

**FEDERAL UNIVERSITY OF SANTA MARIA
CENTER OF RURAL SCIENCES
GRADUATE PROGRAM IN SOIL SCIENCE**

André Carnieletto Dotto

**SOIL VIS-NIR SPECTROSCOPY: PREDICTIVE POTENTIAL AND
THE DEVELOPMENT OF A GRAPHICAL USER INTERFACE IN R**

Santa Maria, RS
2017

André Carnieletto Dotto

**SOIL VIS-NIR SPECTROSCOPY: PREDICTIVE POTENTIAL AND THE
DEVELOPMENT OF A GRAPHICAL USER INTERFACE IN R**

Thesis submitted to the Graduate Program in Soil Science, Area of concentration Physical and Morphogenetic Processes of Soil, at Federal University of Santa Maria (UFSM, RS), as a partial requirement to obtain the degree of **Doctor in Soil Science**.

Advisor: Prof. Dr. Ricardo Simão Diniz Dalmolin

Santa Maria, RS

2017

Ficha catalográfica elaborada através do Programa de Geração Automática da Biblioteca Central da UFSM, com os dados fornecidos pelo(a) autor(a).

Dotto, André Carnieletto
SOIL VIS-NIR SPECTROSCOPY: PREDICTIVE POTENTIAL AND
THE DEVELOPMENT OF A GRAPHICAL USER INTERFACE IN R /
André Carnieletto Dotto.- 2017.
112 p.; 30 cm

Orientador: Ricardo Simão Diniz Dalmolin
Tese (doutorado) - Universidade Federal de Santa
Maria, Centro de Ciências Rurais, Programa de Pós-
Graduação em Ciência do Solo, RS, 2017

1. Alrad Spectra 2. Técnica de espectroscopia 3.
Espectros de solo 4. Análise quimiométrica 5. GUI de
fácil utilização I. Simão Diniz Dalmolin, Ricardo II.
Título.

© 2017

All copyrights reserved to André Carnieletto Dotto. The reproduction of parts or all of this work can only be done by quoting the source.

Address: Universidade Federal de Santa Maria, Centro de Ciências Rurais, Av. Roraima, n. 1000, Prédio 42, sala 3314, Camobi, Santa Maria, RS, Brazil. CEP: 97105-900 Fone +55 (55) 3220-8157;

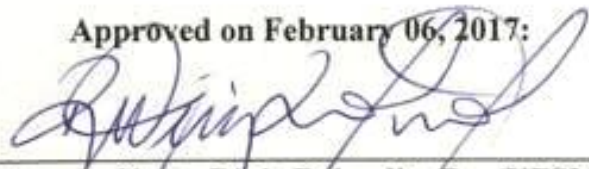
E-mail: andrecdot@gmail.com

André Carnieletto Dotto

**SOIL VIS-NIR SPECTROSCOPY: PREDCTIVE POTENTIAL AND THE
DEVELOPMENT OF A GRAPHICAL USER INTERFACE IN R**

Thesis submitted to the Graduate Program in
Soil Science, Area of concentration Physical
and Morphogenetic Processes of Soil, at
Federal University of Santa Maria (UFSM,
RS), as a partial requirement to obtain the
degree of **Doctor in Soil Science**.

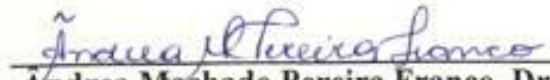
Approved on February 06, 2017:



Ricardo Simão Diniz Dalmolin, Dr. (UFSM)
(President/Advisor)



Alexandre ten Caten, Dr. (UFSC)



Andrea Machado Pereira Franco, Dra (UFSM)



Suzana Romeiro Araújo, Dra (UFRA)



Gustavo de Mattos Vasques, Dr. (EMBRAPA)

Santa Maria, RS
2017

DEDICATION

To my family!
Your support, encouragement and constant love
have sustained me throughout my life!

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Ricardo S. D. Dalmolin and co-advisor Prof. Alexandre ten Caten for the continuous support on my research, for their patience, motivation, and immense knowledge.

Besides my advisors, I would like to thank the rest of my thesis examining committee: Prof. Suzana R. Araujo, Dr. Gustavo de M. Vasques, and Dra Ândrea M. P. Franco for their insightful comments and ideas.

My sincere thank also goes to Prof. Sabine Grunwald, who provided me an opportunity to join their team as exchange student. Without her precious guidance, it would not be possible to conduct my split-site PhD abroad.

Heartfelt thanks go out to my girlfriend Daiana for all your love, support and patience when I needed it most. To our loving dogs, Chokito e Milka, for the comfy company.

Also, I thank my friends in the University of Florida, In particular, Wade for the friendship and fishing, Carla for the beers, Kay for the patience and kindness, Chong for the Chinese food and partnership, Yiming for the advises and chats, Hamza for the wild talks, and Betty for the advices.

I thank my fellow labmates Jean, Gabriel, Pedro, Taciara, Luciano, Ismael, Vicente, Lucas, Walquiria, Boing, Bruno, Luis for the help in soil samples collection and lab analysis, for the stimulating discussions, for the friendship, and for all the fun we have had in the last four years. Especial thanks to my fellow Diego for helping me to develop Alrad Spectra. I also thank my Japanese friend Takao for the partnership overseas and experiences exchanged.

Thank CAPES for funding my Doctoral Scholarship. To the secretary of Graduate Program in Soil Science Heverton for all the assistance. To all those people who were not mentioned here, you have also contributed in different ways to accomplish this thesis.

Last but not the least, I would like to thank my family: my parents, brother, sisters, brother-in-law, sister-in-law, nephew and niece for your love, support and encouragement throughout these years, for giving me strength to continue studying and for being part of my life.

This accomplishment would not have been possible without you all. Thank you.

“Change will not come if we wait for some other person or some other time.
We are the ones we've been waiting for. We are the change that we seek.”
Barack Obama

“Live as if you were to die tomorrow. Learn as if you were to live forever.”
Mahatma Gandhi

RESUMO

ESPECTROSCOPIA DO SOLO NO VIS-IR: POTENCIAL PREDICTIVO E DESENVOLVIMENTO DE UMA INTERFACE GRÁFICA DE USUÁRIO EM R

Autor: André Carnieletto Dotto
Orientador: Ricardo Simão Diniz Dalmolin

Esta tese apresenta um estudo da técnica de espectroscopia do visível ao infravermelho próximo aplicado à predição de propriedades do solo. O propósito foi de desenvolver informações quantitativas sobre o solo, devido à demanda do mapeamento digital de solos, monitoramento ambiental, produção agrícola e aumento das informações espaciais do solo. A espectroscopia surge como uma alternativa para revolucionar a monitorização do solo, permitindo uma amostragem rápida, de baixo custo, não destrutiva, ambientalmente amigável, reproduzível e repetitiva. Para melhorar a eficiência da predição do solo usando dados espectrais, várias técnicas de pré-processamento espectral e modelos multivariados foram explorados. Uma interface gráfica de usuário (GUI) no R, denominada Alrad Spectra, foi desenvolvida para realizar pré-processamento, modelagem multivariada e predição usando dados espectrais. Os objetivos foram: i) prever as propriedades do solo para melhorar a informação do solo usando dados espectrais, ii) comparar os desempenhos dos pré-processamentos espectrais e métodos de calibração multivariada na predição do carbono orgânico do solo, iii) obter predições confiáveis do carbono orgânico do solo, e iv) desenvolver uma interface gráfica de usuário que realize o pré-processamento espectral e a predição do atributo solo usando dados espectroscópicos. Um total de 595 amostras de solo foram coletadas na região central do estado de Santa Catarina, Brasil. A reflectância espectral do solo foi obtida utilizando um espectrorradiômetro FieldSpec 3 com uma alcance espectral de 350-2500 nm com 1 nm de resolução espectral. Os resultados da tese demonstraram o grande desempenho da predição de propriedades do solo usando espectroscopia do visível ao infravermelho próximo. As propriedades do solo que estão diretamente relacionadas aos cromóforos, como o carbono orgânico, apresentaram predições superiores comparados com o tamanho de partículas. O pré-processamento espectral aplicado nos espectros do solo contribuiu para o desenvolvimento de um modelo de predição de alto nível. Comparando diferentes técnicas de pré-processamento espectral para a predição de carbono orgânico revelou que as técnicas de pré-processamento de correção de dispersão apresentaram resultados de predição superiores em comparação com as técnicas de derivação espectrais. Na técnica de correção de dispersão, a remoção do contínuo é o pré-processamento mais adequado a ser usado para a predição de carbono. Na modelagem de calibração, com exceção da floresta aleatória, todos os métodos apresentaram uma elevada predição, sendo destaque o método máquina de vetores de suporte. A metodologia sistemática aplicada neste estudo pode melhorar a confiabilidade da estimativa do carbono orgânico ao examinar como as técnicas de pré-processamento espectral e métodos multivariados afetam a performance da predição usando a análise espectral. O desenvolvimento da GUI de fácil utilização pode beneficiar um grande número de usuários, os quais podem tirar proveito desta análise quimiométrica. Alrad Spectra é a primeira GUI desse tipo e a expectativa é que esta ferramenta possa expandir a aplicação da técnica de espectroscopia.

Palavras-chave: Alrad Spectra, técnica de espectroscopia, espectros de solo, análise quimiométrica, GUI de fácil utilização.

ABSTRACT

SOIL VIS-NIR SPECTROSCOPY: PREDICTIVE POTENTIAL AND THE DEVELOPMENT OF A GRAPHICAL USER INTERFACE IN R

Author: André Carnieletto Dotto
Advisor: Ricardo Simão Diniz Dalmolin

This thesis presents a study of Visible Near-infrared spectroscopy technique applied to predict soil properties. The purpose was to develop quantitative soil information due to the demand of digital soil mapping, environmental monitoring, agricultural production and for increasing spatial information on soil. Soil spectroscopy emerge as an alternative to revolutionize soil monitoring, allowing rapid, low-cost, non-destructive samples sampling, environmental-friendly, reproducible, and repeatable analysis. To improve the efficiency of soil prediction using spectral data, several spectral preprocessing techniques and multivariate models were exploited. A graphical user interface (GUI) in R, named Alrad Spectra, was developed to perform preprocessing, multivariate modeling and prediction using spectral data. Hereby, the objectives were: The objectives were: i) to predict soil properties to improve soil information using spectral data, ii) to compare the performance of spectral preprocessing and multivariate calibration methods in the prediction of soil organic carbon, iii) to obtain reliable soil organic carbon prediction, and iv) to develop a graphical user interface that performs spectral preprocessing and prediction of the soil property using spectroscopic data. A total of 595 soil samples were collected in central region of Santa Catarina State, Brazil. Soil spectral reflectance was obtained using a FieldSpec 3 spectroradiometer with a spectral range of 350–2500 nm with 1 nm of spectral resolution. The outcomes of the thesis have demonstrated the great performance of predicting soil properties using Vis-NIR spectroscopy. Apparently, soil properties that are directly related to the chromophores such as organic carbon presented superior prediction statistics than particle size. Spectral preprocessing applied in the soil spectra contribute to the development of high-level prediction model. Comparing different spectral preprocessing techniques for soil organic carbon (SOC) prediction revealed that the scatter–corrective preprocessing techniques presented superior prediction results compared to spectral derivatives. In scatter–correction technique, continuum removal is the most suitable preprocessing to be used for SOC prediction. In the calibration modeling, excepting for random forest, all of methods presented robust prediction, with emphasis on the support vector machine method. The systematic methodology applied in this study can improve the reliability of SOC estimation by examining how techniques of spectral preprocessing and multivariate methods affect the prediction performance using spectral analysis. The development of easy-to-use graphical user interface may benefit a large number of users, who will take advantage of this useful chemometrics analysis. Alrad Spectra is the first GUI of its kind and the expectation is that this tool can expand the application of the spectroscopy technique.

Keywords: Alrad Spectra, spectroscopy technique, soil spectra, chemometrics analysis, user-friendly GUI.

SUMMARY

1 INTRODUCTION.....	12
2 ARTICLE 1: TWO PREPROCESSING TECHNIQUES TO REDUCE MODEL COVARIABLES IN SOIL PROPERTY PREDICTIONS BY VIS-NIR SPECTROSCOPY.....	16
2.1. INTRODUCTION.....	16
2.2. MATERIAL AND METHODS.....	19
2.2.1. Study site and sample collection.....	19
2.2.2. Soil analysis in the laboratory.....	19
2.2.3. Spectral reflectance measurements.....	20
2.2.4. Spectral preprocessing.....	20
2.2.5. Statistical analysis.....	21
2.2.6. Model training and validation.....	22
2.3. RESULTS AND DISCUSSION.....	22
2.3.1. Exploratory results.....	22
2.3.2. Predictive performance of PLSR and SVM.....	23
2.3.3. Performance of spectral band selection techniques.....	26
2.3.4. Performance of preprocessing techniques.....	28
2.4. CONCLUSIONS.....	28
3 ARTICLE 2: COMPARING THE CAPABILITY OF PREPROCESSING TECHNIQUES AND MULTIVARIATE METHODS TO PREDICT SOIL ORGANIC CARBON USING SPECTROSCOPIC DATA.....	44
3.1. INTRODUCTION.....	44
3.2. MATERIAL AND METHODS.....	47
3.2.1. Study area.....	47
3.2.2. Data collection and soil analysis.....	48
3.2.3. Training and validation sets.....	48
3.2.4. Spectral reflectance measurements.....	49
3.2.5. Spectral preprocessing techniques.....	49
3.2.6. Multivariate methods.....	50
3.3. RESULTS AND DISCUSSION.....	51
3.3.1. Descriptive and inferential statistics.....	51
3.3.2. Characteristics of soil spectral reflectance curves.....	52
3.3.3. Influence of preprocessing techniques in the performance of SOC models.....	55
3.3.3.1. <i>Two groups of preprocessing techniques</i>	55
3.3.3.2. <i>Performance of best preprocessing technique</i>	58
3.3.4. Influence of multivariate methods in the performance of SOC prediction.....	59
3.3.4.1. <i>Partial least-squares regression performance</i>	60
3.3.4.2. <i>Principal component regression performance</i>	61
3.3.4.3. <i>Multiple linear regression performance</i>	62
3.3.4.4. <i>Support vector machine performance</i>	63
3.3.4.5. <i>Random forest performance</i>	64
3.3.4.6. <i>Bayesian model averaging performance</i>	65
3.3.4.7. <i>Weighted average partial least squares performance</i>	65
3.3.4.8. <i>Gaussian process regression performance</i>	66
3.3.4.9. <i>Artificial neural network performance</i>	67
3.3.5. Comparing performances.....	70

3.3.6.	Time to process the models in R	71
3.4.	CONCLUSIONS	72
4	ARTICLE 3: ALRAD SPECTRA: A GRAPHICAL USER INTERFACE IN R TO PERFORM PREPROCESSING, MULTIVARIATE MODELING AND PREDICTION USING SPECTROSCOPIC DATA	82
4.1.	INTRODUCTION	82
4.2.	SOFTWARE.....	83
4.3.	GUI DESCRIPTION	84
4.3.1.	Import Data module	86
4.3.2.	Spectral Preprocessing module	87
4.3.2.1.	<i>Smoothing</i>	88
4.3.2.2.	<i>Binning</i>	88
4.3.2.3.	<i>Absorbance</i>	89
4.3.2.4.	<i>Detrend</i>	89
4.3.2.5.	<i>Continuum removal (CR)</i>	89
4.3.2.6.	<i>Savitzky–Golay derivative (SGD)</i>	90
4.3.2.7.	<i>Standard normal variate (SNV)</i>	90
4.3.2.8.	<i>Multiplicative scatter correction (MSC)</i>	91
4.3.2.9.	<i>Normalization</i>	91
4.3.3.	Modeling module	91
4.3.3.1.	<i>Multiple linear regression (MLR)</i>	92
4.3.3.2.	<i>Partial least squares regression (PLSR)</i>	93
4.3.3.3.	<i>Support vector machines (SVM)</i>	93
4.3.3.4.	<i>Random forest (RF)</i>	93
4.3.3.5.	<i>Artificial neural network (ANN)</i>	94
4.3.3.6.	<i>Gaussian process regression (GPR)</i>	94
4.3.4.	Prediction module	95
4.4.	CASE STUDY.....	96
4.4.1.	Data set.....	96
4.4.2.	Soil spectral preprocessing.....	96
4.4.3.	Modeling and prediction of SOC	98
4.4.4.	Predict unknown SOC	103
4.5.	CONCLUSION	104
5	DISCUSSION	108
6	CONCLUSION.....	110

1 INTRODUCTION

Soil is a natural source of organic and inorganic material that covers the earth's surface being an open and heterogeneous system with complex processes and mechanisms of formation. Due to this, soils present great variability in chemical, physical and biological composition. The soil provides a multiplicity of ecosystem functions, goods and services supporting and regulating life on the planet (MONTANARELLA et al., 2015). Consequently, the preservation and sustainable management of soils is crucial to prevent the major soil threats that endanger humanity such as food security, climate change, environmental degradation, water scarcity, and biodiversity (SANCHEZ et al., 2009).

The preservation and sustainable management of soils involves a number of factors, including access to soil information. The demand of quantitative information for soil mapping purposes, environmental monitoring, agricultural production and especially for increasing spatial information on soil is increasing (HARTEMINK; MINASNY, 2014). For Sanchez et al. (2009) the demand for up-to-date and relevant soil information is growing, but exchanging such information among the science community remains challenging.

The necessity to increase soil information requires complex methodical approaches with an excessive number of parameters to measure. At present, soil analyses, carried out in routine laboratories, are being discussed by soil scientists. This is due to the fact that the methodologies being used exposed problems related to the costs of analysis, production of chemical residues generated by the standard analysis and the time required for the processing of soil samples (SOUSA JUNIOR; DEMATTÊ; ARAÚJO, 2011).

One of the challenges is to propose a technique that has the potential to revolutionize soil monitoring, allowing rapid, low-cost, non-destructive sampling, environmental-friendly, reproducible, and repeatable analysis (VISCARRA ROSSEL et al., 2006). Visible and Near-Infrared (Vis-NIR) reflectance spectroscopy emerges as an alternative method to satisfy these needs (STEVENS et al., 2013). In addition, there is no use of environmentally harmful chemical reagents. The technique is mainly used in the laboratory in controlled environment (VASQUES; GRUNWALD; SICKMAN, 2008), but field measurement has been developed to allow direct and rapid soil information (HARTEMINK; MINASNY, 2014). Soil spectroscopy is about the identification and analysis of the interaction of wavelengths with soil properties. The technique allows the characterization of a series of soil properties simultaneously with only a single spectral sample scan. In this context, Vis-NIR spectroscopy can be used to identify specific soil features in the spectral curves and estimate important soil properties (STONER;

BAUMGARDNER, 1981). Different soil properties such as particle size, moisture, mineralogy and organic matter can influence the absorption of electromagnetic radiation causing variation of the reflectance (DALMOLIN et al., 2005; DEMATTÊ et al., 2004). The technique can be used to predict important soil attributes such as soil organic matter, minerals, texture, nutrients, water, pH, and heavy metals (STENBERG et al., 2010).

To improve the efficiency of soil prediction using spectral data, several spectral preprocessing techniques have been exploited. These techniques have been applied to transform soil spectra, remove noise, emphasize features, and extract useful information for quantitative predictive models. The most utilized spectral preprocessing includes smoothing, normalization, scatter correction, and derivatives (RINNAN; BERG; ENGELSEN, 2009). The selection and performance of these spectral preprocessing in soil prediction are diverse according to many studies. Hence, there is a need to explore and assess a wide range of spectral preprocessing in order to compare their predictive performance in the same soil dataset.

Regarding the predictive performance, a proper modeling approach is needed. Several multivariate calibration methods have been successfully applied with the intention of developing a faster and high-quality model for soil property prediction. Among the methods, partial least-squares regression (PLSR) stands out as the most common calibration method. Moreover, other methods have gained emphasis such as support vector machine (SVM), random forest (RF), artificial neural network (ANN), Bayesian model averaging (BMA), weighted average partial least squares (WAPLS), and Gaussian process regression (GPR). Besides these, multiple linear regression (MLR) and principal components regression (PCR) have presented significant prediction results. The idea to compare well-known and alternatives methods can provide an extensive assessment in the selection of the most accurate model. More efforts should be focused on revealing the potential of these methods in soil analysis. The evaluation of an extensive variety of multivariate statistics would be capable of improving the model prediction based on Vis-NIR spectroscopy and would allow a systematic methodology development for imminent usage in spectral analysis laboratories.

Soil properties prediction studies based on Vis-NIR spectroscopy presented a considerable increase in the last decades (BELLON-MAUREL; MCBRATNEY, 2011). According to Nocita et al. (2015) this growth is due to the minor sample preparation, more applicability under field condition and Vis-NIR instruments are more widespread than mid-infrared. For Viscarra Rossel et al. (2016) soil spectroscopy has grown considerably over the past 30 years because of the development of new spectrometers, new technologies that use microelectromechanical structures, thin film filters, lasers, light emitting diodes, optical fiber

assemblies, high performance detector arrays, producing miniaturized hand-held instruments that are rugged and cheap. The authors suggested that continual improvements in computing and statistics have helped to extract useful information from the spectra and to improve our understanding of soil.

Along with the expansion of this technique came the need to popularize the computational processes involving all stages for soil spectra analyses and simplify the interaction of statistical programs. According to Valero-Mora and Ledesma (2012) graphical user interfaces (GUI) improved the usability of computer program applications and are the most common way of interacting with a computer. The statistical analysis implemented in R programming language (R CORE TEAM, 2016) are operated by a typed language via a command line interface. Writing up commands can be time-consuming and for occasional users of statistical application the amount of effort needed for learning programming language will not pay the price. Therefore, a GUI in R that handles spectral preprocessing and modeling methods in order to predict soil properties can be developed. The GUI may be the cutting-edge for adoption and expansion of soil spectroscopy technique.

Definitively, soil spectroscopy technique can be considered an alternative to improve soil analyses that are currently carried out in routine laboratory by conventional methods (MINASNY; MCBRATNEY, 2008). It seems that the implementation of Vis-NIR spectroscopy in soil laboratories is a matter of time. Thereby, the hypothesis of this study is that the assessment of different methodological procedures can increase the prediction performance of soil properties using Vis-NIR spectroscopy. The objectives are: i) to predict soil properties to improve soil information using spectral data, ii) to compare the performance of spectral preprocessing and multivariate calibration methods in the prediction of soil organic carbon, iii) to obtain reliable soil organic carbon prediction, and iv) to develop a graphical user interface that performs spectral preprocessing and prediction of the soil property using spectroscopic data.

This thesis was submitted to the Graduate Program of Soil Science, Federal University of Santa Maria (UFSM). In the third year, I accomplished the split-site PhD at the University of Florida, USA. This study was financed by three resources. The doctoral scholarship was financed by the Coordination for the Improvement of Higher Education Personnel (CAPES), by the Brazilian National Council for Scientific and Technological Development (CNPq), and by the Foundation for Funding in Research and Innovation of Santa Catarina State (FAPESC), Ministry of Education, Brazil.

Soil samples were collected over an area of about 1,800 km² in central region of Santa Catarina State, Brazil. A total 595 soil samples were collected, wherein 539 followed the depths specifications of 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm from Globalsoilmap.net (ARROUAYS et al., 2014) and 56 samples derived from soil horizons of 11 profiles. Soil samples represented the prominent soil types of the region. The Oxisols are predominant in the area showing an advanced degree of weathering and developing deep soils. Furthermore, in some steep areas, younger and shallower soils, such as Entisols and Inceptisols, are found in a complex relief. The soil chemical (organic carbon) and physical analyzes (particle size) were realized at Pedology Laboratory, UFSM. The spectral scans were carried out at GeoCis Laboratory, Soil Science Department, ESALQ/University of Sao Paulo (USP).

The thesis was elaborated in sections divided into introduction, three scientific articles, discussion and conclusion. The title of the articles are as follows: 1) Two preprocessing techniques to reduce model covariables in soil property predictions by Vis-NIR spectroscopy; 2) Comparing the capability of preprocessing techniques and multivariate methods to predict soil organic carbon using spectroscopic data; and 3) Alrad Spectra: a graphical user interface in R to perform preprocessing, multivariate modeling and prediction using spectroscopic data. The discussion section explores the importance that soil spectroscopy has been gaining among the researchers and shows a perspective of the pathways that this theme should trail, besides guiding future studies and demands.

2 ARTICLE 1: Two preprocessing techniques to reduce model covariables in soil 3 property predictions by Vis-NIR spectroscopy¹

4 Abstract

5 Proximal sensing provides an alternative method to physical and chemical laboratory soil
6 analyses. The aim of this study is to predict SOC, clay, sand, and silt content using reduced
7 spectral features as covariables selected by two spectral preprocessing. A total of 299 soil
8 samples were collected in Santa Catarina state, Brazil. Two preprocessing techniques, detrend
9 transformation and continuum removal (CR), were applied to isolate particular absorption
10 features in the reflectance spectrum. Two techniques were used to select the spectral features
11 in the spectrum: hand and mathematical selection. Partial least squares regression (PLSR) and
12 Support vector machines (SVM) were applied to predict the soil properties. The reduction of
13 predictor covariables by hand selection technique contributed in developing a high-level
14 prediction model for SOC. PLSR and SVM presented no statistical difference between the
15 RMSE results, except for clay content, where SVM presented superior performance. The
16 preprocessing techniques were statistically identical based on RMSE results. Overall, the
17 prediction of SOC, clay, sand and silt presented suitable results using reduced spectral features
18 as covariables in modeling process.

19 **Keywords:** Visible-near infrared spectroscopy, continuum removal, detrend, band ratio.

21 2.1.INTRODUCTION

22
23 Soil is one of the most important components of environmental resources and it has an
24 enormous influence on agricultural productivity (Lal and Moldenhauer, 1987). Soil information
25 is necessary to make decisions concerning management practices, food security (Andrews et
26 al., 2004), and soil security (Koch et al., 2013; McBratney et al., 2014). Soil organic carbon
27 (SOC) and particle size modulate nutrient supply, water holding capacity, soil structure
28 aggregation, and erosion prevention. Moreover, SOC has a significant impact on the global
29 carbon cycle as well as climate change (Janzen, 2004), and is recognized as a key component
30 of well-functioning ecosystems (Stockmann et al., 2015).

¹ Article was submitted to **Soil and Tillage Research**.

31 To develop a faster and more accurate method for SOC and particle size analysis,
32 proximal sensing has been successfully applied to predict these parameters (Conforti et al.,
33 2015; Knox et al., 2015; Ramirez-Lopez et al., 2013). The visible-near infrared (Vis-NIR)
34 reflectance region (350–2500 nm) stands out for its applicability to measure and predict a wide
35 variety of properties of soil samples (Dalmolin et al., 2005; Viscarra Rossel et al., 2006). Vis-
36 NIR uses spectral reflectance to identify properties without any interaction with objects and has
37 the advantages of extensive soil sample volume analysis, non-intrusiveness, timeliness, and
38 affordability (Viscarra Rossel et al., 2006). In addition, soil sample preprocessing is fast,
39 without the use of environmentally harmful chemical reagents (McBratney et al., 2006; Viscarra
40 Rossel and Behrens, 2010). This new soil analysis approach can be considered an alternative to
41 improve the conventional methods of analysis carried out in the laboratory (Minasny and
42 McBratney, 2008).

43 In the new concept of digital soil morphometrics (Hartemink and Minasny, 2014), the
44 application of tools, such as proximal soil sensing and techniques for measuring and quantifying
45 soil attributes, help enhance pedological understanding. Consequently, spectral reflectance has
46 been applied in soil survey, mapping, and quantitative soil property characterization. Various
47 research teams have used preprocessing and regression analysis to predict various soil
48 properties, but no single preprocessing method stood out as the best performing one among
49 these studies (Araújo et al., 2014; Knox et al., 2015; Ramirez-Lopez et al., 2013; Stevens et al.,
50 2010; Terra et al., 2015; Vasques et al., 2008; Viscarra Rossel and Behrens, 2010). Despite
51 these advances, research gaps exist regarding new modeling techniques that have the potential
52 to improve the predictive capabilities using proximal sensing.

53 Relating spectral data to a specific soil property requires a mathematical model. This
54 task is not simple because many factors can influence soil spectroscopy. Soil spectra are
55 complex, and soil attributes interact in complex ways, masking correlations between specific
56 spectral reflectance signatures and a specific soil property. Furthermore, the process is
57 complicated because only overtones of the native chemical structures of soil constituents are
58 found in the Vis-NIR spectrum. According to Wight et al. (2016), impacts from specific soil
59 characteristics on NIR performance are not well understood. These authors created an
60 association of artificial soils based on primary soil characteristics, where a single optimized
61 NIR model's predictive capability was compared by each soil characteristic subset. They
62 concluded that the type of organic matter can affect NIR's predictive ability and, depending
63 upon the accuracy chosen, it may be possible to separate sample populations into categories
64 based on the nature of the organic substrate. In addition, Wight et al. (2016) suggested that

65 texture is the principal characteristic that interferes with the model's accuracy, and it affects the
66 spectral reflectance in the entire region of the Vis-NIR. According to Ben-Dor et al. (1997), soil
67 organic matter influences all of the Vis-NIR spectral region and customizes the shape and the
68 albedo of the spectral curve.

69 Recently, preprocessing techniques have been utilized to transform soil spectral data,
70 remove noise, accentuate features, and detect patterns, including smoothing, detrending,
71 derivatives, averaging, normalization, scatter correction, non-linear transformations, and
72 absorbance transformation. In Vasques et al. (2008), thirty pre-processing transformations were
73 compared to predict soil carbon, e.g., Savitzky–Golay smoothing, averaging, normalization by
74 the range, Norris Gap Derivative, Savitzky–Golay derivatives, and standard normal variate. To
75 select spectral features of interest and make the spectra suitable for modeling by reducing the
76 spectral covariates, detrend, continuum removal (CR) and band ratio (BR) preprocessing
77 techniques were applied. These preprocessing can be used to interpret and extract information
78 from spectral reflectance sets and to identify spectral features related to specific soil properties.
79 Detrend is applied for removing baseline of the signals. CR, proposed by Clark and Roush
80 (1984), consists of removing the continuous features of the spectra and is often used to isolate
81 specific absorption features present in the spectrum to minimize the noise partially. The
82 continuum is represented by a mathematical function used to separate and highlight specific
83 absorption bands of the reflectance spectrum (Mutanga et al., 2005). BR is used to emphasize
84 how two wavelengths affect each other. This preprocessing has the advantage of combining
85 information from two prominent bands and it is an approach used to reduce the size of spectral
86 data.

87 For soil property predictions from Vis-NIR spectra, a mathematical analysis is required
88 to quantify each specific soil property. Generally, the most frequently used multivariate
89 methods are partial least square regression (PLSR) (Chacón Iznaga et al., 2014; Conforti et al.,
90 2015; Knox et al., 2015) and support vector machines (SVM) (Ramirez-Lopez et al., 2013;
91 Terra et al., 2015). One obstacle related to soil spectra and soil property characterization is the
92 complexity of soil components shown in the spectra (Ge et al., 2011; Wight et al., 2016). To
93 solve this problem, SVM and PLSR methods were applied in this study. SVM is a non-
94 parametric data mining method, and PLSR is the most common multivariate calibration model.
95 SVM and PLSR have already shown good results in soil properties predictions (Araújo et al.,
96 2014; Conforti et al., 2015; Knox et al., 2015; Kuang et al., 2015; Nawar et al., 2015; Stevens
97 et al., 2010; Terra et al., 2015). Viscarra Rossel and Behrens (2010) compared the predictions
98 using SVM and PLSR for SOC and clay content ($n = 1104$) using Vis-NIR spectroscopy. The

99 authors presented the best number of wavelet coefficients to use in the regressions, showing
100 that 72 coefficients produced the smallest RMSE when used to predict SOC, and 132
101 coefficients for clay content. The number of coefficients can be reduced based on BR and CR
102 preprocessing to maintaining the robustness of prediction accuracy.

103 The motivation to undertake this study comes from different sources. First, there is a
104 lack of studies applying spectral feature selection in order to reduce spectral covariables and
105 improve soil property prediction. Second, the selection of spectral features facilitates
106 understanding and reduces the multicollinearity of hyperspectral data. Third, there are few soil
107 spectroscopy studies in Brazil. The objective is to predict SOC, clay, sand, and silt content using
108 reduced spectral features as covariables selected by two spectral preprocessing.

109

110 2.2.MATERIAL AND METHODS

111

112 **2.2.1.Study site and sample collection**

113 Soil samples were collected in an area of about 1700 km² in the region within the
114 watershed of the Marombas River in the central region of Santa Catarina state, Brazil. A total
115 of 299 soil samples were collected following the GlobalSoilMap (Arrouays et al., 2014) depths
116 specifications of 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm along with additional
117 samples from profile horizons. The study area presented similar soils due to the homogeneity
118 of the parent material, which were predominantly basalt rocks from a landscape dominated by
119 a smooth relief plateau and few areas with sedimentary rock. According to the Köppen climate
120 classification, the study area has a humid subtropical climate (Cfa). These factors have led to
121 an advanced degree of weathering and the development of deep soils, such as Oxisols, which
122 were predominant in the area and showed high concentrations of iron oxides. Low clay content
123 values were measured in sandy soils, which were often characterized by intense water erosion
124 and low SOC content caused by unsustainable agricultural practices. Moreover, soil samples
125 with very low sand content were mostly associated with Oxisols. Furthermore, in some slope
126 areas, it is possible to find shallow soils, such as Entisols and Inceptisols. The prominent land
127 uses in this region were forest, grassland, and agriculture.

