UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Metodi e analisi statistiche 2020
DIPARTIMENTO DI ECONOMIA E FINANZA

ISBN: 978-88-6629-023-0
pp. 135-162

Comparison of different multivariate calibrations and ensemble methods for estimating selected soil properties with vis-NIR reflectance spectroscopy

Davide Fragnito[1], Natalia Leone[2], Valeria Ancona[2], Vitale Domenico[3], Antonio Lucadamo[4].

[1]Master's Graduate in Statistical, Actuarial and Financial Sciences, e-mail: davide.fragnito.1994@gmail.com,
[2]Water Research Institute, National Research Council, Viale Francesco de Blasio, 5, Bari BA,
[3]CMCC Foundation, Euro-Mediterranean Center on Climatic Change,00100 Viterbo,
[4]DEMM, Department of Law, Economics, Management and quantitative Methods, University of Sannio, Via delle Puglie, 53, Benevento.

Abstract: Sustainable soil management requires a correct assessment of soil chemical and physical properties. Historically, this has been gained through conventional laboratory analyses, which are considered costly and time-consuming, particularly when a large number of soil samples need to be analysed. An alternative, faster and less expensive, approach is based on the use of reflectance spectroscopy in the vis-NIR domain. This approach implies the calibration of predictive models that relate the spectral reflectance to soil properties. The goodness of the models can be particularly influenced by the multivariate methods used. In this article, we compare the performance of different multivariate and statistical ensemble methods for estimating some basic soil properties, such as sand, silt, clay, and organic carbon in the specific pedo-environmental conditions of an important agricultural area in southern Italy.

Keywords: vis-NIR reflectance spectroscopy, prediction of soil properties, multivariate and statistical ensemble methods.

1. Introduction

Soil is one of the main natural resources. It contributes to basic human needs like food, clean water, and clean air, and is a major carrier for biodiversity (Keesstra et al., 2016). From here, the need to preserve this resource (soil) to ensure sustainable and shared prosperity to humanity (FAO, 2015) through sustainable agricultural and non-agricultural uses. Sustainable soil management cannot disregard a correct assessment of its chemical and physical properties and their variability in space and time.

Historically our understanding of the soil system and assessment of their properties has been gained through conventional laboratory analysis (Viscarra-Rossel et al., 2006). The latter, although usefully and practically irreplaceable for detailed investigations, are costly and time-consuming, thus not very suitable when large numbers of soil samples need to be analysed, as, for example, in large soil surveys, or for high-resolution soil mapping and precision agriculture. Hence, the need to develop alternative techniques for soil analyses.

In recent years, vis-NIR reflectance spectroscopy has been shown to be a useful technique for the measurement of various soil properties (Lucadamo and Leone, 2015; Lucadamo et al., 2020). Compared to conventional analytical methods, vis–NIR spectroscopy is faster, cheaper, and non-destructive; it requires less sample preparation, with less or no chemical reagents, is highly adaptable to automated and in situ measurements, and has the potential to analyse various soil properties simultaneously (Viscarra-Rossel et al., 2006; McCarty et al., 2002; Vasques et al., 2008). Reflectance spectroscopy refers to the measure of spectral reflectance (Milton, 1987), i.e., the ratio of the electromagnetic radiation reflected by a soil surface to that which impinges on it (Drury, 1993). Since the characteristics of the radiation reflected from a material are a function of the material's properties, observations of soil reflectance can provide information on the properties and state of the soil (Irons et al., 1989). The reflectance spectra of soil in the vis–NIR are largely non-specific due to the overlapping absorption of soil constituents. This characteristic lack of specificity is compounded by scatter effects, caused by soil structure or specific constituents, such as quartz. All of these factors result in complex absorption patterns that need to be mathematically extracted from the spectra and correlated with soil properties. Therefore, multivariate statistics are required to mathematically extract complex absorption patterns and to correlate these patterns with the measured soil properties for calibration (Martens and Næs, 1989; Stenberg et al., 2010; Araújo et al., 2014; Xu et al., 2018). The selection of the multivariate statistic methods, along with that of proper instrumentation, accessories and optical probe design (Mouazen et al., 2009), improved spectra filtering and pre-processing (Maleki et al., 2008), are essential factors for successful calibration of predictive models (Mouazen et al., 2010; Nawar et al., 2016).

A large number of multivariate calibration methods have been used to relate vis-NIR reflectance spectra with measured soil properties (e.g., Viscarra-Rossel et al., 2006; Janik et al., 2009; Mouazen et al., 2010; Stevens et al., 2010; Viscarra-Rossel and Behrens, 2010; Vohland et al., 2011; Shi et al., 2015; Araújo et al., 2014; Kuang et al., 2015; Were et al., 2015). However, none of these proposed calibration techniques have achieved universal acceptance because a calibration model that works well for one application may be unacceptable for another (Xu, 2018). The specificity of the pedo-environment, besides the choice of the pre-processing methods, may also influence the selection of the statistical calibration methods, being a soil a complex and heterogeneous system.

This study aims to explore the performances of different multivariate and statistical ensemble methods for estimating some basic soil properties, such as sand, silt, clay, and organic carbon (OC) contents, within the specific pedo-environmental conditions of an important, irrigated area of southern Italy. Namely, the compared statistical methods are: Partial Least Squares Regression (PLS), Regression Tree (RT), Bagging and Random Forest algorithm (B, RF), Boosting Regression (BR), Artificial Neural Network (ANN), Multivariate Adaptive Regression Splines (MARS). The remaining part of the article is organized as follows: in section 2, all the statistical methods used for the analysis are introduced; section 3, describes data collection and material; in section 4 the results are synthesized; some concluding remarks are shown in section 5.

