusto 204, 36205-018 Barbacena, Minas Gerais, Brazil
[aa] Federal University of Rio Grande do Norte, R. Joaquim Gregório s/n, 59300-000 Caicó, Rio Grande do Norte, Brazil
[ab] Federal University of Piauí, 64049-550 Teresina, Piauí, Brazil
[ac] EMBRAPA Milho e Sorgo, Rod MG 424 Km 45, 35701-970 Sete Lagoas, Minas Gerais, Brazil
[ad] Institute of Agricultural Sciences, Federal University of Jequitinhonha e Mucuri Valleys, Ave. Ver. João Narciso 1380, 38610-000 Unaí, Minas Gerais, Brazil
[ae] Department of Biosystems Engineering, ESALQ, USP, Ave. Pádua Dias 11, Cx. Postal 9, 13418-900 Piracicaba, SP, Brazil
[af] Federal University of Acre, Rodovia BR 364 Km 04, 69920-900 Rio Branco, Acre, Brazil
[ag] Federal University of Amazonas, Av. General Rodrigo O. J. Ramos 1200, 69067-005 Manaus, Amazonas, Brazil
[ah] EMBRAPA Clima Temperado, BR-392, km 78, 96010-971 Pelotas, Rio Grande do Sul, Brazil
[ai] Department of Agronomy, Federal Rural University of Pernambuco, R. Manuel de Medeiros s/n, 52171-900 Recife, Pernambuco, Brazil
[aj] EMBRAPA Cocais, Quadra 11, Av. São Luís Rei de França 4, 65067-205 São Luís, Maranhão, Brazil
[ak] Paraense Emílio Goeldi Museum, Av. Gov. Magalhães Barata 376, 66040-170 Belém, Pará, Brazil
[al] Exata Laboratory, Rua Silvestre Carvalho Q 11, 75800-000 Jataí, Goiás, Brazil
[am] Federal University of Rondônia, BR 364, Km 9.5, 76801-059 Porto Velho, Rondônia, Brazil
[an] Nacional Institute for Amazonian Research, Ave. André Araújo 2936, 69067-375 Manaus, Amazonas, Brazil
[ao] Department of Forestry Sciences, ESALQ-USP, Ave. Pádua Dias 11, Cx. Postal 9, 13418-900 Piracicaba, São Paulo, Brazil

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
Spectral sensing
Proximal sensing
Vis-NIR-SWIR spectroscopy
Pedometrics

ABSTRACT

The present study was developed in a joint partnership with the Brazilian pedometrics community to standardize and evaluate spectra within the 350–2500 nm range of Brazilian soils. The Brazilian Soil Spectral Library (BSSL) began in 1995, creating a protocol to gather soil samples from different locations in Brazil. The BSSL reached 39,284 soil samples from 65 contributors representing 41 institutions from all 26 states. Through the BSSL spectra database, it was possible to estimate important soil attributes, such as clay, sand, soil organic carbon, cation exchange capacity, pH and base saturation, resulting in differences among the multi-scale models taking Brazil (overall), regional and state scale. In general, spectral descriptive and quantitative behavior indicated important relationship with physical, chemical and mineralogical properties. Statistical analyses showed that six basic patterns of spectral signatures represent the Brazilian soils types and that environmental conditions explain the differences in spectra. This study demonstrates that spectroscopy analyses along with the establishment of soil spectral libraries are a powerful technique for providing information on a national and regional levels. We also developed an interactive online platform showing soil sample locations and their contributors. As soil spectroscopy is considered a fast, simple, accurate and nondestructive analytical procedure, its application may be integrated with wet analysis as an alternative to support the sustainable management of soils.

1. Introduction

Soil is a fundamental natural resource for sustaining life in the planet and economic development, since it provides several ecosystem services and is the basic resource for many human activities (Adhikari and Hartemink, 2016; Jónsson and Davíðsdóttir, 2016). Thus, the knowledge about soil physical and chemical properties and their spatial variabilities are the essence for their sustainable use, planning and adequate management, aiming for greater productivity and conservation (Wall and Nielsen, 2012).

Earth has about 150 million km^2 of land and most of it is not totally known in terms of soil surface composition, which is usually made by traditional wet analysis. Since this technique has been used for > 100 years, it is considered the most relevant to characterize soil properties. However, this approach suffer from the usage of chemical reagents and time consuming (Viscarra Rossel et al., 2016). Besides, there are still uncertainties and discussion of current methods and its results, which frequently lead to difficulties in the interpretation and misleading communication. These issues took research to seek for other strategies on to optimize and/or assist these previous and important wet methods.

Proximal sensing research community has applied spectroscopy techniques systematically on the last 40 years to reach soil properties with important results (Nocita et al., 2014). The spectral range commonly used to study spectral pattern of soils corresponds to 400–700 nm (visible - Vis), 700–1100 nm (near infrared - NIR) and 1100–2500 nm (shortwave infrared - SWIR) which can be obtained by sensors in the field or laboratory and has been the baseline for optical aerial/satellite remote sensing (reflectance spectroscopy, imaging spectroscopy). In this case, from the surface reflectance of the samples measured in laboratory, it is possible to develop models relating the spectral pattern to some soil characteristics which can be extrapolate to satellite spectral data, making possible to map large areas (Demattê, 2016). Since Bowers and Hanks (1965), soil reflectance has been studied and reached a strong background on its interpretation. During this period, sensing data has showed a strong relationship with several soil attributes, i.e., soil organic carbon (Stevens et al., 2008), texture (Brodský et al., 2011), mineral composition (Viscarra Rossel et al., 2006), and others (Nocita et al., 2014). The technique allows the simultaneous characterization of soil attributes with the advantage of being a non-destructive method of in situ observation (Viscarra Rossel et al., 2006).

To make spectral information useful for the soil science community, it is imperative to have reference patterns in a database (Viscarra Rossel and Behrens, 2010), commonly named spectral libraries. A diverse database is fundamental to understand soils spectral behavior and reach its attributes prediction from spectra (Shepherd and Walsh, 2002). After this study, others came along such as Brown et al. (2006) and Viscarra Rossel and McBratney (2008). The ICRAF-ISRIC world soil spectral library (Garrity and Bindraban, 2004), for example, is composed of 785 soil profiles from 58 countries from Africa, Europe, Asia, and the Americas. Viscarra Rossel and Webster (2012) described a large spectral library with ~4000 soil profiles covering the Australian continent. A spectral library covering the United States (US) has been setting on the Rapid Carbon Assessment (RaCA) project (Soil Survey Staff, 2014)

with 144,833 Vis-NIR spectral curves from 32,084 soil profiles. The European spectral library called LUCAS consists of about 20,000 topsoil samples, collected from 23 countries in the European continent, and measured for 13 soil properties in a single laboratory (Stevens et al., 2013). Another important example of soil spectral library (SSL) was the ASTER spectral library (Baldridge et al., 2009), a compilation of 2400 spectra of soils, rocks, minerals and other related materials. SSL initiatives in other countries include: Brazil (Bellinaso et al., 2010), Czech Republic (Brodský et al., 2011), France (Gogé et al., 2012), Denmark (Knadel et al., 2012), Mozambique (Cambule et al., 2012), and China

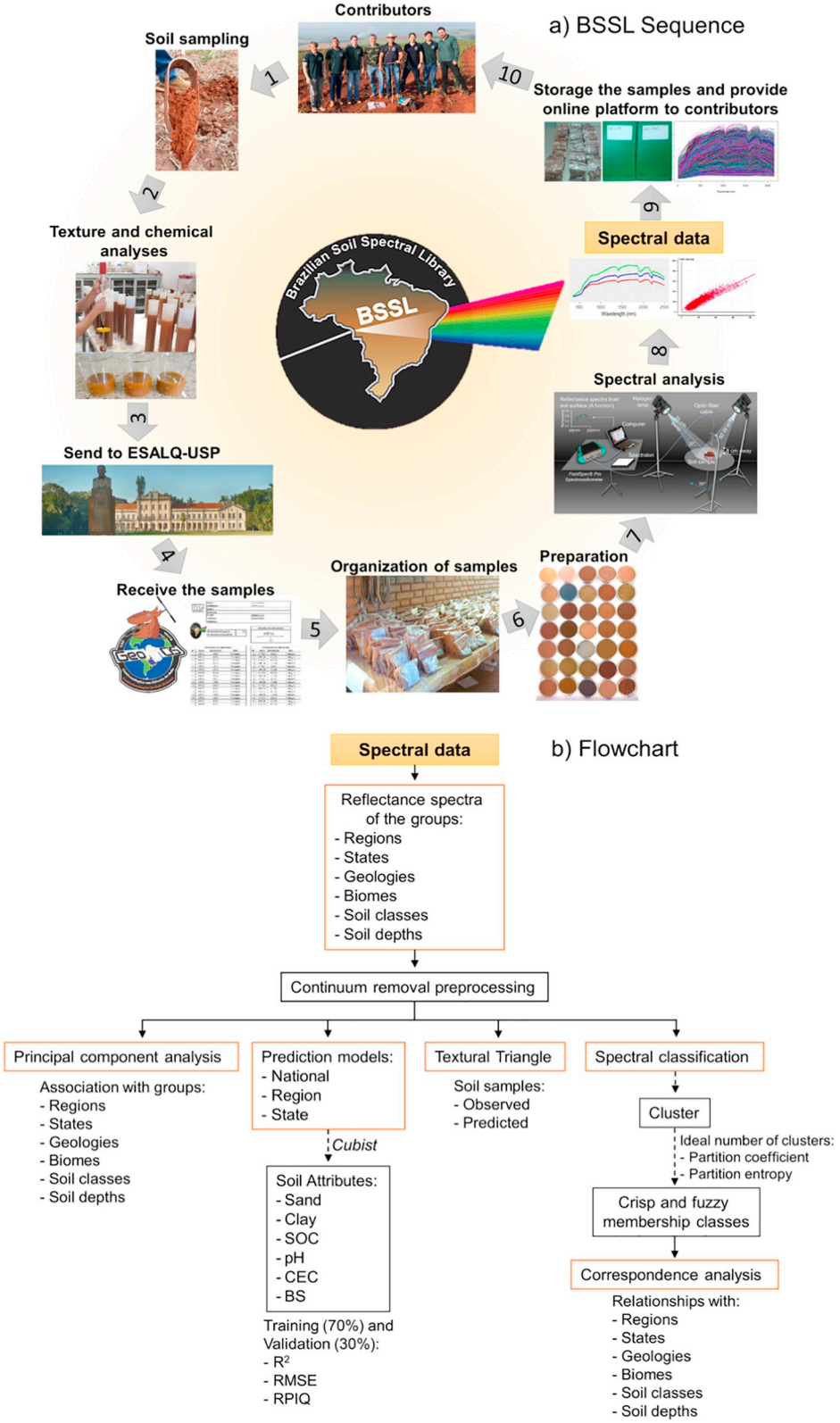

Fig. 1. Methodological sequence for the development of the Brazilian Soil Spectral Library (BSSL) development (a) and flowchart representing the statistical analyses of BSSL(b).

(Ji et al., 2016; Shi et al., 2014). Finally, a world soil spectral library was constructed for soil organic carbon, soil texture, iron, $CaCO_3$, CEC, and pH with 90 participating countries (Viscarra Rossel et al., 2016). Such collaborative initiatives open many doors for its applicability.

Brazil is the largest country in South America, with an area of ~ 851 million ha, and is the fifth largest in the world. It has ~ 152.5 million ha in agricultural land (18% of total). Soil mapping and pedologic properties characterization with conventional survey and laboratory methods is enormous challenge. An example of this effort is the PronaSolos (Polidoro et al., 2016), a forthcoming national program that aims to provide more detailed mapping of soils in Brazil. Thus, soil sensing and the fusion of spectral data are promising allow quick acquisition of information for surveying large areas of soils (Grunwald et al., 2015). The State of São Paulo had its first soil spectral Atlas performed by Epiphanio et al. (1992), which was published afterwards by Formaggio et al. (1996). Bellinaso et al. (2010) and Terra et al. (2015) created soil spectral libraries from states of the South Central of Brazil. However, the country still does not have a standardized SSL to integrate the soil research community and support different applications for studying soil resources.

The objective of this study was to present the first integrated SSL and its relationship with soil attributes and other environmental characteristics covering most of the Brazilian soils. The Brazilian Soil Spectral Library (BSSL) allows the exploration of new approaches on proximal and remote spectral sensing. We hypothesize that spectral data relate to geographical and environmental variables.

2. Material and methods

2.1. The collaborating system

The BSSL started in 1995 with a collection of soil samples from the Department of Soil Science, Luiz de Queiroz College of Agriculture, University of São Paulo (ESALQ-USP), where spectral reflectance was measured and inserted into the database. The collaboration system of the BSSL is shown in Fig. 1a and the flowchart with data description is in Fig. 1b. A dynamic and interactive online platform showing the Brazilian map with all BSSL data was also created. This online platform facilitates the communication between any user who wants to contact the researchers and use their soil spectra dataset. The interactive map can be accessed at < https://bibliotecaespectral.wixsite.com/esalq > .

2.2. Description of the spectral database

The current spectral library contains 39,284 soil samples from 65 contributors representing 41 institutions. The Brazilian spectral database was constructed combining all the soil samples from the collaborators. Fig. 2 shows the maps of Brazil with the region, state, geology, biome, soil class, and sample points. The Brazilian regions are North (N), Northeast (NE), Midwest (MW), Southeast (SE), and South (S) (Fig. 2a). The Brazilian states are Acre (AC), Alagoas (AL), Amapá (AP), Amazonas (AM), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Mato Grosso (MT), Mato Grosso do Sul (MS), Minas Gerais (MG), Pará (PA), Paraíba (PB), Paraná (PR), Pernambuco (PE), Piauí (PI), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rio Grande do Sul (RS), Rondônia (RO), Roraima (RR), Santa Catarina (SC), São Paulo (SP), Sergipe (SE), and Tocantins (TO) (Fig. 2b). The geology is represented by igneous, metamorphic, and sedimentary rocks (Fig. 2c). The biomes are Amazon, Caatinga, Cerrado, Atlantic Forest, Pampa, and Pantanal (Fig. 2d). Only the most representative soil classes are presented in the Brazilian map, which are Lixisols, Ferralsols, and Arenosols (Fig. 2e). The geographic locations of the soil samples are shown in Fig. 2f. When the information provided by the contributors had no geographical coordinates, the points were allocated at the nearest city.