128

129 **2.2.2.Soil analysis in the laboratory**

130 The soil samples were sieved (2 mm) and dried at 45 °C (for 72 h) adopting the standard
131 Brazilian soil analysis method (Donagemma et al., 2011). The soil particle size was determined
132 according to the Pipette method using NaOH dispersant (Donagemma et al., 2011). The SOC

133 was determined by total organic carbon content using the Mebius method in the digestion block
134 (Yeomans and Bremner, 1988). Using this method, the soil organic matter is oxidized with a
135 mixture of $\text{K}_2\text{Cr}_2\text{O}_7$ 0.167 mol L^{-1} and concentrated H_2SO_4 , and the excess of dichromate is
136 titrated with ferrous ammonium sulfate. The reduced dichromate during the reaction with the
137 soil corresponds to organic carbon in the sample.

138

139 **2.2.3.Spectral reflectance measurements**

140 The spectral reflectance of soil samples was obtained using a FieldSpec 3
141 spectroradiometer (Analytical Spectral Devices, Boulder, USA) with a spectral range of 350–
142 2500 nm and a spectral resolution of 1 nm. To carry out the spectral measurements, soil samples
143 were distributed homogeneously in petri dishes. The spectral sensor that was used captured the
144 light through a fiber optic cable allocated 8 cm from the sample surface. The sensor reading
145 area was approximately 2 cm^2 and the lighting was provided by two external halogen lamps of
146 50 W. The lamps were positioned at a distance of 35 cm from the sample (non-collimated rays
147 and zenithal angle of 30°) and between them at an angle of 90° . A Spectralon standard white
148 plate was scanned every 20 min for calibration. For each sample, two replications (one
149 involving a 180° turn of the petri dish) were obtained. Each spectrum was averaged from 100
150 readings over 10 s. Mean values of two replicates were adopted for each subsample.

151

152 **2.2.4.Spectral preprocessing**

153 Soil spectral data were smoothed by the Savitzky–Golay first-order polynomial across
154 a moving window of five bands (Savitzky and Golay, 1964) to reduce the noise. The first order
155 detrending transformation was used to remove the baseline of the signals in the spectral data
156 (Barnes et al., 1989) and isolate particular absorption features. The detrend function, which is
157 recommended only when the overall signal is dominated by backgrounds that are generally of
158 the same shape, is recommended to be utilized prior to the multivariate analysis (Barnes et al.,
159 1989). The CR was used to isolate particular absorption features in the reflectance spectra
160 (Clark and Roush, 1984). CR allowed the normalization of the spectra and thereby facilitated
161 the identification of significant absorption features that ranged across the Vis-NIR spectrum.
162 The CR of the particular absorption feature was calculated by subtracting the band depth (BD)
163 value at a particular wavelength (λ) from 1 (i.e. $\text{CR} = \text{BD}(\lambda) - 1$). Detrend and CR were
164 performed in R programming language (R Core Team, 2016) by prospectr package. BR was
165 determined by the differences between a pair of spectral bands (e.g., first spectral band divided

166 by second, second spectral band divided by third and so on). BR was applied after spectral
167 features selection by detrend (Det+BR) and CR preprocessing (CR+BR).

168 The selection of spectral bands or spectral peaks were achieved by two techniques: hand
169 selection (by observing the shapes, peaks, valleys of the preprocessed spectra with pedological
170 knowledge) and mathematical selection (automated; computerized selection in R). The criterion
171 used to define the spectral features by hand selection came from the need to consider the entire
172 region of the spectrum and to associate the specific spectral bands with the soil characteristics.
173 The hand selection technique elected spectral bands associating the iron oxide features at 412,
174 448, and 476 nm; the water, hydroxyl, and clay mineral absorption at 1400, 1900, and 2200 nm,
175 respectively; additional bands associated with organic matter around 750, 1650, 2200, 2400,
176 2350 nm were also considered.

177 The mathematical selection method looked for peaks in spectrometry data. A peak is a
178 local maximum above a user defined noise threshold. The mathematical selection estimated and
179 removed the baseline of spectrum by applying the ‘SNIP’ method. This baseline estimation is
180 based on the ‘Statistics-sensitive Non-linear Iterative Peak clipping’ algorithm (SNIP)
181 described in Ryan et al. (1988). This technique was applied by detectPeaks function in
182 MALDIquant R package. The whole spectra (entire region of Vis-NIR) of detrend and CR
183 preprocessing were used as control treatment in the modeling process.

184

185 **2.2.5. Statistical analysis**

186 Descriptive statistics (fBasics R package) were calculated to summarize the data set,
187 and the coefficient of variation (CV) provided the variation of the data. The descriptive statistic
188 was performed in the R programming language. The Levene's test (Levene, 1960) (car R
189 package) was used to verify the assumption that variances are equal across training and
190 validations groups with significance level of 5%. The independent t-test was used to determine
191 whether a statistically significant difference exists between the means in the two unrelated
192 groups (training and validation sets). The SVM regression analysis (e1071 R package) applied
193 is a non-parametric statistical data mining method that belongs to the statistical learning theory
194 (Ivanciuc, 2007). In SVM regression analysis, a training model of a sample set (training set) is
195 performed. The procedure is to find a functional model that predicts correctly new cases that
196 are not yet presented with SVM previously. SVM is a group of supervised learning methods
197 that can be applied to classification or regression analysis, with several applications in many
198 scientific areas (Ivanciuc, 2007). PLSR (pls R package) is a method that models linear
199 relationships and is one of the most widely applied methods to predict soil properties from

200 spectral data. PLSR is based on a projection of the predictor x and response y variables into a
201 set of latent variables and corresponding scores, minimizing the dimensionality of the data
202 while maximizing the covariance between x and y variables (Wold et al., 2001). To compare
203 the modeling performance of both spectral bands selection techniques, the Scott Knott test (5%)
204 was applied. The whole spectra were used as control treatment. RMSE values were considered
205 in order to verify the statistical difference of hand selection, mathematical selection techniques,
206 and whole spectra. Scott Knott test was also applied in order to verify the statistical difference
207 between preprocessing. Scott Knott test was carried out by ScottKnott R package.

208

209 **2.2.6. Model training and validation**

210 A total of 299 soil samples were randomly split into training set [$\sim 70\%$] ($n = 209$) and
211 validation set [$\sim 30\%$] ($n = 90$). The fit and accuracy assessment of the models used the
212 following validation parameters: Coefficient of determination (R^2), root mean square error of
213 prediction (RMSE).

214

215 **2.3. RESULTS AND DISCUSSION**

216

217 **2.3.1. Exploratory results**

218 Considering the training and validation set, only clay showed a negatively skewed
219 distribution, with means of 59.56% and 57.53%, respectively (Table 1). The minimum and
220 maximum described the variation in the soil data sets. Generally, higher SOC values appeared
221 in Inceptisols and lower values in Oxisols. In addition, the SOC decreased with increasing
222 depth. The combination of high altitude and low temperature frequently promotes accumulation
223 of carbon in these soils due to the low decomposition of organic matter. The clay content
224 showed the lowest CV, which denotes that the variation from the mean indicates low data
225 dispersion. SOC and silt content showed intermediate dispersion, and sand content exhibited
226 extreme data dispersion (i.e., a high CV). The results of the predictive models confirm the same
227 trend due to the CV in the descriptive statistics. The Levene's test indicated the homogeneity of
228 variance between training and validation sets for SOC (p -value = 0.357), along with clay (p -
229 value = 0.943), sand (p -value = 0.847), and silt (p -value = 0.452). Since p -values are much
230 higher than the significance level of 5%, the variances have no significant difference. This
231 similarity between sets indicates that the random split represents the study population.

232

233

234 **2.3.2. Predictive performance of PLSR and SVM**

235 The predictive statistics of all models for the soil properties are shown in Table 2. In
 236 this table, the models results are placed in ascending order of RMSE. SOC content showed high
 237 accuracy, indicating a strong linear relationship between the measured and predicted variables.
 238 The models of SOC prediction showed a R^2_{val} and RMSE_{val} ranging from 0.68 and 0.56% to
 239 0.90 and 0.32%, respectively. The greater predictive performance was achieved by PLSR with
 240 CR preprocessing using the whole spectra. Among the 20 SOC predictive models, 11 presented
 241 an R^2_{val} higher than 0.81. The statistical difference between the prediction results of PLSR and
 242 SVM are revealed in Table 3. The Scott Knott test (5%) presented the mean comparison test of
 243 RMSE values for both methods. This test showed that there is no statistical difference between
 244 the RMSE values of PLSR and SVM models for SOC prediction. The mean values of RMSE
 245 are practically identical: 0.45% and 0.44%, for PLSR and SVM, respectively. This result
 246 demonstrated that both multivariate methods are suitable for SOC prediction. On the other hand,
 247 there is no right number of spectral bands to estimate soil properties because each soil has a
 248 particular spectral reflectance signature and thereby distinct spectral bands will be selected in
 249 model building.

250 These results are comparable to studies in the literature. Stevens et al. (2010) applied
 251 SVM to predict SOC in Luxembourg using different soil types (clay, silty-clay, silt, sandy-
 252 loam, and sand), and their validation results were slightly higher ($R^2 = 0.84$), but with an
 253 identical RMSE (0.43%). In Australia, a study presented by Viscarra Rossel and Behrens
 254 (2010), the SVM produced the highest fitted model ($R^2 = 0.84$) and lowest error ($\text{RMSE} =$
 255 0.92%) for SOC estimation. The similar performance of the SVM model may be attributed to
 256 the similarity of the sample observations used in their study with a total of 302 (201 for training
 257 and 101 for validation). Chacón Iznaga et al. (2014) used SVM to predict organic matter within
 258 a field in the central region of Cuba and found high $R^2 = 0.92$ and $\text{RMSE} = 0.14\%$. The
 259 performance in the current study showed that SOC can be properly estimated by using
 260 supervised learning models. In Ramirez-Lopez et al. (2013), the SVM prediction results for
 261 modeling organic carbon using Vis-NIR spectra (not continuum removed reflectance) were
 262 moderate ($R^2_{\text{val}} = 0.54$) for a regional soil spectral library with a low $\text{RMSE}_{\text{val}} = 0.27\%$ when
 263 compared to the results in this study. Large datasets have typically larger variances; therefore,
 264 well-performing models are more difficult to develop. According to Guerrero et al. (2015),
 265 small, rather than large, spectral libraries for local scale SOC assessment provide accurate
 266 predictions for effective model performance. Steffens and Buddenbaum (2013) presented SVM

267 models that produced results for a concentration of SOC with $R^2 = 0.97$ and $RMSE = 1.13\%$ to
268 provide laboratory imaging spectroscopy of soil profiles from Munich, Germany.

269 The predictive performance of clay content presented a R^2_{val} and $RMSE_{val}$ ranging from
270 0.42 and 8.96% to 0.62 and 6.84%, respectively (Table 2). The best model was achieved by
271 SVM with detrend preprocessing using the whole spectra. Scott Knott test showed that there is
272 statistical difference between the RMSE values of PLSR and SVM models for clay prediction
273 (Table 3). Clay content was the only soil property where the performances of PLSR and SVM
274 presented statistical difference. SVM presented higher predictive performance for clay
275 compared to PLSR. RMSE mean value for SVM models was 7.68% and 8.58% for PLSR.
276 Higher performances to predict clay content using SVM were achieved by Viscarra Rossel and
277 Behrens (2010) ($R^2 = 0.84$, $RMSE = 7.63\%$). This achievement was attributed to the
278 substantially larger soil sample sets located in different regions in Australia, including a diverse
279 number of soil classes ($n = 1104$), which was three times larger compared to the present study.
280 In addition, Kovačević et al. (2010) achieved high-quality results by applying SVM to predict
281 clay content in eastern Serbia ($R^2 = 0.76$ and normalized root mean squared deviation = 0.11%),
282 although with a small data set ($n = 151$). Terra et al. (2015) used Vis-NIR reflectance and SVM
283 to predict various soil properties in the Midwest and Southeast regions of Brazil, such as particle
284 size, chemical properties that include macro and micronutrients, and iron oxides. The authors
285 achieved high-quality predictions for clay ($R^2_{val} = 0.86$, $RMSE_{val} = 95.34 \text{ g kg}^{-1}$) and sand
286 contents ($R^2_{val} = 0.89$, $RMSE_{val} = 22.16 \text{ g kg}^{-1}$). These high-quality results are associated with
287 the correlation among clay activity and other soil properties. According to Stevens et al. (2013),
288 variations in clay content induce large differences in the spectral shape with non-variation of
289 SOC content.

290 The lowest predictive performance was achieved for sand content. The inferior model
291 result showed a R^2_{val} of 0.13 and $RMSE_{val}$ of 6.97% while the superior showed a R^2_{val} of 0.33
292 and $RMSE_{val}$ of 6.00%, which can be considered a low result (Table 2). Scott Knott test showed
293 that there is no statistical difference between the RMSE values of PLSR and SVM models for
294 sand prediction (Table 3). The mean values of RMSE were 6.44% and 6.64%, for PLSR and
295 SVM, respectively. The R^2 had the lowest value among all four modeled soil properties. This
296 result may have occurred due to the soil classes being mostly composed of Oxisols, which has
297 a relatively low sand fraction (Table 1). Kovačević et al. (2010) applied the SVM to estimate
298 soil properties in eastern Serbia and the performance for sand content was greater compared to

299 this study ($R^2 = 0.59$), with a normalized root mean squared deviation of 0.14%. The high CV
300 value for sand may also explain the relatively large uncertainty in the prediction of sand content.

301 The predictive performance of silt content was considered moderate with a R^2_{val} and
302 RMSE_{val} ranging from 0.40 and 7.67% to 0.56 and 5.26%, respectively (Table 2). The higher
303 model was found applying PLSR with CR preprocessing using hand selected spectral bands. In
304 the Scott Knott test (Table 3) there is no statistical difference between the RMSE values of
305 PLSR and SVM models for silt prediction. The mean values of RMSE were 6.53% and 6.90%,
306 for SVM and PLSR, respectively. There are some caveats to silt content, which is not directly
307 measured by the pipette method, and occasionally, the silt value adds up the clay and sand error
308 measurement.

309 The distinctive parent material found in the area of study (sedimentary and basalt rocks)
310 may have affected the performance for clay, sand, and silt. At sites characterized as Oxisols
311 because of the increased iron oxide concentration, the depth of absorption from 390 to 550 nm
312 also increased (Ben-Dor, 2002; Summers et al., 2011). The influence of iron oxide on the
313 reflectance spectra in the visible spectral region may have masked or decreased the inference
314 of some soil properties, such as particle size content.

315 The SVM acceptance in soil properties estimation has increased in recent years (Araújo
316 et al., 2014; Ramirez-Lopez et al., 2013; Terra et al., 2015) and has generated more accurate
317 calibration results than PLSR in some studies (Thissen et al., 2004; Viscarra Rossel and
318 Behrens, 2010). In Nawar et al. (2015), the results for PLSR with different preprocessing
319 transformation showed a low R^2 between $0.33 \leq 0.52$ ($\text{RMSE } 0.42\% \geq 0.36\%$) for organic
320 matter. In this same study, for clay content the R^2 fluctuated between 0.14 to 0.82. On the other
321 hand, PLSR can also provide satisfactory results. Kuang et al. (2015), compared the
322 performance of PLSR prediction models for SOC and clay content and found that $R^2 \leq 0.81$
323 and $\text{RMSE} \geq 1.46\%$ for SOC and $R^2 \leq 0.81$ and $\text{RMSE} \geq 1.04\%$ for clay.

324 For quite a long time, the most widely used regression method applied to predict soil
325 properties from spectral data was PLSR. Wold et al. (2001) drew our attention to PLSR in
326 handling numerous and collinear variables and to investigate more compounded problems.
327 However, PLSR models are not designed for the complexity of chemical and biological
328 systems. They are also not often used to screen out latent variables that are not useful in
329 explaining the response. In Gomez et al. (2008), PLSR showed better performance when there
330 was no well-identified spectral feature for the property of interest (clay and calcium carbonate).

331

332

333 **2.3.3.Performance of spectral band selection techniques**

334 The two techniques of spectral band selection, represented by hand selection and
335 mathematical selection, were analyzed by its potential to reduce the covariables for modeling
336 procedure. In detrend preprocessing, 13 spectral bands were selected by hand selection and only
337 8 by mathematical selection (Fig. 1). On the other hand, in CR preprocessing, 11 spectral bands
338 were selected by hand selection and by mathematical selection (Fig. 2). For detrend
339 preprocessing the mathematical selection reduced 5 spectral bands and for CR preprocessing
340 the number of spectral bands selected were identical.

341 For SOC prediction, the results of RMSE values showed statistical difference between
342 spectral band selection techniques and whole spectra (Fig. 3). The models applying the whole
343 spectra achieved the best performance in SOC prediction with a mean RMSE value of 0.35%.
344 Among hand and mathematical, the first selection presented lower mean of RMSE value of
345 0.42%, and the second showed mean of RMSE value of 0.52%. The results of RMSE values
346 for clay content were statistically identical regardless the spectral band selection or whole
347 spectra used. The prediction models using whole spectra presented a mean RMSE value of
348 7.93% (Fig. 3). Hand selection showed best performance compared to mathematical selection
349 for clay prediction with a mean RMSE value of 7.99% and 8.36%, respectively. The results of
350 sand content showed that the RMSE value for whole spectra was statistically different from
351 hand and mathematical selection techniques. The mean RMSE values were 6.21%, 6.54% and
352 6.70% for whole spectra, hand and mathematical selection, respectively (Fig. 3). For silt
353 content, hand selection presented statistical difference in RMSE value from whole spectra and
354 mathematical selection technique. The silt content was the only soil property where the hand
355 selection presented the best RMSE results. The mean RMSE value of hand selection, whole
356 spectra and mathematical selection was 6.20%, 6.82% and 7.17%, respectively (Fig.3).

357 The models using all Vis-NIR spectral region (whole spectra) showed superior
358 performance for SOC, clay and silt content. The reduction of spectral bands revealed that the
359 predictive performances of all soil properties were greater for hand selection technique. The
360 mathematical spectral band selection, in which the bands were selected by automated approach
361 presented poor prediction for all soil properties. This is because mathematical selection does
362 not take into consideration the preeminent spectral features to predict soil properties.

363 Selecting the spectral bands by observing the shapes of the preprocessed spectra with
364 pedological knowledge led to better prediction results. The reduction of spectral bands in the
365 Vis-NIR spectrum by hand selection technique increased the predictive performance of models
366 by choosing spectral regions that are associated with specific soil characteristics.

367 Important spectral bands chosen by hand selection were located in the near infrared
368 region. Generally, the spectral features are linked with important spectral active soil
369 components, for example, mineralogy, texture, and iron content (Stevens et al., 2013).
370 Furthermore, the two spectral bands selection techniques shared several wavelengths,
371 particularly near 1400 nm and 1900 to 2400 nm, which confirmed that these wavelengths in the
372 near infrared spectral region provide valuable contribution for soil property estimations.

373 The spectral bands selection techniques reduced the spectral feature space from 2150
374 possible spectral bands to distill distinct spectral features before linking them to the soil
375 property of interest. The spectral features selected by hand selection technique had a
376 considerably higher estimation performance of SOC content compared to textural properties.
377 Many of the spectral bands were likewise selected in the present study in accordance with the
378 spectral bands indicative of specific soil constituents documented in the literature. The spectral
379 bands at 1414 nm and 1920 nm were related to the vibration activity of the hydroxyl group in
380 water molecules (Ben-Dor, 2002). These spectral bands may be indicative of the insufficient
381 air-drying in green houses. According to Ben-Dor (2002), the spectral regions of 1300–1450
382 nm, 1850–1950 nm, and 2200–2400 nm are linked to clay minerals. According to Chang et al.
383 (2001), these are the most predominant spectral bands to predict clay content. However, the
384 equivalent spectral bands selected for both clay and SOC could have under-fitted the model
385 estimation for clay since the SOC could have masked or diminished the clay content. Xie et al.
386 (2012) presented five wavelength ranges that had major contributions to predict organic matter
387 in the NIR region: 1386–1401 nm, 2133–2138 nm, 2175–2194 nm, 2229–2273 nm, and 2315–
388 2327 nm. In addition, soil organic matter also showed correlation bands in the visible region
389 (400–750 nm) (Stenberg et al., 2010), where a total of seven spectral features were selected.
390 Some auxiliary spectral features at near infrared were also selected by hand selection.

391 In hyperspectral data, reduction techniques have promulgated to filter out the most
392 important features. However, the least number of spectral bands have affected model
393 performance. This can be credited to great capacity of multivariate methods, such as PLSR and
394 SVM, in estimating attributes based on the spectral behavior. In the study of Üstün (2003),
395 SVM outperformed PLSR if there is no wavelength selection applied. For Üstün (2003), SVM
396 has some advantages in comparison with PLSR: *i*) it finds a general solution and thus avoids
397 overtraining; *ii*) it gives a solution which is sparse and; *iii*) it is able to model non-linear
398 relations. However, SVM also has a disadvantage such as high computation time in case of a
399 large data set, which leads to a time-consuming optimization.

400

401 2.3.4. Performance of preprocessing techniques

402 The RMSE result for each soil property revealed that there was no statistical difference
403 between the four preprocessing techniques applied (Fig. 4). However, CR preprocessing
404 yielded the lowest RMSE results for SOC, clay and silt content. CR was the most reliable
405 preprocessing method for estimating the soil properties, and overall, provided better estimations
406 than detrend preprocessing.

407 Recent worldwide publications are targeting the CR as a preprocessing technique to
408 estimate soil properties, especially for SOC. The content of SOC had a huge impact on CR
409 absorption feature since soils with high SOC indicate a decrease in albedo across the entire Vis-
410 NIR spectrum (Ben-Dor, 2002). Stenberg (2010) used CR to examine the effect of soil moisture
411 content on Vis-NIR spectra. The results revealed that the CR technique was effective in
412 distinguishing wet and dry soils. In addition, dry soils resulted in deeper absorption features
413 along with high amounts of clay. The CR approach presents the advantage of addressing
414 specific absorptions features as covariables derived from reflectance measurements.
415 Furthermore, preprocessing contributed in the reduction of multicollinearity; otherwise, the
416 variance of the coefficients may be very large and the model might apply unnecessary
417 information. The results in the present study confirmed that CR preprocessing contributed in
418 selecting the most significant bands to estimate the soil properties. Nawar et al. (2016) revealed
419 that for SOC and clay the best predictive results were found by applying continuum removal
420 preprocessing transformation ($R^2 = 0.85$, RMSE = 0.19% for SOC and $R^2 = 0.90$, 5.32% for
421 clay). The appropriate selection of explanatory variables (spectral bands) in the CR
422 preprocessing was essential to improve the modeling performance and reduce the complexity
423 of the models.

424

425 2.4. CONCLUSIONS

426

427 Overall, the prediction of SOC, clay, sand and silt presented suitable results using
428 reduced spectral features as covariables in modeling process. SOC presented a high-level
429 prediction model. The results for clay and silt content showed moderate performances, as
430 opposed to sand content, which showed inferior performance. The hand selection technique
431 showed superior performance in predicting soil properties due to the pedological knowledge,
432 which can associate the spectral features with specific soil characteristics. The predictive
433 performances of PLSR and SVM multivariate methods showed that there was no statistical
434 difference between the RMSE results, except for clay content, where SVM presented superior

435 performance. There was no statistical difference between preprocessing techniques in
436 predicting SOC, clay, sand, and silt. However, CR preprocessing presented the lowest RMSE
437 results compared to detrend, CR+BR and detrend+BR.

438 The main strength of spectral band selection techniques was their effectiveness in
439 reducing predictor covariables, which enhances interpretability and transparency of models.
440 Both techniques contributed by highlighting the bands and features produced by optically active
441 soil components. The selection of spectral bands from entire spectra region accomplished
442 reliable outcomes. This study confirmed the high potential of using spectral preprocessing
443 techniques to estimate soil properties and examining the metrological quality of soil properties
444 from Vis-NIR spectral data. The predictive model performances are influenced by the
445 multivariate method, spectral preprocessing, homogeneity of soil samples, and type of
446 estimated soil properties. The authors suggest that, selecting spectral features is an imminent
447 choice for developing prediction models in upcoming studies.

448 Further studies have to consider that there is no optimal or ‘best’ amount of spectral
449 bands to estimate soil properties because each soil has distinct spectral reflectance signatures.
450 These alternatives for spectral features selection accentuated soil features and detected patterns
451 of individual soil spectral data. Modeling strategies that differ in their capabilities to extract
452 pedological characteristics from the Vis-NIR spectra need to be carefully considered in future
453 studies.

454

455 **Acknowledgements**

456

457 The first author would like to thank the Coordination for the Improvement of Higher
458 Education Personnel (CAPES), the National Council for Scientific and Technological
459 Development (CNPq), and the Foundation for Funding in Research and Innovation of Santa
460 Catarina State (FAPESC, Project no. 2012000094) for providing scholarship and funding to
461 carry out this research. The authors would also like to thank the Geomatics Laboratory at the
462 Federal University of Santa Catarina for collecting the soil samples, the Laboratory of Pedology
463 at Federal University of Santa Maria for their support in the laboratory analysis, the
464 Geotechnology in Soil Science Group (GeoSS) at the University of São Paulo for their support
465 in the spectroscopy analysis.

466

467

468

469 **References**

470

471 Andrews, S.S., Karlen, D.L., Cambardella, C.A., 2004. The soil management assessment
472 framework: A quantitative soil quality evaluation method. *Soil Sci Soc Am J. Soil Sci.*
473 *Soc. Am. J.* 68, 1945–1962.

474 Araújo, S.R., Wetterlind, J., Demattê, J. a. M., Stenberg, B., 2014. Improving the prediction
475 performance of a large tropical vis-NIR spectroscopic soil library from Brazil by
476 clustering into smaller subsets or use of data mining calibration techniques. *Eur. J. Soil*
477 *Sci.* 65, 718–729. doi:10.1111/ejss.12165

478 Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.C.R., McBratney, A.B. (Eds.), 2014.
479 *GlobalSoilMap: Basis of the global spatial soil information system.* CRC
480 Press/Balkema.

481 Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and
482 De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* 43, 772–
483 777.

484 Ben-Dor, E., 2002. Quantitative remote sensing of soil properties, in: *Agronomy, B.-A.* in
485 (Ed.), . Academic Press, pp. 173–243.

486 Ben-Dor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible
487 near-infrared and short wave infrared region (400–2500 nm) during a controlled
488 decomposition process. *Remote Sens. Environ.* 61, 1–15. doi:10.1016/S0034-
489 4257(96)00120-4

490 Carter, M.R., Angers, D.A., Gregorich, E.G., Bolinder, M.A., 1997. Organic carbon and
491 nitrogen stocks and storage profiles in cool, humid soils of eastern Canada. *Can. J. Soil*
492 *Sci.* 77, 205–210. doi:10.4141/S96-111

493 Chacón Iznaga, A., Rodríguez Orozco, M., Aguila Alcantara, E., Carral Pairol, M., Díaz Sicilia,
494 Y.E., de Baerdemaeker, J., Saeys, W., 2014. Vis/NIR spectroscopic measurement of
495 selected soil fertility parameters of Cuban agricultural Cambisols. *Biosyst. Eng.* 125,
496 105–121. doi:10.1016/j.biosystemseng.2014.06.018

497 Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-Infrared Reflectance
498 Spectroscopy–Principal Components Regression Analyses of Soil Properties. *Soil Sci.*
499 *Soc. Am. J.* 65, 480–490. doi:10.2136/sssaj2001.652480x

500 Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: Quantitative analysis techniques for
501 remote sensing applications. *J. Geophys. Res. Solid Earth* 89, 6329–6340.
502 doi:10.1029/JB089iB07p06329

- 503 Conforti, M., Castrignanò, A., Robustelli, G., Scarciglia, F., Stelluti, M., Buttafuoco, G., 2015.
504 Laboratory-based Vis–NIR spectroscopy and partial least square regression with
505 spatially correlated errors for predicting spatial variation of soil organic matter content.
506 CATENA 124, 60–67. doi:10.1016/j.catena.2014.09.004
- 507 Dalmolin, R.S.D., Gonçalves, C.N., Klamt, E., Dick, D.P., 2005. Relationship between the soil
508 constituents and its spectral behavior. *Ciênc. Rural* 35, 481–489. doi:10.1590/S0103-
509 84782005000200042
- 510 Donagemma, G.K., Campos, D.V.B. de, Calderano, S.B., Teixeira, W.G., Viana, J.H.M., 2011.
511 Manual de Métodos de Análise de Solo, Edition 2 rev. 230.
- 512 Ge, Y., Thomasson, J.A., Sui, R., 2011. Remote sensing of soil properties in precision
513 agriculture: A review. *Front. Earth Sci.* 5, 229–238. doi:10.1007/s11707-011-0175-0
- 514 Gomez, C., Lagacherie, P., Coulouma, G., 2008. Continuum removal versus PLSR method for
515 clay and calcium carbonate content estimation from laboratory and airborne
516 hyperspectral measurements. *Geoderma* 148, 141–148.
517 doi:10.1016/j.geoderma.2008.09.016
- 518 Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-
519 Sinoga, J.D., Zornoza, R., Viscarra Rossel, R.A., 2015. Do we really need large spectral
520 libraries for local scale SOC assessment with NIR spectroscopy? *Soil Tillage Res.*
521 doi:10.1016/j.still.2015.07.008
- 522 Hartemink, A.E., Minasny, B., 2014. Towards digital soil morphometrics. *Geoderma* 230–231,
523 305–317. doi:10.1016/j.geoderma.2014.03.008
- 524 Ivanciuc, O., 2007. Applications of Support Vector Machines in Chemistry, in: Lipkowitz,
525 K.B., Cundari, T.R. (Eds.), *Reviews in Computational Chemistry*. John Wiley & Sons,
526 Inc., pp. 291–400.
- 527 Janzen, H.H., 2004. Carbon cycling in earth systems—a soil science perspective. *Agric.*
528 *Ecosyst. Environ.* 104, 399–417. doi:10.1016/j.agee.2004.01.040
- 529 Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B., Harris, W.G., 2015.
530 Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared
531 (MIR) spectroscopy. *Geoderma* 239–240, 229–239.
532 doi:10.1016/j.geoderma.2014.10.019
- 533 Koch, A., McBratney, A., Adams, M., Field, D., Hill, R., Crawford, J., Minasny, B., Lal, R.,
534 Abbott, L., O'Donnell, A., Angers, D., Baldock, J., Barbier, E., Binkley, D., Parton, W.,
535 Wall, D.H., Bird, M., Bouma, J., Chenu, C., Flora, C.B., Goulding, K., Grunwald, S.,
536 Hempel, J., Jastrow, J., Lehmann, J., Lorenz, K., Morgan, C.L., Rice, C.W., Whitehead,

- 537 D., Young, I., Zimmermann, M., 2013. Soil security: Solving the global soil crisis. *Glob.*
538 *Policy* 4, 434–441. doi:10.1111/1758-5899.12096
- 539 Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil
540 properties using support vector machines. *Geoderma* 154, 340–347.
541 doi:10.1016/j.geoderma.2009.11.005
- 542 Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and
543 partial least squares for on-line visible and near infrared spectroscopy measurement of
544 soil organic carbon, pH and clay content. *Soil Tillage Res.* 146, Part B, 243–252.
545 doi:10.1016/j.still.2014.11.002
- 546 Lal, R., Moldenhauer, W.C., 1987. Effects of soil erosion on crop productivity. *Crit. Rev. Plant*
547 *Sci.* 5, 303–367. doi:10.1080/07352688709382244
- 548 Levene, H., 1960. Robust tests for equality of variances. *Robust Tests Equal. Var.* 278–292.
- 549 McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. *Geoderma* 213,
550 203–213. doi:10.1016/j.geoderma.2013.08.013
- 551 McBratney, A.B., Minasny, B., Viscarra Rossel, R., 2006. Spectral soil analysis and inference
552 systems: A powerful combination for solving the soil data crisis. *Geoderma* 136, 272–
553 278. doi:10.1016/j.geoderma.2006.03.051
- 554 McDowell, M.L., Bruland, G.L., Deenik, J.L., Grunwald, S., Knox, N.M., 2012. Soil total
555 carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse
556 reflectance spectroscopy. *Geoderma* 189–190, 312–320.
557 doi:10.1016/j.geoderma.2012.06.009
- 558 Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties
559 from infrared reflectance spectroscopy. *Chemom. Intell. Lab. Syst.* 94, 72–79.
560 doi:10.1016/j.chemolab.2008.06.003
- 561 Mutanga, O.M.C., Skidmore, A.K., Kumar, L., Ferwerda, J., 2005. Estimating tropical pasture
562 quality at canopy level using band depth analysis with continuum removal in the visible
563 domain. *Int. J. Remote Sens.* 26, 1093–1108. doi:10.1080/01431160512331326738
- 564 Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., Mouazen, A.M., 2015. Estimating the soil clay
565 content and organic matter by means of different calibration methods of vis-NIR diffuse
566 reflectance spectroscopy. *Soil Tillage Res.* doi:10.1016/j.still.2015.07.021
- 567 R Core Team, 2016. *R: A Language and Environment for Statistical Computing.*
- 568 Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013.
569 The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra

- 570 of complex datasets. *Geoderma* 195–196, 268–279.
571 doi:10.1016/j.geoderma.2012.12.014
- 572 Ryan, C., Clayton, E., Griffin, W.L., Cousens, D.R., 1988. SNIP, a statistics-sensitive
573 background treatment for the quantitative analysis of PIXE spectra in geoscience
574 applications. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* 34,
575 396–402.
- 576 Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least
577 Squares Procedures. *Anal. Chem.* 36, 1627–1639. doi:10.1021/ac60214a047
- 578 Steffens, M., Buddenbaum, H., 2013. Laboratory imaging spectroscopy of a stagnic Luvisol
579 profile — High resolution soil characterisation, classification and mapping of elemental
580 concentrations. *Geoderma* 195–196, 122–132. doi:10.1016/j.geoderma.2012.11.011
- 581 Stenberg, B., 2010. Effects of soil sample pretreatments and standardised rewetting as
582 interacted with sand classes on Vis-NIR predictions of clay and soil organic carbon.
583 *Geoderma, Diffuse reflectance spectroscopy in soil science and land resource*
584 *assessment* 158, 15–22. doi:10.1016/j.geoderma.2010.04.008
- 585 Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter Five -
586 Visible and Near Infrared Spectroscopy in Soil Science, in: Sparks, D.L. (Ed.),
587 *Advances in Agronomy*. Academic Press, pp. 163–215.
- 588 Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of Soil
589 Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance
590 Spectroscopy. *PLoS ONE* 8, e66409. doi:10.1371/journal.pone.0066409
- 591 Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., van Wesemael, B.,
592 2010. Measuring soil organic carbon in croplands at regional scale using airborne
593 imaging spectroscopy. *Geoderma, Diffuse reflectance spectroscopy in soil science and*
594 *land resource assessment* 158, 32–45. doi:10.1016/j.geoderma.2009.11.032
- 595 Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L.,
596 Hong, S.Y., Rawlins, B.G., Field, D.J., 2015. Global soil organic carbon assessment.
597 *Glob. Food Secur.* 6, 9–16. doi:10.1016/j.gfs.2015.07.001
- 598 Summers, D., Lewis, M., Ostendorf, B., Chittleborough, D., 2011. Visible near-infrared
599 reflectance spectroscopy as a predictive indicator of soil properties. *Ecol. Indic., Spatial*
600 *information and indicators for sustainable management of natural resources* 11, 123–
601 131. doi:10.1016/j.ecolind.2009.05.001

- 602 Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2015. Spectral libraries for quantitative
603 analyses of tropical Brazilian soils: Comparing vis-NIR and mid-IR reflectance data.
604 *Geoderma* 255–256, 81–93. doi:10.1016/j.geoderma.2015.04.017
- 605 Thissen, U., Pepers, M., Üstün, B., Melssen, W.J., Buydens, L.M.C., 2004. Comparing support
606 vector machines to PLS for spectral regression applications. *Chemom. Intell. Lab. Syst.*
607 73, 169–179. doi:10.1016/j.chemolab.2004.01.002
- 608 Üstün, B., 2003. A Comparison of Support Vector Machines and Partial Least Squares
609 regression on spectral data.
- 610 Vasques, G.M., Grunwald, S., Sickman, J.O., 2008. Comparison of multivariate methods for
611 inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma* 146,
612 14–25. doi:10.1016/j.geoderma.2008.04.007
- 613 Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse
614 reflectance spectra. *Geoderma, Diffuse reflectance spectroscopy in soil science and land
615 resource assessment* 158, 46–54. doi:10.1016/j.geoderma.2009.12.025
- 616 Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006.
617 Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for
618 simultaneous assessment of various soil properties. *Geoderma* 131, 59–75.
619 doi:10.1016/j.geoderma.2005.03.007
- 620 Wight, J.P., Ashworth, A.J., Allen, F.L., 2016. Organic substrate, clay type, texture, and water
621 influence on NIR carbon measurements. *Geoderma* 261, 36–43.
622 doi:10.1016/j.geoderma.2015.06.021
- 623 Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics.
624 *Chemom. Intell. Lab. Syst., PLS Methods* 58, 109–130. doi:10.1016/S0169-
625 7439(01)00155-1
- 626 Xie, X.-L., Pan, X.-Z., Sun, B., 2012. Visible and Near-Infrared Diffuse Reflectance
627 Spectroscopy for Prediction of Soil Properties near a Copper Smelter. *Pedosphere* 22,
628 351–366. doi:10.1016/S1002-0160(12)60022-8
- 629 Yeomans, J.C., Bremner, J.M., 1988. A rapid and precise method for routine determination of
630 organic carbon in soil. *Commun. Soil Sci. Plant Anal.* 19, 1467–1476.
631 doi:10.1080/00103628809368027
- 632

633 **Table 1**

634 Descriptive statistics of soil properties for the training and validation.

	Training set (%)				Validation set (%)			
	SOC	Clay	Sand	Silt	SOC	Clay	Sand	Silt
Observations	209	209	209	209	90	90	90	90
Minimum	0.17	20.94	1.00	16.54	0.38	25.41	1.56	18.39
Maximum	4.83	78.48	35.48	77.99	4.21	75.85	32.15	72.94
1st quartile	1.06	53.63	2.98	26.61	1.35	51.35	3.58	29.71
3rd quartile	2.46	68.28	9.95	38.06	2.66	66.22	8.43	39.74
Mean	1.84	59.56	7.51	32.94	2.04	57.53	7.70	34.77
Median	1.68	59.57	4.80	31.04	2.20	57.89	5.37	34.47
St. error of mean	0.07	0.77	0.46	0.67	0.10	1.17	0.76	0.92
Skewness	0.44	-0.65	1.80	1.47	0.00	-0.39	2.02	0.84
Kurtosis	-0.39	0.44	3.11	3.96	-0.74	-0.35	3.50	2.61
CV (%)	55	19	89	30	46	19	93	25

635

636 **Table 2**

637 Predictive performance of soil properties for the validation set.

Soil		Technique of spectral			
Property	Method	Preprocessing	band selection	R^2_{val}	RMSE _{val} (%)*
SOC	PLSR	CR	W.S.	0.90	0.32
	SVM	CR	H.S.	0.87	0.35
	PLSR	Det	W.S.	0.86	0.36
	SVM	Det	W.S.	0.86	0.36
	SVM	CR	W.S.	0.86	0.36
	PLSR	CR	H.S.	0.83	0.41
	SVM	CR+BR	H.S.	0.81	0.42
	PLSR	Det+BR	H.S.	0.81	0.42
	SVM	Det	H.S.	0.81	0.42
	PLSR	Det	H.S.	0.81	0.42
	SVM	Det+BR	H.S.	0.81	0.43
	PLSR	CR+BR	H.S.	0.79	0.46
	SVM	Det+BR	M.	0.76	0.48
	SVM	Det	M.	0.73	0.50
	PLSR	Det+BR	M.	0.72	0.50
	PLSR	Det	M.	0.72	0.51
	SVM	CR+BR	M.	0.69	0.53
	PLSR	CR+BR	M.	0.69	0.54
PLSR	CR	M.	0.68	0.54	
SVM	CR	M.	0.68	0.56	
Clay	SVM	Det	W.S.	0.62	6.84
	SVM	CR	W.S.	0.58	7.18
	SVM	CR	H.S.	0.56	7.21
	SVM	CR+BR	H.S.	0.56	7.30
	PLSR	CR	H.S.	0.52	7.46
	SVM	CR	M.	0.52	7.70
	SVM	CR+BR	M.	0.52	8.03
	SVM	Det	H.S.	0.47	8.04
	SVM	Det+BR	H.S.	0.48	8.08

Clay	SVM	Det	M.	0.47	8.08
	SVM	Det+BR	M.	0.44	8.31
	PLSR	CR+BR	H.S.	0.45	8.33
	PLSR	CR	M.	0.42	8.45
	PLSR	CR+BR	M.	0.42	8.47
	PLSR	Det+BR	H.S.	0.40	8.72
	PLSR	Det	W.S.	0.41	8.74
	PLSR	Det	H.S.	0.40	8.75
	PLSR	Det	M.	0.35	8.93
	PLSR	Det+BR	M.	0.35	8.94
	PLSR	CR	W.S.	0.42	8.96
	Sand	PLSR	CR	W.S.	0.33
PLSR		Det	W.S.	0.26	6.15
SVM		CR+BR	H.S.	0.25	6.26
SVM		CR	W.S.	0.25	6.28
PLSR		Det+BR	H.S.	0.22	6.36
SVM		Det	W.S.	0.25	6.41
PLSR		Det	H.S.	0.19	6.45
PLSR		CR+BR	H.S.	0.18	6.46
PLSR		CR	H.S.	0.17	6.50
SVM		CR+BR	M.	0.20	6.52
PLSR		Det+BR	M.	0.16	6.57
PLSR		CR+BR	M.	0.14	6.62
PLSR		CR	M.	0.13	6.66
PLSR		Det	M.	0.13	6.67
SVM		CR	H.S.	0.17	6.68
SVM		CR	M.	0.14	6.70
SVM		Det	H.S.	0.16	6.79
SVM		Det+BR	H.S.	0.16	6.81
SVM		Det	M.	0.13	6.93
SVM		Det+BR	M.	0.13	6.97

	PLSR	CR	H.S.	0.56	5.26
	SVM	CR	H.S.	0.57	5.35
	SVM	CR+BR	H.S.	0.54	6.06
	SVM	Det	W.S.	0.50	6.17
	SVM	CR	W.S.	0.50	6.20
	PLSR	Det+BR	H.S.	0.46	6.51
	SVM	Det+BR	H.S.	0.45	6.54
	SVM	Det	H.S.	0.44	6.54
	PLSR	CR+BR	H.S.	0.44	6.67
Silt	PLSR	Det	H.S.	0.44	6.71
	SVM	CR+BR	M.	0.40	6.82
	SVM	CR	M.	0.39	6.92
	PLSR	CR	M.	0.34	7.05
	PLSR	CR+BR	M.	0.32	7.16
	PLSR	Det	W.S.	0.41	7.23
	PLSR	Det+BR	M.	0.31	7.23
	SVM	Det+BR	M.	0.32	7.28
	SVM	Det	M.	0.31	7.41
	PLSR	Det	M.	0.28	7.46
	PLSR	CR	W.S.	0.40	7.67

638 *Sorted by ascending order of RMSE. M: mathematical selection, H.S.: hand selection, W.S:
639 whole spectra, CR: continuum removal, Det: detrend, BR: band ratio, PLSR: Partial least square
640 regression, SVM: Support vector machine.

641

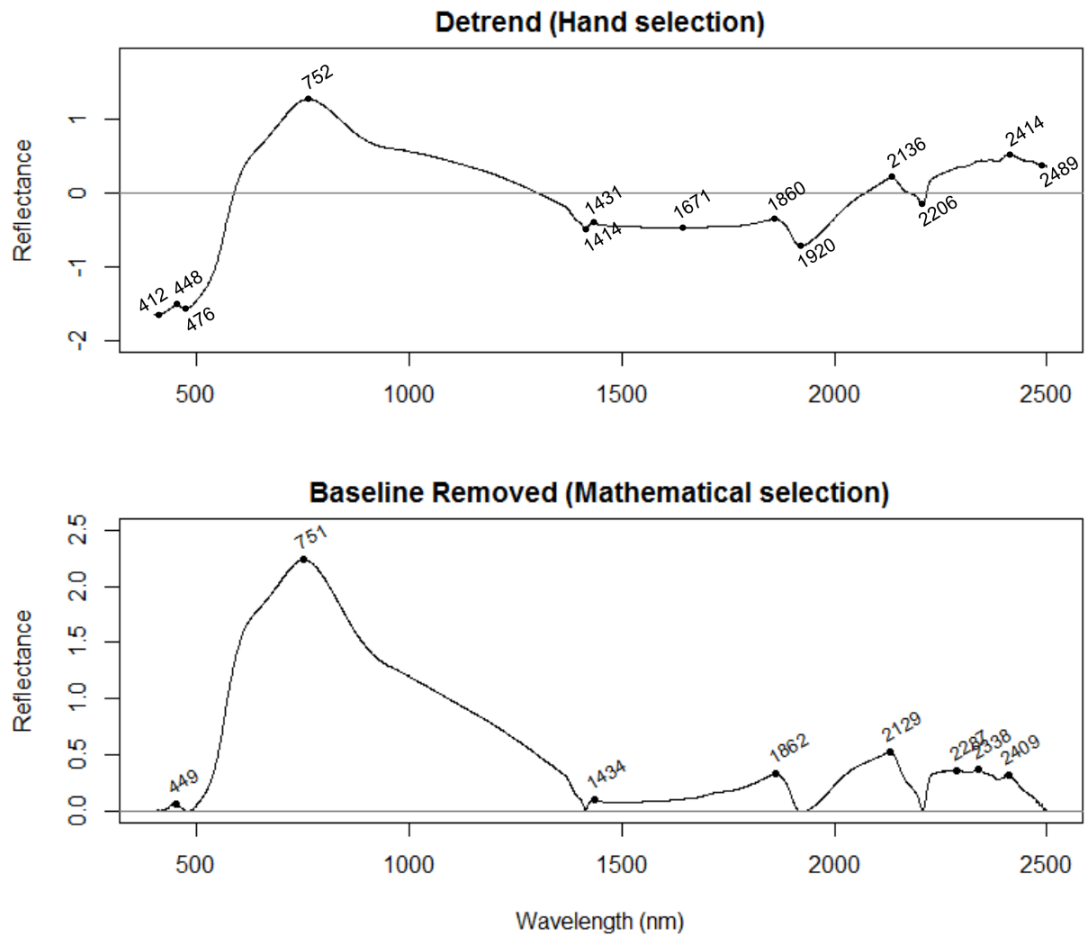
642 **Table 3**

643 Statistical difference between the prediction results of PLSR and SVM methods for each soil
 644 property.

	Method	Mean of RMSE _{val} (%)	Scott Knott test (5%)
SOC	SVM*	0.44	a
	PLSR*	0.45	a
Clay	SVM	7.68	a
	PLSR	8.58	b
Sand	PLSR	6.44	a
	SVM	6.64	a
Silt	SVM	6.53	a
	PLSR	6.90	a

645 *PLSR: Partial Least Square Regression, SVM: Support Vector Machine.

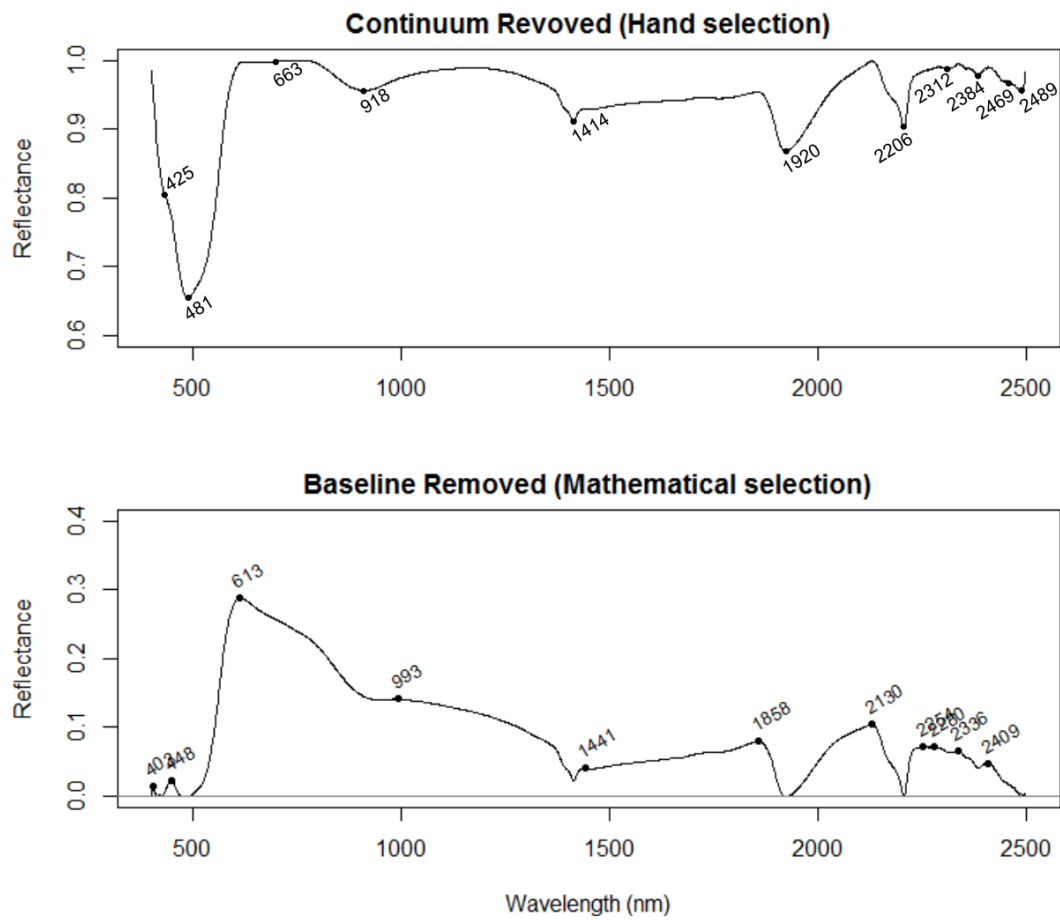
646



647

648 **Fig. 1.** Spectral curves of detrending transformation and its baseline removed in the visible-
 649 near infrared spectrum (average of 299 soil samples). Hand selection has 13 spectral bands and
 650 mathematical selection has 8 spectral bands.

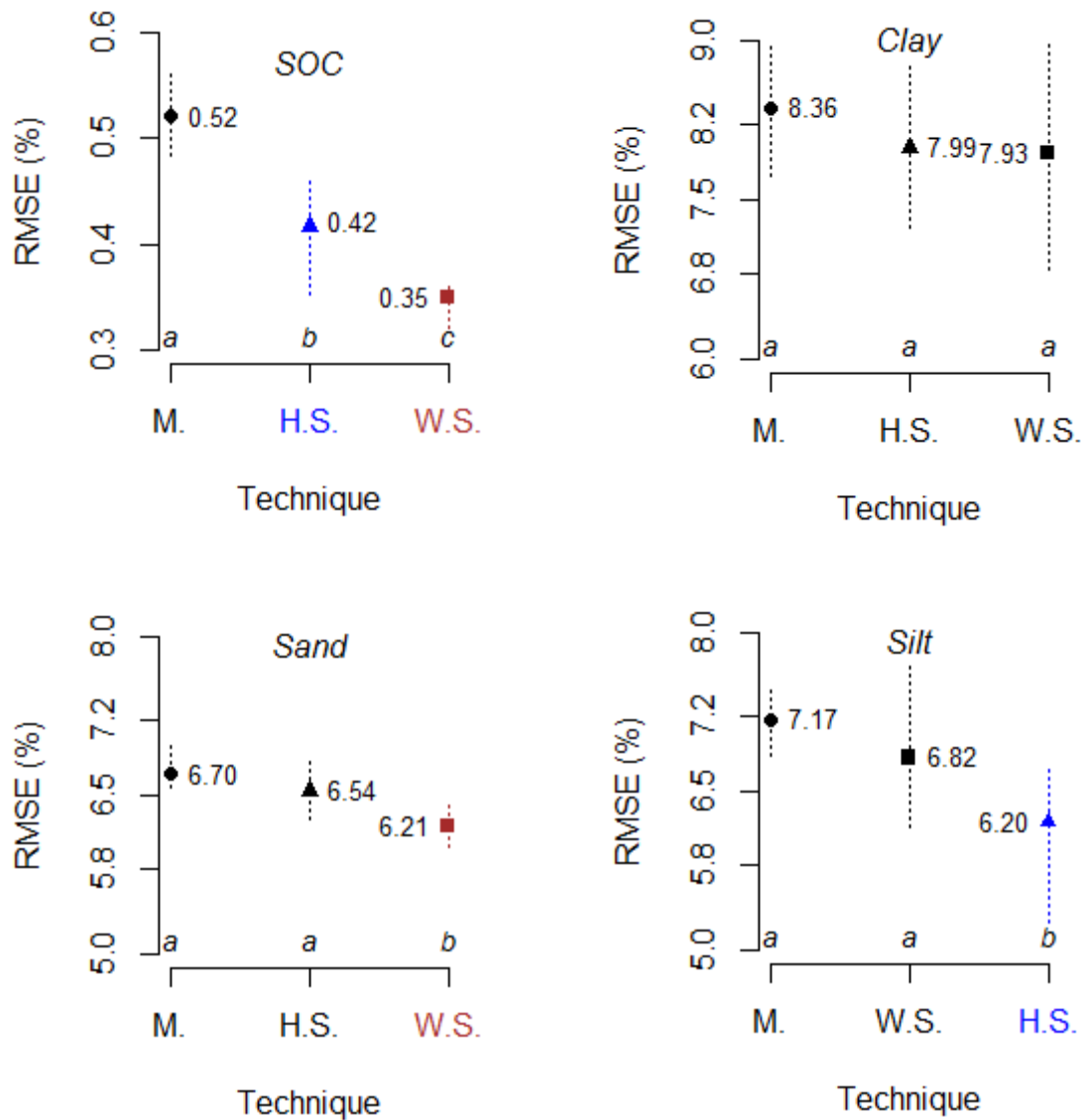
651



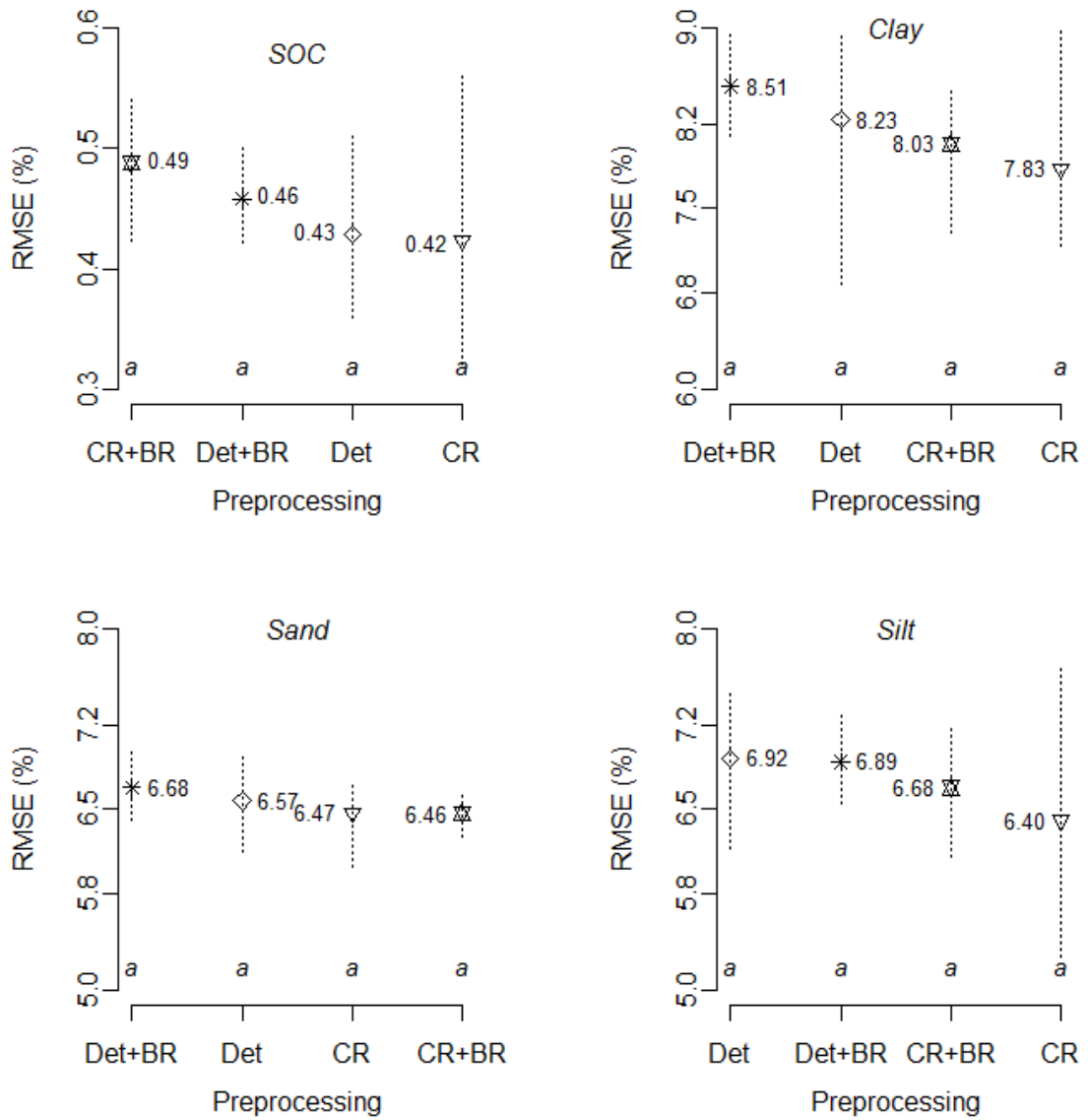
652

653 **Fig. 2.** Spectral curves of continuum removed preprocessing and its baseline removed in the
 654 visible-near infrared spectrum (average of 299 soil samples). Hand selection has 11 spectral
 655 bands and mathematical selection has 11 spectral bands.

656



657 **Fig. 3.** Statistical difference between spectral band selection techniques. In the graphics are the
 658 mean, maximum and minimum values of RMSE. Letters represent the results of Scott Knott
 659 test (significance level of 5%). M: mathematical selection, H.S.: hand selected, W.S: whole
 660 spectra.
 661



662 **Fig. 4.** Statistical difference between preprocessing techniques. In the graphics are the mean,
 663 maximum and minimum values of RMSE. Letters represent the results of Scott Knott test
 664 (significance level of 5%). CR: continuum removal, Det: detrend, BR: band ratio.
 665

3 ARTICLE 2: Comparing the capability of preprocessing techniques and multivariate 2 methods to predict soil organic carbon using spectroscopic data ²

4 Abstract

5 Soil organic carbon (SOC) represents a crucial role as an ecosystems indicator and are
6 recognized as a source in the global carbon cycle. Its quantification requires a method a non-
7 intrusiveness, affordable, and less time-consuming. Visible and near infrared (Vis-NIR)
8 reflectance spectroscopy has demonstrated its applicability to predict SOC over the years. There
9 is a need to assess the predictive performance of SOC combining linear modeling,
10 nonparametric, data mining and learning algorithms approaches all in a single study with
11 several preprocessing as input data. The aims of study are: i) to evaluate the potential of Vis-
12 NIR spectroscopy to predict SOC, ii) to compare the predictive capability between the
13 preprocessing techniques, and iii) to assess the modeling performance of wide range of
14 multivariate methods. Soil sampling was conducted over an area of about 1,800 km² in central
15 region of Santa Catarina State, Brazil, where a total of 595 soil samples were collected. Based
16 on the SOC prediction performance of preprocessing techniques, they can be divided into two
17 categories: scatter-correction techniques and spectral derivatives. Models using scatter-
18 corrective preprocessing presented superior prediction compared from spectral derivatives
19 group. In scatter-correction group, continuum removal is the most suitable preprocessing to be
20 used for SOC prediction. In the modeling performance, excepting for RF, all of methods
21 presented robust prediction. The highest model accuracy for SOC prediction was found
22 applying WAPLS method and NBR preprocessing ($R^2 = 0.82$, RMSE = 0.48%, RPIQ = 3.18).
23 The systematic methodology applied in this study can improve reliability for SOC
24 determinations by examining how techniques of preprocessing and multivariate methods affect
25 spectral analyses.

26 **Keywords:** Spectroscopy technique, modeling, prediction, soil property.

28 3.1.INTRODUCTION

30 Soil organic carbon (SOC) represents a fundamental and crucial role as an ecosystems
31 indicator and it is a key component of Soil Quality concept (Andrews et al., 2004) and more

² Article was submitted to **Geoderma Regional**.

32 recently for Soil Security framework (McBratney et al., 2014). Additionally, SOC pools are
33 recognized as a source in the global carbon cycle (Lal, 2004). This soil property is one of the
34 most important constituents of the soil due to its capacity to affect plant growth as a font of
35 energy and nutrients. SOC is effective to make management decisions and to inspect the
36 changes in different land use. Due to the importance of SOC, the digital soil mapping (DSM)
37 approach has given considerable attention to this soil fraction (Grimm et al., 2008; Grunwald,
38 2009). DSM requires high accuracy, sample density and promptness of SOC measurement.

39 However, accurate estimations in a complex environment are not easy to make.
40 Quantification of SOC demands an alternative technique, which should be capable of dealing
41 with extensive volume analysis, non-intrusiveness, affordable, and less time-consuming
42 (Minasny and McBratney, 2008; Viscarra Rossel et al., 2006). Visible and near infrared (Vis-
43 NIR) reflectance spectroscopy has been applied frequently in soil analysis and has demonstrated
44 its applicability to predict SOC and a variety of other soil properties accurately over the last
45 years (Bellon-Maurel and McBratney, 2011; Viscarra Rossel et al., 2006).

46 To improve the efficiency of SOC prediction using Vis-NIR spectral data, several
47 spectral preprocessing techniques have been introduced. Spectral preprocessing techniques
48 have been used to transform soil spectra, remove noise, emphasize features, and extract useful
49 information for quantitative predictive models. Preprocessing of the spectra include smoothing,
50 normalization, scatter correction, continuum removal, and derivatives. The preprocessing
51 techniques can be divided into two groups, scatter-corrections and spectral derivatives (Rinnan
52 et al., 2009). Scatter-corrections group is represented by continuum removal, normalization by
53 range, standard normal variate, and multiplicative scatter correction. Spectral derivatives
54 preprocessing includes Savitzky-Golay and Norris-Williams derivatives. The performances of
55 both preprocessing groups in soil properties prediction are varied according to the studies. For
56 instance, Ben-Dor et al. (1997) applied first and second derivative to investigate the reflectance
57 spectra of organic matter regarding the possible changes occurred during a biological
58 decomposition process. The authors assumed the use of spectral derivation enhanced weak
59 spectral features and extracted hidden information. Vasques et al. (2008) compared thirty
60 preprocessing including Savitzky-Golay and Norris-Williams derivatives, Kubelka-Munk
61 transformation, reflectance to absorbance transformation, baseline offset, standardizations, and
62 normalizations. Overall, the authors found the results considering Savitzky-Golay derivatives
63 consistently improved the SOC prediction. Similar outcome was achieved in Peng et al. (2014),
64 exploring the effects of eight spectra preprocessing techniques in 298 heterogeneous soil
65 samples from different Provinces in China. Their results indicated that the selection and

66 distribution of the model variables were affected by different preprocessing and Savitzky–
67 Golay derivative obtained a better result in the model development. Stevens et al. (2010) applied
68 absorbance, first and second Norris–Williams derivatives, Savitzky–Golay smoothing and
69 derivatives, Whittaker smoothing, standard normal variate, detrending, and a combination of
70 the previous with the objective to map SOC. Muñoz and Kravchenko (2011) included Savitzky–
71 Golay derivatives, standard normal variate and mean centering preprocessing to predict SOC
72 using three sources of auxiliary information under low carbon contents from Alfisols located in
73 southeast Michigan. The authors concluded no improvements in calibration accuracy were
74 observed when using preprocessing transformations. Nawar et al. (2016) compared the
75 performance of three regression methods subjecting the spectra to seven preprocessing
76 techniques to assess organic matter and clay content in the salt-affected soils from northern
77 Sinai, Egypt, and the best predictions were obtained with continuum removed preprocessing.

78 Besides finding the best preprocessing, another choice facing researches is regarding the
79 proper multivariate modeling approach. In fact, a satisfactory preprocessing technique should
80 always be considered in relation to the forthcoming modeling stage. In order to develop a faster
81 and high-quality model, several multivariate methods for SOC prediction have been
82 successfully utilized. Partial least-squares regression (PLSR) (Wold et al., 1984) is, by far, the
83 most common multivariate calibration method. PLSR has been applied in the SOC prediction
84 by many studies. (Conforti et al., 2015; Knox et al., 2015; Kuang et al., 2015; Viscarra Rossel
85 and Behrens, 2010). Moreover, other methods have shown important results as in principal
86 components regression (PCR) (Kendall, 1957) and multiple linear regression (MLR). Non-
87 parametric data mining method, such as, support vector machine (SVM) (Cortes and Vapnik,
88 1995) and the ensemble learning method, random forest (RF) (Breiman, 2001) are recently
89 gaining ground as multivariate methods to predict SOC. Besides these methods, a new set of
90 machine learning algorithms are being introduced into pedometric approach. Bayesian model
91 averaging (BMA) (Raftery, 1995) is a probabilistic model that represents a set of random
92 variables and their conditional independencies, and has been applied in the study of Leon and
93 Gonzalez (2009) for SOC prediction. Ramirez-Lopez et al. (2013) and Gholizadeh et al. (2016)
94 suggested the weighted average partial least squares (WAPLS) (Shenk et al., 1998) as a
95 memory-based learning multivariate method to prediction SOC. WAPLS remind the human
96 cognitive process, remembering and memorizing previous situations, adapting them for solving
97 the problem by examining the probability. Another learning machine approach is Gaussian
98 process regression (GPR) (Williams and Barber, 1998), which operates the input data into a
99 high dimensional feature space defined by a kernel function. Artificial neural network (ANN)

100 (McCulloch and Pitts, 1943) is a learning algorithm that is inspired by the structure and
101 functional aspects of biological neural networks. ANN has prominent studies on soil properties
102 prediction and some on SOC prediction, for instance, in Kuang et al. (2015) and Were et al.
103 (2015). These data mining approaches have been underutilized and for this reason more efforts
104 should be taken to reveal the potential of these methods in soil applications.