2. Some theoretical aspects

2.1 Partial Least Squares Regression (PLSR)

Partial least squares regression is by far the most used multivariate statistical method in the field of vis-NIR reflectance spectroscopy (Gholizadeh et al., 2016; Leone et al., 2012; Leone et al., 2019; Vibhute et al., 2018; Viscarra-Rossel et al., 2006; Cozzolino and Moron, 2003; Wang et al., 2013; Volkan Bilgili et al., 2010; Lee et al., 2009; Viscarra-Rossel and Behrens, 2010; Kuang et al., 2015; Wetterlind et al., 2008; Viscarra-Rossel and Lark, 2009; Brown et al., 2006; Stevens et al., 2013; Dunn et al., 2002; Fystro, 2002; Mouazen et al., 2007). This method was proposed by H. Wold for the modeling of data sets in terms of chains of matrices (path models), suggesting a procedure named NIPALS (Non-linear Iterative Partial Least Squares) to estimate the parameters (Wold, 1973). Later, other groups led by S. Wold and H. Martens popularized the use of this method for chemical applications by slightly modifying the PLS model with only two matrices containing the explanatory variables (X) and the response variables (Y) to deal with complicated data sets where ordinary regression was difficult or impossible to apply. Several authors (Wold et al., 1993; Wold et al., 2004) started to interpret PLS as the Projection to Latent Structures, providing a more descriptive meaning.

There are two basic approaches, named PLSR1 and PLSR2. In PLSR1, one calibration model is considered for y or separate calibration models are built for each column in Y. With PLS2, one calibration model is built for all columns of Y simultaneously. PLSR was first proposed for analysing NIR spectra by Wold et al. (Wold et al., 1983), who derived an algorithm with orthogonal scores. Successively, Martens (Martens, 1985) and Martens and Naes (Martens and Naes, 1989) proposed a PLSR algorithm with orthogonal loadings. Moreover, Helland (Helland, 1998) showed the equivalence between these proposals for the PLSR1 algorithms, while the geometry of PLS has been explored in depth by Phatak and de Jong (Phatak and De Jong, 1997). For the purpose of this paper, we refer only to the PLSR1 algorithm.

PLSR finds the linear (or polynomial) relationships between a centred response variable vector y and a matrix of centred predictors X expressed as y=f(X)+E. PLS regression seeks then to provide a statistical model based on the reduction of the space spanned by the often-large number of correlated predictors in a lower-dimensional space generated by derived PLS components. These components reflect the information in the X-variables that are of relevance for modelling and predicting the response variable y. The link is then obtained by the following decompositions that lead to orthogonal scores and non-orthogonal loading vectors (Wold's algorithm):

$$\mathbf{X} = \mathbf{t_1 p'_1} + \mathbf{t_2 p'_2} + \cdots \mathbf{t_K p'_K} + \mathbf{E}_{\boldsymbol{K}} = \mathbf{TP'} + \mathbf{E_K}$$

$$\mathbf{y} = \mathbf{t_1 q_1} + \mathbf{t_2 q_2} + \cdots + \mathbf{t_K q_K} + \mathbf{f_K} = \mathbf{Tq} + \mathbf{f_K}$$

where *t* is a vector of scores calculated by $\mathrm{t_k = X_{k-1} w_k}$ with scaled weights $\mathrm{w_k}$ and $\mathrm{T = [t_1, \dots, t_k]}$, *p* are the spectral loadings, *q* the chemical loadings and *E* and *f* are the predictor and response variable residuals, respectively, of the estimated effect for the k-th factor (k=1,…,K). Wold et al. use, for achieving these solutions, the well-known non-linear iterative partial least squares (NIPALS) algorithm for centred X and y data: let $\mathrm{X_0 = X}$ and $\mathrm{Y_0 = Y}$, the orthogonal scores $\{\mathrm{t_1, \dots, t_k}\}$ are then iteratively obtained, where the basic k-th step of the algorithm is given by:

1. Compute the scaled weight vector $\mathrm{w_k = cX'_{k-1} y_{k-1} / y'_{k-1} y_{k-1}}$ with c scaling factor;
2. Compute the orthogonal score $\mathrm{t_k = X_{k-1} w_k}$;
3. Compute the residuals $\mathrm{X_k = \left(I - P_{t_k}\right) X_{k-1} = \left(I - P_{t_k}\right) X}$ and

$y_k = (I - P_{t_k})y_{k-1} = (I - P_{t_k})y$ where $P_{t_k} = t_k(t'_k t_k)^{-1}t'_k$ and $P_{T_k} = T_k(T'_k T_k)^{-1}T_k$ are the orthogonal projection operators onto t_k and the subspace spanned by $\{t_1, \dots, t_k\}$, respectively.

The number of factors to use in PLSR model may be determined through leave-one-out cross-validation. The optimal number of factors should allow the modelling of as much as possible of the correlation between X and y without overfitting y. Then, for the selected number of factors, one calculates the final linear regression coefficients, $b = W(W'P)^{-1}q$ (where W is the weight matrix) and $b_0 = \bar{y} - \bar{x}'b$ to be used in the predictor $\hat{y}_i = b_0 + x_i b$ where x_i is the new spectrum. The well-known Martens algorithm is instead based on the factorization:

$$\mathbf{X} = \tilde{\mathbf{t}}_\mathbf{k}\tilde{\mathbf{w}}'_\mathbf{1} + \cdots + \tilde{\mathbf{t}}_\mathbf{K}\tilde{\mathbf{w}}'_\mathbf{K} + \mathbf{E}_\mathbf{K} = \tilde{\mathbf{T}}\tilde{\mathbf{W}}' + \mathbf{E}_\mathbf{K}$$

$$\mathbf{y} = \tilde{\mathbf{t}}_\mathbf{1}\tilde{\mathbf{q}}_\mathbf{1} + \cdots + \tilde{\mathbf{t}}_\mathbf{K}\tilde{\mathbf{q}}_\mathbf{K} + \mathbf{f}_\mathbf{K} = \tilde{\mathbf{T}}\tilde{\mathbf{q}} + \mathbf{f}_\mathbf{K}$$

which uses a non-orthogonal score matrix $\tilde{T}$, i.e. $\tilde{T}'\,\tilde{T}$ is a non-diagonal matrix, with orthogonal loadings and where the scores $\{t_1, \dots, t_k\}$ are iteratively obtained. The basic k-th step of this algorithm (with $X_0 = X$ and $Y_0 = Y$) is given by:

