Revista Brasileira de Ciência do Solo

Division – Soil in Space and Time | Commission – Pedometric

AlradSpectra: a Quantification Tool for Soil Properties Using Spectroscopic Data in R

André Carnieletto Dotto[(1)], Ricardo Simão Diniz Dalmolin[(1)*], Alexandre ten Caten[(2)], Diego José Gris[(1)] and Luis Fernando Chimelo Ruiz[(3)]

[(1)] Universidade Federal de Santa Maria, Departamento de Solos, Santa Maria, Rio Grande do Sul, Brasil.
[(2)] Universidade Federal de Santa Catarina, Departamento de Agricultura, Biodiversidade e Florestas, Curitibanos, Santa Catarina, Brasil.
[(3)] Universidade Federal do Rio Grande do Sul, Centro Estadual de Pesquisa em Sensoriamento Remoto e Meteorologia, Porto Alegre, Rio Grande do Sul, Brasil.

ABSTRACT: Soil reflectance spectroscopy has become an innovative method for soil property quantification supplying data for studies in soil fertility, soil classification, digital soil mapping, while reducing laboratory time and applying a clean technology. This paper describes the implementation of a Graphical User Interface (GUI) using R named AlradSpectra. It contains several tools to process spectroscopic data and generate models to predict soil properties. The GUI was developed to accomplish tasks such as perform a large range of spectral preprocessing techniques, implement several multivariate calibration methods, generate statistics assessment and graphical output, validate the models using independent dataset, and predict unknown variables using soil spectral data. AlradSpectra has four main modules: Import Data, Spectral Preprocessing, Modeling, and Prediction. The implementation of AlradSpectra is demonstrated by applying visible near-infrared reflectance spectroscopy for soil organic carbon (SOC) prediction. The data contains the value of SOC and Vis-NIR reflectance for 595 soil samples. The prediction statistic assessment of SOC was performed applying all spectral preprocessing and methods. The R^2 considering all models ranged from 0.54 to 0.80. In the partial least squares regression (PLSR) models, the performances were similar to multiple linear regression (MLR) and support vector machines (SVM). The lowest error in the SOC prediction was achieved by PLSR method with standard normal variate (SNV) preprocessing reaching an R^2 of 0.80, the smallest root mean square error (RMSE) of 0.47 %, and ratio of performance to inter-quartile distance (RPIQ) of 3.12. 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 AlradSpectra a useful tool to perform different modeling approaches and predict the desired soil variable.

Keywords: GUI, R environment, multivariate calibration, spectral preprocessing, Pedometrics.

*** Corresponding author:**
E-mail: dalmolin@ufsm.br

Received: December 17, 2018
Approved: April 24, 2019

How to cite: Dotto AC, Dalmolin RSD, ten Caten A, Gris DJ, Ruiz LFC. AlradSpectra: a quantification tool for soil properties using spectroscopic data in R. Rev Bras Cienc Solo. 2019;43:e0180263.
https://doi.org/10.1590/18069657rbcs20180263

INTRODUCTION

Soil reflectance spectroscopy has made it possible to study soil fertility, granulometry quantification, soil class discrimination, while reducing laboratory time, as well as the use of chemical products (Demattê et al., 2019; Moura-Bueno et al., 2019). Soil spectroscopy is a proximal sensing technique based on the detection of the electromagnetic radiation reflected by the soil. In addition, spectroscopy in the visible (Vis: 400-700 nm), near infrared (NIR: 701-1100 nm), and short-wave infrared (SWIR: 1101-2500 nm) regions of the electromagnetic spectrum associated with chemometric methods has allowed the quantification of physical, chemical, and mineralogical soil properties (Viscarra Rossel and Behrens, 2010). This technique has become a well-established method to assess soil properties rapidly and accurately in the laboratory (Ben Dor et al., 2015), with the possibility of predicting several properties in just one spectral reading, facilitating data acquisition from large amounts of samples, and without the use of environmentally hazardous chemicals (Dotto et al., 2016, 2018; Demattê et al., 2019).

In general, to improve the accuracy of subsequent quantitative soil analysis, it is necessary to apply techniques of spectral preprocessing to standardize, transform, remove noise, and emphasize features (Rinnan et al., 2009). Spectral preprocessing has been identified as an indispensable step of spectral data analysis and has shown its importance on subsequent modeling tasks. The modeling procedure is accomplished by applying multivariate calibration methods. They have been commonly used to construct well-fitted models to determine the soil chemical and physical components of interest. The application of linear regression, ordinary least-squares regression, data mining, and machine learning algorithms are examples of modeling methods.

These methods could be systematized by a tool/program to facilitate the processes of spectral preprocessing, modeling, and prediction generation. Automated tools, free and with a user-friendly interface have contributed to the training of qualified professionals and researches in digital soil mapping (DSM) approach. Rizzo et al. (2015) observed the lack of qualified professionals who have knowledge about pedological mapping, computational tools, and remote sensing data. One program that has gained ground in the computational tools is the R programing language (R Development Core Team, 2018). The R community is massive and has growing importance in the last years in terms of process chemometric analysis. For Tippmann (2015), there is a trend for many academics to dive themselves off commercial software and dive in the free, open-source, and popular data-analysis tool. The software R has become one of the most requested statistical computing language and programming environment (Tippmann, 2015). The graphical user interface (GUI) in R came to supply users' needs by incorporating a user-friendly interface, in which there is no need to spend time learning how to deal with functions and its arguments and remembering a lot of commands.

For some users, the limitation of R is the implementation of functions, which must be called as text commands, and the user is required to find the proper packages that will accomplish specific tasks, recall the operations, and its argument options. To facilitate the routines for users, AlradSpectra was developed to eliminate this requirement. It has the advantages of providing a user-friendly GUI, being free and easy to operate, and it requires no initial knowledge of R.