105 Comparing the performances of preprocessing techniques and multivariate methods
106 become complicated and disorganized by the fact of the studies are spread and conducted in
107 dissimilar areas, with distinct soil samples, soil types, spectral range, spectral data acquisition,
108 and different measurement units. Few studies have explored simultaneously in the same
109 database many forms of preprocessing and modeling methods. There is a need to assess the
110 predictive performance of SOC combining linear modeling, nonparametric, data mining and
111 learning algorithms approaches all in a single study with several preprocessing as input data.

112 The combination of a wide variety of preprocessing and multivariate statistics will allow
113 a systematic methodology for SOC prediction, with the advance of comparing the predictive
114 performances in the same dataset. The aims of the present study are: i) to evaluate the potential
115 of Vis–NIR spectroscopy to predict SOC, ii) to compare the predictive capability between the
116 preprocessing techniques, and iii) to assess the modeling performance of wide range of
117 multivariate methods.

118

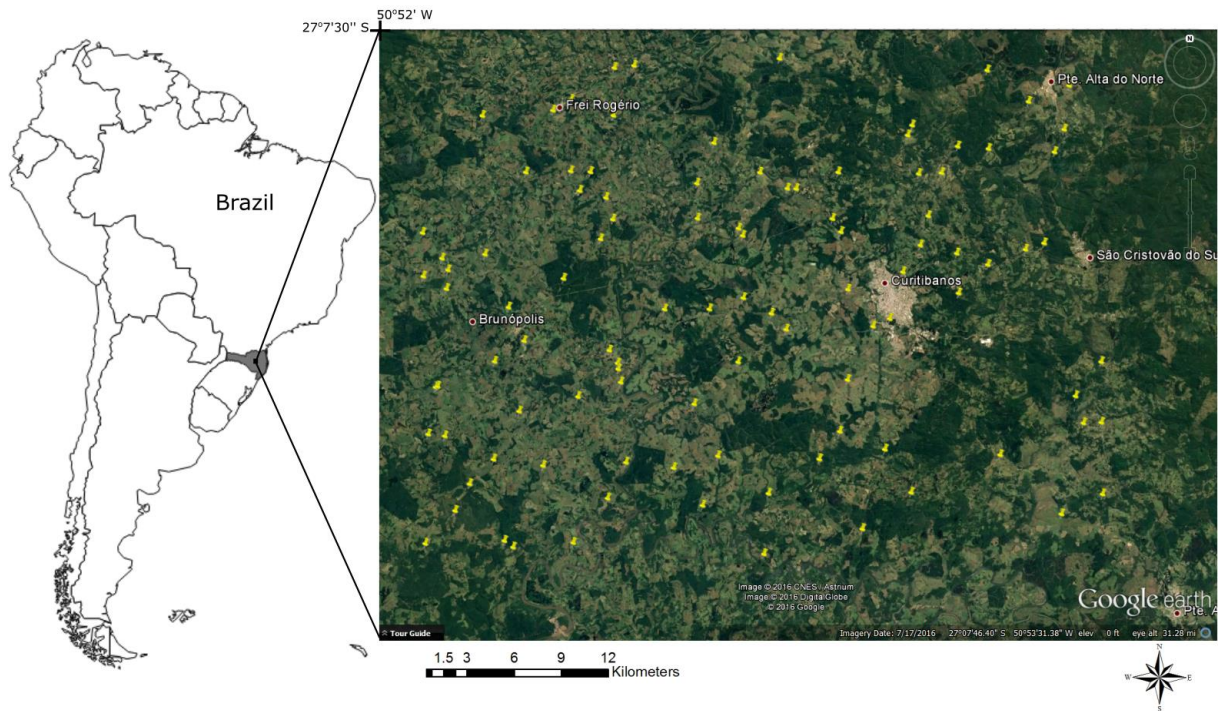
119 3.2.MATERIAL AND METHODS

120

121 3.2.1.Study area

122 Soil samples represented the prominent soil types extending over an area of about 1,800
123 km² in central region of Santa Catarina State, Brazil (Fig. 1). The study area presents similar
124 soils due to the homogeneity of parental material which is predominantly basalt from a
125 landscape dominated by a smooth relief plateau. According to the Köppen climate
126 classification, the study area has a humid subtropical climate (Cfa) with an elevation around
127 1,000 meters. The Oxisols are predominant in the area showing an advanced degree of
128 weathering and developing deep soils. Furthermore, in some steep areas, younger and shallower
129 soils, such as Entisols and Inceptisols, are found in a complex relief.

130



131

132 **Figure 1.** Soil sampling sites and municipalities located in central region of Santa Catarina
 133 State, Brazil.

134

135 3.2.2. Data collection and soil analysis

136 A total of 595 soil samples were collected, wherein 539 followed the depths
 137 specifications of 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm from Globalsoilmap.net
 138 (Arrouays et al., 2014) and 56 samples derived from 11 profiles. Soil samples were dried (at
 139 45°C for 72 hours) and then grounded and sieved (2 mm mesh). Total organic carbon content
 140 was determined by wet combustion method using the Mebius method in the digestion block
 141 (Yeomans and Bremner, 1988). Using this method, soil organic matter is oxidized with a
 142 mixture of $K_2Cr_2O_7$ 0.167 mol L⁻¹ and concentrated H_2SO_4 , and the excess of dichromate is
 143 titrated with ferrous ammonium sulfate. The reduced dichromate during reaction with soil
 144 corresponds to organic carbon in the sample.

145

146 3.2.3. Training and validation sets

147 Seventy percent of the dataset has been chosen by random sampling and used into
 148 training set (n = 417). The remainder thirty percent was used into validation set (n = 178). To
 149 not put in doubt the reliability of splitting sets, the homogeneity of these two sets were assessed
 150 by Levene's test. The Levene's test was applied to verify the assumption of variances were equal

151 across random selection of training and validations groups. Violin graphic showed a density
152 and descriptive statistics of SOC for training and validation sets.

153

154 **3.2.4.Spectral reflectance measurements**

155 Spectral reflectance of soil samples was obtained using a FildSpec 3 spectroradiometer
156 (Analytical Spectral Devices, Boulder, USA) with a spectral range of 350–2500 nm and a
157 spectral resolution of 1 nm. To carry out the spectral measurements, soil samples were
158 distributed homogeneously in petri dishes. The spectral sensor, which was used captured the
159 light through a fiber optic cable allocated 8 cm from the sample surface. The sensor scans an
160 area of approximately 2 cm² and light source was provided by two external halogen lamps of
161 50 W. Lamps were positioned with a distance of 35 cm from the sample (non–collimated rays
162 and zenithal angle of 30°) and between them an angle of 90°. A Spectralon® standard white
163 plate was scanned every 20 minutes for the calibration. For each sample, two replications (one
164 involving a 180° turn of the petri dish) were obtained. Each spectrum was averaged from 100
165 readings over 10 seconds. Mean values of two replicates were used for each sample.

166

167 **3.2.5.Spectral preprocessing techniques**

168 Spectral preprocessing techniques consist in a variety of mathematical procedures for
169 transforming the reflectance measurements before the usage in calibration models. The spectra
170 preprocessing has potential to remove physical variability due to light scattering and enhance
171 features of interest (Rinnan et al., 2009). The preprocessing techniques were selected following
172 the best results from Cambule et al. (2012), Knox et al. (2015), McDowell et al. (2012), Nawar
173 et al. (2016), Peng et al. (2014), Stevens et al. (2013), and Vasques et al. (2008). The
174 preprocessing includes smoothing, averaging, derivatives, normalizations, scatter corrections,
175 and absorbance transformations. Preprocessing techniques were applied to the soil reflectance
176 curves in the range of 350–2500 nm. Seven forms of spectra preprocessing were used to develop
177 models for SOC predicting. The first one was used as ‘control treatment’, where the raw
178 reflectances were only smoothed (SMO) across a moving window of 9 nm. SMO was
179 considered here as a preprocessing even if no transformation was implemented in spectral data.
180 Subsequent, the following preprocessing were applied into raw reflectance. Next six
181 preprocessing were Savitzky–Golay first derivative using a first order polynomial with a search
182 window of 9 nm (SGD), normalization by range (NBR), standard normal variate (SNV),
183 multiplicative scatter correction (MSC), continuum removed reflectance (CRR), and lastly,
184 transformation to absorbance and then application of Savitzky–Golay first derivative using a

185 first order polynomial with a search window of 5 nm (ASG). The SMO, CRR, SGD, ASG, and
 186 SNV preprocessing were carried out using *prospectr* package (Stevens and Ramirez-Lopez,
 187 2013). MSC and NBR were carried out using *pls* (Mevik et al., 2013) and *clusterSim* package
 188 (Walesiak and Dudek, 2016), respectively. Principal component analysis (PCA) (*stats* package,
 189 R Core Team, 2016) was used as a tool to explore the preprocessing and discover important
 190 characteristics of the spectral preprocessing. To compare the treatment means the Scott–Knott
 191 test (Scott and Knott, 1974) was applied. It is a hierarchical clustering algorithm used as an
 192 exploratory data analysis tool. Scott–Knott test was carried out by *ScottKnott* package
 193 (Jelihovschi et al., 2014).

194

195 **3.2.6. Multivariate methods**

196 In order to evaluate the predictive performance of the preprocessing, nine multivariate
 197 methods were implemented. Each type of method (e.g., PLSR, WAPLS) has specific and
 198 different required parameters that control how the relationship between input variables and
 199 outcomes is defined. These parameters were manually optimized to generate the best fit possible
 200 between variables and outcomes. All modeling were conducted using R programming language
 201 (R Core Team, 2016). Following are the multivariate methods and the corresponding R package
 202 applied: PLSR and PCR implemented in the *pls* package (Mevik et al., 2013), MLR in *stats*
 203 package (R Core Team, 2016), SVM in *e1071* package (Meyer, 2001), RF in *randomForest*
 204 package (Liaw and Wiener, 2002), BMA in *BMA* package (Raftery et al., 2015), WAPLS in
 205 *resemble* package (Ramirez-Lopez and Stevens, 2016), GPR in *kernelab* package (Karatzoglou
 206 et al., 2004), and ANN in *elmNN* package (Gosso, 2012). The seven spectral preprocessing
 207 were used as independent variable for each model developed.

208 In order to illustrate the total number of publications, considering the nine multivariate
 209 methods in the last ten years, a search was made into Scopus database selecting articles that
 210 have applied spectroscopy to predict soil properties. To support the selection of the best choice
 211 method to SOC prediction the time–consuming (in minutes) to generate each model was
 212 assessed. Running time for each model was calculated in R by *system.time* command and then
 213 the average for each method was considered. Personal computer with a 3.60 GHz Intel Core i7
 214 processor, 16 GB RAM, and Windows 10 operating system was used to run the models.

215 Three statistics measure were used in the multivariate methods to evaluate the fitted
 216 model: Coefficient of determination (R^2) (Eq. 1), root mean square error (RMSE) (Eq. 2), and
 217 ratio of performance to interquartile range (RPIQ) (Eq. 3). R^2 is the percent of variance
 218 explained by the model. R^2 measure is, by far, the most widely used and reported measure of

219 error and goodness of fit. RMSE is commonly used to measure the difference between predicted
 220 and observed values from the fitted model. It is easily interpreted statistic, since it has same
 221 data units. RPIQ is based on quartiles, which better represents the spread of the population.
 222 According to Bellon-Maurel et al. (2010), soil sample sets often show a skewed distribution,
 223 and not a normal distribution. For this reason, the RPIQ index explains the spread of the dataset
 224 better by using interquartile distance.

225

$$226 \quad R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$227 \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (2)$$

$$228 \quad \text{RPIQ} = \frac{(Q3 - Q1)}{\text{RMSE}} \quad (3)$$

229

230 where \hat{y} is the predicted values, \bar{y} is the mean of observed values, y is the observed values, n is
 231 the number of samples with i equal to 1, 2, ... n , IQ is the difference between the third and first
 232 quartiles ($Q3 - Q1$), $Q1$ is the value found in 25% of the samples, and $Q3$ is the value found in
 233 75% of the samples.

234

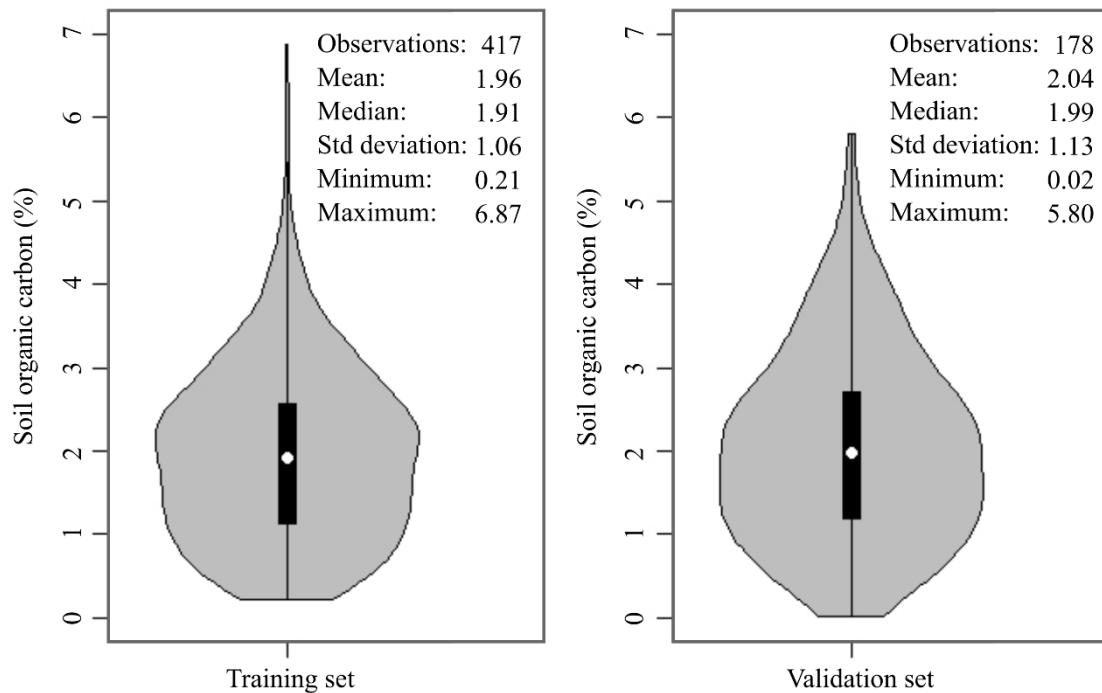
235 3.3.RESULTS AND DISCUSSION

236

237 3.3.1.Descriptive and inferential statistics

238 Considering the density of training and validation sets, more than 50% of total SOC
 239 values is placed among 1% to 3% (Fig. 2). The data presented a widespread variation with
 240 maximum and minimum SOC values of 0.02 and 6.87%, respectively. Model prediction is
 241 potentially influenced by the high variation of data. Standard deviation indicated this tendency.
 242 The large variation of SOC content was expected based on wide depths layers collected in this
 243 study ranging from 0–5 to 100–200 cm. The highest SOC values occurred in soils with forest
 244 at upper depth of 0–5 cm. These soils constantly receive replacement of organic material, which
 245 promotes the accumulation of carbon due to low decomposition of organic matter conditioned
 246 by high altitude and low temperature of the area. The lowest SOC values were found in the
 247 100–200 cm depths, where storage of carbon in soils is reduced. Levene's test achieved a p -
 248 value of 0.205 for the homogeneity of variances tests between training and validation datasets.

249 Since p-value is much higher than significance level of $\alpha = 0.05$, the hypothesis of equal
 250 variance is not rejected, and so there was no significant difference between variances. This
 251 similarity between the training and validation sets is revealing the randomly split groups are
 252 statistically similar and further multivariate analysis is suitable.
 253



254
 255 **Figure 2.** Density of training and validation sets. Dark area indicates inter-quartile range and
 256 white dot indicates median value of dataset. The p-value of Levene's test = 0.205 (significance
 257 level of $\alpha = 0.05$).

258

259 3.3.2.Characteristics of soil spectral reflectance curves

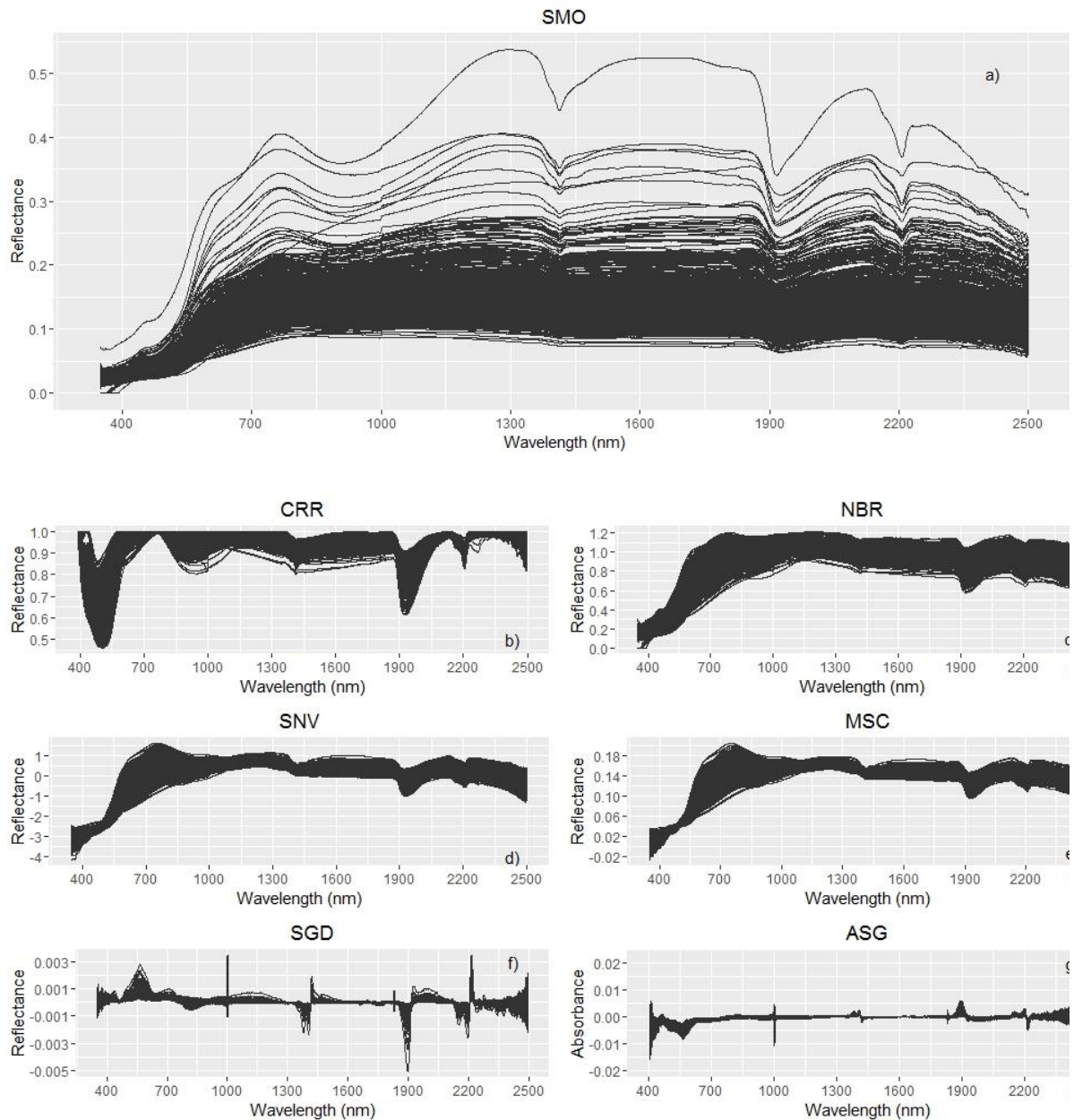
260 Diversity of soils is represented by spectral reflectance curve forms. The raw spectral
 261 reflectance (Fig. 3a) illustrated the curves by its shape and the presence or absence of absorption
 262 bands. Categorization of soil reflectance has important implications for soil genesis,
 263 classification, and survey (Stoner and Baumgardner, 1981). Assessment of spectral curves
 264 provides a tool for qualitative description of Vis-NIR soil reflectance. This descriptive soil
 265 information is important for initial characterization and discrimination.

266 The spectral reflectance curve of each soil sample is characterized by the variability of
 267 its soil properties. Soil samples showed the presence of distinguished soil reflectance curve
 268 forms that were associated with different shapes and absorption bands. This distinction is
 269 mainly due to organic matter content and iron oxides content in these soils. Observing Fig. 3a,

270 there is a predominance of very low reflectance. The majority of soils samples present high
271 content of iron oxides conducting to a low reflectance. According to Stoner and Baumgardner
272 (1981), characteristic shape between 450 and 850 nm indicated the presence of iron oxides
273 (mainly goethite and hematite), absorption at 1400 nm and 1900 nm was due to water molecule
274 vibrations and OH groups, and absorption at 2200 nm indicated the presence of kaolinite.

275 The large amount of soil samples exhibited a low overall reflectance (Fig. 3a). Based
276 on Stoner and Baumgardner (1981), these soils belong to a particular type of spectral curves
277 designated iron-dominated form with high iron content and fine texture. This trend was found
278 in the current study, where reflectance decreases in wavelength beyond 750 nm and absorption
279 in the middle infrared wavelengths is so strong that the water absorption bands are almost
280 undetectable. Few amount of soil samples presented the type of spectral curves called minimally
281 altered with low organic and medium iron content, according to Stoner and Baumgardner
282 (1981). These curves are characterized by overall high reflectance and a convex curve shape.
283 Moreover, strong water absorption bands at 1400 and 1900 nm are noticeable.

284 Characteristic soil spectral reflectance curves influence the subsequent model
285 prediction. Large number of spectral curves with low reflectance intensity can reproduce a more
286 efficient model performance for soils with high iron content and fine texture. On the other hand,
287 little amount of high reflectance soils, in which present low organic and medium iron content,
288 can lead to poor performances for this soil types.



289 **Figure 3.** Illustration of Vis–NIR spectral curves of preprocessing for all soil samples. a) SMO:
 290 smoothed across a moving window of 9 nm, b) CRR: continuum removed reflectance, c) NBR:
 291 normalization by range, d) SNV: standard normal variate, e) MSC: multiplicative scatter
 292 correction, f) SGD: Savitzky–Golay first derivative using a first order polynomial with a search
 293 window of 9 nm, g) ASG: transformation to absorbance and then application of Savitzky–Golay
 294 first derivative using a first order polynomial with a search window of 5 nm.

295

296

297

298

299 **3.3.3. Influence of preprocessing techniques in the performance of SOC models**

300

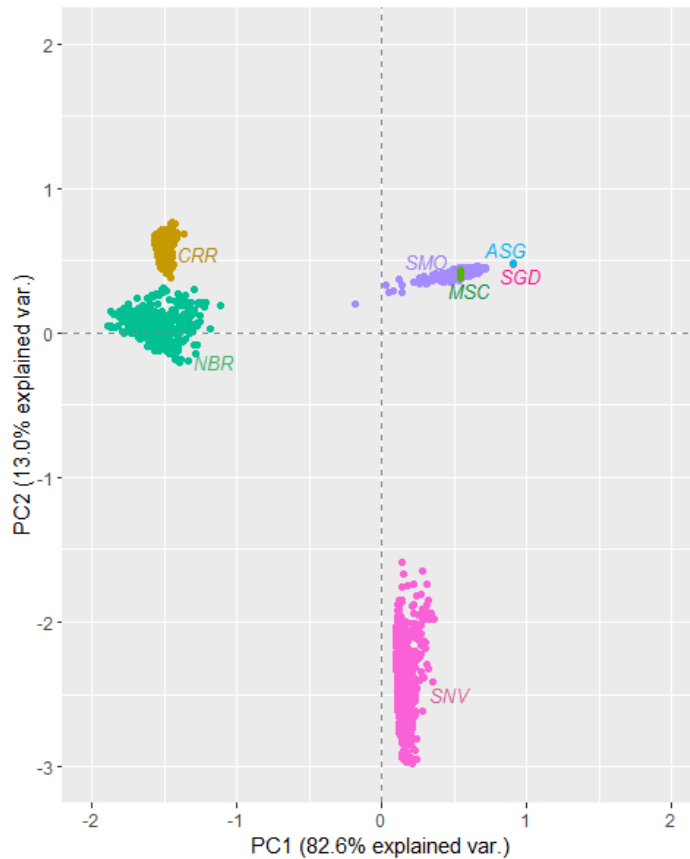
301 *3.3.3.1. Two groups of preprocessing techniques*

302 Spectroscopic measurements can be used to provide a quantitative estimate of most
303 abundant minerals present in soil. In order to enhance the spectral features and fitting the best
304 relationship with a soil property of interest preprocessing techniques were employed. Although,
305 a vast diversity in preprocessing techniques and variations between selected preprocessing can
306 be observed. In order to emphasize the variation and reveal strong patterns in the seven
307 preprocessing principal component analysis (PCA) was utilized. PCA is a technique often used
308 to explore and visualize correlated data. In Fig. 4, each color represents the seven spectral
309 preprocessing in a multidimensional space projected by first and second principal components
310 (PC1 and PC2, respectively). PCA captured the variation occurred in preprocessing. SGD and
311 ASG were grouped together while CRR, NBR and SNV were far-off the symmetric center. The
312 PC1 explain 82.6% of total variance and certain preprocessing are associated suggesting SGD,
313 ASG, MSC, SMO, and SNV are correlated. However, in PC1 is noticeable the SGD and ASG
314 are almost in same position, different from other preprocesses. NBR and CRR are grouped
315 together while SNV are separated in PC2. The finding supports that there are two different
316 preprocessing groups and modeling performance of SOC are affected by this grouping.

317 Preprocessing techniques are divided into two categories: scatter-correction techniques
318 and spectral derivatives. First group of scatter-corrective preprocessing techniques includes
319 CRR, MSC, SNV, and NBR. Spectral derivatives group is represented by SGD and ASG. The
320 performance of models obtained by methods using scatter-corrective preprocessing was
321 superior compared from spectral derivatives group. The scatter-correction preprocessing
322 techniques are designed to reduce physical variability (undesirable scatter effect) and to
323 compare individual features of each element from a common baseline (Rinnan et al., 2009).
324 This group represents powerful preprocessing techniques, which isolates and removes
325 complicated effects caused by physical phenomena, where soil chemical effects can be more
326 easily modeled.

327 Regarding the spectral derivatives group, these preprocessing have the ability to remove
328 both additive and multiplicative effects in spectra. First derivative, applied in the current study,
329 removes the baseline and is estimated by the difference between two subsequent spectral
330 measurement points (Rinnan et al., 2009). Two different preprocessing, SGD and ASG, were
331 used to reduce the signal-to-noise ratio in spectra using Savitzky-Golay derivation. Derivative
332 is calculated at the center of each point fitting a polynomial in a symmetric window on raw

333 spectra. This operation is applied to all points in spectra, sequentially. Estimation of derivatives
 334 operates by a moving-window, where only a local part of spectra is used at time to compute
 335 the derivative. That is one distinction from scatter-corrective preprocessing, which can be
 336 performed on entire window.
 337



338
 339 **Figure 4.** Principal component analysis of seven preprocessing techniques.

340

341 Scatter-correction preprocessing group has developed significant improvement over
 342 Vis-NIR spectral models. The performance assessment of scatter-correction preprocessing
 343 fluctuated within models. In validation, values R^2 fitted varied from 0.54 up to 0.82, while
 344 RMSE varied from 0.77% to 0.48% (Table 1). CRR was ranked the best preprocessing in three
 345 multivariate methods (PLSR, PCR, and RF) achieving the lowest RMSE values. SNV
 346 preprocessing produced highest performance for two methods (MLR and GPR), along with
 347 NBR (WAPLS and ANN). Following, MSC appeared ranked as best preprocessing for only
 348 one method (BMA).

349 The best performance was found for CCR regarding the performance of models using
 350 scatter-correction preprocessing to predict SOC. CRR technique proposed by Clark and Roush
 351 (1984), consists of removing the continuous features of spectra and is often used to isolate

352 specific absorption features present in spectrum. Continuum is represented by a mathematical
353 function used to separate and highlight specific absorption bands of reflectance spectrum
354 (Mutanga et al., 2005). The technique of making a continuum, or hull, is similar to fitting a
355 rubber band over the original spectrum. The spectrum is normalized by setting the value of the
356 hull to 100% reflection, where first and last values of continuum removed spectrum are equal
357 to 1. The strength of CRR is to enhance absorption depths by correcting apparent shifts caused
358 by wavelength dependent scattering.

359 Subsequent scatter-correction technique is SNV preprocessing. SNV achieved the
360 higher prediction for two methods, which were MLR and GPR. This preprocessing has been
361 proposed for removing the multiplicative interference of particle size by simple rotation and
362 offset correction of spectra (Barnes et al., 1989). As observed in Fig. 3d, the similarity between
363 SNV and MSC is obvious. Signal-correction concepts behind SNV are the same as for MSC
364 except, where a common reference signal is not required, which is observed in reflectance
365 values. SNV is designed to operate based on centering the underlying linear slope of each
366 individual sample spectrum (Barnes et al., 1989). Moreover, SNV can be noisy sensitive in
367 spectrum. Instead of using average and standard deviation as correction parameters, it considers
368 to use each observation on its own isolated from remainder dataset.

369 NBR preprocessing presented the best model result for WAPLS and ANN methods,
370 both machine learning algorithms. In NBR, normalization means adjusting values measured on
371 different scales to a common scale. Simple normalization of each sample is a common approach
372 to multiplicative scaling problem. NBR preprocessing refers to the creation of shifted and scaled
373 versions of spectral data, where these normalized values eliminate scattering effects (Rinnan et
374 al., 2009). If the relationship between variables is the most important aspect of spectral data,
375 then normalization is recommended.

376 The final scatter-corrections addressed is MSC preprocessing. MSC achieved the best
377 prediction result only for BMA method. Nonetheless, in BMA method four scatter-corrections
378 preprocessing presented a concentrated performance with a slight higher result for MSC. The
379 purpose of MSC is to eliminate scatter errors, in order to linearize spectral data and decrease
380 noise variance (Geladi et al., 1985). In MSC each spectrum is corrected so that all spectral
381 samples appear to have the same scatter level. It has been demonstrated that, MSC and SNV
382 spectra preprocessing are closely related and differences in prediction ability between these
383 methods seems to be quite small.

384 Spectral derivatives preprocessing achieved greatest performance only for SGD in SVM
385 method. ASG and SMO preprocessing never attended the best model performance in any

386 method. Besides that, SMO preprocessing figured in the lowest model performance for three
387 methods (SVM, RF, and GPR). Interesting finding occurred in the results of SVM and RF
388 modeling. In both methods, spectral derivatives preprocessing (SGD and ASG) reached the
389 highest performances. The results obtained by two spectral derivatives preprocessing
390 performances with SVM and RF are in accordance with Vasques et al. (2008). The authors
391 investigated several multivariate methods including two supervised machine learning
392 (committee trees and regression trees) to assess soil carbon in Florida, USA. Among thirty
393 spectral preprocessing tested, spectral derivatives preprocessing presented the highest
394 predictive performance for both SVM and RF. Both methods are classified as supervised
395 learning algorithms, which SVM is machine learning and RF is ensemble learning. In addition,
396 the two algorithms demonstrate efficiently modeling on large datasets, model accuracy is
397 maintained when there is missing data or outliers, in regression they do not predict beyond the
398 range of response values in training data, they underestimate the high values and overestimate
399 the low values, and they are theoretically difficult to analyze (Breiman, 2001; Ivanciuc, 2007;
400 Mountrakis et al., 2011; Viscarra Rossel and Behrens, 2010).

401 SMO preprocessing frequently generated low accuracy performance regardless of
402 method employed. SMO always figured in the bottom three lowest preprocessing (Table 1).
403 Nawar et al. (2016) obtained similar results where no preprocessing was used for organic matter
404 prediction. Earlier studies has shown calibration models, in which spectra were not
405 preprocessed, are more sensitive to changes compared to models for which preprocessing was
406 applied (Moros et al., 2009).