1) Compute the weight vector $\tilde{w}_k = \tilde{X}'_{k-1}\tilde{y}'_{k-1}$;
2) Compute the non-orthogonal score $\tilde{t}_k = \tilde{X}'_{k-1}\tilde{w}'_k/\tilde{w}'_k\tilde{w}_k$ and set $\tilde{T}_k = [\tilde{t}_1, \dots, \tilde{t}_k]$;
3) Compute the regression coefficients $\tilde{q}_k$ of y in $\tilde{T}_k$ given by $\tilde{q}_k = \left(\tilde{T}'_k\tilde{T}_k\right)^{-1}\tilde{T}'_k\tilde{y}_k$;
4) Compute the residuals $\tilde{X}_k = \tilde{X}_{k-1} - \tilde{t}_k\tilde{w}'_k$ and $\tilde{y}_k = y - \sum_{j=1}^{k}\tilde{q}_{kj}\tilde{t}_k$;

such to obtain the above X decomposition $X = \tilde{T}\tilde{W}' + \tilde{E}_k$, and where the score vectors $\tilde{t}_k = X\tilde{w}_k \,/\, \tilde{w}'_k\tilde{w}_k$ and t_k spam the same vectorial space. The regression coefficient vector is finally given as a simple least square solution $\tilde{b} = \tilde{W}\left(\tilde{W}'X'X\tilde{W}\right)^{-1}\tilde{W}'X'y$ providing the same coefficients as the previous PLS1 formula.

We remark that the latter algorithm, giving the non-orthogonal score vectors, does not provide the problems that Pell (Pell et al., 2007) has recently highlighted for the NIPALS results about their possible inconsistence with respect to model spaces for residual-based outlier detection and prediction purpose. See Ergon (Ergon, 2009) for a re-interpretation of the NIPALS results, which solves the PLSR inconsistency problem.

We highlight that in this paper all computations have been performed by using the software for the chemometric analysis of spectroscopic data called “ParLeS”

(Viscarra-Rossel, 2008). This software implements the most used form of the PLSR1 algorithm, which produces orthogonal scores, and provides several statistical tools to assist the researcher in performing and interpreting the analysis results. For example, the number of samples (rows) to leave out in the cross-validation may be any integer selected by the user and the accuracy of the cross-validation is given by the root-mean-square error (RMSE). Moreover, the goodness of fit is given by R^2 and Q^2 statistics, which give the upper and lower bounds, of how the model well explains the data and predicts new observations. For the selection of an optimal parsimonious PLSR model (i.e., one that represents the variability in the data without causing it to overfit) the Akaike Information Criterion (AIC) (Akaike, 1973) is also provided by ParLeS where N is the sample size and m is the number of model parameters, in this case, the number of factors. A sorted VIP (Variable Importance for Projection) data table, and the percent variation in each of the x and y-data that is explained by each of the PLSR factors, are also given, where the VIP index is computed as:

$$\mathbf{VIP_j(k)} = \mathbf{K}\sum_{k} \mathbf{w_{jk}^2} \left({}^{\mathbf{SSY_k}}\!/_{\mathbf{SSY_{tot}}} \right)$$

where $\mathrm{VIP_j(k)}$ is the importance of the j-th predictor variable based on a model with k factors, $\mathrm{w_{jk}}$ is the corresponding loading weight of the j-th variable in the k-th PLSR factor, $\mathrm{SSY_k}$ is the explained sum of squares of y by a PLSR performed with the only *k-th* factor, $\mathrm{SSY_{tot}}$ is the total sum of squares of y, and K is the total number of predictor variables. The reader is directed to Viscarra-Rossel (Viscarra-Rossel, 2008) for a full description of ParLeS and the algorithms it implements.

2.2 Regression Trees (RT)

An alternative algorithm for analysing the relationship between variables is the "Classification And Regression Trees" (CART). Even if the basic idea is the same, in this paper it has been preferred to separate the classification tree treatment from the tree regression, also by virtue of the fact that in the dataset in our possession, the variable y is a continuous random variable.

The regression tree constructs an H tree from the root node h1, by performing a succession of splits, or divisions, of the full set of observations, to make the units more homogeneous in terms of response variable y. The algorithm used to build the tree follow an approach of step-by-step optimization. To understand how it works, we must break down the deviance as follows:

$$\boldsymbol{D} = \sum_{i=1}^{n}\left[\boldsymbol{y_i} - \boldsymbol{\hat{f}}(\boldsymbol{x_i})\right]^2 = \sum_{h=1}^{J}\{\sum_{i\in \boldsymbol{R_h}}(\boldsymbol{y_i} - \widehat{\boldsymbol{c_h}})^2\} = \sum_{h}\boldsymbol{D_h} \qquad (1)$$

where $\hat{f}(x_i)$ is the predicted value for response variable y; c_h is the arithmetic mean of the observed y_i having component x_i falling in the subinterval; J is the global number of nodes and $R_1, \dots, R_j$ are rectangles in the p-dimensional sense. The growth process of the tree starts with the root node h1 (so J=1; $R_j = \mathbb{R}^p$; $D = \sum_{i=1}^{n}[y_i - M(y)]^2$) with $M(\cdot)$ average operator). And proceeds iteratively according to the following scheme:

- Once a rectangle R_h is chose, the appropriate value of c_h is the average of the corresponding values $\widehat{c_h} = M(y_i\colon x_i \ \in R_h)$;
- If we subdivide the region into two parts the deviance is replaced by $D_h^* = \sum_{i \in R_{h'}}^{n}(y_i - \widehat{c_h}')^2 + \sum_{i \in R_{h''}}^{n}(y_i - \widehat{c_h}'')^2$ with a gain of $g_h = D_h - D_h^*$.
- We can inspect all p explanatory variables and, for each of them, all the possible points of subdivision, selecting the variable and its point of subdivision that maximize g_h.