Most of the samples that compose the database came from the SE and MW regions, with 19,429 and 9391 samples, 50% and 24% of the samples, respectively (Fig. 3a). São Paulo (SP), Mato Grosso do Sul (MS) and Goiás (GO) states (26,474 samples total) correspond to 68% of all samples (Fig. 3b). Samples are from soils formed mainly in three lithologic groups, igneous and sedimentary with 10,621 and 10,409 samples, respectively 21,030 in total (Fig. 3c). The most represented biomes are the Atlantic Forest and Cerrado, with 19,248 (53%) and 12,468 (34%), respectively (Fig. 3d). The soil class with most samples is the Ferralsols (22,674 samples, equivalent to 63% of all samples) located mainly in the SE and MW regions (Fig. 3e). Other soil classes represented in the BSSL are Arenosols. Ferralsols and Lixisols represent the two most important soil classes in Brazil, covering about 31.5 and 26.8% of the Brazilian territory, respectively. These two classes represent 86% of all samples in the database (31,551 samples) (Fig. 3e). Considering the total database, 79% of the samples present A (0–20 cm), B (40–60 cm), C (80–100 cm), and D (100–120 cm) layers (Fig. 3f). Approximately 43% of all samples have soil organic carbon (SOC), 85% have granulometry, 35% have cation exchange capacity (CEC), 67% have values of pH in water and 72% have base saturation (BS) ($BS = [Ca + Mg + K + Na]/CEC \times 100$) (Donagemma et al., 2011) measurements.

2.3. Spectral data, preprocessing and transformations

All soil samples from the database were previously dried at 45 °C, ground and sieved with 2 mm mesh and then homogeneously distributed in Petri dishes prior the measurement of the spectra. The spectral data were acquired by the Geotechnologies in Soil Science group (GeoSS), São Paulo, Brazil, using the Fieldspec 3 spectroradiometer (Analytical Spectral Devices, ASD, Boulder, CO), which has a spectral range from visible to shortwave infrared (350–2500 nm) and spectral resolution of 1 nm from 350 to 700 nm, 3 nm from 700 to 1400 nm, and 10 nm from 1400 to 2500 nm. The sampling interval of data output is 1 nm reporting 2151 channels. One of the strengths of the database is that all spectral analyses followed the standardized spectral library analysis protocol.

The spectral sensor, which was used to capture light through a fiber-optic cable, was allocated at 8 cm from the sample surface. The sensor scanned an area of approximately 2 cm^2, and a light source was provided by two external 50-W halogen lamps. These lamps were positioned at a distance of 35 cm from the sample (non-collimated rays and a zenithal angle of 30°) with an angle of 90° between them. A *Spectralon* standard white plate was scanned every 20 min during scanning. Two replicates (one involving a 180° turn of the petri dish) were obtained for each sample. Each spectrum was averaged from 100 readings over 10 s. The mean values of two replicates were used for each sample. Ninety-eight percent (98%) of soil spectra were measured in GeoSS Lab, following the protocol proposed by Ben-Dor et al. (2015). Although the other 2% were not measured by the same equipment, protocols for spectra acquirement were strictly followed. Considering that practically all the spectral library was built with the same protocol, performing a calibration transfer function would demand time and resources, while the improvements would most likely be fairly small.

The spectral reflectance was transformed to continuum removal (CR) (Clark and Roush, 1984). This preprocessing removes the continuous features of spectra and is often used to isolate specific absorption features. The CR creates a continuum or hull similar to fitting a rubber band over the original spectrum. The spectrum is normalized by setting the value of the hull to 100% reflection, where the first and last values of the continuum-removed spectrum equal 1. We applied CR preprocessing because of its strength and ability to enhance absorption depths by correcting apparent shifts from wavelength-dependent scattering, highlighting specific absorption bands of a reflectance spectrum (Mutanga et al., 2005).Besides that, the CR preprocessing is capable of providing calibration models with high accuracy.

We used heuristically testing to optimize the clustering procedure,

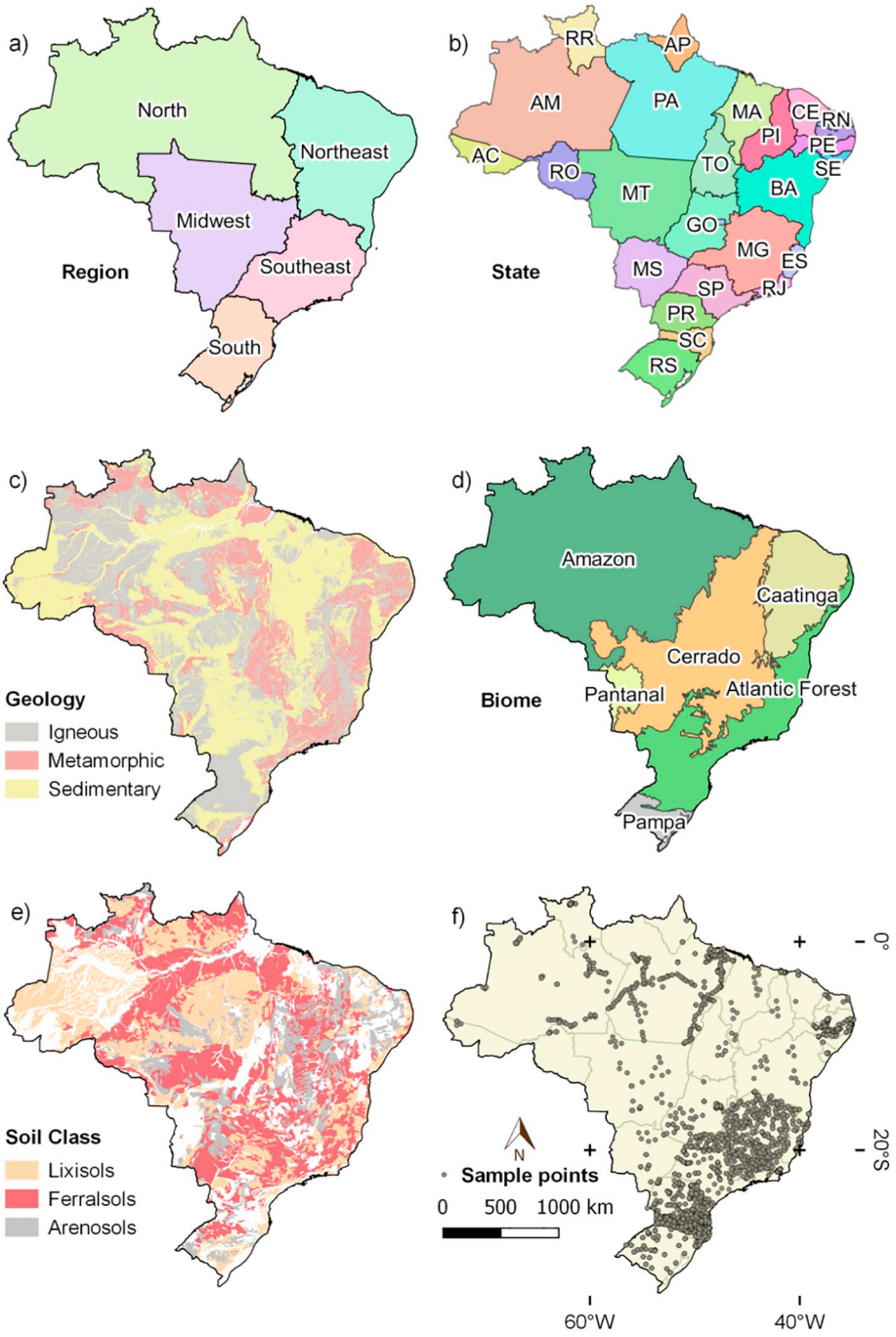

Fig. 2. Maps representing the Brazilian regions (a), states (b), geology (c), biomes (d), main soil classes (e), and sampling locations of the Brazilian Soil Spectral Library (f).

which involved grouping the samples by their reflectance spectra. The reflectance intensity provides important information in the spectral characterization of soils, but in our case, it did not provide good results. Initially, we had clustered the samples based on the reflectance spectra, but the fuzzy performance indicators were not satisfactory. On the other hand, when employing the CR spectra, we not only produced reasonable performance indices but also had results similar to other studies (e.g. Demattê et al., 2016; Terra et al., 2018). Although the reflectance intensity corresponds to a large share of spectral variance, there are other information in the spectrum which are extremely important to soils discrimination (e.g. features related to clay minerals at around 1200, 1900 and 2200 nm). The potential of such information should not be underestimated.

2.4. Principal component analysis

The CR spectra were analyzed by principal component analysis (PCA) to reduce dimensionality and improve computational efficiency. The data was not standardized to make easier the interpretation of absorption features in continuum spectra. We used both the scores and eigenvectors of PCA to assist in the interpretation of BSSL data.

Geographical and environmental characteristics were associated with the spectral data and the Brazilian spectral samples were separated according to 5 regions, 26 states, 3 geologies, 6 biomes, 11 soil classes and 4 soil depths (Fig. 3). The PCA was used to investigate the associations between groups and spectral data. The PCA correlates the average soil spectral reflectances with regions, states, geology, biomes,

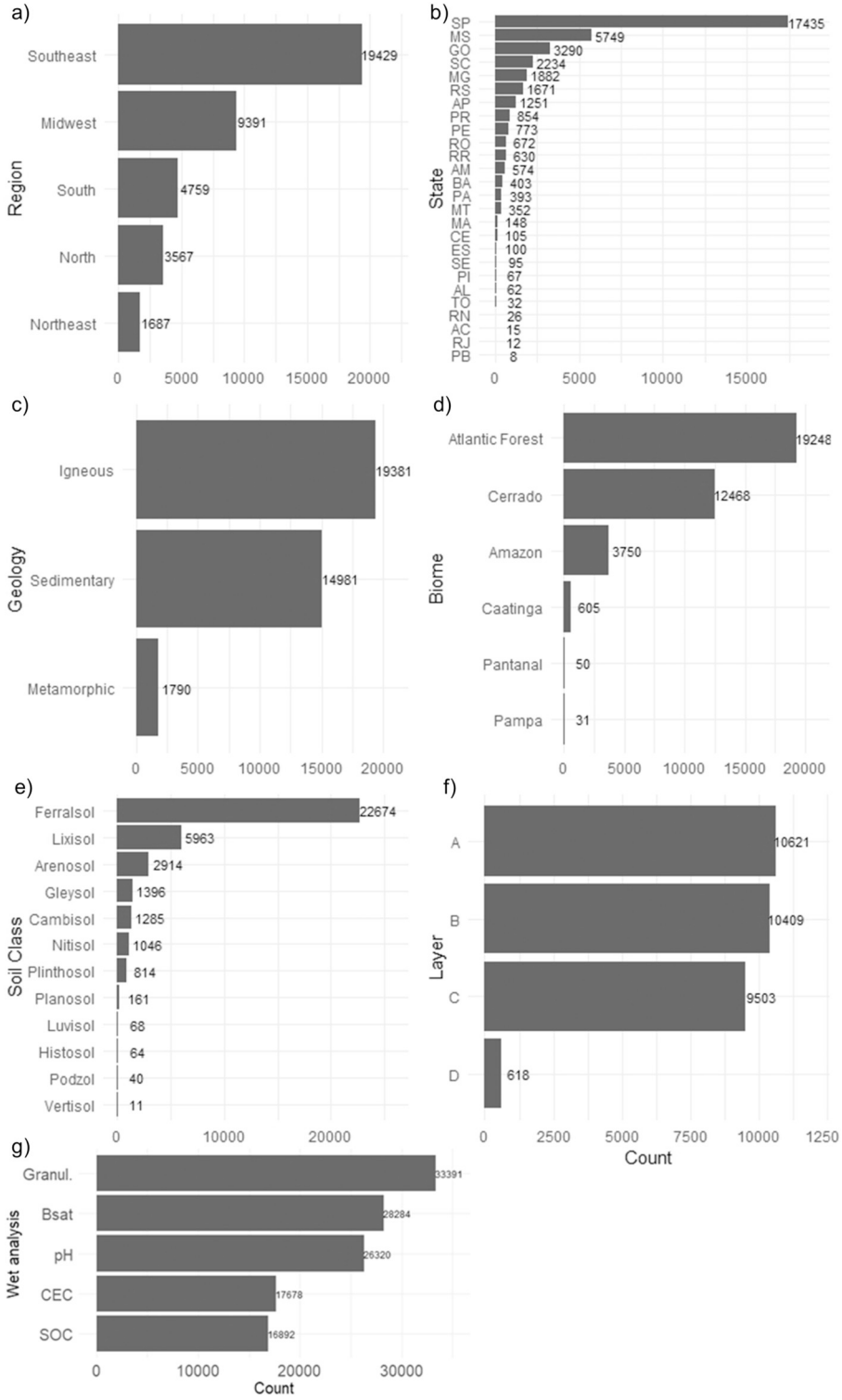

Fig. 3. Distribution of soil samples according to Brazilian regions (a); states (b); geology (c); biomes (d); soil classes (considering layers A and B) (e); soil layers (f); and soil attributes (g). The number of samples varies for each group depending on the available information. Soil classes were defined according to World Reference Base (International Union of Soil Science Working Group WRB, 2015). Soil layers corresponded to A (0–20 cm), B (40–60 cm), C (80–100 cm), and D (100–120 cm).

soil classes, and layers. Soil classes in the Brazilian Soil Classification System were correlated with the WRB classification (IUSS Working Group WRB, 2015). The BSSL presents a large variation of samples considering layers, surface and subsurface horizons, and complete soil profiles. However, only soil samples that had the following depths were selected for PCA with layers' data: A (0–20 cm), B (40–60 cm), C (80–100 cm), and D (100–120 cm). Considering all samples from the spectral database, 84% were taken collected with auger, 12% from complete profiles and 4% only from the surface layer (0–20 cm).

2.5. Spectroscopic modeling of soil attributes

The soil attributes selected for predictive modeling were sand, clay, SOC, pH, CEC, and BS. Several strategies of modeling were performed

to predict these attributes. First, national, regional and state models were developed for each attribute, where the national model included the complete database. The datasets for each soil attribute were separated into training and independent validation by a 70:30 split. This separation was carried out using random division, which was able to separate the groups homogeneously. The cubist method (Quinlan, 1992) was applied to train the spectroscopic models. Cubist applies the M5 (Model Tree approach) to grow categorical decision trees to handle continuous classes by placing a multivariate linear model at each leaf. The model building and estimation process were achieved by the *caret* package (Kuhn et al., 2017) in R (R Core Team, 2018). This package has a set of functions that attempt to streamline the process for creating predictive models. The calibration function was applied to adjust the best fitted model using optimal tuning parameters as follows: cross-validation resampling, committees, and neighbors.