407

408 *3.3.3.2. Performance of best preprocessing technique*

409 CRR was considered the most robust spectral preprocessing based on predictive
410 performance for SOC. CRR presented the higher performance for PLSR, PCR, and RF methods
411 (Table 1). Considering all prediction methods, this preprocessing always appeared among the
412 top four best results. This result demonstrates CRR is suitable preprocessing for SOC prediction
413 with Vis–NIR spectral data. CRR has also been successfully used in some other studies, for
414 instance, to estimate soil color (Viscarra Rossel et al., 2009), clay content (Lagacherie et al.,
415 2008; Nawar et al., 2016; Viscarra Rossel et al., 2009), organic matter (Nawar et al., 2016; Xie
416 et al., 2012), soil organic carbon (Nocita et al., 2014), soil heavy metals (Gholizadeh et al.,
417 2015; Vašát et al., 2014; Xie et al., 2012), soil macro and micro nutrients (Vašát et al., 2014),
418 and soil nitrogen content (Zhang et al., 2016). Application of CRR preprocessing also can be
419 found to characterize world's soil in global spectral library (Viscarra Rossel et al., 2016), to

420 estimate tropical pasture quality (Mutanga et al., 2005), and even in another planet, as in
421 elemental concentration estimation on Mars (spectrometer installed at the robotic rover
422 Curiosity) (Wang et al., 2014). Nocita et al. (2014) applied CRR to predict SOC content by
423 diffuse reflectance spectroscopy from soil samples collected all over the European Union. The
424 authors conclude SOC predictions of mineral soils were more accurate when sand content was
425 added to soil spectra as covariables. Nawar et al. (2016) found similar trend results for CRR
426 preprocessing considering organic matter prediction. In this study, the authors tested different
427 multivariate approaches (e.g. PLSR and SVM) in seven types of spectra preprocessing and
428 results are shown as follows. For PLSR, validation models applying CRR were the best among
429 all preprocessing ($R^2 = 0.79$, RMSE = 0.28%) followed by SMO ($R^2 = 0.59$, RMSE = 0.38%)
430 and Savitzky–Golay first derivative preprocessing ($R^2 = 0.50$, RMSE = 0.42%). These
431 outcomes are similar in pattern to those obtained by the current study for PLSR modeling (Table
432 1). In SVM modeling, Savitzky–Golay first derivative preprocessing generated the highest
433 prediction result ($R^2 = 0.75$, RMSE = 0.26%) followed by CRR ($R^2 = 0.65$, RMSE = 0.29%)
434 and SMO ($R^2 = 0.51$, RMSE = 0.35%). These tendency is in accordance with this study, which
435 for SVM result the best preprocessing was found for SGD and ASG (both applied Savitzky–
436 Golay first derivative) followed by CRR.

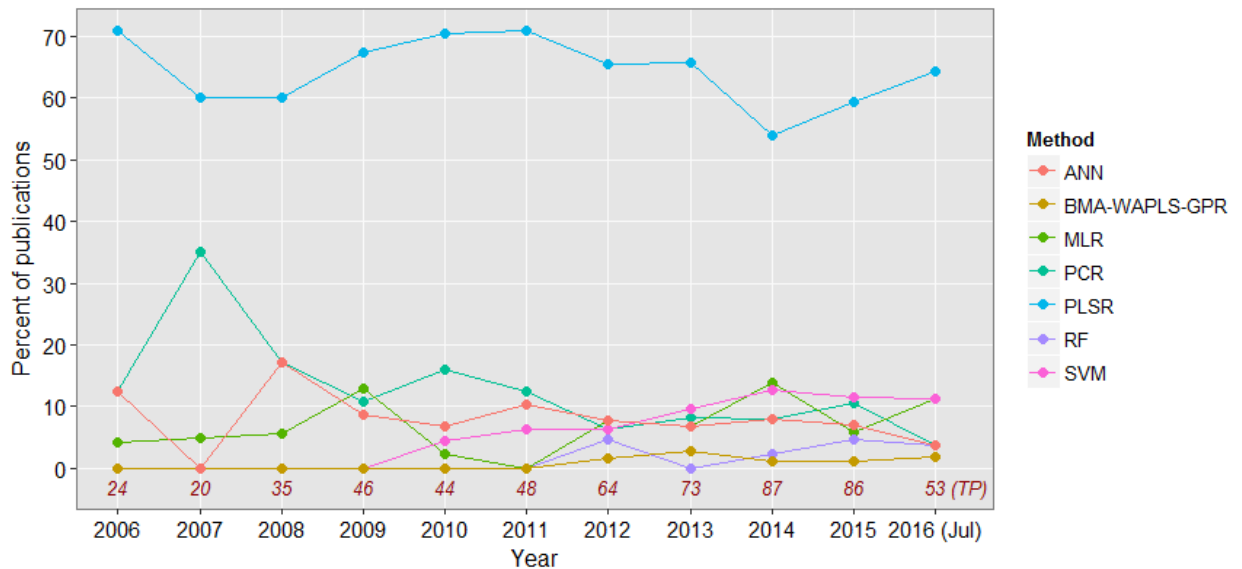
437 The improved performance of CRR preprocessing technique can be attributed to
438 effective noise removal, reduction of physical variability between samples, providing a more
439 consistent definition of band depth (Clark and Roush, 1984). Further advantages of continuum
440 removal are that this technique can be used to analyze the absorption features and to correct
441 band minimum to the true band center (Clark and Roush, 1984). This technique can be used to
442 normalize absorption features and to emphasize reflectance features of spectrum curves. CRR
443 preprocessing should be taken into consideration for SOC content prediction regardless the
444 multivariate method applied to model adjustment.

445

446 **3.3.4. Influence of multivariate methods in the performance of SOC prediction**

447 PLSR is the most suitable method for spectral modeling (Viscarra Rossel et al., 2009).
448 To demonstrate that, a search was conducted into a scientific citation database to compare the
449 volume of multivariate methods published in the last ten years applying spectroscopy to predict
450 soil properties (Fig. 5). The high frequency of publications with PLSR method, over the years,
451 has proven its application in predicting soil properties keeping its use around 65% of all
452 published papers in the last ten years. PCR method has shown a good amount of publications,
453 especially between 2006 until 2011, and is the second most used method over the last years.

454 The remaining methods exhibited quite a few volumes of publications, particularly for data
 455 mining algorithms. Over the last five years, the usage of these methods has growing and
 456 attracting attention of pedometric community. A positive aspect that drew attention was the
 457 quantitatively increase of total publications regarding soil property prediction by spectral data,
 458 confirming the growth of chemometrics prediction in a recent period.
 459



460

461 **Figure 5.** Publications of multivariate methods in the last ten years applying spectroscopy to
 462 predict soil properties. TP is the total number of publications per year. BMA, WAPLS and GPR
 463 were grouped due to the low volume of publications.

464

465 3.3.4.1. Partial least-squares regression performance

466 The dominance of PLSR is remarkable and is an indicative of its strength in SOC
 467 prediction. Prediction accuracy and model performance from PLSR along with the eight
 468 methods are presented in Table 1. In the current study, the performance of models revealed why
 469 PLSR is the most common method. Its predictive power had a satisfactory outcome. In fact, for
 470 PLSR models, R^2 values ranged between 0.67 to 0.81, RMSE values ranged from 0.67% to
 471 0.49%, and RPIQ values was ≥ 2.30 . Regarding the prediction of seven preprocessing with
 472 PLSR, CRR showed reduced RMSE (0.49%) and superior R^2 (0.81) and RPIQ (3.12). The
 473 results are comparable to prediction accuracy established in literature. Viscarra Rossel and
 474 Behrens (2010) applied PLSR method, amongst others, for the prediction of SOC, based on
 475 Vis-NIR spectra using a large spectral library with 1104 soil samples. Compared to this study,
 476 in Viscarra Rossel and Behrens (2010) the PLSR model prediction showed only slightly higher
 477 results ($R^2 = 0.82$, RMSE = 0.96%). Vasques et al. (2008) compared multivariate methods for

478 inferential modeling of soil total carbon and the PLSR models achieving a R_v^2 of 0.82 on average
479 of 30 spectral preprocessing. This performance is considered slightly better comparing the
480 current PLSR result, since in Vasques et al. (2008) the 554 soil samples were transformed into
481 logarithms before modeling. Attempting to improve the prediction performance of a large
482 tropical Vis–NIR spectroscopic soil library from Brazil, Araújo et al. (2014) achieved a R_v^2 of
483 0.60 and $RMSE_v$ of 0.55% for organic matter applying PLSR in 7172 soil samples. Knox et al.
484 (2015) modeled soil carbon fractions with Vis–NIR spectroscopy in a set of 1014 soil samples
485 collected across the state of Florida, USA. The authors applied 10 different spectral
486 preprocessing techniques resulting in a R_v^2 of 0.80, on average, and $RMSE_v$ of 0.48 $\log \text{g}\cdot\text{kg}^{-1}$
487 for PLSR modeling. To compare the calibration of Vis–NIR spectroscopy for on–line
488 measurement of SOC, Kuang et al. (2015) achieved similar R^2 performance with PLSR in
489 cross–validation and inferior RMSE (R_v^2 of 0.81, $RMSE_v$ of 1.99%).

490 These literature results revealed, once more, the better performance ability of linear
491 algorithm PLSR. The low and high results obtained in this study for SOC measurement with
492 PLSR model was consistent and comparable to those reported above. PLSR presented suitable
493 outcomes providing a quantitative modeling that can handle complicated relationships between
494 predictors and responses, and moreover it can deal with complex modeling problems (Wold et
495 al., 2001). PLSR is considerable a popular regression method applied in chemometrics since
496 the emphasis is on predicting responses and not necessarily on trying to understand the
497 underlying relationship between variables (Wold et al., 2001). Additionally, PLSR is a method
498 for constructing predictive models when the factors are many and highly collinear (Wold et al.,
499 1984), which is the case of hyperspectral data.

500 Considering PLSR is the most common method, there is a lack of studies comparing
501 alternatives approaches. Therefore, eight additional methods were applied in order to assess the
502 performances on SOC prediction. Each of methods achievement are discussed individually in
503 the next sections.

504

505 *3.3.4.2. Principal component regression performance*

506 As previously discussed, PCR is the second most frequently method used in
507 chemometrics predictions applying Vis–NIR spectroscopy (Fig. 5). PCR produced results
508 equivalent to PLSR with a R^2 varying from 0.66 to 0.80, RMSE from 0.66% to 0.51%, and
509 $RPIQ \geq 2.31$ (Table 1). Chang et al. (2001) achieved superior result applying PCR. The authors
510 found a R^2 of 0.87 and a RMSE of 0.78% using 726 soil samples to predict total soil carbon

511 from USA. Wang et al. (2015) used optical diffuse reflectance spectroscopy to predict organic
 512 matter with 155 soil samples from China. The authors adopted different spectral preprocessing
 513 from two spectrometers to find R_v^2 results ranging between 0.79 to 0.86 for organic matter
 514 prediction. PCR method indicated prominent results in mentioned literature by the fact that PCR
 515 and PLSR techniques are similar in many ways. PCR and PLSR are both methods to model
 516 response variable when there are a large number of predictor variables, and those predictors are
 517 highly correlated (Wold et al., 1984). Both methods construct new predictor variables, known
 518 as components as linear combinations of original predictor variables. Wentzell and Vega
 519 Montoto (2003) found that there were a few cases indicating higher results for PLSR over PCR,
 520 and a larger number of studies indicating no real difference performances. In their survey, the
 521 results of PCR and PLSR showed their prediction errors and number of latent variables differed.
 522 They concluded that PLSR almost always required fewer latent variables than PCR, but this did
 523 not appear to influence predictive ability. For Hemmateenejad et al. (2007) the successful of
 524 PCR and PLSR methods are related to their ability to overcome problems common to spectral
 525 data, such as collinearity, and their easy implementation due to the availability of software.

526

527 *3.3.4.3. Multiple linear regression performance*

528 The following method reported is MLR. Comparing multivariate methods attended,
 529 MLR accomplished fair performance for SOC prediction with a R^2 ranging from 0.69 to 0.79
 530 and RMSE between 0.64% to 0.52% (Table 1). The highest model was reached with SNV
 531 preprocessing. As MLR is considered the most common form of linear regression analysis,
 532 various studies have been applying it in soil properties prediction. Comparing regression
 533 methods for the prediction of SOC in a degraded south African ecosystem, Bayer et al. (2012)
 534 achieved a R_v^2 of 0.74 and $RMSE_v$ of 0.36% with MLR model in 164 soil samples. This results
 535 are slight inferior based on the best model result for MLR by current study. Viscarra Rossel and
 536 Behrens (2010) compared different data mining algorithms for modeling soil Vis–NIR with a
 537 dataset of 1104 soil samples from Australia. The authors reached higher results with MLR
 538 predicting SOC (R_v^2 ranging between 0.81 to 0.84). One evidence that guided to increase the
 539 model performance was the large number of soil samples. Vasques et al. (2008) achieved a R_v^2
 540 ranging between 0.66 to 0.85 for MLR modeling.

541 The results described are an indicative that MLR is still a beneficial method for SOC
 542 prediction when the choice is a statistical method that uses several explanatory variables to
 543 predict the outcome of a response variable in a simple linear model. MLR assumes the

544 relationships between independent variables and dependent variable are linear. Another
 545 important assumption is absence of multicollinearity thus the independent variables are not
 546 highly correlated. Further suppositions include homoscedasticity and normality. Presuming
 547 these linear regression assumptions, a robust prediction can be achieved using relatively simple
 548 algorithm.

549

550 *3.3.4.4. Support vector machine performance*

551 Starting from SVM, all the following methods are data mining approaches. Data mining
 552 involves methods that extract patterns from a data set applying artificial intelligence and
 553 machine learning. SVM produced a R^2 and RMSE ranging from 0.74 to 0.80 and 0.59% to
 554 0.52%, respectively (Table 1). SGD preprocessing achieved the highest prediction assessment
 555 for SVM. This method has been widely implemented for solving complex regression
 556 assignments (Ramirez-Lopez et al., 2013; Terra et al., 2015; Viscarra Rossel and Behrens,
 557 2010). Viscarra Rossel and Behrens (2010) reported SVM produced a similar result compared
 558 to PLSR, whereas Stevens et al. (2013) presented higher SOC predictions for SVM (R^2 from
 559 0.67 to 0.86) evaluating several data mining calibration methods on a diverse sample set of soil
 560 types in EU. Comparing spectral libraries (Vis–NIR spectroscopy) for quantitative analyses of
 561 tropical Brazilian soils, Terra et al. (2015) found low predictive result ($R^2_v = 0.65$, $RMSE_v =$
 562 0.16 g kg^{-1} , $RPIQ_v = 2.49$) for SOC applying SVM. Ramirez-Lopez et al. (2013) compared a
 563 regional (validation set = 1050) and global soil spectral library (validation set = 900) to predict
 564 SOC with different approaches. Models with SVM obtained prediction results of $R^2 = 0.54$ and
 565 0.57 , $RMSE = 0.27\%$ and 0.93% , for regional and global soil spectral libraries, respectively.
 566 Their results showed slightly higher R^2 was found for global soil spectral library. On the other
 567 hand, prediction error, RMSE, was lower for regional soil spectral library, which is attributed
 568 to the small SOC variation in regional spectral library. Araújo et al. (2014) compared the ability
 569 of multivariate models to determine organic matter from 7172 samples of seven different soil
 570 types collected from several areas of Brazil. The authors found that SVM ($R^2 = 0.69$, $RMSE =$
 571 0.48%) outperformed PLSR ($R^2 = 0.60$, $RMSE = 0.55\%$) for organic matter prediction. They
 572 mentioned SVM managed the capability of reducing problems with heterogeneity and
 573 nonlinearity of spectral data.

574 Results observed in literature corroborate the SVM as a very promising method for the
 575 estimation of SOC content. The greatest performance of SVM can be explained by the fact of
 576 SVM are a group of supervised learning methods, which represent an extension to nonlinear
 577 models of generalized algorithm with the capability of training nonlinear classifiers (Ivanciuc,

2007). Associated with SVM algorithm are the criteria of smaller number of support vectors yield a better model performance (Loosli et al., 2007). The reason for high performance of SVM models are related to the efficiency in modeling linear or nonlinear relationships and handling large databases.

3.3.4.5. Random forest performance

RF is an ensemble learning method for regression modeling. The overall predictive ability of RF models for SOC content was considered inferior. The prediction accuracy expressed a R^2 ranging from 0.47 to 0.77, and RMSE ranging from 0.84% to 0.55% (Table 1). RF approach exposed the lowest model prediction compared with the other methods. To compare different algorithms for modeling soil Vis–NIR spectra, Viscarra Rossel and Behrens (2010) reached lowest results for SOC estimation with RF ($R^2 = 0.71$, RMSE = 1.23%), which the best prediction was found for ANN. Knox et al. (2015) evaluated the potential of Vis–NIR–MIR spectroscopy to predict soil carbon fractions contained 1014 soil samples collected across the state of Florida, USA. RF validation produced a R^2 and RMSE ranging from 0.63 to 0.88 and 0.70 to 0.38 $\log \text{ g} \cdot \text{kg}^{-1}$, respectively, using different spectral preprocessing applied only at Vis–NIR range. Feng et al. (2014) drew attention to the difficulty of interpreting model estimates from log–transformed data. The authors stated that estimating original observation using exponent or anti–log of sample log–transformed data can generate inaccurate estimates of the true population of original data. They suggested for many applications, rather than trying to find an appropriate statistical distribution or transformation to model the observed data, it would probably be better to abandon the classic approach and switch to modern distribution–free methods.

According to Hastie et al. (2009), predictive learning is an important aspect of data mining methods, which are invariant under transformations. As a result, scaling or general transformations are not an issue, and they are immune to the effects of predictor outliers. RF tends to be versatile and flexible with small or large datasets and has becoming an effective tool in prediction (Breiman, 2001). RF can be very fast to train, but quite slow to create predictions once trained. For more accurate ensembles is required more trees, which means the development of model becomes slower. In certain situations, where run–time performance is important other approaches would be preferred. Model interpretability is another issue when compared to linear models. RF models are black boxes approach that are very hard to interpret. One reason for the poor performance of RF models might be based on the high number of trees

611 to fit the model might cause a risk of over correlating the ensemble and causing an overfit
612 problem.

613

614 *3.3.4.6. Bayesian model averaging performance*

615 BMA method provided a new approach regarding SOC prediction. Predictive
616 performance of BMA presented a R^2 and RMSE ranging from 0.68 to 0.80 and 0.65% to 0.51%,
617 respectively (Table 1). BMA has increasingly its applications across many diverse science
618 contexts. BMA was first used in sociology in early 80s as a model selection criterion, and since
619 then it has been widely applied. In soil science community its applications have scarce studies,
620 particularly for soil property prediction. Leon and Gonzalez (2009) predicted SOC using BMA
621 considering several predictors as: loss on ignition, parent material, drainage status, type of soil
622 horizon, clay content, and pH. Their validation analysis showed prediction accuracy for SOC
623 was improved with the BMA approach compared to ordinary least-squares approach. Malone
624 et al. (2014) applied BMA approach for combining digital soil property maps derived from
625 disaggregated legacy soil class maps. The authors determined the efficacy of ensemble
626 modeling as an useful combinatorial approach for combining digital soil property maps from
627 Australia. Poggio et al. (2016) assessed the spatial uncertainty with the Bayesian approach
628 modeling soil organic matter content in the Grampian region of Scotland. Similarly, Xiong et
629 al. (2015) applied Bayesian geostatistics to assess uncertainty associated with the predictive
630 models of SOC in Florida, USA.

631 BMA approach are able to extract empirical relevant relationships calculating a set of
632 ‘models’ assuming that there is no single ‘model’ that describes the data process, instead keeps
633 all ‘models’ and assigns each a weight, respectively. BMA refers to the process of averaging
634 estimates according probability distributions, where all ‘models’ can be interpreted as proxies
635 for some unknown underlying model (Brandl, 2008). BMA approach provided a quantitative
636 explicit tool that can be adjustable and flexible regarding the efficiency of inputs variables to
637 estimate SOC. The distinct advantage of BMA is express which input variable most influenced
638 the ‘models’ via prior specification (probability) (Raftery, 1995). Additionally, the benefit of
639 using BMA for spectral data was to access the uncertainty of each predictive variable.

640

641 *3.3.4.7. Weighted average partial least squares performance*

642 Overall, WAPLS produced the highest accuracy prediction model for SOC ($R^2 = 0.82$,
643 RPIQ = 3.18) (Table 1). The best WAPLS model returned the lowest RMSE value (0.48%)
644 observing all RMSE returned by remainder algorithms. Ramirez-Lopez et al. (2013) drew

645 attention to the great potential of WAPLS in predicting soil properties in large and diverse Vis–
 646 NIR datasets. The authors introduced the spectrum–based learner (SBL) technique, which is a
 647 category of WAPLS, and compared the predictive performance of this technique with other
 648 approaches including SVM and PLSR. SBL outperformed other approaches in both dataset
 649 (regional and global soil spectral libraries) producing the lowest RMSE and the highest R^2
 650 prediction (RMSE = 0.25% and 0.80%, $R^2 = 0.59$ and 0.68 , for regional and global soil spectral
 651 library, respectively). The low predictive performance, compared to this study, was attributed
 652 to large spectral variation as consequence of the diversity of soil formation environments where
 653 samples were collected. Gholizadeh et al. (2016) applied WAPLS approach and other data
 654 mining algorithms (PLSR and SVM) for the prediction of soil texture using Vis–NIR spectra
 655 from Czech Republic (total of 264 samples). The results of WAPLS model outperformed
 656 predictions accuracy of three soil fractions. The authors concluded WAPLS has not yet been
 657 commonly used to predict soil properties, and such statistical method with high prediction
 658 efficiency are the ones that have the best adaptability to analyze the structure of soil data. The
 659 highest performance of WAPLS result is related to important characteristics such as, it uses
 660 multiple models generated by multiple pls components and the final predicted value is a
 661 weighted average of all the predicted values generated by the multiple pls models (Ramirez-
 662 Lopez and Stevens, 2016).

663

664 *3.3.4.8. Gaussian process regression performance*

665 GPR is a machine learning algorithm applying the kernel function to training and
 666 predicting. The accuracy performance of GPR models produced a R^2 and RMSE values ranging
 667 from 0.65 to 0.79, and 0.69% to 0.52%, respectively (Table 1). In literature, there are a lack of
 668 studies addressing GPR method for SOC prediction. Numerous applications of kernel–based
 669 algorithms have been reported in the context of optical pattern and object recognition, text
 670 categorization, time–series prediction, gene expression profile analysis (Muller et al., 2001). In
 671 machine learning, kernel methods are a class of algorithms for pattern analysis. For many
 672 algorithms that solve regression problems, the data have to be explicitly transformed into
 673 feature vector representations, in contrast, kernel methods require only a user–specified kernel.
 674 This is called ‘kernel trick’ replacing its features (predictors) by a kernel function. Several
 675 classes of kernels can be used for machine learning and the selection of kernel is critical to the
 676 success of these algorithms (Karatzoglou et al., 2004).

677 One benefit of this algorithm is often computationally faster than the specific memory
 678 learning method. That means, applying highly complex data input should be efficient to

679 compute and revealing high performing kernel. Interesting research gaps in GPR method have
680 not been sufficiently explored yet making use of kernels for regression problems. The GPR
681 method is an alternative when working with learning algorithms, and results achieved in the
682 current study demonstrated GPR needs to be considered as prediction method for SOC using
683 Vis–NIR spectral data.

684

685 *3.3.4.9. Artificial neural network performance*

686 The final data mining approaches is ANN. For SOC prediction this method produced R^2
687 ranging from 0.64 to 0.80 and RMSE oscillated from 0.69% to 0.51% (Table 1). Evaluating
688 prediction accuracy between all methods ANN produced reasonable outcomes. The highest
689 model achievement (R^2 of 0.80, RMSE of 0.51%, and RPIQ of 3.01), ANN cannot be
690 considered an inferior or inaccurate result. Besides, this statement is corroborated by the
691 suitable performances of ANN models targeting SOC prediction in several studies. According
692 to Viscarra Rossel and Behrens (2010), ANN model returned the best prediction results for
693 SOC ($R^2 = 0.89$, RMSE = 0.75%) compared to PLSR, MLR, SVM, and RF, among others.
694 However, ANN model was implemented on a reduced number of wavelet coefficients. They
695 concluded the study by stating ANN was able to extract more relevant information when more
696 features are used. As ANN are called ‘black box’ systems, the combination of feature selection
697 and nonlinear modeling helped to achieve good predictions. Were et al. (2015) applied ANN
698 algorithm for spatial prediction of SOC stocks in Eastern Mau Forest Reserve, Kenya. The
699 authors found prediction accuracy for ANN model with R^2 value of 0.61 and a RMSE value of
700 15.46 Mg ha⁻¹. They suggested machine learning techniques should be applied for spatial
701 prediction of target soil variables. Kuang et al. (2015) compared ANN and PLSR model
702 performance in cross-validation, laboratory independent validation, on-line validation and on-
703 line independent validation for SOC prediction in two farm fields in Viborg, Denmark. Models
704 based on ANN algorithm showed a stronger prediction capability than those based on PLSR in
705 both fields, which the highest performance was produced by ANN in cross-validation model
706 ($R^2 = 0.90$, RMSE = 1.50%). ANN calibration model for SOC prediction reported in Mouazen
707 et al. (2010), with 133 soil samples collected from Belgium and northern France, produced
708 superior accuracy ($R^2 = 0.84$, RMSE = 0.68%) than the model obtained in the current study (R^2
709 = 0.80, RMSE = 0.51). Daniel et al. (2003) assessed the potential of ANN modeling soil organic
710 matter from spectral range of 400 to 1100 nm in 41 soil samples located in Thailand. ANN
711 models presented increased performance under laboratory ($R^2 = 0.86$) then field based
712 assessments ($R^2 = 0.84$).

713 The suitable performances of ANN models might be attributed to the nature of ANN in
 714 solving nonlinear problems (Kuang et al., 2015). In ANN, the mathematical model assign
 715 weights between elements, and network structure are adjusted depending on the inputs
 716 (McBratney et al., 2003).

717

718 **Table 1.** Performance of SOC predictive models from nine multivariate methods with the
 719 corresponding spectral preprocessing techniques.

Method	Preprocessing	Validation set		
		R ²	RMSE (%) [†]	RPIQ
PLSR	CRR	0.81	0.49	3.12
	NBR	0.80	0.52	2.94
	SNV	0.79	0.52	2.94
	MSC	0.78	0.54	2.84
	ASG	0.71	0.62	2.49
	SMO	0.70	0.63	2.42
	SGD	0.67	0.67	2.30
PCR	CRR	0.80	0.51	3.00
	NBR	0.79	0.52	2.95
	SNV	0.79	0.52	2.92
	MSC	0.78	0.54	2.86
	SMO	0.70	0.62	2.47
	ASG	0.68	0.64	2.39
	SGD	0.66	0.66	2.31
MLR	SNV	0.79	0.52	2.93
	CRR	0.78	0.53	2.88
	MSC	0.78	0.54	2.84
	NBR	0.77	0.56	2.75
	SMO	0.73	0.60	2.56
	ASG	0.71	0.61	2.50
	SGD	0.69	0.64	2.41
SVM	SGD	0.80	0.52	2.94
	ASG	0.80	0.53	2.90
	CRR	0.78	0.53	2.87

SVM	NBR	0.77	0.54	2.82
	MSC	0.76	0.56	2.73
	SNV	0.75	0.56	2.72
	SMO	0.74	0.59	2.59
RF	CRR	0.77	0.55	2.77
	SGD	0.74	0.60	2.58
	ASG	0.72	0.61	2.51
	SNV	0.67	0.66	2.31
	MSC	0.65	0.67	2.27
	NBR	0.54	0.77	1.99
	SMO	0.47	0.84	1.83
BMA	MSC	0.80	0.51	3.03
	SNV	0.79	0.52	2.97
	CRR	0.79	0.52	2.96
	NBR	0.78	0.54	2.85
	SMO	0.72	0.61	2.52
	ASG	0.71	0.61	2.51
	SGD	0.68	0.65	2.36
WAPLS	NBR	0.82	0.48	3.18
	CRR	0.81	0.49	3.10
	SNV	0.80	0.51	2.99
	MSC	0.80	0.51	2.98
	SMO	0.79	0.52	2.96
	ASG	0.71	0.62	2.47
	SGD	0.48	0.74	2.10
GPR	SNV	0.79	0.52	2.96
	MSC	0.79	0.52	2.94
	CRR	0.79	0.53	2.90
	NBR	0.78	0.53	2.89
	ASG	0.69	0.66	2.34
	SGD	0.65	0.69	2.21
	SMO	0.65	0.69	2.21

	NBR	0.80	0.51	3.01
	SNV	0.79	0.52	2.92
	MSC	0.75	0.56	2.73
ANN	CRR	0.73	0.59	2.61
	ASG	0.70	0.63	2.44
	SMO	0.66	0.66	2.32
	SGD	0.64	0.69	2.22

720 † Preprocessing column are ordered by decreasing predictive performance in each multivariate
 721 method. R^2 : coefficient of determination, RMSE: root mean square error, and RPIQ: ratio of
 722 performance to interquartile range.

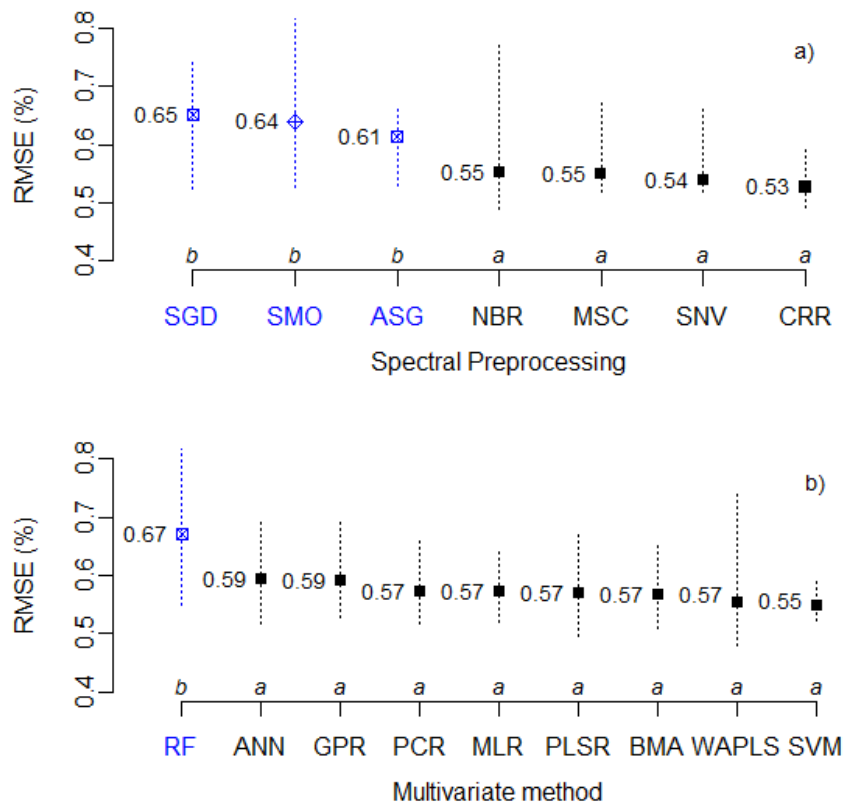
723

724 3.3.5. Comparing performances

725 Comparing the RMSE means of preprocessing techniques (Fig. 6a), the Scott–Knott test
 726 showed a significant difference between two groups. First group is composed by NBR, MSC,
 727 SNV, and CRR, which are the preprocessing belonging to scatter–corrections. According to
 728 Scott–Knott test, all four scatter–corrections preprocessing presented statistically identical
 729 RMSE results. In this group, CRR achieved the best performance. Besides, CRR showed the
 730 smallest variation in maximum and minimum RMSE values, which is another indicator of great
 731 performance of this preprocessing in SOC prediction. The second group is formed by SGD and
 732 ASG (spectral derivatives group) plus the SMO preprocessing. This group presented inferior
 733 results. The poorest result was achieved by SGD, which presented the highest RMSE value, in
 734 average (0.65%).

735 The comparison of multivariate methods is shown in Fig. 6b. The methods were divided
 736 in two groups. Excepting for RF, all of methods were classified into the same group, which
 737 were marked with the letter ‘a’ in the Scott Knott test. According to this, any of the methods
 738 classified in group are suitable and can be applied in SOC prediction, since statistically they
 739 were exactly the same. This result makes very difficult to decide which method showed better
 740 SOC predictive performance. SVM presented the lowest RMSE value in average and the
 741 maximum and minimum RMSE had the smallest scattering. This result is an indication of the
 742 great performance of SVM in predict SOC.

743



744

745 **Figure 6.** Comparison between means of preprocessing techniques (a) and multivariate
 746 methods (b). Dotted line represents maximum and minimum RMSE values. Letters represent
 747 the results of Scott–Knott test (significance level of $\alpha = 0.1$).