The algorithm stops when all the leaves contain a number of sample elements that is less than a preassigned value, or when the relative fall of deviance is less than a prefixed threshold. A large tree is obviously not useful, so the branches of little importance have to be pruned. For this reason, a cost-complexity function can be considered:

$$\boldsymbol{C_\alpha(J) = \sum_{h=1}^{J} D_h + \alpha J}$$

where α is a non-negative penalty parameter. For each α there is a unique smallest tree minimizing $C_\alpha(J)$. The algorithm sequentially eliminates one leaf at a time. At each step the leaf for which elimination causes the smallest increase in $\sum_h D_h$ is selected. The question is to choosing α: generally the cross-validation is used. Trees are frequently used in practice, but it is important to underline their advantages and disadvantages. They have in fact a logical simplicity and are easy to communicate; the step function has a simple, compact mathematical formulation in terms of information to be stored; there is a speed of computation and the possibility to use discrete and categorical variables; a robust forms of deviance can be used; not particularly complicated variations can be introduced, which allow for missing values, in both tree construction and prediction; the method automatically selects the important variables. On the other hand, there is instability of results and difficulty in upgrading the tree; difficulty of approximating some mathematically simple function; procedures of statistical inference are not available and it is not simple to evaluate the order of importance of variables remaining in the pruned tree. (Breiman et al., 1984; Ripley, 1996; Venables & Ripley, 1997; AA.VV., 1995).

2.3 Bootstrap Aggregating (B)

Regression trees suffer from high variance, which means that if you were to divide the calibration dataset into two random parts and fit a regression tree to both halves, the results that could be achieved would be quite different. Conversely, a procedure with low variance will produce similar results when applied repeatedly to separate datasets; linear regression tends to have a low variance if the ratio of n (number of observations) to "p" (number of predictors) is moderately large.

Bootstrap AGGregatING, or BAGGING, is a general procedure for reducing the variance of a statistical learning method and is particularly useful, and often used, in the context of regression trees, although applied in different works (Gholizadeh et al., 2016; Viscarra-Rossel and Behrens, 2010). Briefly remembering that the variance of the mean of observations $\bar{Z}$ of a set of n independent observations $Z_1, \dots, Z_n$, each with variance σ^2, is equal to the ratio of variance to the number of observations, it is statistically valid to say that the average of a series of observations reduces variance.

Hence a natural way to reduce variance and simultaneously increase the accuracy of predictions of a statistical learning method is to take many sub-samples of the "training" of the model from the population, build a separate forecasting model using each train-set, set and calculate the average of the resulting predictions. In other words, you could calculate $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ using separate B training sets, mediate them into $\hat{f}_{avg}(x)$ obtaining a single low-variance statistical learning model, expressed by:

$$\hat{\mathbf{f}}_{\mathbf{avg}}(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{B}} \sum_{\mathbf{b=1}}^{\mathbf{B}} \hat{\mathbf{f}}^{\mathbf{b}}(\mathbf{x})$$

In general, this method is not widely applied, considering that everyone does not have easy access to multiple training datasets, for various reasons, such as the impossibility of replicating that phenomenon or for purely economic issues or lack of time available. In contrast, the bootstrap technique obtains reproductions of samples from the individual training dataset, generating different B "bootstrapped" train samples. Finally, the statistical method is trained through the bootstrap set b-th in such a way as to obtain $\hat{f}^{*b}(x)$, and mediate all the predictions to obtain $\hat{f}_{bag}(x)$ like:

$$\hat{\mathbf{f}}_{\mathbf{bag}}(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{B}} \sum_{\mathbf{b=1}}^{\mathbf{B}} \hat{\mathbf{f}}^{*\mathbf{b}}(\mathbf{x})$$

So, this is the bagging procedure and to apply this algorithm to regression trees, you simply generate B trees using B bootstrapped train samples and mediate the resulting predictions. These trees are allowed to grow and are not subject to

"Pruning" procedures, i.e. pruning final observations. It has also been shown in other works (James et al., 2013) that to observe significant improvements in accuracy, the bagging procedure requires hundreds of tree replications in a single procedure.

2.4 Random Forest (RF)

Random forest is a method also applied in Viscarra-Rossel and Antoine Stevens (Viscarra-Rossel and Behrens, 2010; Stevens et al., 2013) and provides an improvement over trees developed with the bagging procedure through a small modification that decorrelates the trees. As in bagging, we build a series of decision trees on booted training samples. But when you build these decision trees, whenever a division in a tree is considered, a random sample of m predictors is chosen as candidates apart from the complete set of predictors. A new sample of m predictors is taken at each division, generally $m \approx \sqrt{p}$, which means the number of predictors considered in each division is roughly equal to the square root of the total number of predictors. In other words, in the construction of a random forest, with each division of the tree, the algorithm is not even allowed to consider most of the available predictors.

This may sound crazy, but it has intelligent logic. Suppose you have a very powerful predictor in your dataset, along with a number of other moderately strong predictors. So, in the collection of bagging trees, most or all trees will use this strong predictor in the upper-division (James et al., 2013). As a result, all bagging trees will look quite similar to each other, resulting in highly correlated predictions. Unfortunately, the average of many highly correlated quantities does not lead to a large reduction in variance as the average of many unrelated quantities. In particular, this means that the bagging algorithm will not result in a substantial reduction in variance on a single tree in this setting. Random forests overcome this problem by forcing each division to consider only a subset of the predictors.