For each soil attribute the performance of the models were assessed by comparing the predicted and observed values based on the independent validation data set. The coefficient of determination (R^2) (Eq. (1)), root mean squared error (RMSE) (Eq. (2)), and ratio of performance to interquartile distance (RPIQ) (Eq. (3)) were assessed to

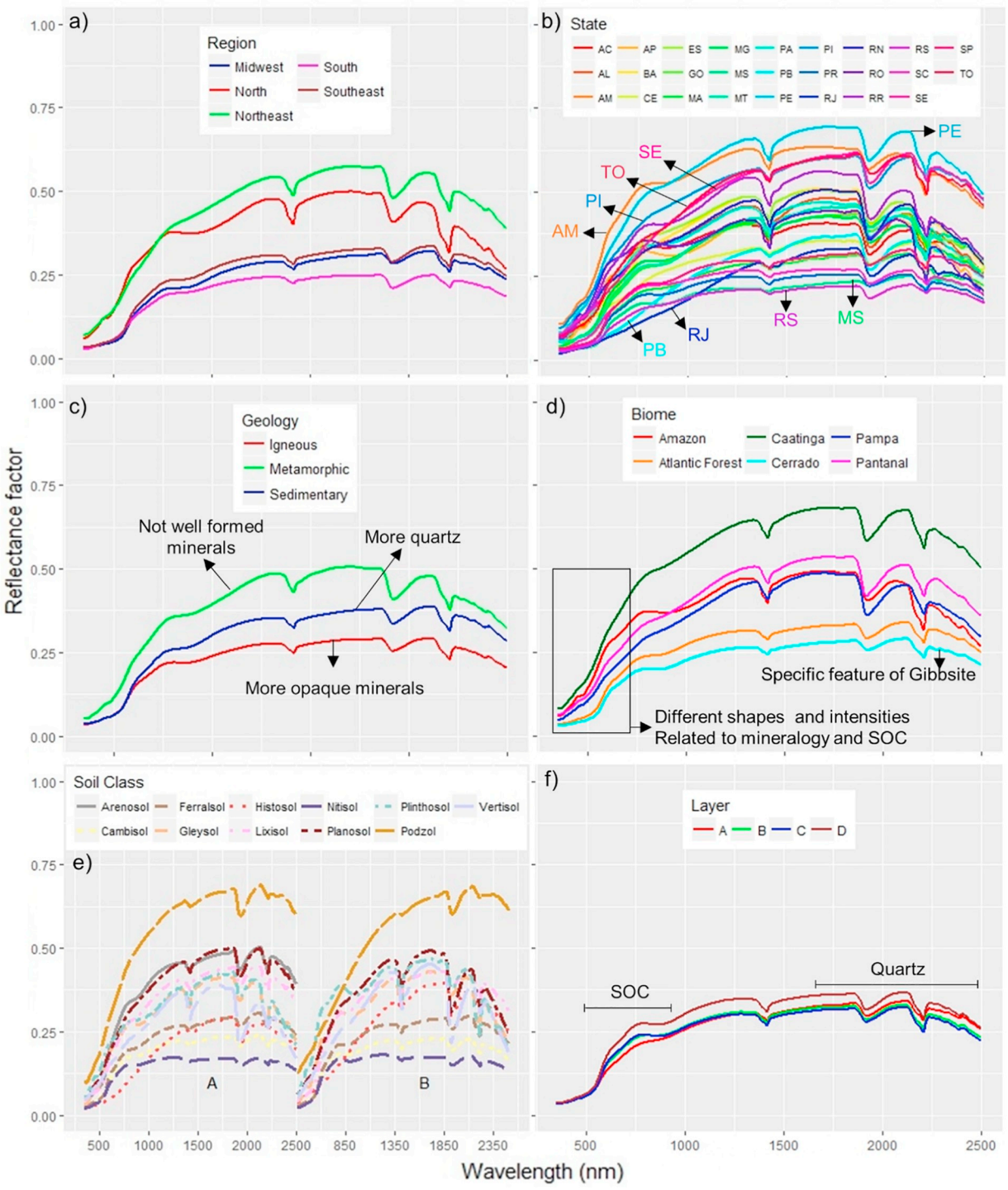

Fig. 4. Soil reflectance spectra averaged according to each region (a); state (b); geology (c); biome (d); soil class (considering layers A and B) (e); and layer (f). The number of samples considered in each group are shown in Fig. 3.

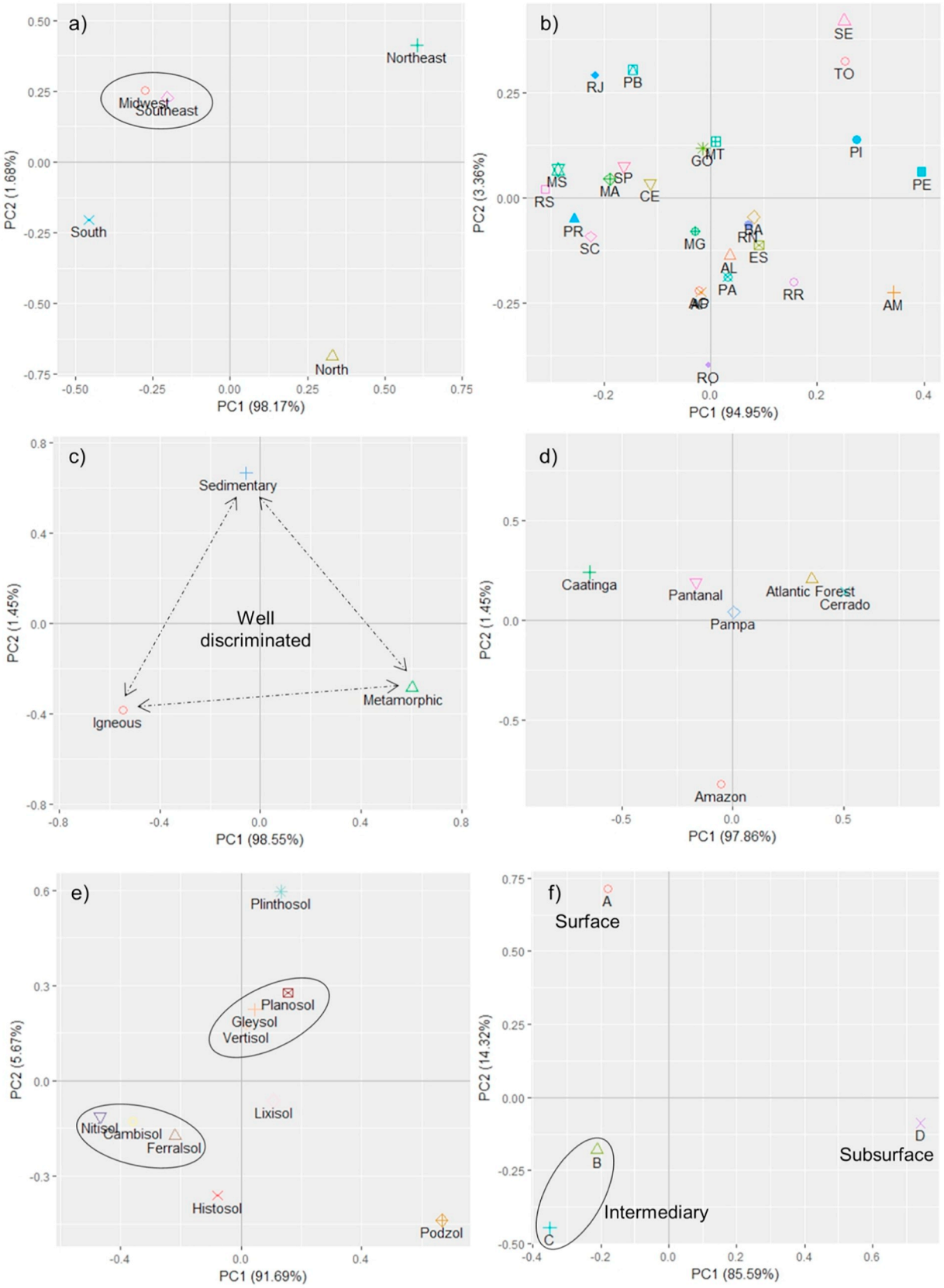

Fig. 5. Principal component scores 1 (PC1) and 2 (PC2) calculated from the average reflectance spectra of each region (a), state (b), geology (c), biome (d), soil class (e), and layer (f).

quantify the inaccuracy of the estimates.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad (2)$$

$$RPIQ = \frac{(Q3 - Q1)}{RMSE} \quad (3)$$

where $\hat{y}$ is the predicted value, $\bar{y}$ is the mean of the observed value, y is the observed values, n is the number of samples with *i* equal to 1, 2, … n, IQ is the difference between the third and first quartiles (Q3 - Q1).

The textural triangle was developed using the reference values for clay, sand and silt, called observed, and using the predicted values for clay and sand obtained by each model, i.e., the predicted values used in the triangle of BSSL were originated from the national model, and the predicted values of S region were originated from the S model. The silt content was calculated by the difference of clay plus sand content. The textural triangle was carried out in R using the *soiltexture* package (Moeys, 2016).

2.6. Spectral patterns classification

The classification of spectral patterns was performed aiming to represent Brazilian soils. The reflectance was transformed to CR, which was used to classify spectra into general groups. The question was: How many classes of spectra were necessary to represent the Brazilian soils? To answer that, we classified the spectra by clustering via similarity measurements. The first three principal component scores were classified by the fuzzy *c*-means algorithm (Bezdek et al., 1984). This approach produces two methods of classification: crisp and fuzzy membership degrees. The first produces the crisp or hard (no-fuzzy) membership degrees of the objects in order to place them into only one discrete cluster. The fuzzy *c*-means technique assigns a fuzzy membership degree to each data point based on its distances to the cluster centers. The fuzzy approach is based on the distance between various input data points (PCA scores). This algorithm assigns a fuzzy membership degree to each data point based on its distances to the cluster centers. The farther from the center of the cluster the smaller the probability of the point being classified in the respective class. The fuzzy membership degrees are continuous and range from 0 to 1. Each sample has a membership in every cluster, where close to 1 indicates a high degree of similarity between the sample and a cluster, while close to 0 implies a low similarity (Bezdek et al., 1984).

Fuzzy clustering requires the user to predefine the number of clusters (*c*), but it is not always possible to know this number in advance. To obtain the ideal number of c-means cluster two validation functions were performed. The two most important validity index functions to determine the optimal number of clusters are as follow: (a) partition coefficient (pC) and(b) partition entropy (pE) (Bezdek et al., 1984). The best performance is achieved when the pC achieves its maximum value or pE obtains its minimum. All analyses and statistical procedures described above were performed by the R programming (R Core Team, 2018). The crisp and fuzzy *c*-means clustering was carried out using the

Table 1
Cubist model parameters, descriptive statistics, and results of prediction models of Soil Organic Carbon (SOC).

SOC (g kg^{-1})		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		8.9	11.9	0.0	431.1	18076	12653	5423	0.82	5.07	1.38	0.78	6.89	0.94
Regions	South	12.0	15.1	0.0	141.4	1833	1283	550	0.82	6.66	2.69	0.71	8.12	2.05
	Southeast	8.0	5.2	0.0	54.9	9252	6476	2776	0.72	2.82	2.06	0.74	2.75	2.11
	Midwest	8.6	5.3	0.6	57.0	3104	2173	931	0.84	2.10	3.04	0.84	2.28	2.55
	Northeast	13.4	35.2	0.0	431.1	1309	916	393	0.87	14.64	0.69	0.79	9.97	1.15
	North	8.2	6.7	0.0	105.6	2578	1805	773	0.64	4.20	1.66	0.58	4.41	1.-80
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	1.5	1.0	0.7	4.1	32	19	13	0.29	0.92	1.49	0.43	0.76	1.22
	AM	8.0	8.1	0.1	105.6	435	304	131	0.78	3.20	1.64	0.72	5.92	1.15
	AP	12.4	6.6	2.0	56.0	817	571	246	0.21	5.70	1.31	0.32	6.09	1.64
	BA	1.8	1.7	0.3	20.2	242	169	73	0.87	0.75	1.64	0.84	0.39	2.31
	CE	40.5	55.7	0.5	310.8	105	73	32	0.98	6.55	5.54	0.93	25.93	0.91
	ES	-	-	-	-	-	-	-	-	-	-	-	-	-
	GO	11.5	6.7	1.7	66.0	618	432	186	0.84	2.70	3.66	0.83	2.73	2.98
	MA	1.5	0.8	0.2	4.0	74	51	23	0.14	0.77	1.25	0.40	0.75	0.96
	MG	11.8	7.6	0.0	59.9	1065	745	320	0.88	2.74	3.81	0.89	2.45	3.85
	MS	7.5	4.3	0.6	32.6	2269	1588	681	0.85	1.73	2.68	0.85	1.72	2.71
	MT	12.8	5.0	4.1	25.6	217	151	66	0.90	1.57	5.38	0.89	1.65	5.11
	PA	6.3	5.3	0.1	38.9	305	213	92	0.61	3.60	1.42	0.69	3.24	1.73
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	10.8	8.7	0.0	70.0	773	541	232	0.84	3.45	2.32	0.84	4.02	2.10
	PI	8.5	7.9	0.1	33.3	67	34	33	0.46	5.88	1.13	0.41	6.57	1.70
	PR	-	-	-	-	-	-	-	-	-	-	-	-	-
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-	-
	RO	7.7	4.3	1.2	20.9	642	449	193	0.66	2.51	1.85	0.65	2.51	1.62
	RR	1.7	0.9	0.0	8.0	379	265	114	0.17	0.87	1.09	0.24	0.73	1.32
	RS	8.2	15.2	0.0	141.4	1238	866	372	0.81	6.86	0.57	0.76	8.57	0.44
	SC	20.0	11.2	0.2	93.2	595	416	179	0.74	5.92	2.51	0.86	3.89	3.86
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	7.5	4.6	0.0	48.0	8185	5729	2456	0.66	2.80	1.90	0.66	2.63	1.99
	TO	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2
Cubist model parameters, descriptive statistics, and results of prediction models of clay.