748

749 3.3.6. Time to process the models in R

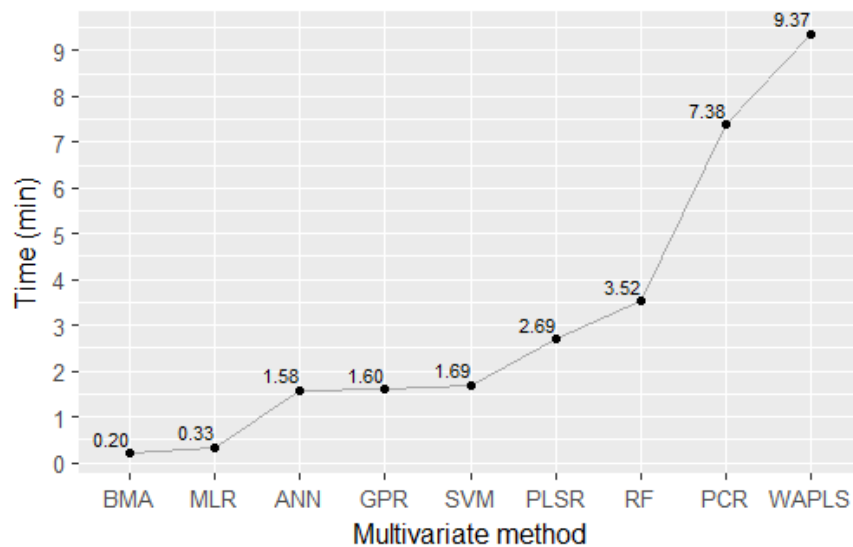
750

751 The best multivariate method is, presumably, the one that produces the best predictive
 752 ability with a robust accuracy result. Nonetheless, the rules to decide which method is better or
 753 which algorithm is more likely to use, it seems to be a tough decision. To complement this
 754 assessment, since the methods revealed prominent results, the time to process each model in R
 755 was calculate. In order to find which of the nine methods indicate the lowest time–consuming,
 756 the averages of seven preprocessing models were determined (Fig. 7). This procedure required
 757 to run the models in the same computer. The time to process the modeling are influenced by
 758 several factors such as, computation system, number of observations, number of variables in
 759 the prediction model, method used, etc. BMA and MLR were the more efficient being the
 760 lowest time–consuming methods, where the average modeling was processed in 0.20 and 0.33
 761 min, respectively. The next three methods, ANN, GPR, and SVM, required around 1.58 and
 1.69 min to process the modeling. PLSR and RF started to increase the time, requiring 2.69 and

762 3.52 min, respectively to process the models. The least efficient methods were PCR and
 763 WAPLS with exceedingly long computation time (7.38 and 9.37 min, respectively).

764 The evaluation of the time consumed revealed that BMA was the most efficient method.
 765 As SVM produced great predictive performance overall (Table 1) and its time-consuming was
 766 acceptable (1.69 min), it can be considered a solid method to SOC prediction. WAPLS and
 767 PCR were the less efficient methods. However, PCR can be replaced by PLSR method since
 768 they showed similar performance for SOC prediction and PLSR took less time to process the
 769 models in R. An alternative, instead of using WAPLS, is to apply GPR, since kernel function
 770 speeds the process and the performance of models are not significantly diminished.

771



772

773 **Figure 7.** Time to process the models in R. For each method, the average of seven
 774 preprocessing models was considered.

775

776 3.4.CONCLUSIONS

777

778 The study explored a systematic methodology in SOC prediction using Vis–NIR
 779 spectroscopic data to support the choices of spectral preprocessing and multivariate method.
 780 Regarding the preprocessing techniques, scatter–correction group (NBR, MSC, SNV, and
 781 CRR) showed improved prediction capability. Overall, continuum removal preprocessing
 782 produced the greatest predictive result, which confirms the potential of this preprocessing in
 783 predicting SOC. However, spectral derivatives preprocessing group, which include SGD and
 784 ASG, showed superior results for SVM and RF methods revealing their capability to better
 785 handle derivative transformation. In the multivariate methods, excepting for RF, all of methods

786 presented robust prediction. The highest model accuracy for SOC prediction was found
 787 applying WAPLS method and NBR preprocessing ($R^2 = 0.82$, RMSE = 0.48%, RPIQ = 3.18).
 788 The systematic methodology applied in this study can improve reliability for SOC
 789 determinations by examining how techniques of preprocessing and multivariate methods affect
 790 spectral analyses. The quantification of SOC is able to boost up soil properties information and
 791 supply digital soil mapping approach into developing soil properties maps.

792

793 **Acknowledgements**

794

795 The authors would like to thank the reviewers for their insightful comments on the
 796 paper. This research was funded by Coordination for the Improvement of Higher Education
 797 Personnel (CAPES). Second and third authors thank the National Council for Scientific and
 798 Technological Development (CNPq) (project nº442718/2014-4) and Foundation for Funding in
 799 Research and Innovation of Santa Catarina State (FAPESC) (project no. 2012000094), Ministry
 800 of Education, Brazil, for the financial support. The authors are grateful to GeoCis Laboratory,
 801 Soil Department, ESALQ/University of Sao Paulo, Brazil for Vis–NIR spectral measurement.

802

803 **References**

804

- 805 Andrews, S.S., Karlen, D.L., Cambardella, C.A., 2004. The soil management assessment
 806 framework: A quantitative soil quality evaluation method. *Soil Sci Soc Am J. Soil Sci.*
 807 *Soc. Am. J.* 68, 1945–1962.
- 808 Araújo, S.R., Wetterlind, J., Demattê, J. a. M., Stenberg, B., 2014. Improving the prediction
 809 performance of a large tropical vis-NIR spectroscopic soil library from Brazil by
 810 clustering into smaller subsets or use of data mining calibration techniques. *Eur. J. Soil*
 811 *Sci.* 65, 718–729. doi:10.1111/ejss.12165
- 812 Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.C.R., McBratney, A.B. (Eds.), 2014.
 813 *GlobalSoilMap: Basis of the global spatial soil information system.* CRC
 814 Press/Balkema.
- 815 Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and
 816 De-trending of Near-Infrared Diffuse Reflectance Spectra. *Appl. Spectrosc.* 43, 772–
 817 777.
- 818 Bayer, A., Bachmann, M., Muller, A., Kaufmann, H., 2012. A Comparison of Feature-Based
 819 MLR and PLS Regression Techniques for the Prediction of Three Soil Constituents in

- 820 a Degraded South African Ecosystem. *Appl. Environ. Soil Sci.* 2012, e971252.
821 doi:10.1155/2012/971252
- 822 Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., McBratney, A., 2010.
823 Critical review of chemometric indicators commonly used for assessing the quality of
824 the prediction of soil attributes by NIR spectroscopy. *TrAC Trends Anal. Chem.* 29,
825 1073–1081. doi:10.1016/j.trac.2010.05.006
- 826 Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (NIR) and mid-infrared (MIR)
827 spectroscopic techniques for assessing the amount of carbon stock in soils – Critical
828 review and research perspectives. *Soil Biol. Biochem.* 43, 1398–1410.
829 doi:10.1016/j.soilbio.2011.02.019
- 830 Ben-Dor, E., Inbar, Y., Chen, Y., 1997. The reflectance spectra of organic matter in the visible
831 near-infrared and short wave infrared region (400–2500 nm) during a controlled
832 decomposition process. *Remote Sens. Environ.* 61, 1–15. doi:10.1016/S0034-
833 4257(96)00120-4
- 834 Brandl, B., 2008. Bayesian model averaging and model selection: two sides of the same coin
835 when identifying the determinants of trade union density? *Cent. Eur. J. Oper. Res.* 17,
836 13–29. doi:10.1007/s10100-008-0072-0
- 837 Breiman, L., 2001. Random Forests. *Mach Learn* 45, 5–32. doi:10.1023/A:1010933404324
- 838 Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., Smaling, E.M.A., 2012. Building a near
839 infrared spectral library for soil organic carbon estimation in the Limpopo National
840 Park, Mozambique. *Geoderma* 183–184, 41–48. doi:10.1016/j.geoderma.2012.03.011
- 841 Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-Infrared Reflectance
842 Spectroscopy–Principal Components Regression Analyses of Soil Properties. *Soil Sci.*
843 *Soc. Am. J.* 65, 480–490. doi:10.2136/sssaj2001.652480x
- 844 Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: Quantitative analysis techniques for
845 remote sensing applications. *J. Geophys. Res. Solid Earth* 89, 6329–6340.
846 doi:10.1029/JB089iB07p06329
- 847 Conforti, M., Castrignanò, A., Robustelli, G., Scarciglia, F., Stelluti, M., Buttafuoco, G., 2015.
848 Laboratory-based Vis–NIR spectroscopy and partial least square regression with
849 spatially correlated errors for predicting spatial variation of soil organic matter content.
850 *CATENA* 124, 60–67. doi:10.1016/j.catena.2014.09.004
- 851 Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20, 273–297.
852 doi:10.1007/BF00994018

- 853 Daniel, K.W., Tripathi, N.K., Honda, K., 2003. Artificial neural network analysis of laboratory
854 and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand).
855 Soil Res. 41, 47–59.
- 856 Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M., 2014. Log-transformation and
857 its implications for data analysis. Shanghai Arch. Psychiatry 26, 105–109.
858 doi:10.3969/j.issn.1002-0829.2014.02.009
- 859 Geladi, P., MacDougall, D., Martens, H., 1985. Linearization and Scatter-Correction for Near-
860 Infrared Reflectance Spectra of Meat. Appl. Spectrosc. 39, 491–500.
- 861 Gholizadeh, A., Borůvka, L., Saberioon, M., Vašát, R., 2016. A Memory-Based Learning
862 Approach as Compared to Other Data Mining Algorithms for the Prediction of Soil
863 Texture Using Diffuse Reflectance Spectra. Remote Sens. 8, 341.
864 doi:10.3390/rs8040341
- 865 Gholizadeh, A., Borůvka, L., Saberioon, M.M., Kozák, J., Vašát, R., Němeček, K., 2015.
866 Comparing different data preprocessing methods for monitoring soil heavy metals based
867 on soil spectral features. Soil Water Res. 10, 218–227. doi:10.17221/113/2015-SWR
- 868 Gosso, A., 2012. elmNN: Implementation of ELM (Extreme Learning Machine) algorithm for
869 SLFN (Single Hidden Layer Feedforward Neural Networks). R Package Version 1.
- 870 Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations
871 and stocks on Barro Colorado Island — Digital soil mapping using Random Forests
872 analysis. Geoderma 146, 102–113. doi:10.1016/j.geoderma.2008.05.008
- 873 Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling
874 approaches. Geoderma 152, 195–207. doi:10.1016/j.geoderma.2009.06.003
- 875 Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer
876 Series in Statistics. Springer New York, New York, NY.
- 877 Hemmateenejad, B., Akhond, M., Samari, F., 2007. A comparative study between PCR and
878 PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and
879 phenol: Effect of wavelength selection. Spectrochim. Acta. A. Mol. Biomol. Spectrosc.
880 67, 958–965. doi:10.1016/j.saa.2006.09.014
- 881 Ivanciuc, O., 2007. Applications of Support Vector Machines in Chemistry, in: Lipkowitz,
882 K.B., Cundari, T.R. (Eds.), Reviews in Computational Chemistry. John Wiley & Sons,
883 Inc., pp. 291–400.
- 884 Jelihovschi, E.G., Faria, J.C., Allaman, I.B., 2014. ScottKnott: A Package for Performing the
885 Scott-Knott Clustering Algorithm in R. Trends Appl. Comput. Math. 15, 3–17.

- 886 Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab - An S4 Package for Kernel
 887 Methods in R. *J. Stat. Softw.* 11, 1–20. doi:10.18637/jss.v011.i09
- 888 Kendall, M.G., 1957. *A Course in Multivariate Analysis*. Griffin, London.
- 889 Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B., Harris, W.G., 2015.
 890 Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared
 891 (MIR) spectroscopy. *Geoderma* 239–240, 229–239.
 892 doi:10.1016/j.geoderma.2014.10.019
- 893 Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and
 894 partial least squares for on-line visible and near infrared spectroscopy measurement of
 895 soil organic carbon, pH and clay content. *Soil Tillage Res.* 146, Part B, 243–252.
 896 doi:10.1016/j.still.2014.11.002
- 897 Lagacherie, P., Baret, F., Feret, J.-B., Madeira Netto, J., Robbez-Masson, J.M., 2008.
 898 Estimation of soil clay and calcium carbonate using laboratory, field and airborne
 899 hyperspectral measurements. *Remote Sens. Environ.* 112, 825–835.
 900 doi:10.1016/j.rse.2007.06.014
- 901 Lal, R., 2004. Soil carbon sequestration to mitigate climate change. *Geoderma* 123, 1–22.
 902 doi:10.1016/j.geoderma.2004.01.032
- 903 Leon, A., Gonzalez, R.L., 2009. Predicting soil organic carbon percentage from loss-on-ignition
 904 using Bayesian Model Averaging. *Aust. J. Soil Res.* 47, 763–769. doi:10.1071/SR08119
- 905 Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. *R News* 2, 18–
 906 22.
- 907 Loosli, G., Canu, S., Bottou, L., 2007. Training Invariant Support Vector Machines using
 908 Selective Sampling, in: Bottou, L., Chapelle, O., DeCoste, D., Weston, J. (Eds.), *Large*
 909 *Scale Kernel Machines*. MIT Press, Cambridge, MA, pp. 301–320.
- 910 Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to
 911 combine soil property rasters from legacy soil maps and from point data. *Geoderma*
 912 232–234, 34–44. doi:10.1016/j.geoderma.2014.04.033
- 913 McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. *Geoderma* 213,
 914 203–213. doi:10.1016/j.geoderma.2013.08.013
- 915 McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping.
 916 *Geoderma* 117, 3–52. doi:10.1016/S0016-7061(03)00223-4
- 917 McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity.
 918 *Bull. Math. Biophys.* 5, 115–133. doi:10.1007/BF02478259

- 919 McDowell, M.L., Bruland, G.L., Deenik, J.L., Grunwald, S., Knox, N.M., 2012. Soil total
920 carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse
921 reflectance spectroscopy. *Geoderma* 189–190, 312–320.
922 doi:10.1016/j.geoderma.2012.06.009
- 923 Mevik, B.-H., Wehrens, R., Liland, K.H., 2013. pls: Partial Least Squares and Principal
924 Component Regression. R Package Version 2.5-0.2.
- 925 Meyer, D., 2001. Support Vector Machines The Interface to libsvm in package e1071. R News
926 1.
- 927 Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties
928 from infrared reflectance spectroscopy. *Chemom. Intell. Lab. Syst.* 94, 72–79.
929 doi:10.1016/j.chemolab.2008.06.003
- 930 Moros, J., Vallejuelo, S.F.-O. de, Gredilla, A., Diego, A. de, Madariaga, J.M., Garrigues, S.,
931 Guardia, M. de la, 2009. Use of Reflectance Infrared Spectroscopy for Monitoring the
932 Metal Content of the Estuarine Sediments of the Nerbioi-Ibaizabal River (Metropolitan
933 Bilbao, Bay of Biscay, Basque Country). *Environ. Sci. Technol.* 43, 9314–9320.
934 doi:10.1021/es9005898
- 935 Mouazen, A.M., Kuang, B., De Baerdemaeker, J., Ramon, H., 2010. Comparison among
936 principal component, partial least squares and back propagation neural network analyses
937 for accuracy of measurement of selected soil properties with visible and near infrared
938 spectroscopy. *Geoderma, Diffuse reflectance spectroscopy in soil science and land
939 resource assessment* 158, 23–31. doi:10.1016/j.geoderma.2010.03.001
- 940 Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review.
941 *ISPRS J. Photogramm. Remote Sens.* 66, 247–259. doi:10.1016/j.isprsjprs.2010.11.001
- 942 Muller, K.R., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B., 2001. An introduction to kernel-
943 based learning algorithms. *IEEE Trans. Neural Netw.* 12, 181–201.
944 doi:10.1109/72.914517
- 945 Muñoz, J.D., Kravchenko, A., 2011. Soil carbon mapping using on-the-go near infrared
946 spectroscopy, topography and aerial photographs. *Geoderma* 166, 102–110.
947 doi:10.1016/j.geoderma.2011.07.017
- 948 Mutanga, O.M.C., Skidmore, A.K., Kumar, L., Ferwerda, J., 2005. Estimating tropical pasture
949 quality at canopy level using band depth analysis with continuum removal in the visible
950 domain. *Int. J. Remote Sens.* 26, 1093–1108. doi:10.1080/01431160512331326738
- 951 Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., Mouazen, A.M., 2016. Estimating the soil clay
952 content and organic matter by means of different calibration methods of vis-NIR diffuse

- 953 reflectance spectroscopy. *Soil Tillage Res.* 155, 510–522.
 954 doi:10.1016/j.still.2015.07.021
- 955 Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014.
 956 Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a
 957 local partial least square regression approach. *Soil Biol. Biochem.* 68, 337–347.
 958 doi:10.1016/j.soilbio.2013.10.022
- 959 Peng, X., Shi, T., Song, A., Chen, Y., Gao, W., 2014. Estimating Soil Organic Carbon Using
 960 VIS/NIR Spectroscopy with SVMR and SPA Methods. *Remote Sens.* 6, 2699–2717.
 961 doi:10.3390/rs6042699
- 962 Poggio, L., Gimona, A., Spezia, L., Brewer, M.J., 2016. Bayesian spatial modelling of soil
 963 properties and their uncertainty: The example of soil organic matter in Scotland using
 964 R-INLA. *Geoderma* 277, 69–82. doi:10.1016/j.geoderma.2016.04.026
- 965 R Core Team, 2016. R: A Language and Environment for Statistical Computing.
- 966 Raftery, A., Hoeting, J., Volinsky, C., Painter, I., Yeung, K.Y., 2015. BMA: Bayesian Model
 967 Averaging. R Package Version 3186.
- 968 Raftery, A.E., 1995. Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–
 969 164.
- 970 Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013.
 971 The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra
 972 of complex datasets. *Geoderma* 195–196, 268–279.
 973 doi:10.1016/j.geoderma.2012.12.014
- 974 Ramirez-Lopez, L., Stevens, A., 2016. resemble: Regression and similarity evaluation for
 975 memory-based learning in spectral chemometrics. R Package Version 122.
- 976 Rinnan, Å., Berg, F. van den, Engelsen, S.B., 2009. Review of the most common pre-processing
 977 techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* 28, 1201–1222.
 978 doi:10.1016/j.trac.2009.07.007
- 979 Scott, A.J., Knott, M., 1974. A Cluster Analysis Method for Grouping Means in the Analysis
 980 of Variance. *Biometrics* 30, 507–512. doi:10.2307/2529204
- 981 Shenk, J.S., Westerhaus, M.O., Berzaghi, P., 1998. Investigation of a LOCAL calibration
 982 procedure for near infrared instruments. *J. Infrared Spectrosc.* 5, 223–232.
 983 doi:10.1255/jnirs.115
- 984 Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of Soil
 985 Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance
 986 Spectroscopy. *PLoS ONE* 8, e66409. doi:10.1371/journal.pone.0066409

- 987 Stevens, A., Ramirez-Lopez, L., 2013. An introduction to the prospectr package. R Package
988 Vignette.
- 989 Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Liroy, R., Hoffmann, L., van Wesemael, B.,
990 2010. Measuring soil organic carbon in croplands at regional scale using airborne
991 imaging spectroscopy. *Geoderma, Diffuse reflectance spectroscopy in soil science and*
992 *land resource assessment* 158, 32–45. doi:10.1016/j.geoderma.2009.11.032
- 993 Stoner, E.R., Baumgardner, M.F., 1981. Characteristic variations in reflectance of surface soils.
994 ResearchGate 45.
- 995 Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2015. Spectral libraries for quantitative
996 analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data.
997 *Geoderma* 255–256, 81–93. doi:10.1016/j.geoderma.2015.04.017
- 998 Vašát, R., Kodešová, R., Borůvka, L., Klement, A., Jakšík, O., Gholizadeh, A., 2014.
999 Consideration of peak parameters derived from continuum-removed spectra to predict
1000 extractable nutrients in soils with visible and near-infrared diffuse reflectance
1001 spectroscopy (VNIR-DRS). *Geoderma* 232–234, 208–218.
1002 doi:10.1016/j.geoderma.2014.05.012
- 1003 Vasques, G.M., Grunwald, S., Sickman, J.O., 2008. Comparison of multivariate methods for
1004 inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma* 146,
1005 14–25. doi:10.1016/j.geoderma.2008.04.007
- 1006 Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse
1007 reflectance spectra. *Geoderma, Diffuse reflectance spectroscopy in soil science and land*
1008 *resource assessment* 158, 46–54. doi:10.1016/j.geoderma.2009.12.025
- 1009 Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd,
1010 K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aichi, H., Barthès, B.G.,
1011 Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W.,
1012 Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C.,
1013 Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P.,
1014 Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C.,
1015 Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize
1016 the world's soil. *Earth-Sci. Rev.* 155, 198–230. doi:10.1016/j.earscirev.2016.01.012
- 1017 Viscarra Rossel, R.A., Cattle, S.R., Ortega, A., Fouad, Y., 2009. In situ measurements of soil
1018 colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma* 150,
1019 253–266. doi:10.1016/j.geoderma.2009.01.025

- 1020 Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006.
1021 Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for
1022 simultaneous assessment of various soil properties. *Geoderma* 131, 59–75.
1023 doi:10.1016/j.geoderma.2005.03.007
- 1024 Walesiak, M., Dudek, A., 2016. clusterSim: Searching for Optimal Clustering Procedure for a
1025 Data Set.
- 1026 Wang, W., Li, S., Qi, H., Ayhan, B., Kwan, C., Vance, S., 2014. Revisiting the Preprocessing
1027 Procedures for Elemental Concentration Estimation based on CHEMCAM LIBS on
1028 MARS Rover, in: ResearchGate. Presented at the 6th Workshop on Hyperspectral Image
1029 and Signal Processing: Evolution in Remote Sensing.
- 1030 Wang, Y., Huang, T., Liu, J., Lin, Z., Li, S., Wang, R., Ge, Y., 2015. Soil pH value, organic
1031 matter and macronutrients contents prediction using optical diffuse reflectance
1032 spectroscopy. *Comput. Electron. Agric.* 111, 69–77. doi:10.1016/j.compag.2014.11.019
- 1033 Wentzell, P.D., Vega Montoto, L., 2003. Comparison of principal components regression and
1034 partial least squares regression through generic simulations of complex mixtures.
1035 *Chemom. Intell. Lab. Syst.* 65, 257–279. doi:10.1016/S0169-7439(02)00138-7
- 1036 Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector
1037 regression, artificial neural networks, and random forests for predicting and mapping
1038 soil organic carbon stocks across an Afromontane landscape. *Ecol. Indic.* 52, 394–403.
1039 doi:10.1016/j.ecolind.2014.12.028
- 1040 Williams, C.K.I., Barber, D., 1998. Bayesian classification with Gaussian processes. *IEEE*
1041 *Trans. Pattern Anal. Mach. Intell.* 20, 1342–1351. doi:10.1109/34.735807
- 1042 Wold, S., Ruhe, A., Wold, H., Dunn, I., W., 1984. The Collinearity Problem in Linear
1043 Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM*
1044 *J. Sci. Stat. Comput.* 5, 735–743. doi:10.1137/0905052
- 1045 Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics.
1046 *Chemom. Intell. Lab. Syst., PLS Methods* 58, 109–130. doi:10.1016/S0169-
1047 7439(01)00155-1
- 1048 Xie, X.-L., Pan, X.-Z., Sun, B., 2012. Visible and Near-Infrared Diffuse Reflectance
1049 Spectroscopy for Prediction of Soil Properties near a Copper Smelter. *Pedosphere* 22,
1050 351–366. doi:10.1016/S1002-0160(12)60022-8
- 1051 Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Bliznyuk, N., 2015. Assessing
1052 uncertainty in soil organic carbon modeling across a highly heterogeneous landscape.
1053 *Geoderma* 251–252, 105–116. doi:10.1016/j.geoderma.2015.03.028

- 1054 Yeomans, J.C., Bremner, J.M., 1988. A rapid and precise method for routine determination of
1055 organic carbon in soil. *Commun. Soil Sci. Plant Anal.* 19, 1467–1476.
1056 doi:10.1080/00103628809368027
- 1057 Zhang, Y., Li, M., Zheng, L., Zhao, Y., Pei, X., 2016. Soil nitrogen content forecasting based
1058 on real-time NIR spectroscopy. *Comput. Electron. Agric.* 124, 29–36.
1059 doi:10.1016/j.compag.2016.03.016
1060

4 ARTICLE 3: Alrad Spectra: a graphical user interface in R to perform preprocessing, multivariate modeling and prediction using spectroscopic data³

Abstract

This paper describes the implementation of a R graphical user interface (GUI) named Alrad Spectra. It uses spectroscopic data to process the spectra and then generate models to predict the Y variable. The GUI was developed to accomplish tasks such as perform a large range of spectral preprocessing techniques, implement several multivariate calibration methods, statistics assessment, graphical output, validate the models using independent data sets, and predict unknown Y variables. Alrad Spectra has four main modules: Import Data, Spectral Preprocessing, Modeling, and Prediction. The capacity of performing multiple tasks, being free and open-source, easy to operate, and requiring no initial knowledge of R programming language are features that make Alrad Spectra an useful tool for general public, researches, precision agriculture managers, and for the usage in analytical laboratories. The implementation of Alrad Spectra is demonstrated by applying visible near-infrared reflectance spectroscopy for soil organic carbon prediction.

Keywords: GUI; R environment; multivariate calibration; spectral preprocessing.

4.1. INTRODUCTION

Alrad Spectra is a graphical user interface (GUI) implemented in R programming language [1] that was developed to perform preprocessing, multivariate modeling and prediction using spectroscopic data. The features of Alrad Spectra include: i) import large database files; ii) perform a large range of spectral preprocessing and transformation techniques; iii) implement several multivariate calibration methods, which can provide well-fitted and accurate models; iv) provide statistics assessment; v) deliver graphical output; vi) validate the models using independent data sets; and vii) predict unknown Y variables.

Alrad Spectra encompasses the following steps: import data file, data exploration, spectral preprocessing, modeling, and prediction. Variations in the spectral data, which are caused by chemical and physical characteristics, can be modeled in conjunction with the target information. Spectral data preprocessing is an important step in the spectra analysis, which involves specific processing on the raw data. To standardize and transform spectra, remove

³ Article was submitted to **Chemometrics and Intelligent Laboratory Systems**.

32 noise, emphasize features, and improve accuracy of subsequent quantitative analysis [2], in
33 general, it is necessary to apply techniques of preprocessing. Spectral data preprocessing has
34 been identified as an indispensable part of spectral data analysis and has shown its importance
35 on subsequent modeling tasks. The modeling step is accomplished by applying multivariate
36 calibration methods. They have been commonly used to construct well-fitted models to
37 determine the chemical components of interest. The application of linear regression, ordinary
38 least-squares regression, data mining and machine learning algorithms are examples of
39 modeling methods used in Alrad Spectra.

40 Alrad Spectra runs in R, which is an open-source, powerful statistical programming
41 language that has the latest statistical techniques with thousands of add-on packages available
42 on the download servers. The growing importance of R has been huge in the last years. For
43 Tippmann [3], there is a trend for many academics to wean themselves off commercial software
44 and dive in the free, open-source, and popular data-analysis tool. R has become one of the most
45 requested statistical computing language and programming environment. The GUI in R came
46 to supply users' needs by incorporating a user-friendly interface, in which there is no need to
47 spend time learning how to deal with functions and its arguments, and remembering a lot of
48 commands.

49 For some users, the limitation of R is the implementation of functions, which must be
50 called as text commands, and the user is required to find the proper packages that will
51 accomplish specific tasks, recall the operations, and its argument options. To facilitate the
52 routines for users, Alrad Spectra was developed to compensate this requirement. It has the
53 advantages of providing a user-friendly GUI, being free and easy to operate, it requires no initial
54 knowledge of R programming language, and it is the first of its kind in R. Plus, Alrad Spectra
55 can process spectroscopic data from soils, water, grains, food, vegetation, etc.

56 The aim of this paper is to describe the development of Alrad Spectra by performing
57 spectral preprocessing, utilizing multivariate calibration modeling to predict the Y variable
58 using spectroscopic data. The implementation of Alrad Spectra is demonstrated by using soil
59 Visible Near-infrared (Vis-NIR) reflectance spectroscopy data to predict of soil organic carbon
60 (SOC). The description includes data entry procedure, spectral data preprocessing, modeling
61 process, prediction statistics assessment, and SOC prediction.

62

63 4.2.SOFTWARE

64 The Alrad Spectra runs under R version 3.0 or higher. The AlradSpectra package is
65 sited at open source community *github.com* repository (github.com/AlradSpectra). The

66 devtools package is required to download and install Alrad Spectra from the source-website.
 67 The commands to install, load, and initialize AlradSpectra package in R are shown in Fig. 1.
 68 As Alrad Spectra is operated in a user-friendly graphical interface, all of the operations and
 69 parameters required for chemometric analysis can be set through the GUI. Spectral data can be
 70 loaded, saved, processed and analyzed through GUI components. Alrad Spectra combines
 71 commands, functions and packages creating an easy and interactive application freely accessed
 72 by the public (GPL-3 License).

73

```

1  ### Installing Alrad Spectra ###
2
3  install.packages("devtools")           # You need to install devtools package
4                                         # only for the first time.
5  devtools::install_github("AlradSpectra/AlradSpectra") # Now, install Alrad Spectra from GitHub.
6  library(AlradSpectra)                 # Load Alrad Spectra.
7  AlradSpectra()                         # Initialize AlradSpectra.

```

74

75 Fig. 1. Commands to install, load and initialize Alrad Spectra in R.

76

77 4.3.GUI DESCRIPTION

78 Alrad Spectra was designed using the toolkit implementation of RGtk2 package [4],
 79 which facilities the R language for programming graphical interfaces using Gtk (Gimp Tool
 80 Kit). The required packages to build Alrad Spectra for each stage are listed in Table 1.

81

82 Table 1. Packages required to implement Alrad Spectra.

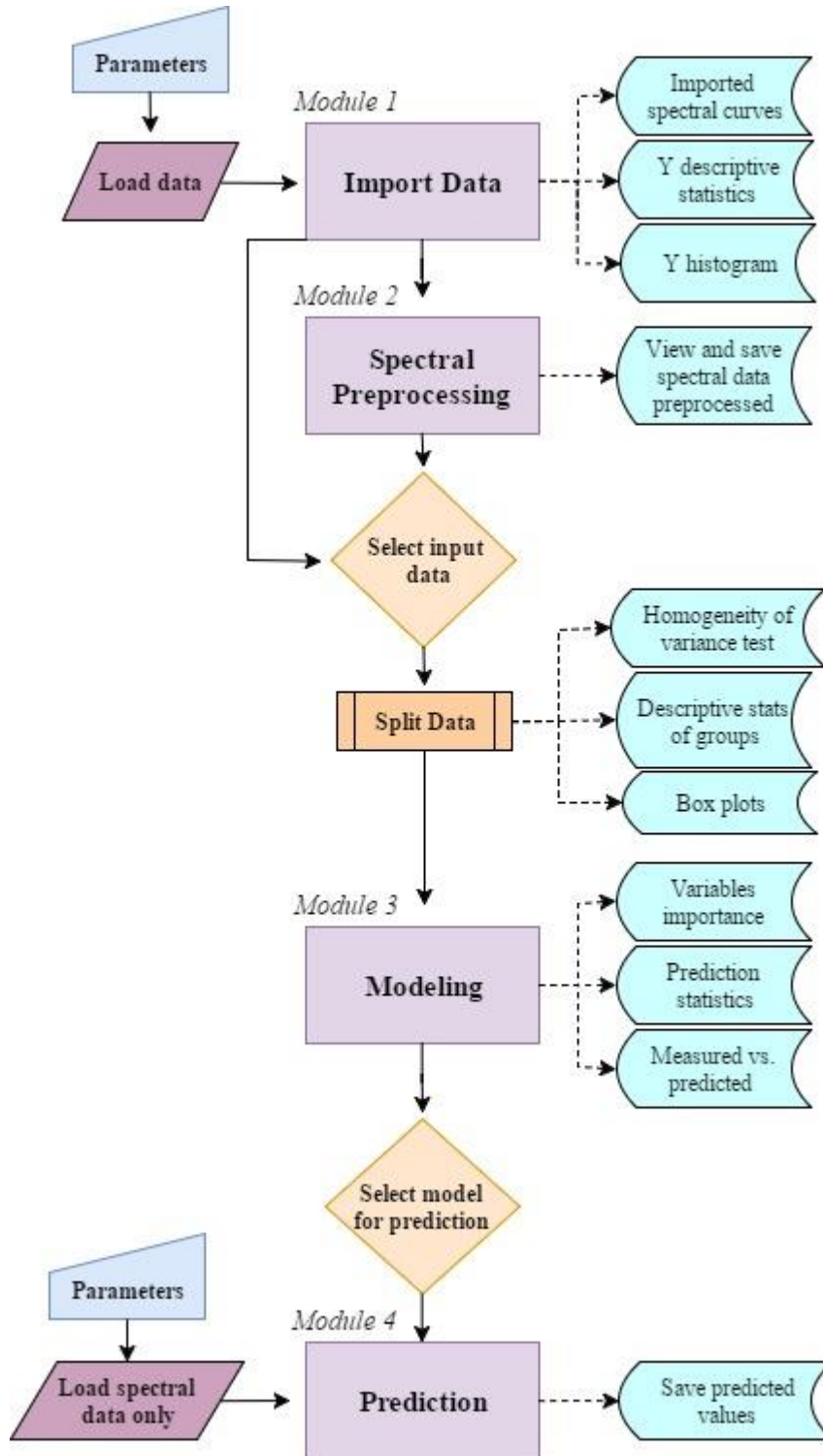
Component	R Package*	Reference
Graphical Integration	devtools	[5]
	gWidgetsRGtk2	[6]
Descriptive statistics	fitdistrplus	[7]
Levene's Test	car	[8]
Plots	ggplot2	[9]
	graphics	[1]
	gridExtra	[10]
Spectral Preprocessing	clusterSim	[11]
	pls	[12]
	prospectr	[13]
Modeling and Prediction	caret	[14]
	e1071	[15]

	<code>elmNN</code>	[16]
	<code>kernlab</code>	[17]
Modeling and Prediction	<code>pls</code>	[12]
	<code>randomForest</code>	[18]

83 * Package dependencies are also installed.