Variations in the spectral data, which are caused by chemical and physical properties, can be modeled in conjunction with the target information. In this sense, we propose to develop and evaluate an automated, friendly, and free tool for spectral transformation, multivariate modeling, and prediction using spectroscopic data, denominated as AlradSpectra. The aim of this study was to perform spectral preprocessing and multivariate calibration

modeling to predict soil organic carbon (SOC) applying visible near-infrared (Vis-NIR) reflectance spectroscopy using the AlradSpectra.

MATERIALS AND METHODS

Implementation of AlradSpectra

AlradSpectra was implemented in R to perform spectral preprocessing, multivariate modeling, and prediction using spectroscopic data. The features include: i) import large database files; ii) perform nine types of spectral preprocessing and transformation techniques; iii) implement five multivariate calibration methods, which can provide well-fitted and accurate models; iv) provide statistics assessment; v) deliver graphical outputs; vi) validate the model using independent dataset; and vii) predict soil properties.

The AlradSpectra package is sited at the open source community GitHub repository (external link: *github.com/AlradSpectra/AlradSpectra*) (GPL-3 License). The devtools package is required to download and install AlradSpectra from the source website. During installation, if an error occurs, it is due to the old versions of packages already installed on your computer. The best way to solve this would be to uninstall the outdated ones and run the GUI installation commands again. As the program is operated in a user-friendly graphical interface, all of the operations and parameters required for chemometric analysis can be set through the GUI. The required packages to build AlradSpectra for each stage are listed in table 1. However, the user does not need to install these packages to open the program, but only needs to run the following commands in R console.

```
# Opening AlradSpectra
> install.packages("devtools") # Install devtools package only for the first time
> devtools::install_github("AlradSpectra/AlradSpectra") # Installation
> AlradSpectra::AlradSpectra() # Loading and Initialization
```

AlradSpectra interface was developed with a main menu with four different components, which are titled Import Data, Spectral Preprocessing, Modeling, and Prediction (Figure 1). The diagram illustrated in figure 2 is showing the workflow in sequential order. The first module is used to import data, view the imported data in tabular form, view the imported spectral curves, and view the descriptive statistics and histogram of the dependent variable. The next module performs the desired spectral preprocessing. The Modeling module allows the selection of input data and performs the modeling. The Prediction module can validate the models using an independent data set and predict the soil property using spectroscopic data only. The four main modules are described individually in the subsequent sections.

Import data module

The Import Data module enables to load data in text file (.txt) or comma-delimited values (.csv) file formats by browsing the file interactively or typing the file path. The samples have to be placed in rows and the variables in columns. The following file parameters needed to be set: file separator (usually comma, semicolon, or tab), decimal separator (dot or comma), whether the file has a header (the first row has column names), inform in which column the spectral data starts and ends, the first and last wavelength of the spectrum, and lastly, indicate the column that contains the soil variable and give it a name (not necessarily the same as the column name). These parameters will be required in preprocessing and modeling processes. The 'Import file' runs the commands to load the data, the 'View data' shows the data as a table, and the 'View imported spectra' shows the original spectral curves, while the 'View Y descriptive statistics' shows the descriptive statistics of the dependent variable in

a text dialog. The 'View Y histogram' displays a colorful histogram of the dependent variable. All images can be saved using the 'Save plot' button.

Table 1. Packages required to implement AlradSpectra that will be installed automatically

Component	R Package[(1)]	Reference
Graphical Integration	devtools	Wickham et al. (2016)
	gWidgetsRGtk2	Lawrence and Verzani (2014)
Descriptive statistics	fitdistrplus	Delignette-Muller and Dutang (2015)
Levene's Test	car	Fox and Weisberg (2011)
Plots	ggplot2	Wickham (2009)
	graphics	R Development Core Team (2018)
	gridExtra	Auguie (2016)
Spectral Preprocessing	clusterSim	Walesiak and Dudek (2016)
	pls	Mevik et al. (2016)
	prospectr	Stevens and Ramirez-Lopez (2013)
Modeling and Prediction	caret	Kuhn (2017)
	e1071	Dimitriadou et al. (2017)
	kernlab	Karatzoglou et al. (2004)
	pls	Mevik et al. (2016)
	randomForest	Liaw and Wiener (2002)

[(1)] Package dependencies are also installed.