Clay (g kg^{-1})		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		327.5	217.6	0	986.8	32350	22645	9705	0.88	77	4.35	0.88	75.93	4.36
Regions	South	448.6	197.4	0	880	4059	2841	1218	0.83	82.02	3.91	0.83	81.12	4.2
	Southeast	284.5	201.6	5	960	17448	12214	5234	0.91	59.89	3.76	0.92	58.82	3.71
	Midwest	352.5	242.4	0	910	7656	5359	2297	0.94	59.25	7.75	0.94	58.61	7.68
	Northeast	222.4	136.9	3	629	609	426	183	0.75	70.13	3.06	0.78	59.28	2.99
	North	378.6	190.8	10	986.8	2578	1805	773	0.71	102.67	2.62	0.74	96.85	2.-48
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	130.9	76.7	30	300	32	22	10	0.72	45.38	2.2	0.71	38.71	2.07
	AM	315.8	224	10	986.8	550	385	165	0.85	86.37	2.83	0.8	101.5	2.18
	AP	469.8	172.4	80	920	432	302	130	0.57	112.74	2.31	0.61	109.45	2.54
	BA	230.2	136	11	626	402	281	121	0.83	56.45	3.6	0.84	52.85	4.01
	CE	256.8	144.4	20	534	23	16	7	0.49	105.19	1.96	0.42	107.84	1.77
	ES	209.9	130.7	10	600	100	70	30	0.92	37.6	3.99	0.92	38.53	5.97
	GO	311.6	253	20	890	2148	1504	644	0.97	47.24	9.97	0.96	52.15	8.99
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	550.7	244.1	10	960	1729	1210	519	0.88	83.95	5.1	0.88	87.79	5.01
	MS	372	236.7	0	910	5350	3745	1605	0.94	55.81	7.84	0.93	61.27	7.41
	MT	245.3	182.5	20	840	158	111	47	0.81	88.04	1.14	0.83	53.89	1
	PA	341.4	205	15.8	931.4	296	207	89	0.63	124.26	1.93	0.69	116.23	2.38
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	268.2	102.7	90	490	69	48	21	0.67	58.7	3.34	0.63	63.54	2.44
	PI	147.2	148.8	3	595	66	46	20	0.6	93.72	2.15	0.5	105.55	1.58
	PR	555.7	217.1	0	880	299	209	90	0.86	83.06	4.33	0.81	88.3	3.45
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	268.2	138.2	98	533	17	9	8	0.52	100.94	2.16	0.43	98.51	1.49
	RO	451.3	161.1	80	880	642	449	193	0.89	53.96	4.45	0.89	55.85	4.65
	RR	327.9	129	48	800	626	438	188	0.69	73.49	2.18	0.68	69.38	2.33
	RS	445.6	207.8	0	837	1642	1149	493	0.84	84.21	4.37	0.85	78.8	4.6
	SC	435.7	181.1	0	800	2118	1483	635	0.84	72.58	4	0.85	70.18	4.13
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	255.5	173.2	5	910	15617	10932	4685	0.9	54.24	3.13	0.9	54.33	3.17
	TO	106.3	124.8	10	530	32	19	13	0.82	34.39	2.04	0.8	86.52	0.69

ppclust package (Cebeci et al., 2018).

The SSL was clustered by fuzzy c-means and crisp fuzzy techniques to evaluate the potential of spectral data in the discrimination of soil types. By testing the SSL with two different clustering methods, we intended to evaluate if the soil spectral groups were robust and capable of being replicated by different clustering methods. The inter-comparison between c-means and crisp methods was done using Sankey diagram (Schmidt, 2008), which assesses the relative associations between memberships. The diagram represents the relationship between groups by linking the clusters from the c-means technique with the ones from crisp fuzzy analysis. Higher associations between two groups are represented by proportionally larger links, while poor relation between a pair of clusters is represented by thin links.

2.7. Correspondence analysis

The associations between geological and environmental characteristics with soil spectral classes defined by cluster analysis were identified through correspondence analysis (CA). This technique is designed specifically for the analysis of categorical variables, and its primary goal is to illustrate the most important relationships among the variables' response categories using a graphical representation (Benzécri, 1992). The concept is similar to PCA but applies to categorical rather than continuous data. It summarizes the associations between the spectral soil classes and the variables (regions, states, geology, biomes, soil classes, and layers) in two-dimensional graphical forms. The CA plots are derived from a table where the rows are the characteristics (e.g. states of Brazil, biome, etc.) and the columns are the six spectral classes. The CA was applied using the *FactoMineR* package (Lê et al., 2008).

3. Results and discussion

3.1. Soil reflectance spectra vs physiographic and soil characteristics

The higher sand content in NE region caused an increase in reflectance compared to other regions (Fig. 4a). On the average, the sand content of NE region is 651 g kg^{-1}, while the S region is 264 g kg^{-1}. Soils from the S region of Brazil, where predominantly formed under the influence of basalt or related to igneous rocks (Fig. 2c), which presented low reflectance values due to iron oxides and opaque minerals (Fig. 4a). This region also has specific temperate climate, which favors the preservation of SOC, and this agree with its lower spectra. In general, in relation to the geology, spectral signatures from igneous rocks are usually rich in calcium and iron and had low reflectance (Fig. 4c), while soils formed in metamorphic parent material showed high reflectance values. In these soils, reflectance features were mainly linked to orthoclase, quartz and plagioclase minerals. The low reflectance values found in soils formed in igneous rocks, such as basalt and diabase with high amounts of iron, were correlated with high clay contents and consequently higher influence of scattering. The soils from the Cerrado biome revealed the lowest spectral reflectance and has a spectral feature located at 2265 nm related to gibbsite (Fig. 4d). In fact, most of Cerrado present high weathered soils (Ferralsols) and agrees with the indicated spectral feature (Madeira Netto, 2001) (Fig. 4d). The Atlantic Forest also presented a low reflectance curve. The higher

Table 3
Cubist model parameters, descriptive statistics, and results of prediction models of sand.

Sand (g kg^{-1})		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		529.7	284.1	0.0	990.0	33481	23437	10044	0.87	102.06	5.17	0.87	103.03	5.08
Regions	South	242.9	213.9	10.0	990.0	4059	2841	1218	0.78	101.82	2.65	0.77	101.54	2.65
	Southeast	606.2	252.4	0.0	970.0	17687	12381	5306	0.90	80.67	4.34	0.90	80.02	4.35
	Midwest	567.8	278.9	0.0	966.0	7656	5359	2297	0.94	66.10	8.05	0.94	68.04	7.90
	Northeast	651.9	237.3	26.0	988.0	682	477	205	0.80	106.18	2.74	0.83	102.00	2.98
	North	363.0	244.2	0.0	980.0	3397	2378	1019	0.79	113.73	4.04	0.78	114.95	3.6-8
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	796.9	87.9	610.0	940.0	32	19	13	0.40	64.16	1.44	0.38	78.69	1.21
	AM	373.6	197.6	0.0	910.0	551	386	165	0.71	108.92	2.31	0.71	119.88	2.41
	AP	203.2	223.4	0.0	808.0	1250	875	375	0.84	90.47	3.54	0.83	93.95	3.30
	BA	734.6	150.7	96.0	988.0	402	281	121	0.78	69.31	3.13	0.80	71.66	2.51
	CE	434.1	280.9	26.0	977.0	95	67	29	0.88	107.93	5.22	0.90	87.27	5.43
	ES	747.4	136.1	360.0	950.0	100	70	30	0.86	46.28	3.35	0.87	57.95	4.23
	GO	626.9	294.5	36.9	951.0	2148	1504	644	0.96	55.65	10.26	0.97	53.94	10.31
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	271.2	215.7	0.0	970.0	1729	1210	519	0.80	98.88	3.20	0.78	102.97	3.25
	MS	542.0	269.5	0.0	966.0	5350	3745	1605	0.94	64.73	8.03	0.95	61.89	8.11
	MT	638.6	237.0	20.0	960.0	158	111	47	0.87	81.81	1.92	0.83	110.60	1.81
	PA	474.8	260.8	13.0	945.0	296	207	89	0.67	152.30	2.83	0.73	134.56	3.09
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	467.07	166.3	93.0	897.0	69	48	21	0.21	147.06	1.40	0.30	141.66	1.38
	PI	610.2	358.2	27.0	985.0	67	47	20	0.83	153.65	4.56	0.80	169.41	3.83
	PR	298.6	247.4	10.0	960.0	299	209	90	0.91	74.91	5.07	0.90	72.17	3.50
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-	-
	RO	508.2	162.9	100.0	900.0	642	449	193	0.88	54.50	4.40	0.88	59.80	4.52
	RR	446.1	177.7	6.0	910.0	626	438	188	0.55	121.23	1.90	0.58	117.05	1.95
	RS	241.7	229.5	10.0	928.0	1642	1149	493	0.78	111.46	2.39	0.81	99.93	2.43
	SC	235.9	194.2	10.0	990.0	2118	1483	635	0.82	82.92	3.27	0.82	83.12	3.01
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	641.9	228.6	0.0	969.0	15856	11099	4757	0.89	76.13	2.89	0.89	75.58	3.02
	TO	846.6	164.9	290.0	980.0	32	22	10	0.91	48.14	1.77	0.88	73.29	1.50

reflectance was found in Caatinga biome (Fig. 4d), which can be explained by the predominance of sandy particles in soils, besides the high temperatures that have accelerate the decomposition of soil organic matter resulting on relatively low SOC. The representation of the spectral curves of each biome is a generalization, because within each of them there is a great complexity of soil types. For instance, the Atlantic Forest biome extends from southern to northern Brazil with different soil types. Among the soil classes, the spectral curve of the Nitisols showed the lowest reflectance (Fig. 4e). This soil originated mainly from mafic rocks such as basalt and diabase, present high amounts of clay, iron oxide and opaque minerals (IUSS Working Group WRB, 2015). The spectral curve of the Histosols presented low reflectance in the visible spectral region due to the high content of SOC. Contrarily, the Podzols had the highest reflectance, followed by the Cambisols rich in quartz (Fig. 4e). The Arenosols, for example, showed greater reflectance in the A layer due to higher sand content in relation to subsurface (Fig. 4f). The average reflectance curves of the four layers (soil depths) (Fig. 4f) indicate the differences in SOC content: from 550 to 900 nm the reflectance increases while SOC decreases. The spectral range from 1500 to 2500 nm was influenced by quartz, due to the high reflectance values in all four soil layers. The B and C layers were identical, which is attributed to low variation in mineralogy and texture. In general, features (shapes and intensities) are related with several soil properties and assist on to reach a comprehensive information of the soil sample (Fig. 4d), despite statistical models. An interesting observation is that spectra from GO and MT (MW Brazil) presents gibbsite oxide (Fig. 4b), seen at 2265 nm, as the ones from RN is completely absent of this feature, which agrees with the more and less weathered soils of these two regions, respectively.

The PCA revealed that the first principal component (PC1) accounts explained 85% of the variance in the data (Fig. 5). The PCA of spectra averaged by region (Fig. 5a) was able to detect that the N, NE and S regions presented different soil spectral patterns. However, soil spectra from the MW and SE were grouped together, showing nearly the same spectral pattern in these regions. Among soils from 26 Brazilian states some showed similar factor loadings (Fig. 5b). For example, the average soil spectra from MT and GO states as well as for AP and AC; MS and RS; PR, and SC; BA, RN, and ES; MA, and SP; AL, and PA; RJ, and PB grouped together. Soil spectra from AM, PE, PI, TO, SE, and RO states showed different patterns and were not grouped. For the three geological classes, the PCA discriminated them by showing separated data distributions (Fig. 5c). The PCA result for the biomes indicated that Caatinga and Amazon present distinct spectral curves (Fig. 5d). Indeed, these are two important and very distinct environments (Amazon: tropical humid soils; Caatinga: semiarid soils). This finding is corroborated by the result in Fig. 4d, where the Caatinga showed a spectral curve with high reflectance and the Amazon presented higher intensity reflectance in the visible region. This interpretation agrees with spectra by regions (Fig. 5a). However, soils from Pantanal and Pampa presented small differences in spectral pattern similar to the Atlantic Forest and Cerrado (Fig. 5d). The principal component scores discriminated well among Podzols, Plinthosols, Histosols, and Lixisols, using only B layer (40–60 cm depth) (Fig. 5e). Conversely, Nitisols, Cambisols, and Ferralsols were grouped relatively close together suggesting similarities spectral pattern. The same arrangement was found among Planosols, Gleysols, and Vertisols, which are soils formed under the influence of hydromorphic conditions with more prolonged water saturation typically exhibiting the Fe^{3+} reduction and high SOC. In contrast Podzols,

Table 4
Cubist model parameters, descriptive statistics, and results of prediction models of pH.

pH (H_2O)		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		5.4	0.7	0.0	8.9	26163	18314	7849	0.53	0.40	1.70	0.54	0.39	1.66
Regions	South	4.7	0.5	3.8	6.5	328	229	99	0.35	0.44	1.42	0.34	0.47	1.19
	Southeast	5.5	0.6	0.6	8.7	17001	11900	5101	0.49	0.42	1.83	0.49	0.42	1.83
	Midwest	5.4	0.7	0.0	8.5	5947	4162	1785	0.51	0.45	1.82	0.52	0.43	1.84
	Northeast	5.5	1.0	2.8	8.9	732	512	220	0.60	0.69	2.26	0.65	0.63	2.01
	North	4.9	0.6	2.5	7.7	2155	1508	647	0.41	0.47	1.69	0.46	0.45	1.76
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	6.0	0.4	5.4	7.0	32	19	13	0.41	0.26	1.48	0.46	0.06	1.54
	AM	4.6	0.5	3.2	7.1	501	350	151	0.35	0.43	1.73	0.43	0.41	1.41
	AP	-	-	-	-	-	-	-	-	-	-	-	-	-
	BA	5.2	0.8	3.6	8.4	403	282	121	0.57	0.53	1.58	0.58	0.50	1.61
	CE	5.7	1.2	3.0	8.3	33	19	14	0.73	0.62	1.23	0.97	0.45	1.31
	ES	-	-	-	-	-	-	-	-	-	-	-	-	-
	GO	5.5	0.6	4.0	8.2	2050	1435	615	0.41	0.57	2.20	0.41	0.54	2.19
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	5.1	0.7	3.0	8.3	1352	946	406	0.50	0.46	1.61	0.55	0.45	1.45
	MS	5.3	0.7	0.0	8.5	3730	2611	1119	0.53	0.42	1.80	0.58	0.42	1.74
	MT	5.6	0.6	3.8	7.4	167	116	51	0.31	0.72	1.69	0.45	0.66	1.79
	PA	4.7	0.7	2.5	7.7	308		93	0.59	0.45	1.42	0.50	0.40	1.88
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	5.1	0.9	3.9	8.2	69	41	28	0.62	0.32	1.49	0.97	0.20	0.72
	PI	6.0	1.7	2.8	8.9	67	40	27	0.54	1.25	2.96	0.29	1.38	1.68
	PR	-	-	-	-	-	-	-	-	-	-	-	-	-
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-	-
	RO	5.1	0.4	3.9	6.4	672	470	202	0.26	0.59	1.94	0.30	0.54	2.00
	RR	4.9	0.6	3.5	7.6	627	438	189	0.53	0.22	1.38	0.55	0.15	1.32
	RS	5.0	0.5	4.4	6.0	23	16	7	0.43	0.26	1.39	0.35	0.07	0.14
	SC	4.7	0.4	3.8	6.5	305	213	92	0.29	0.62	1.40	0.30	0.57	1.33
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	5.5	0.6	0.6	8.1	15547	10882	4665	0.47	0.42	1.90	0.48	0.44	1.89
	TO	-	-	-	-	-	-	-	-	-	-	-	-	-