84

85 The diagram of Alrad Spectra development showing the workflow in sequential order is
86 illustrated in Fig. 2. Alrad Spectra interface has a main menu with four different modules, which
87 are titled: Import Data, Spectral Preprocessing, Modeling and Prediction. The first module is
88 used to import data, view the imported data in tabular form, view the imported spectral curves,
89 and view the descriptive statistics and histogram of the Y variable. After importing the data, the
90 next module performs the desired spectral preprocessing. In Modeling module, the interface
91 automatically loads the original or the preprocessed spectra, when previously executed,
92 allowing the selection of input data for modeling. Next, the size of validation set must be
93 selected. After selecting the preprocessing and setting the validation set size, the user is able to
94 split the data into training and validation groups. After splitting, the user can also test the
95 homogeneity of variances of the groups, view descriptive statistics and view a boxplot of
96 training and validation groups. There are six modeling methods present in Alrad Spectra. Each
97 method offers tuning parameters. The tuning is essentially selecting the best parameters for an
98 algorithm to optimize its modeling performance given a working environment. The Prediction
99 module can validate the models using an independent data set and predict the Y variable using
100 spectroscopic data only. The four main modules are described individually in the subsequent
101 sections.
102



103

104 Fig. 2. Flowchart of Alrad Spectra.

105

106 **4.3.1.Import Data module**

107

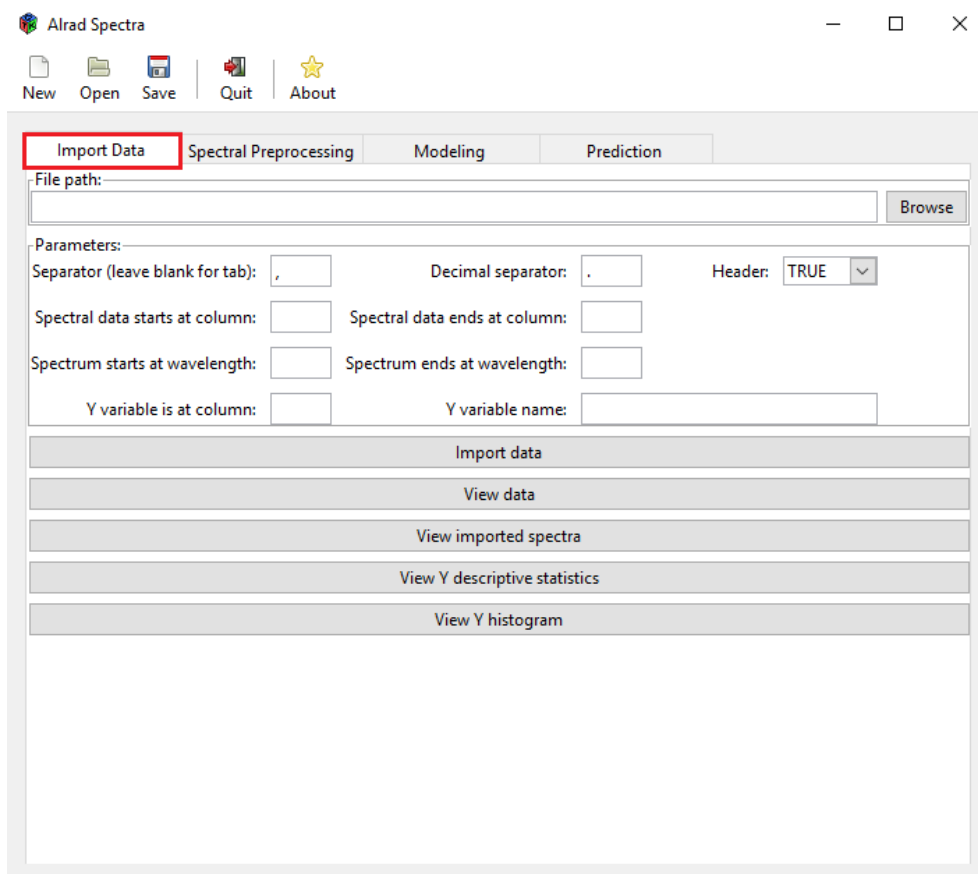
108

109

The graphical interface of Alrad Spectra is presented in Fig. 3. The Import Data module enables the user to load data in text file (.txt) or comma-delimited values (.csv) file formats by clicking the browsing the file or typing the file path. The samples have to be placed in rows and

110 the variables in columns. The user needs to set the file parameters as follow: file separator
 111 (usually comma, semicolon or tab), decimal separator (dot or comma), whether the file has a
 112 header (first row has column names), write in which column the spectral data starts and ends,
 113 write the first and last wavelength of the spectrum, and lastly, indicate the column that contains
 114 the Y variable and give it a name. These parameters will be required in preprocessing and
 115 modeling processes. The ‘Import file’ runs the commands to load the data, the ‘View data’
 116 shows the data as a table, and the ‘View imported spectra’ shows the original spectral curves,
 117 while the ‘View Y descriptive statistics’ shows the descriptive statistics of the Y variable in a
 118 text dialog (`fitdistrplus` package). The ‘View Y histogram’ displays a colorful histogram
 119 of the Y variable (`ggplot2` package). All images can be saved using the ‘Save plot’ the plot
 120 window.

121



122

123 Fig 3. Graphical user interface of Alrad Spectra showing the Import Data module.

124

125 4.3.2.Spectral Preprocessing module

126 The Spectral Preprocessing module will be functional only after properly data file
 127 importation in the first module. Alrad Spectra offers a total of nine preprocessing: smoothing,

128 binning, absorbance, detrend, continuum removal, Savitzky-Golay derivatives (SGD), standard
 129 normal variate (SNV), multiplicative scatter correction (MSC), and normalizations. They are
 130 the most commonly used preprocessing steps of spectra. In each preprocessing tab, there is a
 131 ‘View spectra’ button, which allows to view the preprocessed spectral curves, which can be
 132 saved by ‘Save plot’ in the plot window. The ‘Save preprocessed spectra’ permits to save the
 133 spectral data in text file (.txt) or comma-delimited values (.csv) file formats. A selection of
 134 preprocessing contains parameters to be defined by the user. Spectral preprocessing
 135 descriptions and mathematical procedures are discussed in the next section.

136

137 *4.3.2.1.Smoothing*

138 It is a simple moving average of a spectral data using a convolution function [13]. The
 139 moving average is the most common smoothing in spectral data, mainly because it is the easiest
 140 and comprehensible filter. The moving average smoothing is ideal for reducing random noise
 141 while retaining a sharp step response. In this preprocessing the user is requested to choose the
 142 number of smoothing points. In the `prospectr` package, the smoothing is implemented by the
 143 `movav` function. The equation of moving average filter is written below (Eq. 1).

144

$$S_i = \frac{1}{M} \sum_{j=0}^{M-1} x [i+j] \quad (1)$$

145 where, x is the original reflectance value ($i = 1, 2, \dots, N$), S_i is the output signal, and M is the
 146 number of points used in the moving average.

147

148 *4.3.2.2.Binning*

149 Binning is a preprocessing technique used to reduce the effects of minor observation
 150 errors by computing average values of a spectral data. To perform spectral binning, the bin size
 151 has to be specified. The original spectral data values are replaced by a value representative of
 152 that interval (bin size). Spectral binning is a common technique used for high-throughput data
 153 preprocessing. The binning is implemented by the `binning` function in the `prospectr`
 154 package.

155

156

157

158 *4.3.2.3.Absorbance*

159 Absorbance spectroscopy is an analytical technique based on measuring the amount of
 160 light absorbed by a sample at a given wavelength. The reflectance to absorbance transformation
 161 is obtained by running equation 2 in R console.

162

$$A = \log_{10} 1/R \quad (2)$$

163 where, A is the Absorbance, \log_{10} is the logarithm base 10, and R is the Reflectance.

164

165 *4.3.2.4.Detrend*

166 Detrend normalizes the spectral data by applying a standard normal variate
 167 transformation followed by fitting a second-degree polynomial regression model and returning
 168 the fitted residuals [19]. Detrend is often applied to remove the effects in the cases where a
 169 constant, linear, or curved offset is present in the spectral curve. The effect of detrend is to
 170 remove low-frequency variance. Detrending in essence is equivalent to high-pass filtering. For
 171 example, the variance at low frequencies is diminished relative to variance at high frequencies.
 172 The Detrend preprocessing applies the `detrend` function in the `prospectr` package.

173

174 *4.3.2.5.Continuum removal (CR)*

175 The CR technique, proposed by Clark and Roush [20], consists of removing the
 176 continuous features of the spectra and is often used to isolate specific absorption features
 177 present in the spectrum to minimize the noise. The continuum is represented by a mathematical
 178 function used to separate and highlight specific absorption bands of the reflectance spectrum
 179 [21]. The CR is computed by identifying the local reflectance spectrum maxima points, and
 180 then, these points are connected by linear interpolation to form the continuum reflectance. The
 181 `continuumRemoval` function allows to compute the continuum removed values from
 182 `prospectr` package. The parameters to be defined are the number of smoothing points, order
 183 of polynomial, and order of derivative. The mathematical description of CR is presented below
 184 (Eq. 3).

185

$$\varphi_i = \frac{x_i}{c_i}; i = \{1, \dots, p\} \quad (3)$$

186 where, x_i is the original reflectance values and c_i is the continuum reflectance values at the i^{th}
 187 wavelength of a set of p wavelengths, and φ_i is the final reflectance value after continuum
 188 removed.

189

190 4.3.2.6. *Savitzky–Golay derivative (SGD)*

191 Derivatives are a common technique performed to remove unimportant baseline signal
 192 from samples by taking the derivative of the measured responses with respect to the variable
 193 number (wavelength). This preprocessing has the ability to remove both additive and
 194 multiplicative effects in spectra. The Savitzky-Golay derivatization algorithm [22] requires
 195 selection of smoothing points (filter width), the orders of polynomial and derivative. The SGD
 196 is implemented by the `savitzkyGolay` function in the `prospectr` package. The mathematical
 197 description of SGD is given by equation 4.

198

$$x_j = \frac{1}{N} \sum_{-m}^m c_h x_{j+m} \quad (4)$$

199 where, x_j is the new value, N is a normalizing coefficient, m is the number of neighbor values
 200 at each side of j and c_h are pre-computed coefficients, that depends on the chosen polynomial
 201 and derivative orders.

202

203 4.3.2.7. *Standard normal variate (SNV)*

204 SNV is frequently performed in spectral data to remove scatter. It is applied to every
 205 spectrum individually. The average and standard deviation of all points for the spectrum is
 206 calculated. Every data point of the spectra is subtracted from the mean and divided by the
 207 standard deviation. SNV is designed to operate based on centering the underlying linear slope
 208 of each individual sample spectrum (Eq. 5) [19]. The SNV is implemented by
 209 `standardNormalVariate` function in `prospectr` package.

210

$$SNV = \frac{x_i - \bar{x}_i}{s_i} \quad (5)$$

211 where, x_i is the original reflectance, \bar{x}_i is the mean the original reflectance, s_i is the standard
 212 deviation of the original reflectance.

213

214

215 4.3.2.8. *Multiplicative scatter correction (MSC)*

216 MSC is achieved by regressing a measured spectrum against a reference spectrum and
 217 then correcting the measured spectrum using the slope and intercept of this linear fit. This
 218 preprocessing technique has proven to be effective in minimizing baseline offsets and
 219 multiplicative effect [23]. The outcome of MSC, in many cases, is very similar to SNV, except
 220 SNV corrects each spectrum individually and does not need the entire data set. The `pls` package
 221 includes the `msc` function for MSC preprocessing in R. The mathematical description of MSC
 222 is given by equation 6.

223

$$MSC = \frac{x_i - a_i}{b_i} \quad (6)$$

224 where, x_i is the original reflectance value, a_i and b_i are the regression coefficients for sample i .

225

226 4.3.2.9. *Normalization*

227 Normalization means adjusting values measured on different scales to a common scale.
 228 Simple normalization is a common approach to multiplicative scaling problem. Normalization
 229 preprocessing refers to the creation of shifted and scaled versions of spectral data, where these
 230 normalized values eliminate scattering effects [2]. If the relationship between variables is the
 231 most important aspect of spectral data, then normalization is recommended. Five types of
 232 normalization were included in Alrad Spectra: standardization, normalization in range, quotient
 233 transformation, normalization, and normalization with zero being the central point.
 234 Normalization preprocessing algorithms are implemented by `data.Normalization` function
 235 in `clusterSim` package.

236

237 **4.3.3. Modeling module**

238 In the Modeling module, the first step requires to select the input data for modeling
 239 process. In the combo box, will be display the imported spectral data, called Original and the
 240 spectral preprocessing names if previously performed. When the preprocessing is performed
 241 more than one times (i.e. using different parameters, when available) the preprocessed data
 242 selected in this step corresponds to the last preprocessing. After selecting the input data, the
 243 user chooses the size of the validation set, in percentage. The split data is accomplished by
 244 randomly dividing the observation samples. The selection of validation set ranges from 5% to
 245 50%. The samples that are not included in the validation are used for training the models. Only
 246 after completing the split data, the homogeneity test, descriptive statistics and boxplot can be

247 accomplished and the multivariate methods tab can be manipulated. Levene's test for
248 homogeneity of variances was implemented to verify the assumption that variances are equal
249 across random selection of validation and training groups. The descriptive statistics and the box
250 plot of Y variable can be visualized using their respective buttons. To perform the modeling
251 with different preprocessing, the user must select the preprocessing of interest and repeat the
252 split data by clicking the 'Split data' button. The modeling covers different methods, including
253 multiple linear regression [24], partial least squares regression [25], support vector machines
254 [26], random forest [27], artificial neural network [28], and Gaussian process regression [29].
255 In each method, tuning parameters are presented in order to achieve the best fitted model. The
256 `trainControl` function in `caret` package generates parameters that further control how
257 models are created, with possible values. One of these parameters are the resampling method,
258 which is implemented to adjust the best fitted models. The resampling methods utilized are 'cv'
259 (K-fold cross-validation), 'repeatedcv' (repeated K-fold cross-validation), 'LOOCV' (leave-
260 one-out cross-validation), 'LGOCV' (leave-group-out cross-validation), 'boot' (bootstrap),
261 'boot632' (0.632 bootstrap), 'oob' (out-of-bag error estimates, only for tree models), and
262 'none'. For 'LOOCV', no uncertainty estimates are given for the resampled performance
263 measures. The number of folds and resampling iterations controls the number of folds in 'cv'
264 and number of resampling iterations for 'boot' and 'LGOCV'. The number of repetition applied
265 only to 'repeatedcv'. The model building and estimation process is achieved by the `caret`
266 package. This package has a set of functions that attempt to streamline the process for creating
267 predictive models. The `train` function can be used to evaluate the effect of model performance
268 using optimal tuning parameters [14]. Once the modeling is completed, 'View variables
269 importance' shows the importance of each variable for the model in a scale of 0 to 100. The
270 'Prediction statistics' shows the training and validation statistical assessments, and 'View
271 measured vs. predicted' shows the scatterplot for training and validation groups with its
272 prediction statistics. In PLSR model, the partial least squares (PLS) components vs. RMSE
273 values figure was included. The modeling methods used in Alrad Spectra are discussed in the
274 sections bellow.

275

276 4.3.3.1. Multiple linear regression (MLR)

277 MLR is a statistical method that uses several explanatory variables to predict the outcome
278 of a response variable in a simple linear model [30]. MLR assumes the relationships between
279 independent variables and dependent variable are linear. Another important assumption is
280 absence of multicollinearity, the independent variables are not highly correlated, presence of

281 homoscedasticity and normality. Presuming these assumptions, a robust prediction can be
282 achieved using a relatively simple algorithm. The tuning parameter in MLR method to be
283 defined by the user are the band interval, resampling method, number of folds or resampling
284 iterations, and number of repetitions. MLR is implemented by the generalized linear model with
285 stepwise feature selection and the best fitted model is chosen by Akaike information criterion
286 (AIC) [31]. The `glmStepAIC` function, in the `caret` package, is applied in the context of model
287 selection to find the best fitted model involving a subset of predictors.

288

289 *4.3.3.2. Partial least squares regression (PLSR)*

290 PLSR can handle complicated relationships between predictors and responses, and
291 moreover, can deal with complex modeling problems [25]. Additionally, PLSR is a method for
292 constructing predictive models when the factors are many and highly collinear [32], which is
293 the case of hyperspectral data. PLSR has become a popular technique used in chemometrics
294 that is used for quantitative analysis of d reflectance spectra. [33]. The PLSR model is tuned by
295 the `caret` package and the best parameters are employed to adjust the final model by the `pls`
296 function available in the `pls` package. In the PLSR model, the tuning parameters are resampling
297 method, number of folds or resampling iterations, number of repetitions, and number of
298 components to include in the model.

299

300 *4.3.3.3. Support vector machines (SVM)*

301 SVM are a group of supervised learning methods, which represent an extension to
302 nonlinear models of generalized algorithm with the capability of training nonlinear classifiers
303 [34]. Associated with SVM algorithm is the criteria of smaller number of support vectors yield
304 a better model performance [35]. SVM models are efficient in modeling linear or nonlinear
305 relationships and handling large databases. The `caret` package tunes the SVM model and the
306 best parameters are employed to adjust the final model by `svm` function available in the `e1071`
307 package. The tuning parameters for SVM are resampling method, number of folds or
308 resampling iterations, number of repetitions, and Linear or Radial kernels.

309

310 *4.3.3.4. Random forest (RF)*

311 Random forests are a combination of tree predictors such that each tree depends on the
312 values of a random vector sampled independently and with the same distribution for all trees in
313 the forest [27]. RF are versatile and flexible with small or large data sets and has becoming an
314 effective tool in prediction. Model interpretability is an issue when compared to linear models.

315 RF models are black boxes approach that are very hard to interpret. The tuning model is
316 executed by the `caret` package, while the final model is performed by the `randomForest`
317 function in the `randomForest` package. In Alrad Spectra, the tuning parameters for RF are:
318 resampling method, number of folds or resampling iterations, number of repetitions, randomly
319 selected predictors (`mtry`), and number of trees (`ntree`).

320

321 *4.3.3.5. Artificial neural network (ANN)*

322 In ANN, the mathematical model assigns weights between elements, and a network
323 structure is adjusted depending on the inputs. This method implements the extreme learning
324 machine algorithm for the single hidden layer feedforward neural networks [36]. First, it
325 generates input weights and hidden layer bias, then calculates the output from the hidden layer
326 based on the activation function. At the end, the trained neural network model is returned. The
327 tuning parameters (`caret` package) present in the GUI for the ANN modeling are: resampling
328 method, number of folds or resampling iterations, number of repetitions, activation function,
329 and hidden units (number of hidden neurons). The type of activation function are: ‘`purelin`’
330 (linear), ‘`radbas`’ (radial basis), ‘`sin`’ (sine), and ‘`tansig`’ (tan-sigmoid). The `elmtrain` function
331 in `elmNN` package employs the best tuned parameters and perform the final ANN model.

332

333 *4.3.3.6. Gaussian process regression (GPR)*

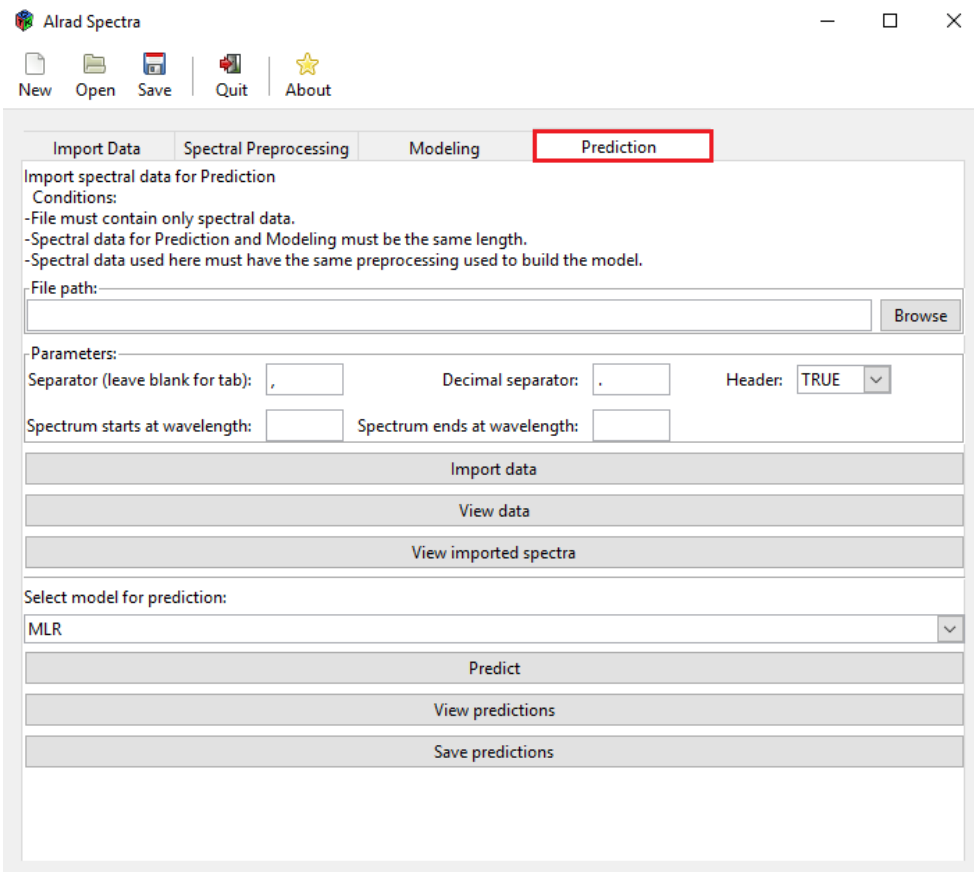
334 When the task is to predict an output value, it is possible to carry out nonparametric
335 regression using Gaussian processes. The solution for the regression problem under a Gaussian
336 process is to place a kernel function on each training data point [29]. Gaussian process applies
337 a kernel function for training and predicting. In machine learning, kernel methods are a class of
338 algorithms for pattern analysis. For many algorithms that solve regression problems, the data
339 have to be explicitly transformed into feature vector representations. In contrast, kernel methods
340 require only a user-specified kernel. This is called ‘kernel trick’ replacing its features
341 (predictors) by a kernel function. Several classes of kernels can be used for machine learning
342 and the selection of kernel is critical to the success of these algorithms [17]. In Alrad Spectra,
343 Linear and Radial kernels are included as tuning parameters. Furthermore, the other tuning
344 parameters are resampling method, number of folds or resampling iterations, number of
345 repetitions, and initial noise variance. The `caret` package was used to train and tune the
346 parameters for the model. The `gausspr` function in `kernlab` package performed the GPR final
347 model.

348

349 **4.3.4.Prediction module**

350 The Prediction module (Fig. 4) is implemented in order to predict the Y variable using
 351 the built models using spectroscopic data only. The prediction process requires the following
 352 conditions: file must contain only spectral data, spectral data for Prediction and Modeling must
 353 be the same length, and spectral data used in Prediction must have the same preprocessing used
 354 to build the model. The first step to perform the Prediction is to import a new data set containing
 355 the spectral data only (samples in rows and spectral variables in columns). It is possible to
 356 observe the imported data as a table in 'View data', and verify the spectral curves by 'View
 357 imported spectra'. The prediction is performed by selecting the model built previously. In
 358 'View predictions' and 'Save predictions' buttons, it is possible to obtain the predicted values
 359 and save the results.

360



361

362 Fig. 4. Graphical user interface of Alrad Spectra showing the Prediction module.

363

364

365

366 4.4.CASE STUDY

367 **4.4.1.Data set**

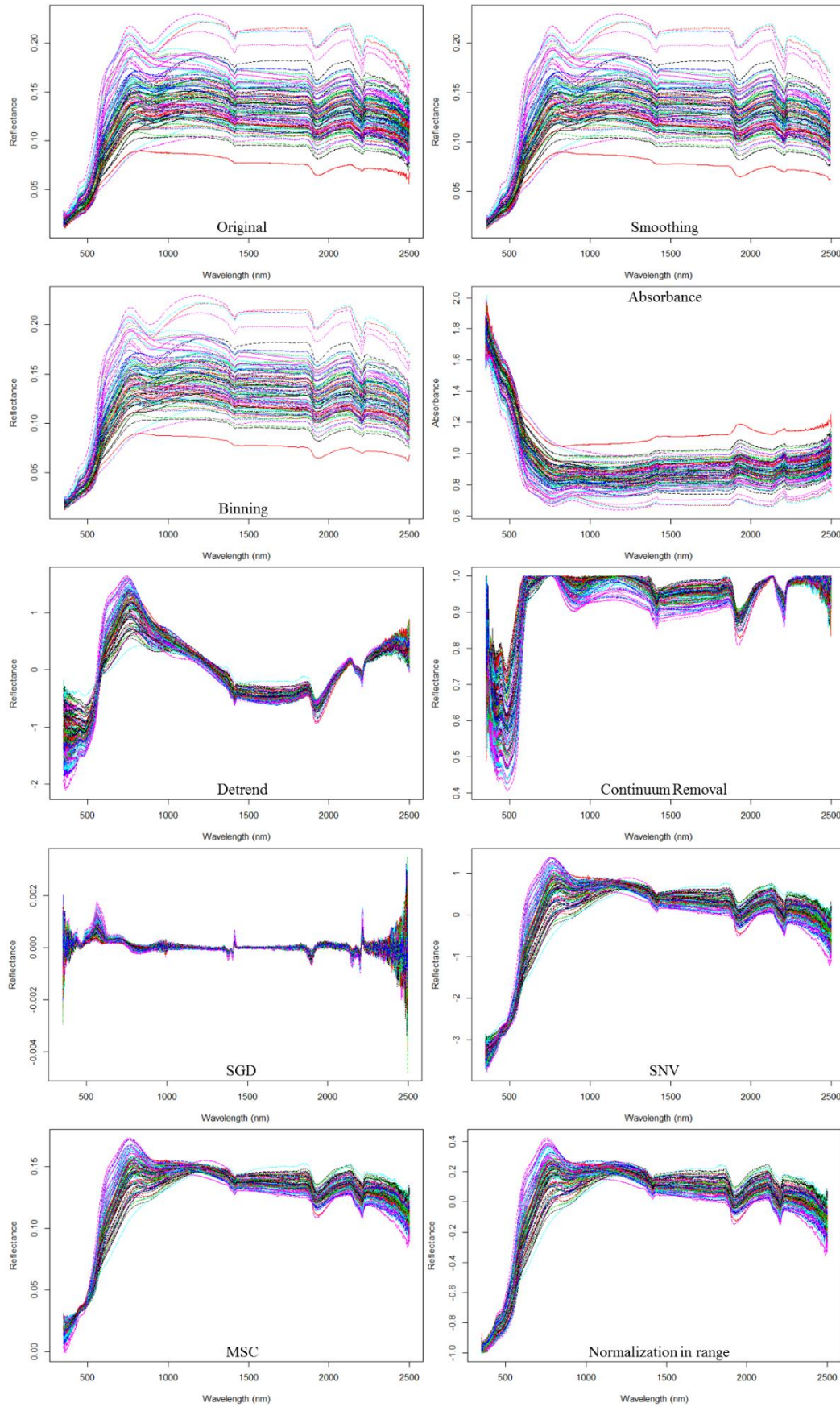
368 The soil spectral data example used to test Alrad Spectra consists of 595 soil samples.
369 The soil samples were located in central region of Santa Catarina State, Brazil. The
370 experimental data contains the value of SOC and Vis-NIR reflectance. SOC content was
371 determined through the traditional laboratory analysis by wet combustion using the Mebius
372 method in the digestion block [37]. Soil spectral reflectance was obtained using a FieldSpec 3
373 spectroradiometer (Analytical Spectral Devices, Boulder, USA) with a spectral range of 400–
374 2500 nm (Vis-NIR) with 1 nm of spectral resolution. The soil data file is placed and available
375 in the user's R library, inside AlradSpectra/exdata directory, e.g.,
376 "C:\Users\UserName\Documents\R\win-library\3.3\AlradSpectra\extdata". The first 95 soil
377 samples were applied in Prediction module as soils with the unknown SOC value and the
378 subsequent 500 soil samples were used in the Modeling process. The 500 samples were
379 randomly split into 70% and 30% to train and validate the models, respectively.

380

381 **4.4.2.Soil spectral preprocessing**

382 The soil spectral data file was imported in Import Data module by establishing the
383 parameters: the file separator was comma, decimal separator was dot, header was true, the
384 spectral data started at column 4 and ended at column 2104, the spectrum number started at 400
385 nm and ended at 2500 nm, and the Y variable was at column number 3, and was named Soil
386 Organic Carbon (%). The Descriptive statistics of whole SOC values are shown in Table 2. The
387 original (initial) spectral curves imported along with all spectral preprocessed curves can be
388 visualized in Fig. 5 and evaluated qualitatively. The spectral reflectance curves showed the
389 diversity of soils by its shape and the presence or absence of absorption bands. Categorization
390 of soil reflectance has important implications for soil genesis, classification, and survey [38].
391 The smoothing preprocessing example was accomplished with 11 smoothing points. For
392 binning preprocessing, it was applied 10 bins size. In the SGD, it was applied 5 smoothing
393 points, first order of polynomial and first order of derivative. Normalization in range was
394 applied in the normalization preprocessing. The absorbance, detrend, CR, SNV, and MSC
395 preprocessing do not have parameters to be set and were also performed.

396



397
398
399

Fig. 5. The original and preprocessed spectral curves performed in Alrad Spectra.

400 4.4.3. Modeling and prediction of SOC

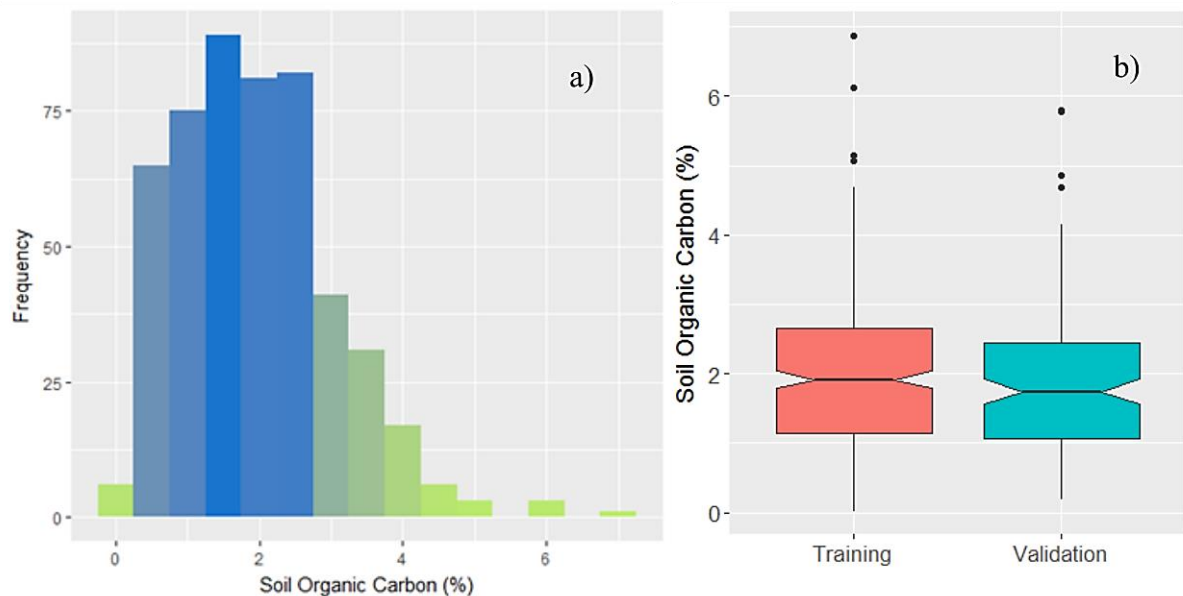
401 A predictive model was built by each of multivariate calibration methods. The original
 402 spectral data without preprocessing plus the nine-spectral preprocessing were used as
 403 independent variables to build the models. The Levene's test for homogeneity of variance
 404 presented a p-value of 0.918 which is greater than significance level of 5%. This result indicate
 405 that the training and validation groups were homogeneous and suitable for the modeling stage.
 406 The descriptive statistics of training and validation groups are presented in Table 2. The
 407 histogram of whole SOC values and the box plot of training and validation groups are presented
 408 in Fig. 6.