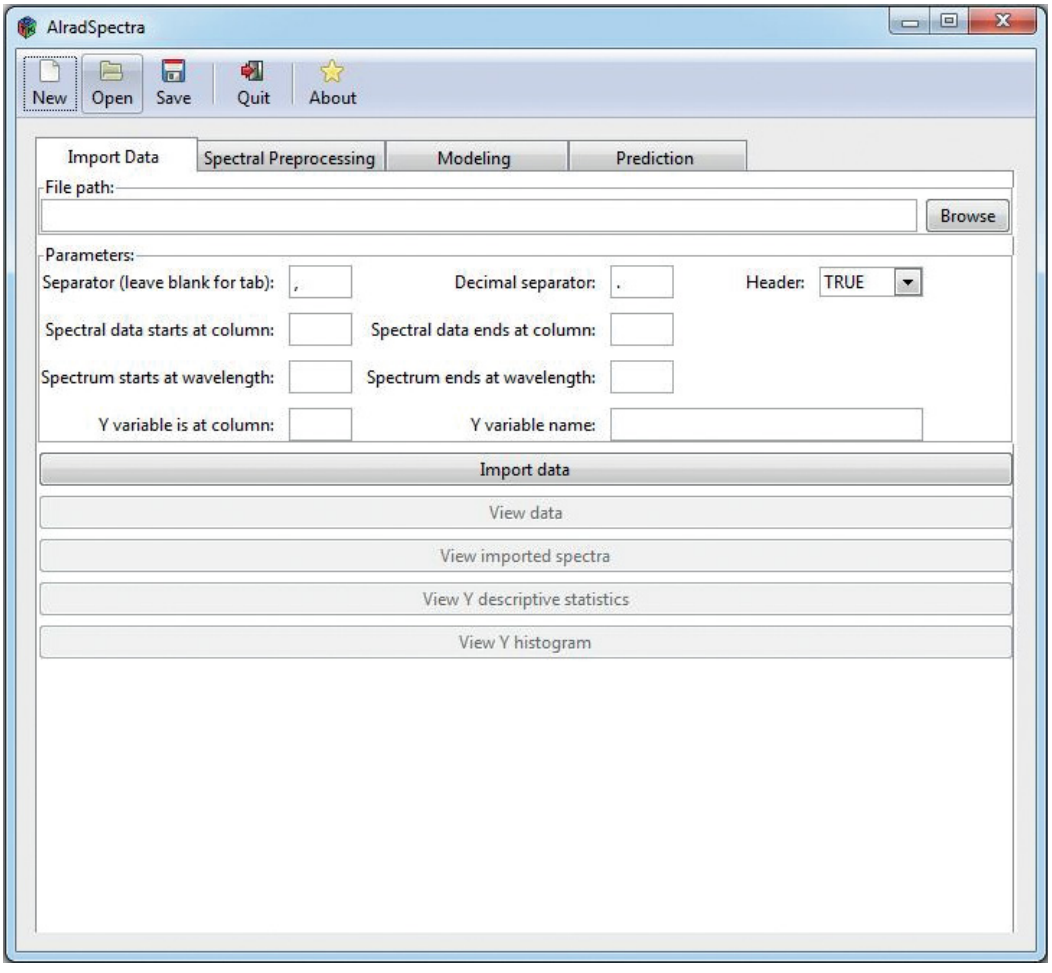

Figure 1. Graphical user interface of AlradSpectra showing the Import Data module.

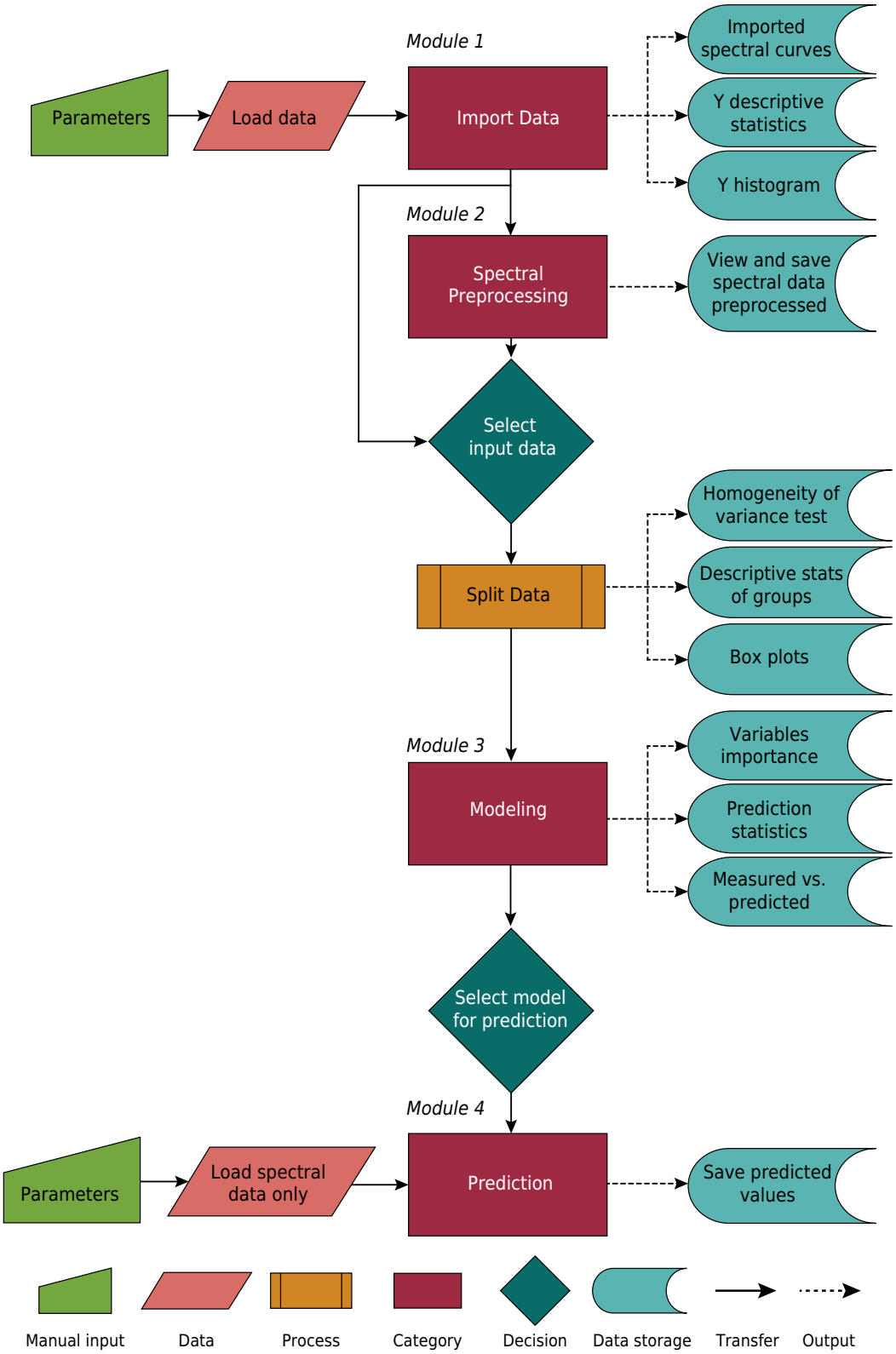

Figure 2. Flowchart of AlradSpectra.