Plinthosols, Histosols and Lixisols show large differences in the content of SOC, iron oxides and texture. For instance, Podzols have SOC mainly associated with a sandy texture, dominantly quartz, and associated with complexes of Al and Fe. Plinthosols are characterized by high iron oxides and low crystallinity degree in the form of nodules and low content of SOC. Histosols have high content of SOC, and low Fe content. Lixisols have kaolinite dominance, variable texture and low SOC content. Another group is represented by soils without texture gradient such as Nitisols, Cambisols, and Ferralsols. On the other hand, Lixisols with textural gradient or Histosols, with very high SOC were discriminated by principal component scores. The PCA for layers detected that A (surface) and D (subsurface) layers were separated suggesting distinct spectral pattern in these two layers (Fig. 5f). The B and C layer were placed in proximity in the PCA graphs indicating that they were similar in relation to the soil spectra (Fig. 5f).

3.2. National, regional, and state prediction of soil attributes

The national model produced R^2 of 0.78 and RMSE of 6.89 g kg^{-1} for SOC prediction in validation model (Table 1). For SOC prediction in validation model, the regional models showed R^2 ranging from 0.58 to 0.84 and RMSE from 2.28 to 9.97 g kg^{-1}. The MW region presented the best results, while the N region the worse. The state that presented the best results was MT with R^2 of 0.89 and RMSE of 1.65 g kg^{-1}. From the 18 states with SOC prediction models, 8 showed a R^2 above 0.80 and only two showed R^2 below 0.32 (AP and RR). The worst R^2 were associated either with a small number of samples or a high variability in SOC.

Among all the soil variables predicted, clay showed the highest coefficient of determination in validation mode at national level with 0.88 and RMSE of 75.93 g kg^{-1} (Table 2). At the regional level, all clay validation models had R^2 higher than 0.71 and the SE and MW regions had R^2 higher than 0.91 and RMSE lesser than 60 g kg^{-1}. At the state level, the best result was found for GO with R^2 of 0.96, RMSE of 52.15 g kg^{-1} and RPIQ of 8.99. From the 21 states with clay prediction, 13 showed R^2 higher than 0.80 and only 3 had R^2 below 0.50.

For sand predictions in validation mode at the national level showed R^2 of 0.87 and RMSE of 103.03 g kg^{-1} (Table 3). At regional scale, the MW region showed the best performing validation sand model with R^2 of 0.94 and the S region the worse, but still with moderate well predictions (R^2 = 0.77 and RMSE = 114.95 g kg^{-1}). At the state level, 20 sand models were generated with GO showing the best results (R^2 = 0.97 and RMSE = 53.94 g kg^{-1}). In 14 states the R^2 were higher than 0.80 and only PE had R^2 below 0.50 for predictive modeling of sand.

At the national level, the validation of the model generated for pH prediction showed a R^2 of 0.54 and RMSE of 0.39 (Table 4). In general, the national pH model was better than the regional ones. Only the NE region presented a higher R^2 (0.65) in validation mode, while the S region showed the smallest R^2 (0.34). The good result for the NE region is related to the well-performing pH models generated for the CE and PE states that belong to this region and had the highest R^2, both with 0.97. From the 16 analyzed states, 9 pH prediction models had poor results with R^2 below 0.50.

The prediction validation of CEC at the national level reached a R^2 of 0.68 and RMSE of 24.02 $cmol_c$ kg^{-1} (Table 5). At the regional level, NE showed the best validation results (R^2 = 0.89 and RMSE = 27.68 $cmol_c$ kg^{-1}) and the S region the worst (R^2 = 0.64 and RMSE = 3.81 $cmol_c$ kg^{-1}). At the state level, MT, RN and SE showed a R^2 above 0.93, and three states showed a R^2 below 0.30 (AL, PA and PE).

Table 5
Cubist model parameters, descriptive statistics, and results of prediction models of Cation Exchange Capacity (CEC).

CEC ($cmol_c\ kg^{-1}$)		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		47.7	41.9	0.0	958.0	17433	12203	5230	0.66	25.78	1.46	0.68	24.02	1.52
Regions	South	12.6	5.9	1.3	35.7	631	442	189	0.60	3.72	1.77	0.64	3.81	1.83
	Southeast	54.2	41.5	0.0	778.7	9896	6927	2969	0.75	21.64	1.63	0.79	20.02	1.73
	Midwest	49.9	32.9	0.0	528.4	3974	2782	1192	0.77	16.06	2.35	0.75	17.29	2.11
	Northeast	53.6	88.6	0.0	958.0	682	477	205	0.82	39.44	1.72	0.89	27.68	2.38
	North	23.3	23.9	0.1	248.0	2250	1575	675	0.74	13.16	1.81	0.72	12.40	1.-95
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	49.7	15.9	27.0	97.0	31	18	13	0.11	20.41	1.10	0.02	10.26	0.78
	AM	22.0	35.8	1.0	248.0	501	351	150	0.91	11.75	0.56	0.84	14.92	0.45
	AP	37.3	16.3	15.1	138.8	432	302	130	0.67	9.70	1.99	0.59	10.17	1.73
	BA	33.5	39.6	0.0	224.0	403	282	121	0.79	19.42	3.08	0.77	19.30	3.01
	CE	25.4	24.5	3.3	117.7	33	23	10	0.70	15.84	0.86	0.74	10.44	1.31
	ES	85.5	28.0	44.7	183.5	100	70	30	0.35	22.94	1.58	0.29	25.23	1.10
	GO	68.3	38.9	3.2	454.8	606	424	182	0.72	23.32	1.71	0.72	17.14	2.34
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	66.9	45.1	0.5	778.7	1745	1222	524	0.68	26.71	1.92	0.67	24.77	1.81
	MS	49.0	30.2	10.1	528.4	3101	2171	930	0.79	13.99	2.50	0.76	14.85	2.32
	MT	18.5	19.3	0.0	75.0	267	187	80	0.97	3.16	12.04	0.93	5.39	6.82
	PA	15.6	22.7	1.4	183.2	302	211	91	0.45	20.01	0.50	0.26	17.87	0.50
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	6.8	2.6	2.7	17.3	69	48	21	0.25	1.95	1.48	0.16	2.71	1.36
	PI	199.8	177.4	17.1	958.0	48	24	24	0.53	88.89	2.37	0.53	211.64	0.95
	PR	-	-	-	-	-	-	-	-	-	-	-	-	-
	RJ	16.6	8.6	7.4	32.9	12	6	6	0.27	10.41	0.50	0.56	6.94	2.07
	RN	36.6	35.0	1.9	102.2	25	13	12	0.95	7.35	5.40	0.99	5.11	11.48
	RO	29.8	13.1	8.1	102.1	638	447	191	0.69	7.30	1.96	0.65	8.06	1.90
	RR	4.5	2.6	0.1	24.0	377	264	113	0.40	1.91	1.41	0.36	2.64	1.11
	RS	13.5	7.2	1.3	35.7	326	228	98	0.55	5.14	1.66	0.50	4.63	1.72
	SC	11.6	4.0	3.8	24.5	305	214	92	0.63	2.43	2.53	0.65	2.32	2.34
	SE	142.4	142.1	23.0	627.4	65	45	20	0.93	41.99	1.23	0.94	37.78	5.30
	SP	51.0	40.1	0.0	564.0	8039	5627	2412	0.82	18.67	1.50	0.79	17.65	1.50
	TO	-	-	-	-	-	-	-		-	-	-	-	-

Of all soil attributes analyzed in this article, BS presented the poorest result at the national level with $R^2 = 0.49$ and RMSE = 17.01% in validation mode (Table 6). However, at the regional level only the SE region showed a R^2 below 0.50 ($R^2 = 0.49$ and RMSE = 16.28%) and the NE showed the best performing BS model (R^2 of 0.79 and RMSE = 13.42%). At the state level, 16 models were generated, of which 14 showed R^2 above 0.65 and AM and GO states showed higher R^2 (0.70 and 0.69, respectively).

Higher errors in the calibration models than in validation are definitely odd, especially when the modeling process uses a machine learning algorithm. In our case, this is related to a limited number of samples for modeling at the state level. The SOC model predictions at the state of AL (Table 1), for example, there were only 32 samples available for calibration and 19 for validation, which resulted in a R^2 of 0.29 (calibration) and 0.43 (validation). Conversely, prediction models for SP state had a total of 8185 samples, 5729 for calibration and 2456 for validation. In this case, the R^2 for calibration and validation were practically the same (0.66). Representative and comprehensive datasets are essential for a robust calibration, because otherwise the errors may be high and results incoherent. The BSSL has been constantly populated with new samples, therefore we believe that soon it will be possible to calibrate good prediction models for all states and regions.

3.3. Synthesis of soil attributes prediction

In summary, we computed the best model performances (i.e., highest model fits and lowest errors) by down scaling results (i.e., from national to state levels). For example, for SOC validation models the R^2 were 0.78–0.84 - 0.93 (Table 1), for clay 0.88–0.94 - 0.96 (Table 2), and for sand 0.87–0.94 - 0.95 for national and best performing regional and state models, respectively (Table 3). However, the performance of state-specific soil models differed widely due to multiple factors including sample size, soil variance, and soil-forming factor differences. The model performance for the same soil attribute in different states differed widely (e.g., R^2 of clay varied from 0.42 to 0.96), hindered detailed discussion on several factors which still need further studies. The variability in statistical metrics of soil attributes assessments are not new. Nocita et al. (2014) found several discrepancies in regard to R^2 for the same soil attribute, i.e., SOC, pH, and others. Zeng et al. (2016) indicated that local predictions can be better modeled by understanding the soil development, i.e., parent material, biome and land use. Shepherd and Walsh (2002), obtained R^2 for CEC from 0.6 (national model) to 0.8 (local models). Grunwald et al. (2018) found that upscaled SOC spectral models performed better in terms of R^2 and RPIQ, whereas the downscaled models showed less bias and smaller RMSE in Florida, USA. This study found no universal trend that could explain the scalability of the models, such as spectral variance, soil attribute variance, methods, and environmental characteristics or diversity.

Overall, SOC models with high model fit ($R^2 > 0.85$ in regions CE, MG, MT and SC) coincided with relatively high mean SOC of > 12 $g\,kg^{-1}$irrespective of SOC variabilities that were very large (e.g., 55.7 $g\,kg^{-1}$ standard deviation, SD, in CE) or low (e.g., only 5.0 $g\,kg^{-1}$SC in MT). Similar trends were discovered for soil texture models. For example, clay models with high model fit ($R^2 > 0.85$ in regions ES, GO, MG, MS, RO, RS, SC and SP) corresponded with high mean clay content of > 210 $g\,kg^{-1}$, though almost all of these models

Table 6
Cubist model parameters, descriptive statistics, and results of prediction models of Base Saturation (BS).

BS (%)		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
Nacional		39.6	23.6	0.0	100.0	28450	19915	8535	0.50	16.93	2.39	0.49	17.01	2.40
Regions	South	44.4	22.9	1.7	89.7	326	228	98	0.56	15.26	2.55	0.54	15.51	2.60
	Southeast	44.9	22.4	0.0	100.0	16981	11887	5094	0.48	16.33	2.20	0.49	16.28	2.24
	Midwest	31.4	21.9	1.0	99.0	7404	5183	2221	0.58	14.62	2.28	0.57	14.72	2.32
	Northeast	37.9	29.3	1.4	100.0	636	445	191	0.81	12.93	3.47	0.79	13.42	3.20
	North	29.8	24.1	0.0	100.0	3103	2172	931	0.70	13.42	2.72	0.69	13.93	2.70
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	41.4	18.1	16.0	74.0	32	19	13	0.14	17.16	1.51	0.33	15.40	2.11
	AM	19.0	14.2	1.0	100.0	501	351	150	0.47	9.29	1.29	0.70	9.91	1.41
	AP	40.7	27.2	0.0	88.0	1249	874	375	0.71	14.75	3.53	0.63	16.63	3.10
	BA	23.1	17.8	1.4	94.4	402	281	121	0.53	12.44	1.72	0.57	12.33	1.70
	CE	86.0	11.1	52.0	100.0	33	23	10	0.69	6.76	1.26	0.60	8.07	0.93
	ES	-	-	-	-	-	-	-	-	-	-	-	-	-
	GO	29.8	24.6	1.0	99.2	2132	1492	640	0.70	13.65	3.01	0.69	13.40	2.87
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	30.2	23.3	0.0	100.0	1745	1222	524	0.57	16.75	2.12	0.55	16.96	2.02
	MS	31.5	20.5	2.0	97.0	5122	3585	1537	0.58	13.73	2.24	0.43	15.98	1.92
	MT	50.3	18.8	9.0	96.0	150	105	45	0.58	11.89	1.77	0.38	16.01	1.12
	PA	26.8	21.1	0.0	95.5	302	211	91	0.54	14.81	1.73	0.52	15.00	1.75
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	41.4	22.0	9.0	100.0	69	48	21	0.59	14.22	1.93	0.28	19.56	1.28
	PI	76.9	20.4	25.0	100.0	67	47	20	0.57	12.37	1.88	0.37	18.68	1.49
	PR	-	-	-	-	-	-	-	-	-	-	-	-	-
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-	-
	RO	15.0	8.6	2.0	63.0	642	449	193	0.46	6.76	1.63	0.45	6.25	1.28
	RR	38.1	22.8	1.1	100.0	377	264	113	0.55	16.13	2.10	0.39	17.88	2.07
	RS	44.4	22.9	1.7	89.7	326	228	98	0.69	13.13	3.02	0.35	18.61	2.11
	SC	-	-	-	-	-	-	-	-	-	-	-	-	-
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	46.8	21.6	0.0	99.8	15124	10587	4537	0.52	15.36	2.22	0.33	17.69	1.97
	TO	-	-	-	-	-	-	-	-	-	-	-	-	-

covered a wide range in clay contents with SD values ranging from 130 to 253 g kg^{-1}. Trends among pH, CEC, and BS models were less clear among regions in terms of underlying factors to explain model performance. The model performance in this study showed comparable results with other spectral library studies (Viscarra Rossel et al., 2016) This suggest that soil spectral models for key indicators on Brazilian soils could be developed with similar quality as documented in other soil spectral studies.