Therefore, on average $\frac{p-m}{p}$ divisions will not even consider the strong predictor, and therefore other predictors will have a better chance. You can think of this process as a decoration of the trees, thus making their average less variable and, therefore, more reliable. The main difference between bagging and random forests is the choice of the size of the predictor subset of m size. For example, if a random forest is constructed using m plus p, this is simply the same as bagging. On the data used, the random forests they use lead to a reduction in error compared to the $m \approx \sqrt{p}$ bagging procedure.

Using a small value of m in building a random forest will generally be useful when we have a large number of related predictors. As with bagging, random forests do not adapt too much to the increase in the number of B iterations, but in practice, a value large enough to allow the error rate to stabilize has stabilized (James et al., 2013).

2.5 Boosting Regression (BR)

Like the bagging algorithm, boosting, or "enhancement," is a general approach that can be applied to many statistical learning methods for regression or classification. Here we limit our discussion on incentive to the context of regression trees, as approached before by Gholizadeh, Brown and Stevens (Gholizadeh et al., 2016; Brown et al., 2006; Stevens et al., 2013).

Remember that bagging involves creating multiple copies of the original training dataset using bootstrap, adapting a separate decision tree to each copy, and then combining all the trees to create a single predictive model. In particular, each tree is based on a bootstrap dataset, independent of other trees. Boosting works in a similar way, except that trees are grown sequentially: each tree is grown using information from previously grown trees. The upgrade does not involve bootstrap sampling, but each tree adapts to a modified version of the original dataset.

Considering the approach of the regressive technique, such as bagging, boosting also involves the combination of a large number of decision-making trees, $\hat{\mathrm{f}}^1(\mathrm{x}), \hat{\mathrm{f}}^2(\mathrm{x}), \dots, \hat{\mathrm{f}}^\mathrm{B}(\mathrm{x})$. The algorithm that governs the regressive approach of enhancement could thus be summarized as follow:

1. Consider the relation between the dependent variable y and the explicative ones: $y = f(x)$.
2. Set $\hat{\mathrm{f}}(\mathrm{x}) = 0$ and $\mathrm{r_i} = \mathrm{y_i}$ for each "*i*" in the train dataset; $\mathrm{r_i}$ and $\mathrm{y_i}$ are the residuals and the value of the response variable for the generic observation i, respectively.
3. For $\mathrm{b} = 1, 2, \dots, \mathrm{B}$ repeat the following sub-process:
   a) Fit a tree $\hat{\mathrm{f}}^\mathrm{b}$ with d splits (i.e. d-1 terminal nodes) to the dataset train.
   b) Update $\hat{\mathrm{f}}$ to add in a reduced version of the new regression tree: $\hat{\mathrm{f}}(\mathrm{x}) \leftarrow \hat{\mathrm{f}}(\mathrm{x}) + \lambda\hat{\mathrm{f}}^\mathrm{b}(\mathrm{x})$ where λ is a shrinking parameter.
   c) Update residuals as follows: $\mathrm{r_i} \leftarrow \mathrm{r_i} - \lambda\hat{\mathrm{f}}^\mathrm{b}(\mathrm{x})$.
4. Get the boosted model, as:

$$\hat{\mathbf{f}}(\mathbf{x}) = \sum_{\mathbf{b}=\mathbf{1}}^{\mathbf{B}} \boldsymbol{\lambda}\hat{\mathbf{f}}^{\mathbf{b}}(\mathbf{x})$$

The idea behind this procedure is this: unlike adapting a single large decision tree to the dataset, which equates to forced and potentially excessive data fitting (it could be affected by "overfit", that is, by overfitting the data by calibrating excessively correctly and validating less than enough), the enhancement approach allows for slow learning. Given the current model, it was preferred to adopt a regression tree to the remnants of the model. That is, we adapt a tree using the current residuals, instead of the variable Y, as an answer. You then add this new decision tree in the adapted function to update the residues. Each of these trees can be quite small, with a few terminal nodes, determined by the d parameter in the algorithm. By adapting small trees to the residues, we notice an improvement, albeit slow, of $\hat{f}$ areas where it does not work well. The shrinking parameter λ further slows down the process, allowing more and different-shaped trees to attack the residues. In general, slow-learning statistical learning approaches tend to work well. Note that in upgrading, the construction of each tree depends heavily on the trees that have already been cultivated. It can then be summarized that boosting has three optimization parameters:

1) The number of B trees. Unlike bagging and random forests, boosting can be oversized to the data if B is too large, although this oversizing tends to occur slowly if at all.
2) The shrinking parameter λ, a small positive number, which controls the speed with which it learns boosting. Typical values are 0.01 or 0.001 and the right choice may be the problem. A very small value of λ requires the use of a very large value of B to achieve good performance.
3) The number of divisions in each tree, which controls the complexity of the boosting set. Often d plus 1 works well, in case each tree is a "stump", consisting of a single division. In this case, the boosting set adapts to an additive model, because each term involves only a single variable. More generally, the d parameter can be interpreted as the depth of interaction and controls the interaction order of the boosting model, as d divisions can involve, at most, d variables. This highlights a difference between enhancement and random forests: in boosting, given that the growth of a particular tree considers others that have already been trained, it can be trusted to trust the condition that smaller trees are sufficiently adequate even in interpretation. For example, the use of stumps, mentioned above, leads to an additive pattern (James et al., 2013).

2.6 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) represents an artificial reproduction of a biological neural network of a human brain, including "neurons", nerve cells that are interconnected in a real network, and applied to predict soil contents, above all, by Kuang (Kuang et al., 2015).

However, it is important to note that in such a widespread network not all interconnections have the same specific weight in terms of importance; in fact, some have a high priority, associated with greater weight, than others. Like biological networks, artificial neural networks also have interconnected neurons and a pattern that faithfully reports the structure of a biological neural network.