Spectral preprocessing module

The Spectral Preprocessing module will be functional only after properly importing the data in the first module. A total of nine preprocessing algorithms were implemented, which are smoothing, binning, absorbance, detrend, continuum removal, Savitzky-Golay derivatives (SGD), standard normal variate (SNV), multiplicative scatter correction (MSC),

and normalizations. They are the most commonly used algorithms for preprocessing spectroscopic data. In each preprocessing tab, there is a 'View spectra' button, which allows to view the preprocessed spectral curves and be saved by 'Save plot' in the plot window. The 'Save preprocessed spectra' saves the spectral data in comma-delimited values (.csv) file format. Some preprocessing techniques have parameters to be defined by the user. The summary of each spectral preprocessing with the function and package used are found in table 2.

Modeling module

In the Modeling module, the first step requires to select the input data for modeling process. A drop-down list will display the imported spectral data, called Original, and the preprocessed spectra, if previously performed. When the same preprocessing is performed more than once (i.e., using different parameters, when available) the preprocessed data selected in this step corresponds to the last preprocessing generated. After selecting the input data, the user chooses the size of the validation set, in percentage. The split data is accomplished by randomly dividing the observation samples. The selection of validation set ranges from 0 to 50 %. The samples that are not included in the validation are used for training the models. Only after completing the split data, the homogeneity test, descriptive statistics, and boxplot are enabled and the multivariate methods tab can be manipulated. Levene's test for homogeneity of variances was implemented to verify the assumption that variances are equal across random selection of validation and training groups. The descriptive statistics and the boxplot of the dependent variable can be visualized using their respective buttons. To perform the modeling with different preprocessing, the user must select the preprocessing of interest and repeat the split data by clicking the 'Split data' button. In addition, in order to obtain the best-fitted model with the same preprocessing and calibration/validation set, simply repeat the slip data procedure.

The modeling covers different methods, including multiple linear regression, partial least squares regression (Wold et al., 1984), support vector machines (Cortes e Vapnik, 1995), random forest (Breiman, 2001), and Gaussian process regression (Williams e Barber, 1998). The glmStepAIC function is applied in the context of model selection to find the best-fitted model involving a subset of predictors for MLR model. The tuning parameter in MLR is the band interval, resampling method, number of folds or resampling iterations, and number of repetitions. The best parameters for PLSR model are employed to adjust the final model by the plsr function available in the pls package. In the PLSR, the tuning parameters are resampling method, number of folds or resampling iterations, number of repetitions, and number of components to include in the model. In PLSR, the Partial Least Squares (PLS) components vs. Root Mean Square Error (RMSE) values graphic was included. The best parameters for SVM model are employed to adjust the final model by svm function available in the e1071 package. The tuning parameters for SVM are resampling method, number of folds or resampling iterations, number of repetitions, and Liner or Radial kernels. The SVM models are efficient in modeling linear or nonlinear relationships and handling large databases. The final RF model is performed by the randomForest function in the randomForest package. The tuning parameters for RF are resampling method, number of folds or resampling iterations, number of repetitions, randomly selected predictors (mtry), and number of trees (ntree). The RF models are black boxes approach, which are very hard to interpret. The gausspr function in kernlab package performed the GPR final model and the tuning parameters are resampling method, number of folds or resampling iterations, number of repetitions, and initial noise variance.

In each method, the caret package was used to train and tune the models. The trainControl function in the caret package generates parameters that further control

Table 2. Spectral preprocessing and methods with the respective function and R package implemented in the AlradSpectra

Spectral Preprocessing (SP) Method (M) Function (F) Package (P)	Summary
SP: Smoothing F: movav P: prospectr	It is a simple moving average of a spectral data using a convolution function.
SP: Binning F: binning P: prospectr	Binning is used to reducing the effects of minor observation errors by computing average values of spectral data. To perform spectral binning, the bin size has to be specified (bin size).
SP: Absorbance F: $A = \log_{10} 1/R$	Absorbance is based on measuring the amount of light absorbed by a sample at a given wavelength.
SP: Detrend F: detrend P: prospectr	Detrend normalizes the spectral data by applying a standard normal variate transformation followed by fitting a second-degree polynomial regression model and returning the fitted residuals.
SP: Continuum Removal (CR) F: continuumRemoval P: prospectr	Continuum Removal remove the continuous features of the spectra and is often used to isolate specific absorption features present in the spectrum to minimize the noise. The continuum is represented by a mathematical function used to separate and highlight specific absorption bands of the reflectance spectrum.
SP: Savitzky–Golay Derivative F: savitzkyGolay P: prospectr	Derivatives are performed to remove unimportant baseline signal from samples by taking the derivative of the measured responses with respect to the variable number (wavelength). The Savitzky-Golay derivatization algorithm requires selection of smoothing points (filter width), the orders of polynomial and derivative.
SP: Standard Normal Variate (SNV) F: standardNormalVariate P: prospectr	Standard Normal Variate is performed in spectral data to remove scatter. It is applied to every spectrum individually. Standard Normal Variate is designed to operate based on centering the underlying linear slope of each individual sample spectrum.
SP: Multiplicative Scatter Correction (MSC) F: msc P: pls	Multiplicative Scatter Correction is achieved by regressing a measured spectrum against a reference spectrum. The MSC is effective in minimizing baseline offsets and multiplicative effect. The outcome of MSC, in many cases, is very similar to SNV, except SNV corrects each spectrum individually and does not need the entire data set.
SP: Normalization F: data.Normalization P: clusterSim	Normalization means adjusting values measured on different scales to a common scale, where these normalized values eliminate scattering effects. Five types of normalization were included in AlradSpectra: standardization, normalization in range, quotient transformation, normalization, and normalization with zero being the central point.
SP: Multiple Linear Regression (MLR) F: glmStepAIC P: caret	Multiple Linear Regression is a statistical method that uses several explanatory variables to predict the outcome of a response variable in a simple linear model (Galton, 1886). The MLR assumes the relationships between independent variables and the dependent variable are linear.
M: Partial Least Squares Regression (PLSR) F: plsr P: pls	Partial Least Squares Regression can handle complicated relationships between predictors and responses and can deal with complex modeling problems. Additionally, PLSR is a method for constructing predictive models when the factors are many and highly collinear (Wold et al., 1984), which is the case of hyperspectral data.
M: Support Vector Machines (SVM) F: svm P: e1071	Support Vector Machines are a group of supervised learning methods, which represent an extension to nonlinear models of generalized algorithm with the capability of training nonlinear classifiers (Ivanciuc, 2007). Associated with SVM algorithm is the criteria of smaller number of support vectors yield a better model performance (Loosli et al., 2007).
M: Random Forest (RF) F: randomForest P: randomForest	Random Forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). The RF is versatile and flexible with a small or large data set. Model interpretability is an issue when compared to linear models.
M: Gaussian Process Regression (GPR) F: gausspr P: kernlab	Gaussian Process Regression is a nonparametric regression using Gaussian processes, which applies a kernel function for training and predicting. In machine learning, kernel methods are a class of algorithms for pattern analysis. This approach replaces the features (predictors) by a kernel function. Several classes of kernels can be used for machine learning, and the selection of kernel is critical to the success of these algorithms (Karatzoglou et al., 2004).