Samples from certain states showed high pedological variation (factors and processes of soil formation). States with pedological complexity and/or less sample size may have contributed to lower model fits and higher errors than more homogeneous and/or more densely sampled states. Though high sample size may not necessarily mean better model performance as indicated by models in SP, which performed excellent for clay, sand, and CEC, moderately for SOC, and less well for pH and BS.

In order to confirm the effectiveness of the spectroscopic method to determine clay and sand contents, the textural triangle was produced showing simultaneously the observed values derived from traditional laboratory analysis and the predicted ones by the spectroscopic method (Fig. 6). The textural triangle of the entire BSSL (Fig. 6a) showed that most soil samples were placed in the soil texture classes of sand, loamy sand, sandy loam, sandy clay loam, sandy clay, and clay. The SE region showed similar trends in soil texture when compared to the whole BSSL database because the SE region contributed 52% of the total samples with textural data (Fig. 6b). However, the samples of SE showed lower silt content than the entire BSSL collection. The predicted soil textures for the S region (Fig. 6c) N region (Fig. 6e), and NE region (Fig. 6f) showed more scatter than the MW region (Fig. 6d). In the S region, most soils belong to clay textural class followed by silt clay and clay loam mainly due to the parent material predominantly formed by igneous rocks (Fig. 2c). The samples with the highest silt contents belong to the S region. This is related to the low temperature in this region on soil formation. In the MW region, the prediction model for clay and sand was the most accurate and this is reflected in the predicted samples, which presented the same trend as the observed ones (Fig. 6d). The vast majority of samples present a soil textural class varying from sand to clay. The percentage of silt in these soils is low. In general, the MW region presents more weathered-leached soils such as Ferralsols, derived from sedimentary and igneous rocks forming loamy sand and clayey soils. Soils from the N region showed a large variation in texture (Fig. 6e) corroborating the lower prediction performance of clay (Table 2) and sand content (Table 3) compared to other models. The NE region present a large amount of sandy soils and for this reason the predicted model had good performance (Fig. 6f). In the NE region, the majority of the soils in this study is formed from sedimentary materials, which is showed by the textural classes with high sand content (Fig. 6f).

The textural variations found in the triangles of each region are due to differences in the geology, climate, and relief. Each region presents textural diversity as there is also a great variation of types of soils (Fig. 2e). It is important to emphasize that the predicted samples had the same tendency of the observed ones. This is a great finding considering the world demand for soil analyses with > 600 million soil samples processed every year which represents a consumption of about 840 thousand kg of dichromate and ammonium ferrous sulfate and 3 million L of sulfuric acid, just for SOC analysis (Demattê et al., 2019). The effectiveness of soil spectroscopic analysis is justified by the fact that it is fast, simple, accurate, cheap, and most importantly non-polluting method. The possibility of predicting several attributes with just one spectral reading, the easy and rapid data acquisition of large

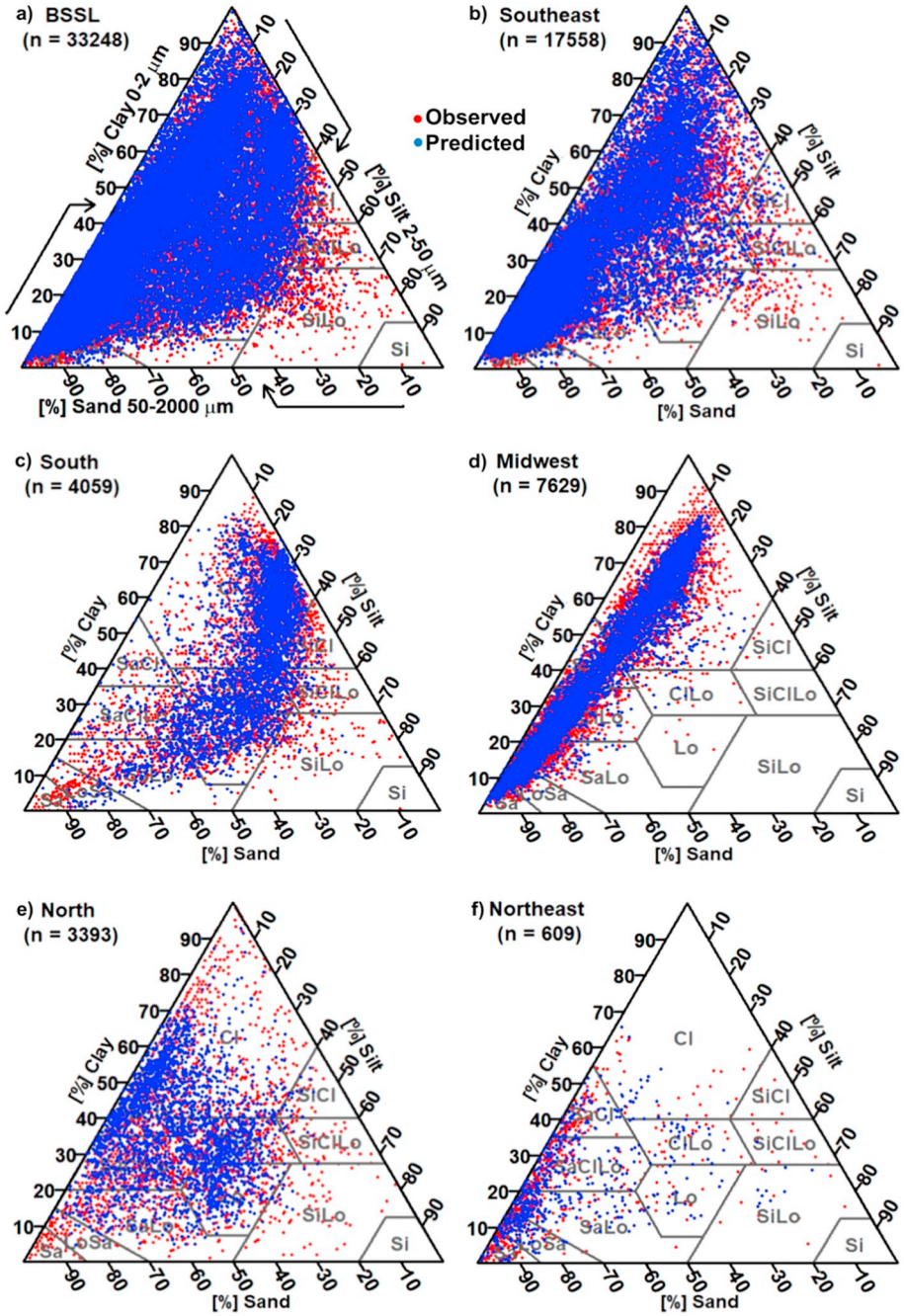

Fig. 6. Soil texture triangle calculated from the entire database (a) and for Brazilian regions, (b) Southeast (SE), (c) South (S), (d) Midwest (MW), (e) North (N), and (f) Northeast (NE) regions. Cl: clay; SiCl: silty clay; SaCl: sandy clay; ClLo: clay loam; SiClLo: silty clay loam; SaClLo: sandy clay loam; Lo: loam; SiLo: silty loam; SaLo: sandy loam; Si: silt; LoSa: loamy sand; and Sa: sand.

amounts of samples without using environmentally hazardous chemicals are the major advantages of the Vis-NIR-SWIR spectroscopy technique for soil analysis (Minasny and McBratney, 2008; Viscarra Rossel and Behrens, 2010).

3.4. Spectral classification

In order to categorize how many spectral patterns are required to represent Brazilian soils according to the shape of the 39,284 spectral signatures, the first three principal component scores (Fig. 7a) were used as variables in the cluster analysis (Terra et al., 2015). The eigenvectors of PC1 are dominated by positive loadings along the wavelengths, which captured 64% of the total variance. The high positive loadings were found in the visible region that showed the characteristic absorptions for iron oxides (Fig. 7a). The eigenvector of the PC2 (10%) showed high negative loadings near at wavelengths for the characteristic absorptions of 2:1 clay mineral (illite and smectite) and possibly organic matter. The PC3 (7.6%) fluctuated between positive and negative loadings.

In order to reduce the dimensionality of the data, the first three principal component scores (Fig. 7a), were applied to determine the optimal number of clusters. We selected six clusters (classes) because, since the pE was maximized and the pC was minimized when six clusters were obtained (Table 7), which was then selected to represent the most satisfactory cluster for the data. In the crisp clustering, each observation receives membership values of 0 or 1 for each cluster. In

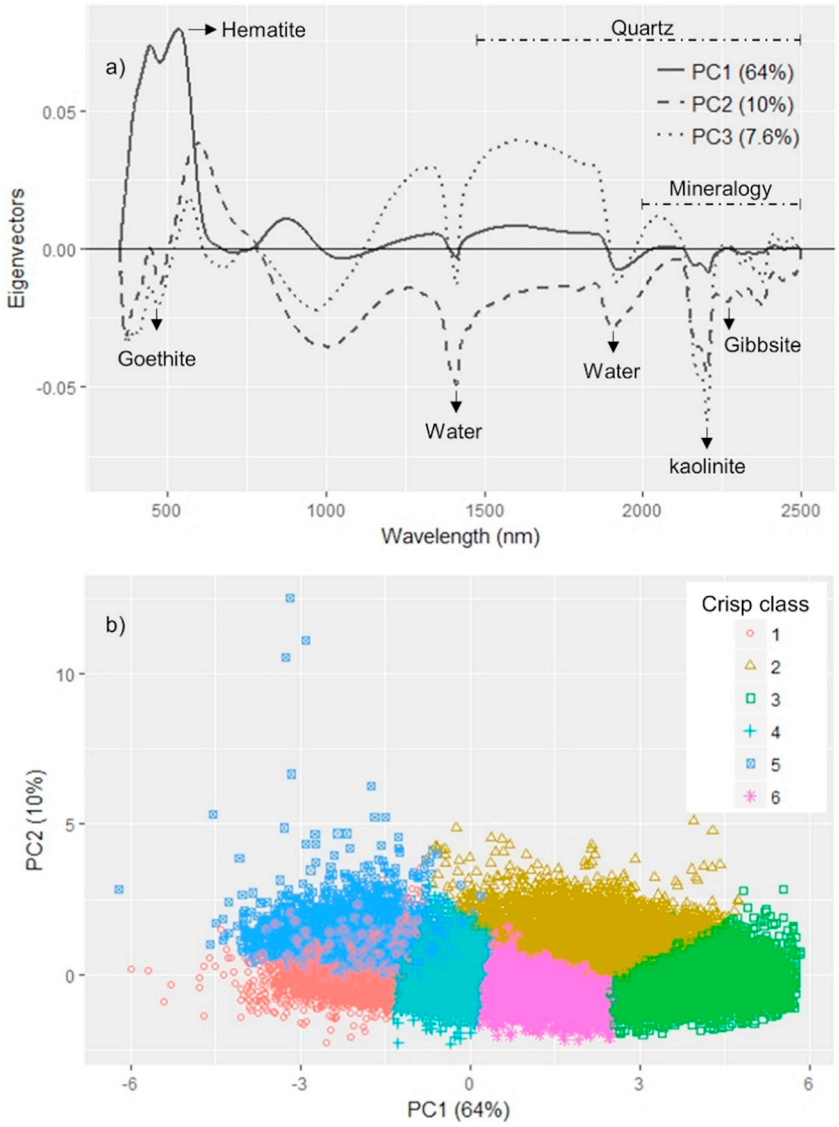

Fig. 7. Principal components eigenvectors of PC 1, 2, and 3, (a) and crisp fuzzy-*c*-means classification, considering six groups (b). Principal components analysis was performed with the continuum removed spectra. Sampling points clustering was based on PC scores.

Table 7
Fuzzy validation indices for the optimum number of clusters, the partition entropy (pE), and the partition coefficient (pC).

Number of clusters[a]	pE	pC
3	0.57	0.68
4	0.63	0.68
5	0.71	0.65
6	**1.06**	**0.48**
7	0.86	0.61
8	0.91	0.60
9	0.97	0.58
10	1.04	0.55
11	1.11	0.53
12	1.05	0.56

[a] In bold is the optimal number of clusters.

the scatter diagram it shows the distribution of all observations colored in the 6 crisp classes (Fig. 7b).

The average CR spectra of 6 classes is presented in Fig. 8. The average spectrum of classes 1, 4, and 5 (Fig. 8a, g, i) were characterized by absorptions representative of soils with abundant iron oxides (400 to 600 nm), while the classes 2, 3, and 6 (Fig. 8c, e, l). Although fairly similar, class 4 may be differentiated from 1 and 5 classes by the absorption features at 500 and 900 nm. The spectrum from class 4 (Fig. 8g) presented features less pronounced than the other two (Fig. 8a, i). These features were related to the crystal field electronic effect of hematite mineral and consequently to the ferric ion (Fe^{+3}) observed in such iron oxides. The interaction between electromagnetic energy and hematite results in electronic transitions, creating the absorption features centered at 530 and 885 nm. Spectra from classes 1 and 5 can be distinguished from each other by the CR reflectance factor at the SWIR-1 range (1000–1800 nm), whereas class 5 presented a lower CR factor. Fuzzy class 2 can be distinguished from the others by the lower CR factor of features centered at 1200, 1900 and 2200 nm. Finally, class 3 spectrum has high CR factor between 350 and 750 nm, which is related to low content of iron oxides in soils.