Historically, the earliest ANNs are The Perceptron, proposed by Rosenblatt (Rosenblatt, 1958) and the Artron due to R. Lee (Lee, 1959). Then the Adaline (Adaptive Linear Neuron) and The Madaline (Many Adaline), due to Widrow et al. (Widrow et al., 1960, 1988). The first one is an artificial neuron also known as the ALC (adaptive linear combiner), the ALC being its principal component. The second one is an ANN (network) formulation based on the Adaline above, but it is a multilayer NN. Principles of the above four neurons are common building blocks in almost all ANN architectures.

Four major multi-layer general-purpose network architectures are:

- The Back-Propagation network: a multi-layer Perceptron-based ANN, giving an elegant solution to hidden-layers learning (Rumelhart et al., 1986). Its computational elegance stems from its mathematical foundation that may be considered as a gradient version of Richard Bellman's Dynamic Programming theory (Bellman, 1954)
- The Hopfield Network (Hopfield, 1982): this network is different from the earlier ANNs in many important aspects, especially in its recurrent feature of employing feedback between neurons. Hence, although several of its principles have been incorporated in ANNs based on the earlier four ANNs, it is to a great extent an ANN-class in itself. Its weight adjustment mechanism is based on the AM principle
- The Counter-Propagation Network (Hecht-Nielsen, 1987): Kohonen's Self-OrganizingMapping (SOM) is employed to facilitate unsupervised learning, utilizing the WTA principle to economize computation and structure.
- The LAMSTAR (LargeMemory Storage And Retrieval) network: a Hebbian network that uses a multitude of Kohonen SOM layers and their WTA principle. It is unique in its employs these by using Kantian-based Link-Weights (Graupe and Lynn, 1969) to link different layers (types of stored information) The link weights allow the network to simultaneously integrate inputs of various dimensions or nature of representation and incorporating correlation between input words. Furthermore, the network incorporates (graduated) forgetting in its learning structure and it can continue running uninterrupted when partial data is missing.

In this paper we use the Back-propagation method.

2.7 Multivariate Adaptive Regression Splines (MARS)

Many of the classic regression models have only linear aspects, but they can be adapted to non-linear models in the data by manually adding nonlinear terms to the model; however, in order to do so, the analyst must know a priori the specific nature of non-linearities and interactions. Alternatively, there are many inherently non-linear algorithms. When using these models, the exact shape of the non-linearity should not be explicitly known or specified before the model is formed. Rather, these algorithms will look for and discover non-linearities and interactions in data that help maximize predictive accuracy.

An example of such algorithms is the Multivariate Adaptive Regression Spline (MARS) (Friedman, 1991), an algorithm that automatically creates a linear pattern that sometimes provides an intuitive approach to the non-linearity after grasping the concept of multiple linear regression. They provide a cost-effective approach to capturing non-linear relationships in your data by evaluating breakpoints (nodes) similar to step functions. The procedure evaluates each data point for each predictor as a node and creates a linear regression model.

The MARS procedure will first search for the single point through the x-value range where two different linear relationships between Y and X reach the smallest error. For a single node, the hinge function is of the type:

$$y = \begin{cases} \boldsymbol{\beta_0 + \beta_1(a_1 - x)} & \mathbf{x < a_1} \\ \boldsymbol{\beta_0 + \beta_1(x - a_1)} & \mathbf{x > a_1} \end{cases}$$

Once the first node is found, the search continues for a second node. This procedure continues until many nodes are found, producing (potentially) a highly non-linear forecast equation. Including many nodes can allow you to adapt a really good relationship with the available training data, but it could lead you not to generalize, and therefore predict, very well with new and/or unknown data. Therefore, once you have identified the complete set of nodes, you can sequentially remove nodes that do not contribute significantly to predictive precision. This process is known as "pruning" and has been used to find the optimal number of nodes. There are two important optimization parameters associated with the MARS model: the maximum degree of interactions and the number of terms maintained in the final model. A grid search is necessary to identify the optimal mix of hyperparameters, i.e., the different combinations of interaction complexity and the number of terms to keep in the final model.

The advantages of MARS models are numerous. First, they naturally manage mixed types of predictors (quantitative and qualitative), consider all possible binary partitions of categories for a quantitative predictor in two groups, thus generating a pair of indicative functions for the two categories. These templates also require minimal functionality design and automatically select the features. For example, because they scan each predictor to identify a subdivision that improves predictive accuracy, non-informational features will not be selected. What is more, highly correlated predictors do not prevent predictive accuracy as much as OLS models.

However, one drawback of MARS models is that they are generally slower to train. Because the algorithm analyses each predictor value for potential breakpoints, computational performance can be affected by both increases in the number of observations and the number of variables. In addition, although related predictors do not necessarily hinder the model's performance, they can make it difficult to interpret. When two features are "almost perfectly" related, the algorithm will essentially select the first one that occurs when scanning features. Therefore, choosing one at random, the related function probably will not be included because it does not add any explanatory power to the analysis. (Gene et al., 1979). In soil environment, MARS are used to predict textures and contents by many authors (Volkan Bilgili et al., 2010; Viscarra-Rossel and Behrens, 2010; Nawar et al., 2016; Stevens et al, 2013).