how models are created, with possible values. One of these parameters is the resampling method, which is implemented to adjust the best-fitted models. Resampling methods involve repeatedly drawing samples from a calibration set and refitting a model of interest on each sample to obtain additional information about the fitted model. Such an approach may allow us to obtain information that would not be available from fitting the model only once using the original calibration samples. The resampling methods utilized are 'cv' (K-fold cross-validation), 'repeatedcv' (repeated K-fold cross-validation), 'LOOCV' (leave-one-out cross-validation), 'LGOCV' (leave-group-out cross-validation), 'boot' (bootstrap), 'boot632' (0.632 bootstrap), 'oob' (out-of-bag error estimates, only for tree models), and 'none'. For 'LOOCV', no uncertainty estimates are given for the resampled performance measures. The number of folds and resampling iterations controls the number of folds in 'cv' and number of resampling iterations for 'boot' and 'LGOCV'. The number of repetitions applies only to 'repeatedcv'. Once the modeling is completed, 'View variables importance' shows the importance of each variable for the model on a scale of 0 to 100. The 'Prediction statistics' shows the training and validation of statistical assessments, and 'View measured vs. predicted' shows the scatterplot for training and validation groups with its prediction statistics. The modeling methods used in AlradSpectra are summarized in table 2.

Prediction module

The Prediction module is implemented to predict the desired soil attribute using the built models. The prediction process requires the following conditions: file must contain only spectral data, spectral data for Prediction and Modeling must be the same length, and spectral data used in Prediction must have the same preprocessing used to build the model. The first step to perform the Prediction is to import a new data set containing only the spectral data (samples in rows and spectral variables in columns). The imported data can be opened in 'View data' and the spectral curves in 'View imported spectra'. The prediction is performed by selecting the model previously built. In 'View predictions' and 'Save predictions' buttons, it is possible to obtain the predicted values and save the results.

Applying AlradSpectra to model and predict SOC

The soil spectral data used in this study consists of 595 soil samples located in the central region of Santa Catarina State, Brazil. The experimental data contains the value of SOC and Vis-NIR reflectance. The SOC content was determined through the traditional laboratory analysis by wet combustion using the Mebius method in the digestion block (Yeomans e Bremner, 1988). Soil spectral reflectance was obtained using a FieldSpec 3 spectroradiometer (ASD Inc.) and was interpolated to 1 nm interval with a spectral range of 400-2.500 nm (Vis-NIR). The soil data file is placed and free available in the user's R library, inside AlradSpectra/extdata directory, e.g., "*C:\Users\UserName\Documents\R\win-library\3.3\ AlradSpectra\extdata*". The first 95 soil samples were applied in Prediction module as soils with unknown SOC values, and the subsequent 500 soil samples were used in the Modeling process. The 500 samples were randomly split into 70/30 % to train and validate the models, respectively. The soil spectral data file was imported in Import Data module by establishing the parameters: the file separator was comma, decimal separator was dot, header was true, the spectral data started at column 4 and ended at column 2104, the spectrum number started at 400 nm and ended at 2500 nm, and the dependent variable was at column number 3, and was named Soil Organic Carbon (%). The smoothing preprocessing example was accomplished with 11 smoothing points. For binning preprocessing, bin size was set to 10. In the SGD, it was applied 5 smoothing points, the first order of polynomial and the first order of derivative. Normalization in range was applied in the normalization preprocessing. The absorbance, detrend, CR, SNV,

and MSC preprocessing do not have parameters to be set and were also performed. A predictive model was built by each of multivariate calibration methods. For MLR, PLSR, SVM, and GPR models, the 'cv' resampling method with 10 folds were set as tuning parameters. For the MLR models, the band interval parameter was 25 for all models. For SVM models, the kernel parameter was Linear Kernel. For RF, the resampling method was 'oob' with 5 random predictors (mtry) and 500 trees (ntree). In GPR, the kernel function for the modeling was Linear kernel.

RESULTS

Categorization of soil reflectance

The original (reflectance) spectral curves imported along with all spectral preprocessed curves can be visualized in figure 3 and evaluated qualitatively. The spectral reflectance curves showed the diversity of soils by its shape and the presence or absence of absorption bands. Categorization of soil reflectance has important implications for soil genesis, classification, and survey (Stoner and Baumgardner, 1981).

Modeling for SOC prediction

The original spectral data without preprocessing plus the nine spectral preprocessing were used as independent variables to build the models. The Levene's test for homogeneity of variance presented a p-value of 0.918, which is greater than the significance level of 5 %. This result indicated that the training (70 % of samples) and validation (30 % of samples) groups were homogeneous and suitable for the modeling stage. The descriptive statistics of training and validation groups are presented in table 3. The histogram is showing the frequency of SOC, in which the blue color represents the higher distribution (Figure 4a). The boxplot of training and validation values express the homogeneity of the groups (Figure 4b).