In the fuzzy-*c*-means clustering, each data point can belong to more than one cluster. The probability of each soil sample being classified in the fuzzy membership class 1 is shown in Fig. 8b. In the center of the fuzzy membership class 1 are the samples with high probability of pertaining to this class (red color). The same analogy applies to the other five classes. These findings suggest that six types of spectra

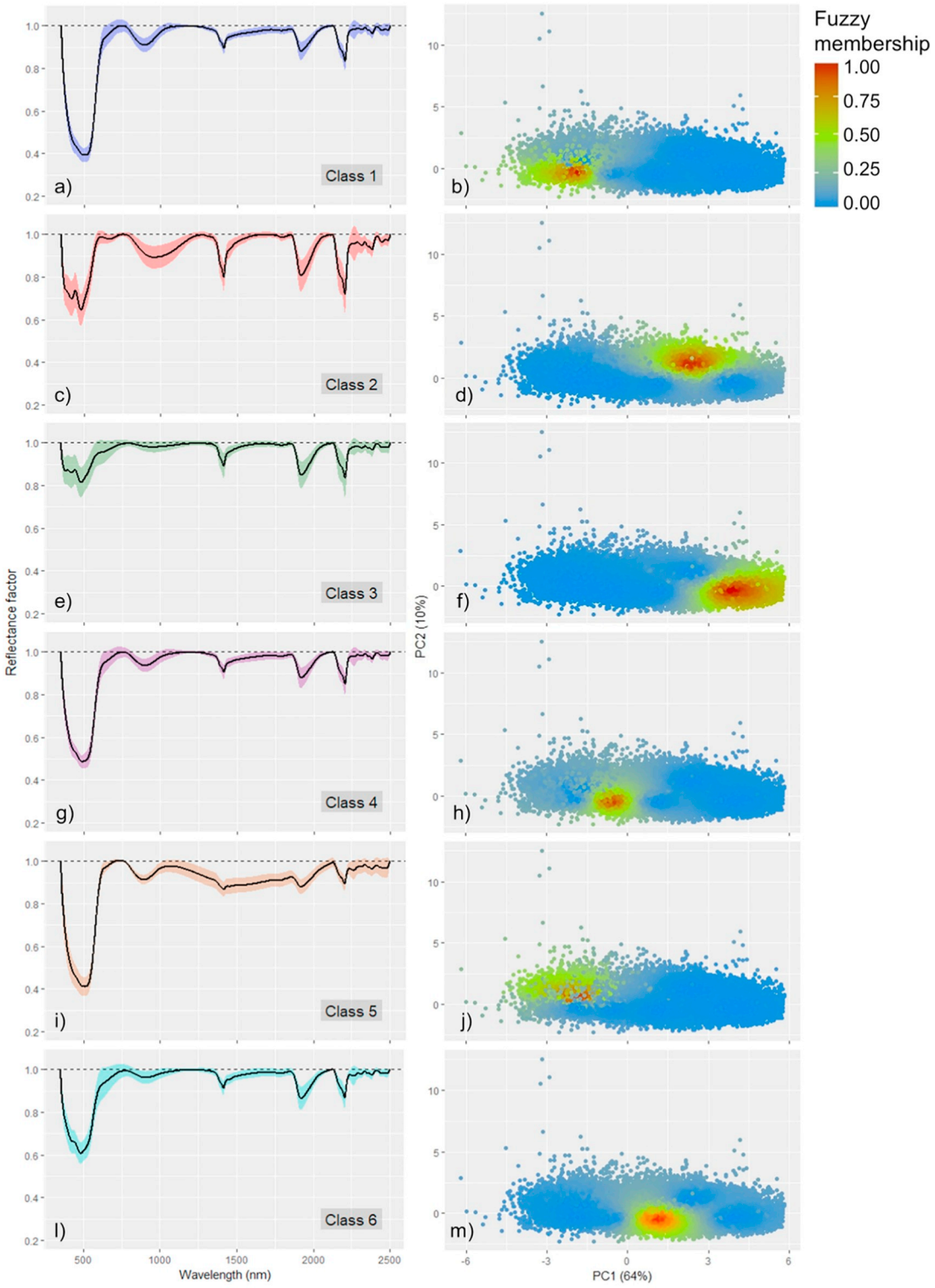

Fig. 8. The average continuum removed spectrum of each fuzzy cluster (a, c, e, g, i, and l) and fuzzy membership values for the 6 clusters (b, d, f, h, j, and m).

represent the whole population of Brazilian soils. The six classes of spectra were discriminated according to the spectral pattern of soils, which is directly linked to intrinsic heterogeneous characteristics, where by contents of SOC, iron oxides, mineralogy of the clay fraction, particle size distribution, and moisture, are the ones that most influence the spectral responses.

Stoner and Baumgardner (1981) found five spectral classes in a large database of the U.S. and Brazil. The authors suggested that five soil spectral reflectance curves could be distinguished as sharing in common certain differentiating characteristics concerning mainly the organic matter and iron oxide contents. One of the classes was detected because it had its origin in Brazil (Paraná state). Formaggio et al. (1996) identified four patterns of spectral curves according to the shape and intensity of the parameters in one state (São Paulo) in Brazil. Viscarra Rossel et al. (2016), used a global spectral library to characterize the world's soil and found six classes of spectra. The authors

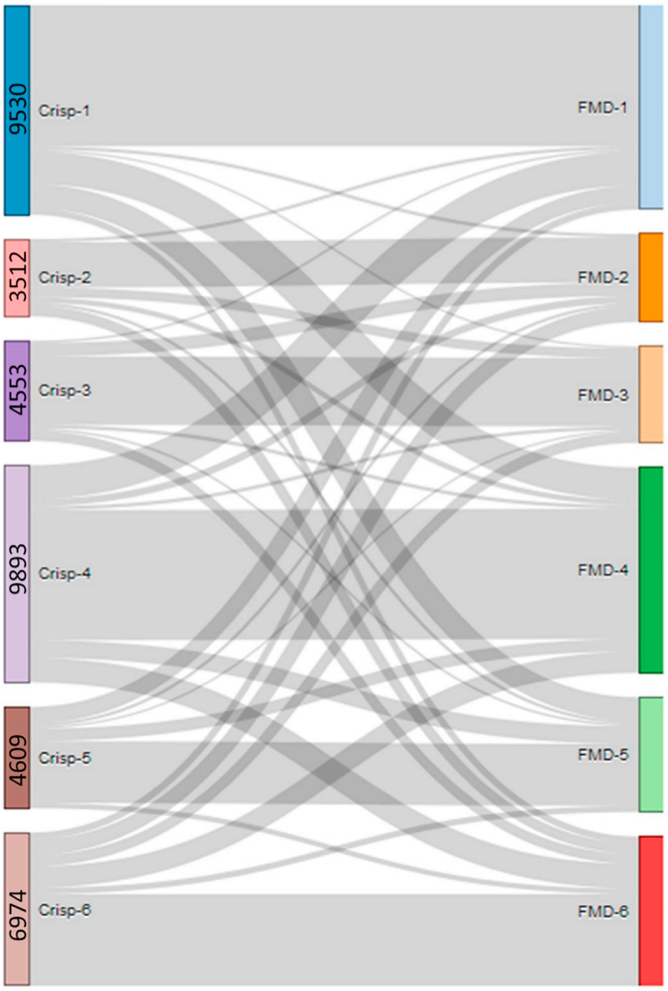

Fig. 9. Sankey diagram showing the relative associations between crisp and Fuzzy Membership Degree (FMD) for each class.

also stated that grouping the spectra into more homogeneous spectral classes can improved the modeling by removing bias in the predictions. Terra et al. (2018) found 6 different patterns of soil spectra based on differences in reflectance intensity and absorption features caused by weathering intensification, which enabled to distinguish soil samples regarding similarity of particle size distribution, mineralogy, and some chemical properties.

From the six classes defined by crisp fuzzy (Crisp-1 to Crisp-6), Crisp-4 presents the largest number of samples (9893 samples), closely followed by Crisp-1 (9530 samples) (Fig. 9). Most of the samples in fuzzy c-means classes (FMD-1 to FMD-6) were correctly assigned to the correspondent crisp classes. More than half of FMD-1 samples were classified as Crisp-1 (Fig. 9), another part was misclassified as Crisp-4 and Crisp-5, while few ones were defined as Crisp-2, Crisp-3, and Crisp-6. This suggests that Crisp-1, 4, and 5 are related to each other, which is corroborated by the similarity in CR spectra of these classes (Fig. 8a, g, i). In FMD-2, the dominant misclassified classes were Crisp-3 and Crisp-6 (Fig. 9). Most of FMD-3 individuals were correctly assigned as Crisp-3, with few samples misclassified as Crisp-2 and Crisp-6. FMD-4 showed higher misclassification with Crisp-1, followed by Crisp-6. As expected, FMD-5 was mostly misclassified as Crisp-1 and Crisp-4, confirming the correlation between their spectral pattern. Finally, FMD-6 was mainly misclassified as Crisp-4, demonstrating the similarity between the spectra of these classes (Fig. 8g, l).

3.5. Correspondence analysis

The CA analysis showed that spectral classes 1 and 5 were correlated with the MW region, while class 4 was similar with region SE, class 6 with region S, class 2 with region N, and class 3 with region NE (Fig. 10a). The spectral classes 5, 1, 4, and 6 resemble MS, SP, PR, GO, MA, RS, RJ, MT, and PB states, that is the points are very close in the simultaneous plot of row and column coordinates (Fig. 10b). The spectral class 3 showed proximity with AL, RN, AM, ES, BA, and PE with most of them from regions N and NE. The spectral class 2 showed some similarity with AP, RO, RR, PA, and AC states. The sedimentary rocks were highly associated with spectral class 1, metamorphic rocks were correlated with classes 2 and 3, and igneous rocks with classes 5 and 6 (Fig. 10c). For the CA between spectral classes and biomes (Fig. 10d), the classes 1, 4, and 6 were related with Atlantic Forest biome, class 5 with Cerrado, class 3 with Pampa, Pantanal and Caatinga, and class 2 with Amazon. For the CA of spectral and soil classes only profile samples that have layer B collected were used. The Gleysols and Plinthosols classes were associated with spectral class 2, Ferralsol was highly associated to classes 1, 4 and 5 (Fig. 10e). Nitosols and Lixisols were associated with class 6, and Cambisols with class 3. Histosols, Arenosols, Podzols, Planosols, and Vertisols were not associated with any particular spectral class but it is worth mentioning that they were closer to class 3. The CA of spectral classes and soil layers showed that classes 6 and 3 were correlated with A layer (Fig. 10f). The B layer showed some association with spectral classes 2, 4, and 5 (Fig. 10f), which is corroborated by the fact that spectral classes 4 and 5 showed similarity in CR spectrum (Fig. 8g, i). The C layer was strongly associated with class 1. It is also important to mention that the spectral class 4 was in the middle of layers A and B, and spectral class 5 was in the middle of layers B and C. The D layer presented low association with classes (Fig. 10f).

4. Conclusions

The BSSL provided strong evidence to be a useful tool to estimate soil attributes such as clay, sand, SOC, CEC, pH, and BS, with variable results. There were differences among models considering national (for all Brazil), regional and state scales. The results were coherent for clay, sand, SOC, and CEC. The attributes with low content in soils are more prone to show high inaccuracy and need further evaluations, such as chemical ones (Ca, K, P, others). In general, spectral signatures also had great relationship with soil mineralogy. Cluster analysis showed that Brazil has six classes of spectral signatures among the studied population and there were clear differences among spectra developed in different geographic (i.e., states) and environmental locations (i.e., geology). The results endorse the importance and relevance of spectral libraries for soil evaluation in support on its quantification, quality and classification. The large spectral database, will be enhanced at scales that meet the users' needs for soil mapping in Brazil. The use of sensors and geotechnologies, allows a higher sample throughput and denser sampling. Both can generate data for soil survey and mapping that will assist in the sustainable management of agriculture and forest systems. We believe that soil spectroscopy in Brazil is on the right track, because it is a fast, simple, accurate and most importantly non-pollutant method. In addition, a strong collaborative network and infrastructure has been formed that supports the expansion of soil spectral soil mapping. With the approval of Brazil's National Soils Program (PronaSolos), which is promising in relation to the mapping of Brazilian soils, we hope that techniques such as soil spectroscopy can be applied.