409

410 Table 2. Descriptive statistics of SOC for whole, training and validation sets.

SOC (%)	N	Min	Max	Mean	Median	Std Dev.	Skewness	Kurtosis
Whole set	500	0.02	6.87	1.95	1.86	1.08	0.79	4.06
Training set	350	0.02	6.87	1.98	1.87	1.11	0.88	4.38
Validation set	150	0.21	4.69	1.86	1.84	1.00	0.46	2.62

411



412

413 Fig. 6. Histogram (a), and box plot of training and validation groups (b) for SOC performed in
 414 Alrad Spectra.

415

416

417 The prediction statistic assessment of SOC models are shown in Table 3. The results are
418 ordered by the smallest RMSE for each method. For all models, the ‘cv’ resampling method
419 with 10 k-folds were set as tuning parameters, except for RF, which the resampling method was
420 ‘oob’. For the MLR models, the band interval parameter was 25 for all models. The outcomes
421 of MLR models showed that the greatest SOC prediction was achieved when SNV
422 preprocessing was applied, reaching a R_{val}^2 of 0.81, RMSE_{val} of 0.51%, and RPIQ_{val} of 3.20.
423 The R_{val}^2 of all models ranged from 0.54 to 0.81. In the PLSR models, the performances were
424 similar than MLR, with the R_{val}^2 ranging from 0.56 to 0.81. The greatest SOC prediction was
425 also achieved by SNV preprocessing once more with a R_{val}^2 of 0.81, RMSE_{val} of 0.51%, and
426 RPIQ_{val} of 2.84. In the validation performance, seven preprocessing exhibited R^2 above 0.75.
427 PLSR obtained the highest R_{val}^2 value over all SOC prediction model. The PLSR models were
428 performed using 30 components.

429 For the training set, several SVM models presented a high performance, in which most
430 of preprocessing are considered well-fitted models with the results in predicted values similar
431 to the observed values. For the validation set, the best performance was achieved by absorbance
432 preprocessing with a R_{val}^2 of 0.78, RMSE_{val} of 0.51%, and RPIQ_{val} of 2.55. The CR
433 preprocessing presented the unreliable performance in SOC prediction with SVM (R_{val}^2 of
434 0.61). However, in the RF models, CR preprocessing showed one of the best SOC prediction
435 performance. The RF method showed a weak performance for original, binning, absorbance
436 preprocessing, with a R_{val}^2 ranging from 0.37 to 0.43. For SVM models, the tuning parameter
437 was Support Vector Machine with Linear Kernel, and for RF were 5 randomly predictors and
438 500 trees.

439 In the validation of ANN models, the preprocessing presented unreliable outcomes. The
440 higher performance in SOC prediction was found for SNV preprocessing (R_{val}^2 of 0.54; RMSE
441 of 0.71%) followed by original preprocessing (R_{val}^2 of 0.54; RMSE of 0.75%). The ANN model
442 with SGD preprocessing presented the smaller SOC predictive performance (R_{val}^2 of 0.15;
443 RMSE of 0.99%). In ANN models, the tuning parameters applied were ‘purelin’ activation
444 function and 10 hidden units. GPR models can lead to substantial improvements in training the
445 models which led to a high accuracy for training samples. However, when the model is
446 validated the prediction statistics showed more sensible outcomes. Observing the results of
447 validation set, the R^2 value oscillated from 0.48 to 0.77, where the higher performance was
448 achieved by absorbance preprocessing. In GPR, the tuning parameters for the modeling was
449 composed of Linear kernel function.

Table 3. The prediction statistics of SOC for each model.

Method	Preprocessing	Training set			Validation set		
		R ²	RMSE (%)	RPIQ	R ²	RMSE (%)*	RPIQ
MLR	SNV	0.84	0.43	3.24	0.81	0.51	3.20
	Smoothing	0.80	0.48	3.07	0.77	0.52	2.77
	Detrend	0.84	0.44	3.37	0.76	0.52	2.76
	CR	0.86	0.41	3.82	0.76	0.53	2.39
	Absorbance	0.84	0.43	3.58	0.76	0.53	2.59
	Normalization	0.84	0.41	3.49	0.78	0.55	2.90
	Original	0.80	0.48	3.00	0.72	0.59	2.68
	MSC	0.85	0.40	3.50	0.75	0.61	2.70
	Binning	0.63	0.65	2.18	0.57	0.71	2.19
	SGD	0.74	0.56	2.75	0.54	0.72	1.73
PLSR	SNV	0.84	0.42	3.47	0.81	0.51	2.84
	Detrend	0.83	0.46	3.24	0.75	0.51	2.83
	CR	0.86	0.40	3.91	0.78	0.53	3.08
	Absorbance	0.84	0.43	3.53	0.76	0.53	2.61
	Normalization	0.82	0.44	3.31	0.79	0.54	2.94
	Original	0.76	0.51	2.77	0.75	0.56	2.83
	MSC	0.85	0.40	3.72	0.76	0.57	2.55
	Binning	0.78	0.50	2.84	0.71	0.59	2.64
	Smoothing	0.79	0.50	2.96	0.70	0.60	2.41
	SGD	0.75	0.54	2.85	0.56	0.71	1.77
SVM	Absorbance	0.86	0.41	3.79	0.78	0.51	2.55
	SNV	0.95	0.26	5.70	0.74	0.52	2.75
	Normalization	0.94	0.26	5.81	0.75	0.53	2.48
	Original	0.80	0.48	2.98	0.74	0.56	2.81
	MSC	0.95	0.24	6.18	0.73	0.61	2.38
	Smoothing	0.80	0.51	3.04	0.68	0.62	2.03
	Binning	0.79	0.49	2.83	0.69	0.63	2.60
	Detrend	0.98	0.15	9.31	0.66	0.72	2.09
	SGD	0.99	0.10	14.16	0.53	0.77	1.93
	CR	0.99	0.10	14.27	0.61	0.85	1.61

	Detrend	0.67	0.66	2.26	0.67	0.57	2.50	
	CR	0.73	0.58	2.70	0.69	0.60	2.13	
	SGD	0.68	0.66	2.32	0.58	0.71	1.76	
	Smoothing	0.38	0.89	1.73	0.44	0.71	1.79	
RF	SNV	0.60	0.70	2.15	0.51	0.72	1.87	
	MSC	0.55	0.70	2.01	0.61	0.77	2.13	
	Normalization	0.55	0.70	2.05	0.60	0.79	2.01	
	Binning	0.39	0.84	1.69	0.40	0.84	1.85	
	Absorbance	0.40	0.84	1.83	0.37	0.85	1.60	
	Original	0.40	0.82	1.72	0.43	0.86	1.88	
	<hr/>							
		SNV	0.58	0.71	2.13	0.54	0.71	1.91
		Original	0.60	0.67	2.10	0.54	0.75	2.11
		Detrend	0.47	0.81	1.83	0.44	0.76	1.91
	Smoothing	0.43	0.86	1.80	0.40	0.76	1.67	
ANN	MSC	0.53	0.71	1.98	0.51	0.83	1.98	
	Normalization	0.50	0.73	1.98	0.49	0.84	1.89	
	CR	0.45	0.81	1.95	0.35	0.85	1.49	
	Absorbance	0.45	0.80	1.91	0.36	0.86	1.59	
	Binning	0.42	0.82	1.74	0.31	0.90	1.72	
	SGD	0.19	0.98	1.57	0.15	0.99	1.27	
	<hr/>							
		Absorbance	0.85	0.42	3.69	0.77	0.52	2.65
		Normalization	0.93	0.27	5.31	0.76	0.57	2.77
		SNV	0.95	0.26	5.84	0.72	0.58	2.34
	Original	0.81	0.46	3.06	0.73	0.58	2.75	
GPR	MSC	0.92	0.29	4.87	0.76	0.59	2.81	
	Detrend	0.97	0.21	7.11	0.69	0.60	2.38	
	Binning	0.72	0.57	2.48	0.64	0.65	2.38	
	Smoothing	0.80	0.50	3.07	0.65	0.65	1.94	
	CR	0.99	0.11	14.08	0.61	0.73	1.75	
	SGD	0.99	0.00	461.00	0.48	0.83	1.51	

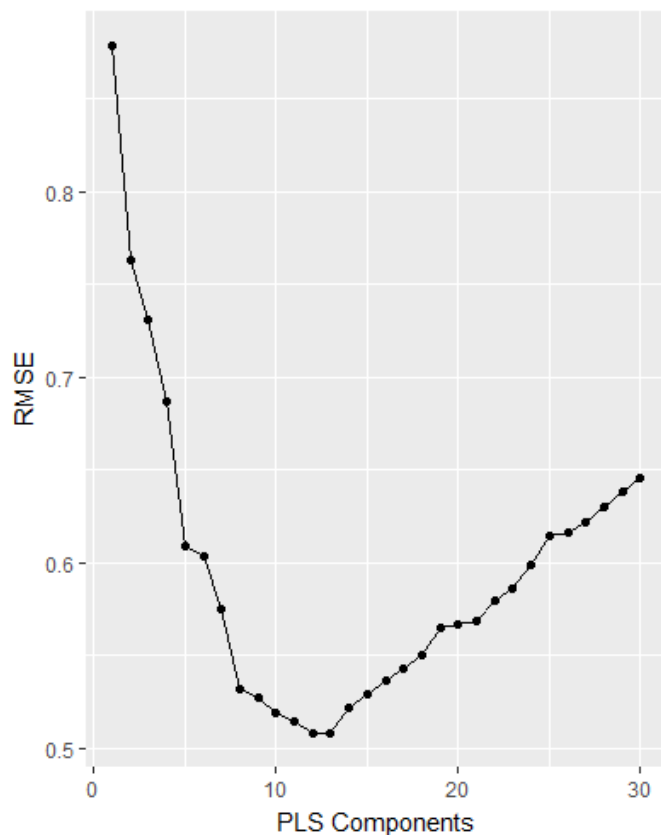
450 * The results are ordered by the smallest $RMSE_{val}$ for each method. Multiple linear regression
451 (MLR), partial least squares regression (PLSR), support vector machines (SVM), random forest
452 (RF), artificial neural network (ANN), Gaussian processes regression (GPR), continuum

453 removal (CR), Savitzky–Golay derivative (SGD), standard normal variate (SNV),
 454 multiplicative scatter correction (MSC).

455

456 The PLSR method with SNV (PLSR-SNV) spectral preprocessing yielded the greatest
 457 SOC prediction performance. The PLSR is able to show the RMSE values of all PLS
 458 components utilized to build the model (Fig. 7). The smallest RMSE was achieved with 13 PLS
 459 components, which means that this model needed 13 PLS components to achieved the best
 460 performance. The variables importance of PLSR-SNV is shown in Fig. 8a. In this figure, the
 461 importance of whole spectral variables was revealed. The most important variables in PLSR
 462 model were around 2200 nm and 1414 nm. However, there were spectral bands in the entire
 463 Vis-NIR range that presented high importance in SOC prediction. In addition, Fig. 8b provides
 464 the measured vs. predicted SOC values, with the 1:1 line for training and validation groups.
 465 The statistics assessment, R^2 , RMSE, and RPIQ, were displayed for training and validation
 466 groups. The closer to the 1:1 line the samples were the better was the prediction. The soil
 467 samples with high SOC content showed high dispersion in relation to 1:1 line.

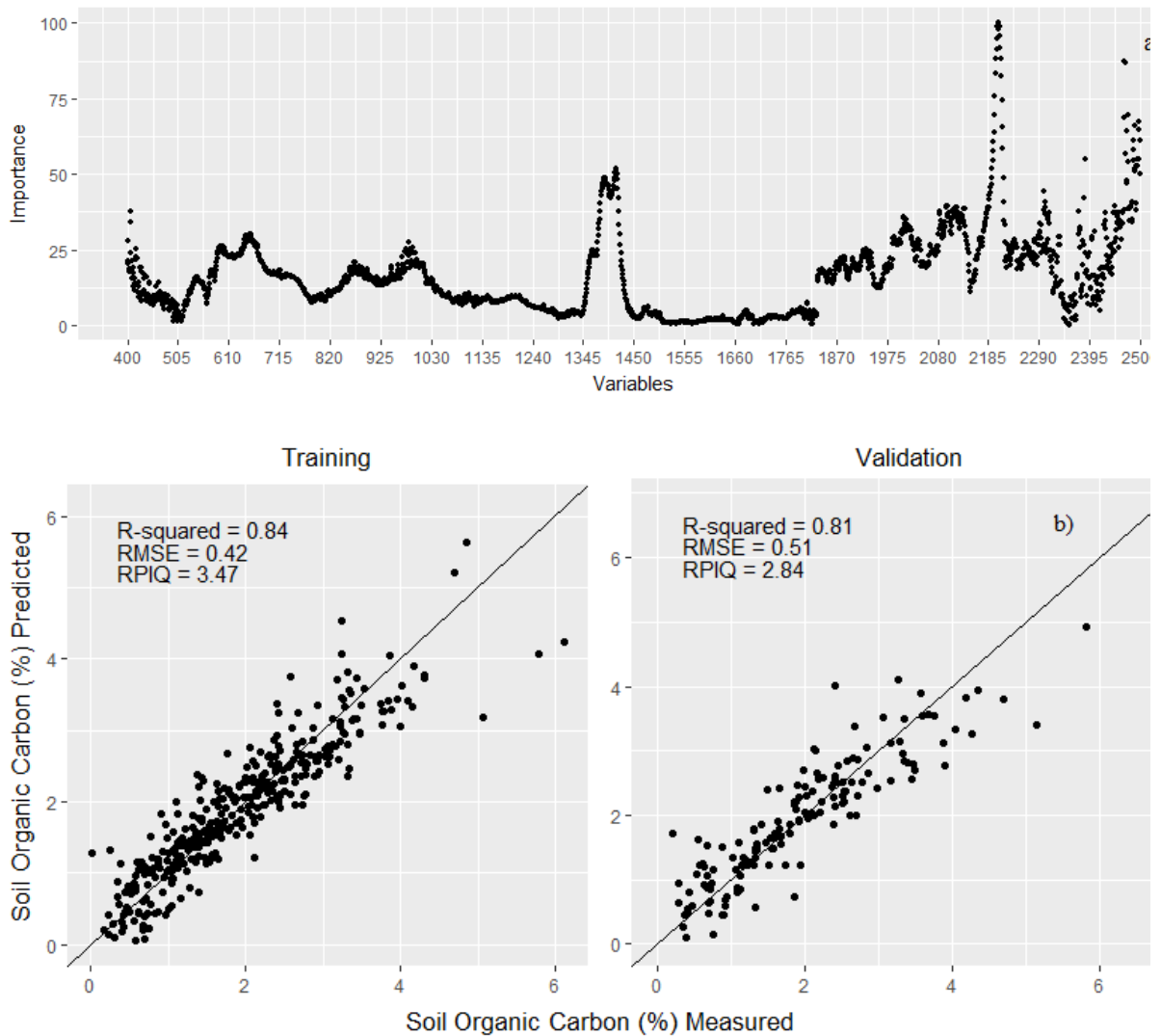
468



469

470 Fig. 7. The 30 partial least squares components vs. RMSE values of PLSR model with SNV
 471 preprocessing performed in Alrad Spectra.

472



473 Fig. 8. The variables importance (a) and measured vs. predicted SOC values and the prediction
 474 statistics for training and validation sets (b) of PLSR model with SNV preprocessing performed
 475 in Alrad Spectra.

476

477 4.4.4. Predict unknown SOC

478 To predict unknown SOC content using spectroscopic data only, a few conditions have to
 479 be accomplished as detailed in the Prediction description. The best SOC predictive model built
 480 in Modeling module was achieved by PLSR-SNV and it was selected to predict SOC of new
 481 soil samples. In this step, the 95 soil samples obtained a predicted SOC content ranging from -
 482 0.21% to 3.79%. The predictions had an average SOC content of 1.88% and a standard
 483 deviation of 0.97. Prediction module offers the advantage of predict SOC using only the spectral
 484 behavior of the soil.

485 4.5.CONCLUSION

486 The GUI described in this study is a user-friendly tool for chemometrics analysis using
487 spectroscopic data. The interface offers the possibility of spectral data preprocessing, perform
488 different modeling algorithms and predict the desired variable. In the case study, Alrad Spectra
489 has proven to be an efficient tool in predicting soil organic carbon. All the operations can be
490 carried out by the user without the need of R programming skills. The intentions of building
491 Alrad Spectra were to facilitate the usage of R programming and to promote and expand the
492 usage of reflectance spectroscopy technique. These characteristics make Alrad Spectra an
493 useful tool for general public, researches, precision agriculture managers, and for the usage in
494 analytical laboratories.

495

496 **Acknowledgements**

497 The authors would like to thank the reviewers for their insightful comments on the paper.
498 This research was funded by Coordination for the Improvement of Higher Education Personnel
499 (CAPES), the National Council for Scientific and Technological Development (CNPq), and
500 Foundation for Funding in Research and Innovation of Santa Catarina State (FAPESC),
501 Ministry of Education, Brazil.

502

503 **References**

- 504 [1] R Core Team, R: A Language and Environment for Statistical Computing, (2016).
505 <http://www.R-project.org/>.
- 506 [2] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing
507 techniques for near-infrared spectra, *TrAC Trends Anal. Chem.* 28 (2009) 1201–1222.
508 doi:10.1016/j.trac.2009.07.007.
- 509 [3] S. Tippmann, Programming tools: Adventures with R, *Nat. News.* 517 (2015) 109.
510 doi:10.1038/517109a.
- 511 [4] M. Lawrence, D.T. Lang, RGtk2: A Graphical User Interface Toolkit for R, *J. Stat. Softw.*
512 37 (2010) 1–52.
- 513 [5] W. Hadley, W. Chang, devtools: Tools to Make Developing R Packages Easier, (2016).
514 <https://CRAN.R-project.org/package=devtools>.
- 515 [6] M. Lawrence, J. Verzani, gWidgetsRGtk2: Toolkit implementation of gWidgets for
516 RGtk2, (2014). <https://CRAN.R-project.org/package=gWidgetsRGtk2>.
- 517 [7] M.L. Delignette-Muller, C. Dutang, fitdistrplus: An R Package for Fitting Distributions,
518 *J. Stat. Softw.* 64 (2015) 1--34.

- 519 [8] J. Fox, S. Weisberg, *An R Companion to Applied Regression*, Second, SAGE
520 Publications, Thousand Oaks, CA, 2011.
521 <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>.
- 522 [9] H. Wickham, *ggplot2 - Elegant Graphics for Data Analysis*, 2009.
523 <http://www.springer.com/la/book/9780387981413> (accessed November 23, 2016).
- 524 [10] B. Auguie, *gridExtra: Miscellaneous Functions for “Grid” Graphics*, R Package Version
525 221. (2016). <https://CRAN.R-project.org/package=gridExtra>.
- 526 [11] M. Walesiak, A. Dudek, *clusterSim: Searching for Optimal Clustering Procedure for a*
527 *Data Set*, (2016). <https://CRAN.R-project.org/package=clusterSim>.
- 528 [12] B.-H. Mevik, R. Wehrens, K.H. Liland, *pls: Partial Least Squares and Principal*
529 *Component Regression*, R Package Version 25-0. 2 (2013). [https://CRAN.R-](https://CRAN.R-project.org/package=pls)
530 [project.org/package=pls](https://CRAN.R-project.org/package=pls).
- 531 [13] A. Stevens, L. Ramirez-Lopez, *An introduction to the prospectr package*, ResearchGate.
532 (2013).
533 [https://www.researchgate.net/publication/255941339_An_introduction_to_the_prospectr](https://www.researchgate.net/publication/255941339_An_introduction_to_the_prospectr_package)
534 [_package](https://www.researchgate.net/publication/255941339_An_introduction_to_the_prospectr_package) (accessed November 23, 2016).
- 535 [14] M. Kuhn et al., *caret: Classification and Regression Training*, (2016). [https://CRAN.R-](https://CRAN.R-project.org/package=caret)
536 [project.org/package=caret](https://CRAN.R-project.org/package=caret).
- 537 [15] D. Meyer, E. Dimitriadou, K. Hornik, F. Leisch, A. Weingessel, *E1071: Misc Functions*
538 *of the Department of Statistics (E1071)*, TU Wien, ResearchGate. 1 (2009).
539 [https://www.researchgate.net/publication/221678005_E1071_Misc_Functions_of_the_D](https://www.researchgate.net/publication/221678005_E1071_Misc_Functions_of_the_Department_of_Statistics_E1071_TU_Wien)
540 [epartment_of_Statistics_E1071_TU_Wien](https://www.researchgate.net/publication/221678005_E1071_Misc_Functions_of_the_Department_of_Statistics_E1071_TU_Wien) (accessed November 23, 2016).
- 541 [16] A. Gosso, *elmNN: Implementation of ELM (Extreme Learning Machine) algorithm for*
542 *SLFN (Single Hidden Layer Feedforward Neural Networks)*, R Package Version 1.
543 (2012).
- 544 [17] A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, *kernlab - An S4 Package for Kernel*
545 *Methods in R*, *J. Stat. Softw.* 11 (2004) 1–20. doi:10.18637/jss.v011.i09.
- 546 [18] A. Liaw, M. Wiener, *Classification and Regression by randomForest*, *R News.* 2 (2002)
547 18–22.
- 548 [19] R.J. Barnes, M.S. Dhanoa, S.J. Lister, *Standard Normal Variate Transformation and De-*
549 *trending of Near-Infrared Diffuse Reflectance Spectra*, *Appl. Spectrosc.* 43 (1989) 772–
550 777.

- 551 [20] R.N. Clark, T.L. Roush, Reflectance spectroscopy: Quantitative analysis techniques for
552 remote sensing applications, *J. Geophys. Res. Solid Earth.* 89 (1984) 6329–6340.
553 doi:10.1029/JB089iB07p06329.
- 554 [21] O.M.C. Mutanga, A.K. Skidmore, L. Kumar, J. Ferwerda, Estimating tropical pasture
555 quality at canopy level using band depth analysis with continuum removal in the visible
556 domain, *Int. J. Remote Sens.* 26 (2005) 1093–1108.
557 doi:10.1080/01431160512331326738.
- 558 [22] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least
559 Squares Procedures., *Anal. Chem.* 36 (1964) 1627–1639. doi:10.1021/ac60214a047.
- 560 [23] H. Martens, S.Å. Jensen, P. Geladi, Multivariate Linearity Transformation for Near-
561 Infrared Reflectance Spectrometry, in: O.H.J. Christie (Ed.), *Proc. Nord. Symp. Appl.*
562 *Stat.*, Stockholm Forlag Publication, Stavanger, Norway, 1983: pp. 205–234.
- 563 [24] F. Galton, Regression towards mediocrity in hereditary stature, *J. Anthropol. Inst. G. B.*
564 *Irel.* 15 (1886) 246–263.
- 565 [25] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics,
566 *Chemom. Intell. Lab. Syst.* 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.
- 567 [26] C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.* 20 (1995) 273–297.
568 doi:10.1007/BF00994018.
- 569 [27] L. Breiman, Random Forests, *Mach Learn.* 45 (2001) 5–32.
570 doi:10.1023/A:1010933404324.
- 571 [28] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity,
572 *Bull. Math. Biophys.* 5 (1943) 115–133. doi:10.1007/BF02478259.
- 573 [29] C.K.I. Williams, D. Barber, Bayesian classification with Gaussian processes, *IEEE Trans.*
574 *Pattern Anal. Mach. Intell.* 20 (1998) 1342–1351. doi:10.1109/34.735807.
- 575 [30] F. Galton, *Natural Inheritance*, 5th ed., Macmillan, New York, 1889.
- 576 [31] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control.*
577 19 (1974) 716–723. doi:10.1109/TAC.1974.1100705.
- 578 [32] S. Wold, A. Ruhe, H. Wold, I. Dunn W., The Collinearity Problem in Linear Regression.
579 The Partial Least Squares (PLS) Approach to Generalized Inverses, *SIAM J. Sci. Stat.*
580 *Comput.* 5 (1984) 735–743. doi:10.1137/0905052.
- 581 [33] R.A. Viscarra Rossel, T. Behrens, Using data mining to model and interpret soil diffuse
582 reflectance spectra, *Geoderma.* 158 (2010) 46–54. doi:10.1016/j.geoderma.2009.12.025.
- 583 [34] O. Ivanciuc, Applications of Support Vector Machines in Chemistry, in: K.B. Lipkowitz,
584 T.R. Cundari (Eds.), *Rev. Comput. Chem.*, John Wiley & Sons, Inc., 2007: pp. 291–400.

- 585 <http://onlinelibrary.wiley.com/doi/10.1002/9780470116449.ch6/summary> (accessed
586 August 7, 2015).
- 587 [35] G. Loosli, S. Canu, L. Bottou, Training Invariant Support Vector Machines using Selective
588 Sampling, in: L. Bottou, O. Chapelle, D. DeCoste, J. Weston (Eds.), *Large Scale Kernel*
589 *Mach.*, MIT Press, Cambridge, MA, 2007: pp. 301–320.
- 590 [36] G.B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme Learning Machine for Regression and
591 Multiclass Classification, *IEEE Trans. Syst. Man Cybern. Part B Cybern.* 42 (2011) 513–
592 529. doi:10.1109/TSMCB.2011.2168604.
- 593 [37] J.C. Yeomans, J.M. Bremner, A rapid and precise method for routine determination of
594 organic carbon in soil, *Commun. Soil Sci. Plant Anal.* 19 (1988) 1467–1476.
595 doi:10.1080/00103628809368027.
- 596 [38] E.R. Stoner, M.F. Baumgardner, Characteristic Variations in Reflectance of Surface Soils,
597 *Soil Sci. Soc. Am. J.* 45 (1981) 1161–1165.
598 doi:10.2136/sssaj1981.03615995004500060031x.
599

5 DISCUSSION

Over the last 30 years, soil reflectance measurement in the laboratory has been increasing substantially. Studies focusing on build robust model for a given soil property has been developed in different regions of the globe covering various soil properties. The findings of articles 1 and 2 contributed to soil spectroscopy development by applying different combinations of preprocessing techniques and multivariate calibration methods.

The effort dedicated over the thesis exposed the enormous potential of spectroscopic technique to quantify soil properties. The advantages of this technique exceeded expectations. Nowadays, soil spectroscopy is facing a remarkable growth. Soil properties assessment using standard methodologies in routine laboratory has become almost unviable. The potential of soil spectroscopy technique is well-known because its faster and cost-effective methods in soil property quantification. The members and commissions of Soil Societies should dedicate due attention to soil spectroscopy analyses potential.

In this context, soil spectroscopy is established an alternative strategy using a chemometrics approach for soil prediction. The unification of a common protocol for soil spectra analyses can increase its reliability and comparability. There are no standards or protocols for uniform laboratory and field reflectance measurements. The lack of standards in this well-recognized tool to assess soil properties can yields significant problems. Consequently, different protocols based on the literature, experience, convenience and infrastructure are been established. This has becoming a considerable issue for comparing and sharing soil spectral data between users. Besides, the construction of soil spectral libraries can be affected. In the study of BEN DOR; ONG; LAU (2015), the authors proposed to establish a standard protocol for soil measurement in the laboratory. They confirm that the any soil reflectance measurement can be corrected to normalize all possible variations to a soil benchmark setup. For GRUNWALD; VASQUES; RIVERO (2015) the need for soil property data leads to a need for integration pathways fusing lab and field based soil measurements, proximal and remote sensor data, environmental covariates, and methods. According to the authors, filling existing gaps in soil data will depend on the fusion of soil environmental, spectral data and methods to estimate soil properties. In addition, this interdependence will produce spatially and temporally continuous soil maps and models across various scales. Initiatives like these can contribute to establish new soil spectral libraries and expand the existing ones.

Nevertheless, there is certainly still room for improvement and expansion. The development and adhesion of soil spectroscopy is rising in a such way that it is moving towards to be establish as a soil analysis technique in routine laboratories for soil management. For this to happen, it is necessary data harmonization addressing methods and protocols.

Another major contribution of the thesis to promote the expansion of soil spectroscopy among scientists involves the development of the graphical user interface called Alrad Spectra. The requirement of massive R commands and codes for implementation of statistical procedures of both articles 1 and 2 led to the creation of this innovative tool to simplify the R activities. The advantages of this graphical interface are that it is a free, user-friendly tool and it is able to process spectral data from soils, water, grains, food, vegetation, etc. Alrad Spectra comes across with the intention to encourage and expand the usage of spectroscopic technique in R.

6 CONCLUSION

The outcomes of the thesis have demonstrated the great performance of predicting soil properties using Vis-NIR spectroscopy. Apparently, soil properties that are directly related to the chromophores such as organic carbon presented superior prediction statistics than particle size. Spectral preprocessing applied in the soil spectra contributed to the development of high-level prediction model. Comparing different spectral preprocessing techniques for SOC prediction revealed that the scatter-corrective preprocessing techniques presented superior prediction results compared to spectral derivatives. In scatter-correction technique, continuum removal is the most suitable preprocessing to be used for SOC prediction. In the calibration modeling, excepting for random forest, all of methods presented robust prediction with emphasis on the support vector machine. The systematic methodology applied in this study can improve the reliability of SOC estimation by examining how techniques of spectral preprocessing and multivariate methods affect the prediction performance using spectral analysis. The development of easy-to-use graphical user interface may benefit a large number of users, who will take advantage of this useful chemometrics analysis. Alrad Spectra is the first GUI of its kind and the expectation is that this tool can expand the application of the spectroscopy technique.

REFERENCES

- ARROUAYS, D. et al. (EDS.). **GlobalSoilMap: Basis of the global spatial soil information system**. CRC Press/Balkema, 2014.
- BELLON-MAUREL, V.; MCBRATNEY, A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. **Soil Biology and Biochemistry**, v. 43, n. 7, p. 1398–1410, jul. 2011.
- BEN DOR, E.; ONG, C.; LAU, I. C. Reflectance measurements of soils in the laboratory: Standards and protocols. **Geoderma**, v. 245–246, p. 112–124, maio 2015.
- DALMOLIN, R. S. D. et al. Relationship between the soil constituents and its spectral behavior. **Ciência Rural**, v. 35, n. 2, p. 481–489, abr. 2005.
- DEMATTE, J. A. M. et al. Visible–NIR reflectance: a new approach on soil evaluation. **Geoderma**, v. 121, n. 1–2, p. 95–112, jul. 2004.
- GRUNWALD, S.; VASQUES, G. M.; RIVERO, R. G. Chapter One - Fusion of soil and remote sensing data to model soil properties. In: SPARKS, D. L. (Ed.). . **Advances in Agronomy**. Academic Press, 2015. v. 131p. 1–109.
- HARTEMINK, A. E.; MINASNY, B. Towards digital soil morphometrics. **Geoderma**, v. 230–231, p. 305–317, out. 2014.
- MINASNY, B.; MCBRATNEY, A. B. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. **Chemometrics and Intelligent Laboratory Systems**, v. 94, n. 1, p. 72–79, 15 nov. 2008.
- MONTANARELLA ET AL., L. **Status of the World’s Soil Resources**. FAO, 2015.
- NOCITA, M. et al. Chapter Four - Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. In: SPARKS, D. L. (Ed.). . **Advances in Agronomy**. Academic Press, 2015. v. 132p. 139–159.
- R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, , 2016. Disponível em: <<http://www.R-project.org/>>.
- RINNAN, Å.; BERG, F. VAN DEN; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC Trends in Analytical Chemistry**, v. 28, n. 10, p. 1201–1222, nov. 2009.
- SANCHEZ, P. A. et al. Digital Soil Map of the World. **Science**, v. 325, n. 5941, p. 680–681, 7 ago. 2009.

SOUSA JUNIOR, J. G.; DEMATTÊ, J. A. M.; ARAÚJO, S. R. Modelos espectrais terrestres e orbitais na determinação de teores de atributos dos solos: potencial e custos. **Bragantia**, v. 70, n. 3, p. 610–621, 2011.

STENBERG, B. et al. Chapter Five - Visible and Near Infrared Spectroscopy in Soil Science. In: SPARKS, D. L. (Ed.). . **Advances in Agronomy**. Academic Press, 2010. v. 107p. 163–215.

STEVENS, A. et al. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. **PLoS ONE**, v. 8, n. 6, p. e66409, 19 jun. 2013.

STONER, E. R.; BAUMGARDNER, M. F. Characteristic Variations in Reflectance of Surface Soils. **Soil Science Society of America Journal**, v. 45, n. 6, p. 1161–1165, 12/01 1981.

VALERO-MORA, P. M.; LEDESMA, R. Graphical User Interfaces for R. **Journal of Statistical Software**, v. 49, n. 1, 2012.

VASQUES, G. M.; GRUNWALD, S.; SICKMAN, J. O. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. **Geoderma**, v. 146, n. 1–2, p. 14–25, 31 jul. 2008.

VISCARRA ROSSEL, R. A. et al. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. **Geoderma**, v. 131, n. 1–2, p. 59–75, mar. 2006.

VISCARRA ROSSEL, R. A. et al. A global spectral library to characterize the world's soil. **Earth-Science Reviews**, v. 155, p. 198–230, Abril 2016.