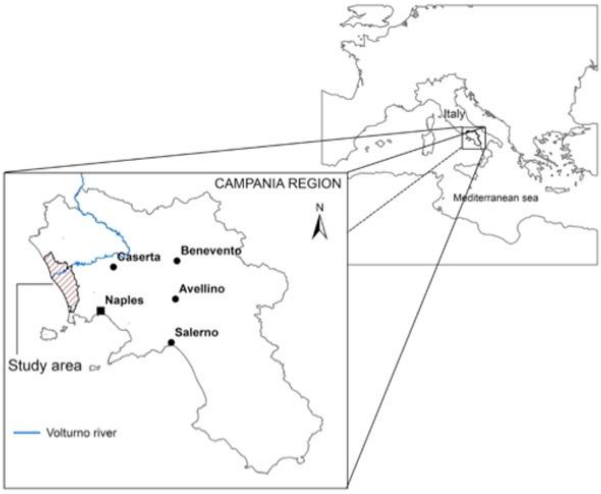
3. Materials and methods

3.1 Study area and soil sampling

The area under investigation (Figure 1) is located in the north-western part of the Campania Region, in southern Italy (Coord. 41°01’00’’ N, 13°58’00’’ E), within a fertile agricultural land, mainly devoted to irrigated vegetal crops and fruit trees (Geoportale Regione Campania, 2019). The climate is typically Mediterranean, with the wettest period between late autumn (October–November) and early spring (March–April). Temperature and potential evapotranspiration temperatures show an inverse trend compared to rainfall, with the highest values during summer (June–August).

The dominant soils types are Gleyic, Gleyic-Vertic, Calcari-Gleyic and Calcari-Fluvic Cambisols, and Calcaric Gleysols (Di Gennaro, 2002). For this study, an existing soil database, made available CNR-ISAFoM, was used. The database contains information on soil organic carbon (OC) and particle size distribution (sand, silt, and clay), used in our application. Information about OC was available for ninety-six samples, while those for sand, silt, and clay contents were available only for eighty-two samples.

Figure 1. Localisation of the study area in Southern Italy.



The analytical data regarded surface soil samples randomly collected in 1999 within the study area, air-dried, and ground to a size fraction passing a 2 mm sieve. Soil organic carbon and texture were determined according to the Italian Official Methods for Soil Analysis (MIPAF, 2000). Namely, total clay (soil separate with < 0.002 mm particle diameter) and silt (soil separate with 0.002 to 0.05 mm particle diameter) contents were determined with the pipet method. Total sand content (soil separate with 0.05 to 2.0 mm particle diameter) was determined by wet sieving; OC content was determined using Walkey-Black methods.

3.2 Vis-NIR spectroscopy

The diffuse vis–NIR spectral reflectance was measured in the laboratory, on a residual fraction of soil samples, under controlled light conditions, using the procedure described in Leone et al. (2019). Noisy portions of the measured reflectance spectra, between 350 and 399 nm and between 2451 and 2500 nm, were removed, leaving spectra in the range of 400-2450 nm for the analysis. The resulting reflectance spectra were normalised, using The continuum removal approach (Clark and Roush, 1984). To this end, a convex hull was fitted over the original spectral curve, then the absorption spectrum was calculated by taking the ratio between the original reflectance spectrum and the enveloping curve (Van der Meer, 1999; De Jong, 1992).

3.3 Statistical calibrations

The selected statistical calibrations were performed to predict the investigated soil properties from reflectance spectra, using both the software “ParLeS” (Viscarra-Rossel, 2008) and “R x64 3.6.3” (R Core Team, 2020). All models performed to calibrate the spectral data with the reference (laboratory) soil data have employed two-thirds of the available samples for calibration and the remaining third for independently validating them. For each variable, the selection of samples was carried out as follows: first, the samples were sorted following ascending order of the variable, then, sequentially, every two samples were taken for calibration and the third for validation**.**

To enhance the predictive power of these statistical calibration models, spectroscopic data were transformed and pre-processed prior to data analysis, with the aim of removing undesired variation in the data (Eriksson et al., 2006). In this study, we assessed all the transformation and pre-processing methods, either alone or in combination, before calibrations.

The combination of the following procedures provided the best results: reflectance (R) to absorbance (A) transformation (A = log 1/R), wavelet detrending, median filtering, second derivative of absorbance, and data enhancement (mean centre). In particular, reflectance to absorbance transformation reduces nonlinearities (Viscarra-Rossel, 2008), while wavelet detrending (Daubechies, 1992) corrects light scattering variation and baseline. The median filter, in addition (Viscarra-Rossel, 2008), reduces the effects of random spectral noise, thereby providing smoother spectra. Lastly, the second derivative removes additive and linear baseline effects (Burger and Geladi, 2007), while amplifying absorption features, which are indicative of the contents of the soil materials. Mean centring is a commonly used method of data enhancement to reduce redundant information and better evaluate differences.

Leave-one-out cross-validation (Efron and Tibshir, 1994) was then used to determine the number of factors to retain in the calibration models. To select the optimal cross-validated calibration model, we computed the root mean square error (RMSE) of predictions:

$$\mathbf{RMSE} = \sqrt{\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{y_{pred}} - \mathbf{y_{ob}})^{2}}$$

in which N is the sample population size, $\mathbf{y_{pred}}$ is the predicted value, and $\mathbf{y_{ob}}$ is the observed value. In this case, the model with the lowest RMSE is selected. However, a more parsimonious model, i.e., a model with fewer factors representing the variability in the data set, without causing overfitting, is preferred. For that purpose, the optimal selection of factors can be based on the penalizing Akaike Information Criterion (AIC) (Akaike, 1969; Li et al., 2002):

$$\mathbf{AIC= N\ log\ (RMSE) + 2m.}$$

in which N is the sample population size and m is the number of model parameters (i.e., the number of factors). This criterion is applied after have verified that residuals have zero mean normal distribution.

To evaluate the accuracy of models, the adjusted coefficient of determination ($\mathbf{R^2_{adj}}$) and the relative percent deviation (RPD), i.e., the ratio of the standard deviation of analysed data (i.e., the soil properties) to RMSE, was performed. In accordance with previous studies (Williams, 1987; Viscarra-Rossel, 2007) the quality of predictions expressed by RPD was classified as follows: RPD < 1.0 indicates very poor model/predictions and their use is not recommended; RPD between 1.0 and 1.4 indicates poor model/predictions where only high and low values are distinguishable; RPD between 1.4 and 1.8 indicates fair model/ predictions which may be used for assessment and correlation; RPD values between 1.8 and 2.0 indicates good model/ predictions where quantitative predictions are possible; RPD between 2.0 and 2.5 indicates very good, quantitative model/ predictions, and RPD > 2.5 indicates excellent model/predictions. RPD statistic is also carried out to assess the performance of validation using the independent data set.