The prediction statistic assessment for SOC models is ordered by the smallest RMSE value for each method (Table 4). The outcomes of MLR models showed that the greatest SOC prediction was achieved when SNV preprocessing was applied, reaching an R^2_{val} of 0.80, $RMSE_{val}$ of 0.51 %, and $RPIQ_{val}$ of 3.20. The R^2_{val} of all models ranged from 0.54 to 0.80. In the PLSR models, the performances were similar than MLR, with the R^2_{val} ranging from 0.56 to 0.80. The lowest error in the SOC prediction was achieved by PLSR method with SNV (PLSR-SNV) with an R^2_{val} of 0.80, $RMSE_{val}$ of 0.47 %, and $RPIQ_{val}$ of 3.12. In the validation performance, seven preprocessing exhibited R^2 above 0.75. The PLSR obtained the highest R^2_{val} value overall SOC prediction model. For the training set, several SVM models presented high performance, in which most of preprocessing are considered well-fitted models with the results in predicted values similar to the observed values. For the validation set, the best performance was achieved by absorbance preprocessing with an R^2_{val} of 0.78, $RMSE_{val}$ of 0.51 %, and $RPIQ_{val}$ of 2.55. The CR preprocessing presented the unreliable performance in SOC prediction with SVM (R^2_{val} = 0.61). However, in the RF models, CR preprocessing showed one of the best SOC prediction performance. The RF method showed a weak performance for original, binning, absorbance preprocessing, with an R^2_{val} ranging from 0.37 to 0.43. The higher performance in SOC prediction was found for RF-SNV preprocessing (R^2_{val} of 0.54; $RMSE_{val}$ of 0.71 %) followed by original preprocessing (R^2_{val} of 0.54; $RMSE_{val}$ of 0.75 %). The GPR models can lead to substantial improvements in training the models which led to a high accuracy for training samples. However, when the model is validated the prediction statistics showed more sensible outcomes. Observing the results of the validation set, the R^2 value oscillated from 0.48 to 0.77, where the higher performance was achieved by absorbance preprocessing.

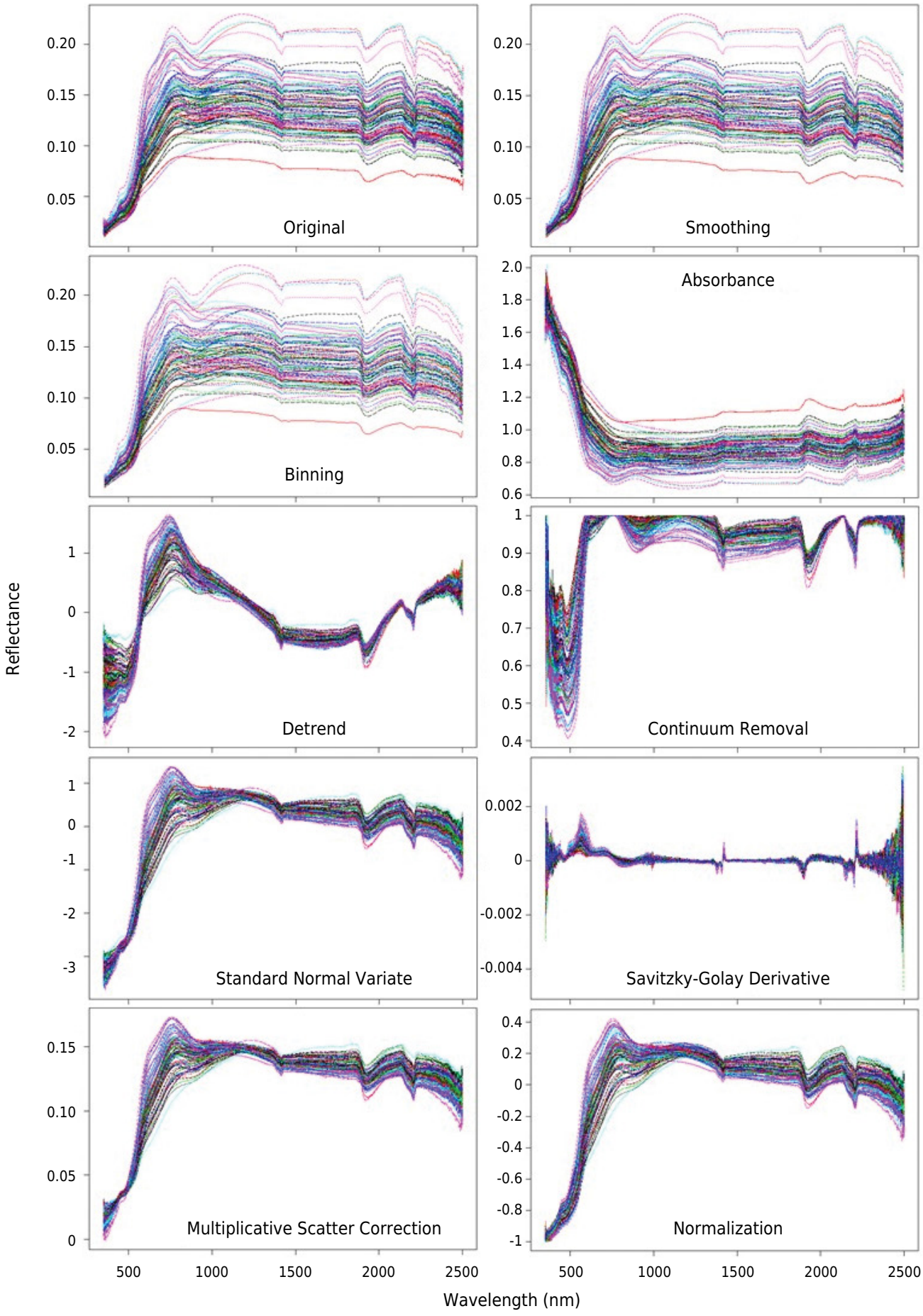

Figure 3. The original and preprocessed spectral curves performed in AlradSpectra.