Besides the importance of spectral standardization, spectral libraries

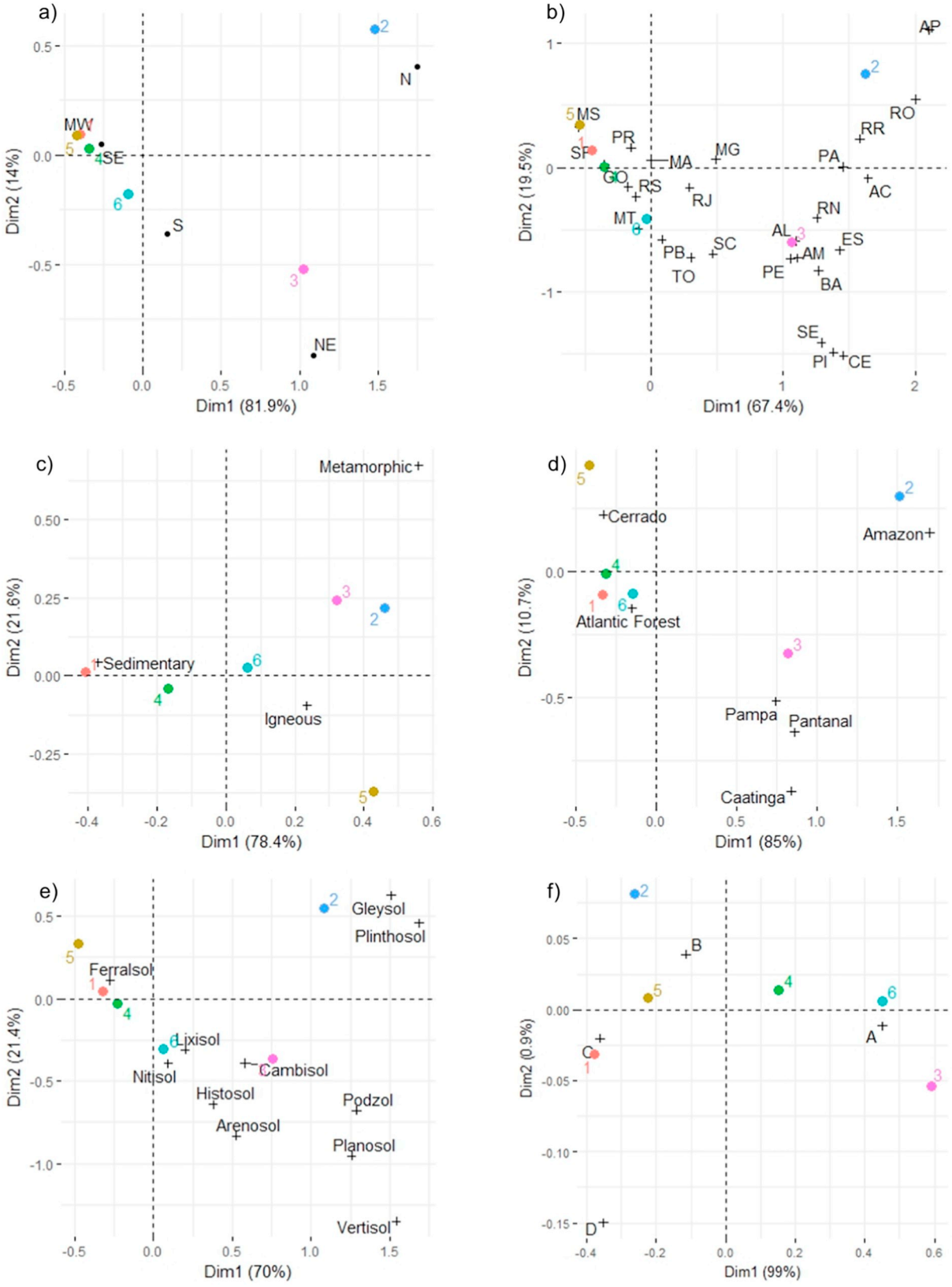

Fig. 10. Ordination diagrams from the correspondence analysis (CA) between the 6 spectral classes and Brazilian regions (a), states (b), geology (c), biomes (d), soil classes (only profile samples that have layer B were used) (e), and layers (f).

must be accompanied by chemical and physical characterization of soil attributes. Depending on the volume of soil samples, the standard procedures applied to spectral measurements can be complex and time-consuming. Despite the spectral variation can also be an issue to be faced, the incorporation of soil chemical and physical data is crucial. The spectral acquisition is no more an obstacle to the organization of spectral libraries but the collection of reliable and consistent soil chemical and physical data poses challenges. Soil spectral information is dependent on wet analysis. In some cases, the soil sample collection and wet chemical/physical analyses were conceived before the BSSL initiative. Therefore, standardized collection and analyses of samples were not possible in all circumstances. We agreed that it can create a margin of error in both calibration and validation processes, but the novelty of this database and the difficulty to gather information must be taken into consideration. Furthermore, having a SSL that represent Brazilian soils is just as important as defining the degree of uncertainty. Therefore, in the first stage of the SSL's development we decided to include samples from many wet chemistry and physical laboratories.

The results are a first step towards the establishment of the BSSL. The database is being continuously increased with new information, consequently increasing its representativeness along the Brazilian territory. Improvements in the frame-work have been conduced, including computational routines implementing sophisticated statistical procedures, which will reduce the uncertainties in the calibration procedure. Among them, data will be filtered and standardized with wavelets, which will help to account for the inconsistencies in sample preparation, different measurement protocols and instruments that were used. In parallel, different machine learning algorithms have been evaluated, aiming to define the most suitable data mining method for our dataset. These mining procedures account for local relationships in the data providing to the models a wide usability at different spatial scales (local, regional and national). We also developed an interactive online platform to disseminate the use of spectroscopy in soil science, and to interact with the database administrators. The soil dataset and their contributors can be accessed at < https://bibliotecaespectral.wixsite.com/esalq > . By increasing the number of users, the data available and knowledge will also increase and consequently the BSSL will be constantly improved to represent the variability of Brazilian soils. We have to keep in mind that the importance of spectra is not only concentrated on chemical agriculture information (i.e., Ca, Mg, K, others). From one measurement, we can achieve chemical, physical and mineralogical information, which are also important for soil mapping and agriculture as well. Finally, in our vision, wet analysis is an important method and now has the great opportunity to merge knowledge (and aggregate information) with proximal sensing, to evolve on soil analysis to a new generation and its benefits.

Acknowledgements

We would like to thank the São Paulo Research Foundation (FAPESP) for the financial support for first author (Project grant n. 2014/22262-0), second author (Project grant n. 2017/03207-6) and third author (Project grant n. 2016/26176-6) and the National Council for Scientific and Technological Development (CNPq). Also, to thank the Geotechnologies on Soil Science group - GeoSS (*esalqgeocis.wixsite.com/english*) and to everybody that directly or indirectly assisted on publishing this study.

References

Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services — a global review. Geoderma 262, 101–111. https://doi.org/10.1016/J.GEODERMA.2015.08.009.

Baldridge, A.M., Hook, S.J., Grove, C.I., Rivera, G., 2009. The ASTER spectral library version 2.0. Remote Sens. Environ. 113, 711–715. https://doi.org/10.1016/J.RSE.2008.11.007.

Bellinaso, H., Demattê, J.A.M., Romeiro, S.A., 2010. Soil spectral library and its use in soil classification. R. Bras. Ci. Solo 34, 861–870. https://doi.org/10.1590/S0100-06832010000300027.

Ben-Dor, E., Ong, C., Lau, I.C., 2015. Reflectance Measurements of soils in the laboratory: Standards ans protocols. Geoderma 245, 112–124. https://doi.org/10.1016/j.geoderma.2015.01.002.

Benzécri, J.P., 1992. Correspondence Analysis Handbook. Marcel Dekker, New York, NY.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. 10, 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.

Bowers, S.A., Hanks, R.J., 1965. Reflection of radiant energy from soils. Soil Sci. 100, 130–138.

Brodský, L., Klement, A., Penížek, V., Kodešová, R., Borůvka, L., 2011. Building soil spectral library of the Czech soils for quantitative digital soil mapping. Soil Water Res 6, 165–172.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025.

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., Smaling, E.M.A., 2012. Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. Geoderma 183–184, 41–48. https://doi.org/10.1016/j.geoderma.2012.03.011.

Cebeci, Z., Yildiz, F., Kavlak, A.T., Cebeci, C., Onder, H., 2018. ppclust: Probabilistic and Possibilistic Cluster Analysis. R Packag. (version 0.1.1).

Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J. Geophys. Res. 89, 6329–6340. https://doi.org/10.1029/JB089iB07p06329.

Demattê, J.A.M., 2016. From profile Morphometrics to digital soil mapping. In: Digital Soil Morphometrics. Springer International Publishing, pp. 383–399. https://doi.org/10.1007/978-3-319-28295-4_24.

Demattê, J.A.M., Oliveira, J. de C., Tavares, T.R., Lopez, L.R., Terra, F. da S., Araújo, S.R., Fongaro, C.T., Maia, S.M.F., Mello, F.F. de C., Rizzo, R., Vicente, S., de Melo Bortolleto, M.A., Cerqueira, P.H.R., 2016. Soil chemical alteration due to slaughterhouse waste application as identified by spectral reflectance in São Paulo State, Brazil: an environmental monitoring useful tool. Environ. Earth Sci. 75. https://doi.org/10.1007/s12665-016-6042-2.

Demattê, J.A.M., Dotto, A.C., Bedin, L.G., Sayão, V.M., Souza, A.B. e, 2019. Soil analytical quality control by traditional and spectroscopy techniques: constructing the future of a hybrid laboratory for low environmental impact. Geoderma 337, 111–121. https://doi.org/10.1016/J.GEODERMA.2018.09.010.

Donagemma, G.K., Campos, D.V.B. de, Calderano, S.B., Teixeira, W.G., Viana, J.H.M., 2011. Manual de métodos de análise de solo, 2 rev. ed, Embrapa Solos.

Epiphanio, J.C.N., Formaggio, A.R., Valeriano, M.D.M., Oliveira, J.B., 1992. Comportamento espectral de solos do Estado de São Paulo. Instituto Nacional de Pesquisas Espaciais. São José dos Campos, São Paulo.

Formaggio, A.R., Epiphanio, J.C.N., Valeriano, M.M., Oliveira, J.B., 1996. Comportamento espectral (450-2.450 nm) de solos tropicais de São Paulo. Rev. Bras. Ciência do Solo 20, 467–474.

Garrity, D., Bindraban, P., 2004. A Globally Distributed Soil Spectral Library Visible Near Infrared Diffuse Reflectance Spectra. ICRAF (World Agroforestry Centre)/ISRIC (World Soil Information) Spectral Library, Nairobi, Kenya.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemom. Intell. Lab. Syst. 110, 168–176. https://doi.org/10.1016/J.CHEMOLAB.2011.11.003.

Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Fusion of soil and remote sensing data to model soil properties. In: Sparks, D.L. Adv. Agron. 131, 1–109.

Grunwald, S., Yu, C., Xiong, X., 2018. Transferability and scalability of total soil carbon prediction models in Florida, USA. Pedosphere J. 28 (6), 856–872.

IUSS Working Group WRB, 2015. World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports No. 106. FAO, Rome.

Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R.A., Mouazen, A.M., 2016. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. Soil Tillage Res. 155. https://doi.org/10.1016/j.still.2015.06.004.

Jónsson, J.Ö.G., Davíðsdóttir, B., 2016. Classification and valuation of soil ecosystem services. Agric. Syst. 145, 24–38. https://doi.org/10.1016/j.agsy.2016.02.010.

Knadel, M., Deng, F., Thomsen, A., Greve, M., 2012. Development of a Danish national Vis–NIR soil spectral library for soil organic carbon determination. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney, Australia, pp. 403–408.

Kuhn, M., et al., 2017. Caret: Classification and Regression Training. R Package Version 6.0-73.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: an R package for multivariate analysis. J. Stat. Softw. 25, 1–18. https://doi.org/10.18637/jss.v025.i01.

Madeira Netto, J.S., 2001. Comportamento espectral dos solos. In: Meneses, P.R., Madeira Netto, J.S. (Eds.), Sensoriamento remoto - reflectância dos alvos naturais. Brasília, Brazil, pp. 127–147.

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemom. Intell. Lab. Syst. 94, 72–79. https://doi.org/10.1016/j.chemolab.2008.06.003.
Moeys, J., 2016. soiltexture: Functions for Soil Texture Plot, Classification and Transformation. R package version 1.4.1.
Mutanga, O.M.C., Skidmore, A.K., Kumar, L., Ferwerda, J., 2005. Estimating tropical pasture quality at canopy level using band depth analysis with continuum removal in the visible domain. Int. J. Remote Sens. 26, 1093–1108. https://doi.org/10.1080/01431160512331326738.
Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022.
Polidoro, J.C., Mendonça-Santos, M.D.L., Lumbreras, J.F., Coelho, M.R., Carvalho, Filho, De, A., Da Motta, P.E.F., Carvalho, Junior, De, W., Araujo, Filho, De, J.C., Curcio, G.R., Correia, J.R., Martins, E.D.S., Spera, S.T., Oliveira, S.R.D.M., Bolfe, E.L., Manzatto, C.V., Tosto, S.G., Venturieri, A., Sa, I.B., De Oliveira, V.A., Shinzato, E., Anjos, L.H.C. Dos, Valladares, G.S., Ribeiro, J.L., De Medeiros, P.S.C., Moreira, F.M.D.S., Silva, L.S.L., Sequinatto, L., Aglio, M.L.D., Dart, R.D.O., 2016. Programa Nacional de Solos do Brasil (PronaSolos), 1st ed. Embrapa Solos, Rio de Janeiro, RJ.
Quinlan, J., 1992. Learning with continuous classes. In: Adams, A., Sterling, L. (Eds.), Proceedings AI'92, 5th Australian Conference on Artificial Intelligence. World Scientific, Singapure, pp. 343–348.
R Core Team, 2018. R: A language and environment for statistical computing.
Schmidt, M., 2008. The Sankey diagram in energy and material flow management. J. Ind. Ecol. 12, 82–94. https://doi.org/10.1111/j.1530-9290.2008.00004.x.
Shepherd, K.D., Walsh, M.G., 2002. & SOIL & PLANT ANALYSIS Development of Reflectance Spectral Libraries for Characterization of Soil Properties 988–998.
Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., Viscarra Rossel, R.A., 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Sci. China Earth Sci. 57, 1671–1680. https://doi.org/10.1007/s11430-013-4808-x.
Soil Survey Staff, 2014. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys, Twelfth Ed. ed. Natural Resources Conservation Service. U. S. Department of Agriculture Handbook.
Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma 144, 395–404. https://doi.org/10.1016/J.GEODERMA.2007.12.009.
Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near InfraRed reflectance spectroscopy. PLoS One 8, e66409. https://doi.org/10.1371/journal.pone.0066409.
Stoner, E.R., Baumgardner, M.F., 1981. Characteristic variations in reflectance of surface soils. Soil Sci. Soc. Am. J. 45, 1161. https://doi.org/10.2136/sssaj1981.03615995004500060031x.
Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: comparing vis–NIR and mid-IR reflectance data. Geoderma 255–256, 81–93. https://doi.org/10.1016/j.geoderma.2015.04.017.
Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2018. Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. Geoderma 318, 123–136. https://doi.org/10.1016/J.GEODERMA.2017.10.053.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, Diffuse reflectance spectroscopy in soil science and land resource assessment 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.
Viscarra Rossel, R.A., McBratney, A.B., 2008. Diffuse reflectance spectroscopy as a tool for digital soil mapping. In: Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht, pp. 165–172. https://doi.org/10.1007/978-1-4020-8592-5_13.
Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63, 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Science Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012.
Viscarra Rossel, R.A.A., McGlynn, R.N.N., McBratney, A.B.B., 2006. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. Geoderma 137, 70–82. https://doi.org/10.1016/j.geoderma.2006.07.004.
Wall, D.H., Nielsen, U.N., 2012. Biodiversity and ecosystem services: is it the same below ground? Nat. Educ. Knowl. 3, 8.
Zeng, R., Zhao, Y.-G., Li, D.-C., Wu, D.-W., Wei, C.-L., Zhang, G.-L., 2016. Selection of "local" models for prediction of soil organic matter using a regional soil Vis-NIR spectral library. Soil Sci. 181, 13–19. https://doi.org/10.1097/SS.0000000000000132.