Table 3. Descriptive statistics of SOC for whole, training and validation sets

	Set	Observation	Minimum	Maximum	Mean	Median	Standard deviation	Skewness	Kurtosis
Soil organic carbon (%)	Whole	500	0.02	6.87	1.95	1.86	1.08	0.79	4.06
	Training	350	0.02	6.87	1.98	1.87	1.11	0.88	4.38
	Validation	150	0.21	4.69	1.86	1.84	1.00	0.46	2.62
	Prediction	95	0.34	4.83	2.18	2.19	1.07	0.24	-0.59

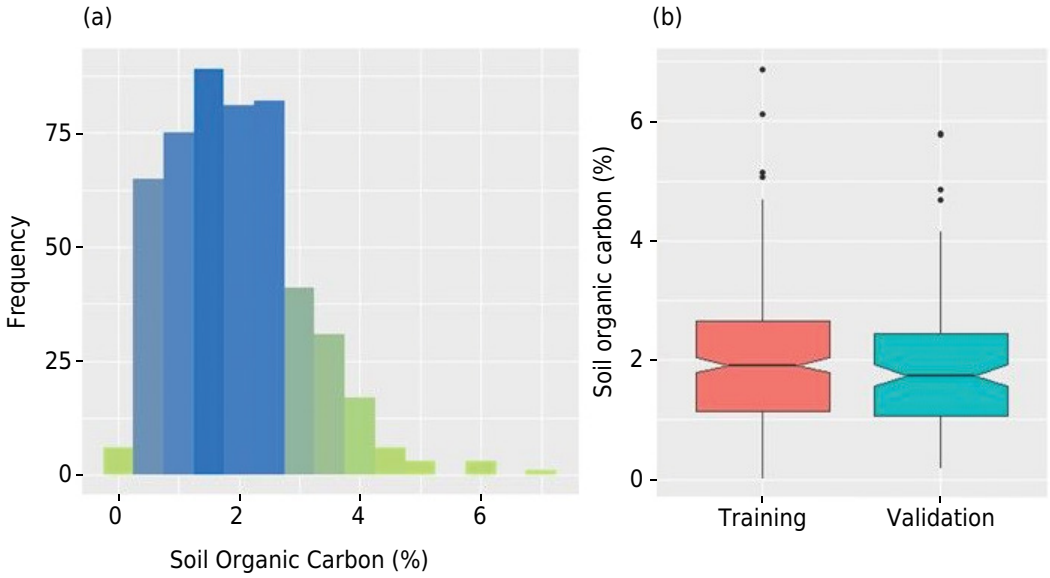

Figure 4. Histogram (a) and boxplot of training and validation groups (b) for soil organic carbon performed in AlradSpectra.

DISCUSSION

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

Soil spectroscopy has shown capability in providing a rapid assessment of various physical and chemical soil properties (Demattê et al., 2019). This technology can maximize fertilizer application and has gained prominence due to fast information and the environmental appeal regarding the use of a clean methodology. In addition, the quantification of soil properties via spectroscopy is generated by building calibration models that correlate the spectra with the reference analytical values. The types of spectral preprocessing and multivariate methods influence the quantification of soil properties (Dotto et al., 2018). In this study, we intended to develop the AlradSpectra to facilitate and disseminate the use of soil spectroscopy technique. The GUI presented in this study will assist other studies to select appropriate preprocessing and methods to quantify the soil attributes. Furthermore, AlradSpectra can process spectroscopic data from soils, water, grains, food, and vegetation.

CONCLUSION

AlradSpectra has proven to be an efficient tool in predicting soil organic carbon. The AlradSpectra described in this study is a user-friendly tool for chemometrics analysis using

Table 4. The prediction statistics of SOC for each model

Method	Preprocessing	Training set			Validation set		
		R^2	RMSE	RPIQ	R^2	RMSE[(1)]	RPIQ
			%			%	
MLR	SNV	0.84	0.43	3.24	0.80	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	0.44	3.34	0.80	0.47	3.12
	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
RF	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
	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

Continue

Continuation

GPR	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
	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

(1) The results are ordered by the smallest root mean square error (RMSE) for each method. RPIQ: ratio of performance to inter-quartile distance; MLR: multiple linear regression; PLSR: partial least squares regression; SVM: support vector machines; RF: random forest; GPR: Gaussian process regression; CR: continuum removal; SGD: Savitzky-Golay derivative; SNV: standard normal variate; MSC: multiplicative scatter correction.

spectroscopic data. The interface offers the possibility of spectral data preprocessing, run different modeling algorithms, and predict the desired soil variable. All the operations can be carried out by the user without the need of R programming skills. These characteristics make AlradSpectra a useful tool for predicting soil properties. The intentions of building AlradSpectra were to facilitate the usage of R and to promote and expand the usage of reflectance spectroscopy technique in the soil science community.

ACKNOWLEDGMENTS

This research was funded by Coordination for the Improvement of Higher Education Personnel (CAPES), the National Council for Scientific and Technological Development (CNPq), Ministry of Education, Brazil, and Foundation for Funding in Research and Innovation of Santa Catarina State (FAPESC).

AUTHOR CONTRIBUTIONS

Conceptualization: Andre Carnieletto Dotto, Ricardo Simão Diniz Dalmolin, and Alexandre ten Caten.

Methodology: Andre Carnieletto Dotto and Diego Jose Gris.

Software: Andre Carnieletto Dotto and Diego Jose Gris.

Validation: Luis Fernando Chimelo Ruiz.

Data Curation: Diego Jose Gris and Luis Fernando Chimelo Ruiz.

Writing – Original Draft: Andre Carnieletto Dotto and Diego Jose Gris.

Writing – Review & Editing: Ricardo Simão Diniz Dalmolin and Alexandre ten Caten.

Supervision: Ricardo Simão Diniz Dalmolin.

REFERENCES

Auguie B. gridExtra: miscellaneous functions for “Grid” graphics. R package version 2.2.1; 2016. Available from: https://mran.microsoft.com/snapshot/2016-11-19/web/packages/gridExtra/index.html.

Ben Dor E, Ong C, Lau IC. Reflectance measurements of soils in the laboratory: standards and protocols. Geoderma. 2015;245-46:112-24. https://doi.org/10.1016/j.geoderma.2015.01.002

Breiman L. Random forests. Mach Learn. 2001;45:5-32. https://doi.org/10.1023/A:1010933404324

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273-97. https://doi.org/10.1007/BF00994018

Delignette-Muller ML, Dutang C. fitdistrplus: an R package for fitting distributions. J Stat Softw. 2015;64:1-34. https://doi.org/10.18637/jss.v064.i04

Demattê JAM, Dotto AC, Bedin LG, Sayão VM, Souza AB. Soil analytical quality control by traditional and spectroscopy techniques: constructing the future of a hybrid laboratory for low environmental impact. Geoderma. 2019;337:111-21. https://doi.org/10.1016/j.geoderma.2018.09.010

Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A, Meyer D, Weingessel A. Misc functions of the department of statistics (e1071), TU Wien. R Packag version 1.6-8. 2017. Available from: https://www.researchgate.net/profile/Friedrich_Leisch/publication/221678005_E1071_Misc_Functions_of_the_Department_of_Statistics_E1071_TU_Wien/links/547305880cf24bc8ea19ad1d/E1071-Misc-Functions-of-the-Department-of-Statistics-E1071-TU-Wien.pdf

Dotto AC, Dalmolin RSD, ten Caten A, Grunwald S. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma. 2018;314:262-74. https://doi.org/10.1016/j.geoderma.2017.11.006

Dotto AC, Dalmolin RSD, ten Caten A, Moura-Bueno JM. Potential of spectroradiometry to classify soil clay content. Rev Bras Cienc Solo. 2016;40:e0151105. https://doi.org/10.1590/18069657rbcs20151105

Fox J, Weisberg S. An R companion to applied regression. 2nd ed. Thousand Oaks: Sage Publications; 2011.

Ivanciuc O. Applications of support vector machines in chemistry. In: Lipkowitz KB, Cundari TR, editors. Reviews in computational chemistry. New York: John Wiley & Sons, Inc.; 2007. vol. 23. p. 291-400.

Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab - an S4 package for kernel methods in R. J Stat Softw. 2004;11:1-20. https://doi.org/10.1016/j.csda.2009.09.023

Kuhn M. caret: classification and regression training. R package version 6.0-73; 2017. Available from: http://adsabs.harvard.edu/abs/2015ascl.soft05003K

Lawrence M, Verzani J. gWidgetsRGtk2: Toolkit implementation of gWidgets for RGtk2. R package version 0.0-83; 2014. Available from: https://rdrr.io/cran/gWidgetsRGtk2/

Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2:18-22.

Loosli G, Bottou L, Canu S. Training invariant SVMs using selective sampling. In: Bottou L, Chapelle O, DeCoste D, Weston J, editors. Large-scale kernel mach. London: The MIT Press; 2007. p. 301-20.

Mevik B-H, Wehrens R, Liland KH. pls: partial least squares and principal component regression. R Packag version 2.6-0; 2016. Available from: http://mevik.net/work/software/pls.html

Moura-Bueno JM, Dalmolin RSD, ten Caten A, Dotto AC, Demattê JAM. Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. Geoderma. 2019;337:565-81. https://doi.org/10.1016/j.geoderma.2018.10.015

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2018. Available from: http://www.R-project.org/.

Rinnan Å, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. Trac-Trends Anal Chem. 2009;28:1201-22. https://doi.org/10.1016/j.trac.2009.07.007

Rizzo R, Demattê JAM, Lacerda MPC. Espectros Vis-NIR do solo e *Fuzzy* k-médias aplicados na delimitação de unidades de mapeamento de solos em topossequências. Rev Bras Cienc Solo. 2015;39:1533-43. https://doi.org/10.1590/01000683rbcs20140694

Stevens A, Ramirez-Lopez L. An introduction to the prospectr package. R package Vignette; 2013. Available from: ftp://200.236.31.2/CRAN/web/packages/prospectr/vignettes/prospectr-intro.pdf

Stoner ER, Baumgardner MF. Characteristic variations in reflectance of surface soils. Soil Sci Soc Am J. 1981;45:1161-5. https://doi.org/10.2136/sssaj1981.03615995004500060031x

Tippmann S. Programming tools: Adventures with R. Nature. 2015;517:109-10. https://doi.org/10.1038/517109a

Viscarra Rossel RA, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma. 2010;158:46-54. https://doi.org/10.1016/j.geoderma.2009.12.025

Walesiak M, Dudek A. clusterSim: searching for optimal clustering procedure for a data set. R package version 0.45-1; 2016. Available from: https://rdrr.io/cran/clusterSim/

Wickham H. ggplot2: elegant graphics for data analysis. London: Springer; 2009.

Wickham H, Hester J, Chang W. devtools: Tools to Make Developing R Packages Easier. R package version 1.12.0; 2016. Available from: https://rdrr.io/cran/devtools/

Williams CKI, Barber D. Bayesian classification with Gaussian processes. IEEE T Pattern Anal Mach Intell. 1998;20:1342-51. https://doi.org/10.1109/34.735807

Wold S, Ruhe A, Wold H, Dunn III WJ. The collinearity problem in linear regression. the Partial Least Squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comp. 1984;5:735-43. https://doi.org/10.1137/0905052

Yeomans JC, Bremner JM. A rapid and precise method for routine determination of organic carbon in soil. Commun Soil Sci Plant Anal. 1988;19:1467-76. https://doi.org/10.1080/00103628809368027