4. Results and discussion

4.1 Descriptive statistics of soil properties

The investigated soil variables were statistically described in terms of minimum, maximum, mean, coefficient of variation (CV), and skewness. Furthermore, a log-transformation was performed for those variables that did not follow a normal distribution. Summary statistics of calibration and validation subsets are reported in Table 1. Organic carbon content ranges from 2.71 to 215.6 g $\mathrm{Kg^{-1}}$, and is on average moderate (21.5 g $\mathrm{Kg^{-1}}$). A slight difference in the average values can be observed between the calibration (22.5 g $\mathrm{Kg^{-1}}$) and validation (19.6 g $\mathrm{Kg^{-1}}$) sub-sets. Skewness always exhibits high values: 4.57 g $\mathrm{Kg^{-1}}$ for the whole dataset; 4.52 and 3.74 g $\mathrm{Kg^{-1}}$, for the calibration and validation sub-stets, respectively, thus indicating a significant deviation from the normal distribution.

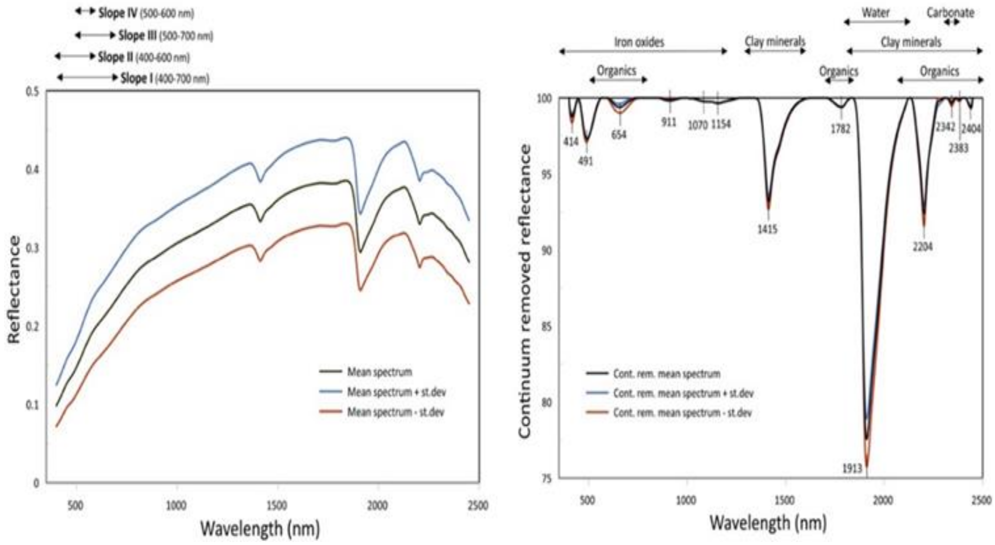
Table 1. Descriptive statistics of the selected soil properties for calibration and validation datasets.

		n	**Mean**	**Range**	**CV**	**Skewness**
OC	Calibration	64	22.5	4.4 – 215.6	1.43	4.52
(g Kg^{-1})	Validation	32	19.6	2.8 – 121.4	1.12	3.74
Sand	Calibration	54	419.8	80.0 – 940.0	0.53	0.59
(g Kg^{-1})	Validation	28	423.6	70.0 – 950.0	0.56	0.64
Silt	Calibration	54	201.1	10.0 – 370.0	0.39	-0.35
(g Kg^{-1})	Validation	28	201.1	10.0 – 390.0	0.43	-0.23
Clay	Calibration	54	377.4	50.0 – 730.0	0.46	-0.14
(g Kg^{-1})	Validation	28	378.9	10.0 – 770.0	0.50	-0.09

Soil separates, i.e., the size groups of mineral particles, is dominated by the sand, (421.1 g Kg^{-1}), on average), followed by clay (377.7 g Kg^{-1}) and silt (201.1 g Kg^{-1}) fractions. The dominant, basic soil textural classes are: clay, clay-loam, sandy-clay-loam, and sandy-loam. Extreme and mean values for all, sand, silt, and clay calibration and validation subsets are similar. Skewness was consistently low, thus indicating, for these variables, a frequency distribution close to the normal distribution. Differences between calibration and validation sub-sets are minimal, and the CV is moderate for both these variables. Skewness is consistently low. Considering that the mean and coefficient of variation (CV) for the calibration and validation sets are comparable for all the considered soil properties, the selection of both datasets can be considered representative (Ding et al., 2018). Figure 2 shows the average soil spectrum and the relative continuum removed reflectance of the investigated soil samples and their standard deviation.

The average spectrum (Figure 2a) shows a typical convex shape and a moderate overall reflectance. The dispersion of the spectral intensity, as measured by the standard deviation, was evident. Many studies demonstrated that different soil properties, especially particle size distribution and organic carbon content, may affect the overall reflectance (Stenberg et al., 2010). Changes in slopes of different ranges in the visible region are also observed. Various studies have related visible reflectance slope to soil organic matter content (Summers et al., 2011). The average continuum removed spectrum (Figure 2b) shows several absorption bands across the entire vis-NIR region, which can be related to clay minerals, organic matter, iron oxides, water, and carbonate contents (Stenberg et al., 2010; Leone A.P., 2000).

Figure 2. Mean of spectral reflectance **(a)** and continuum removed spectral reflectance **(b)** of sampled soils. In **(a)** the position of spectral ranges where the various visible reflectance slopes were calculated; in **(b)** the approximate positions of some fundamental soil constituents are shown.

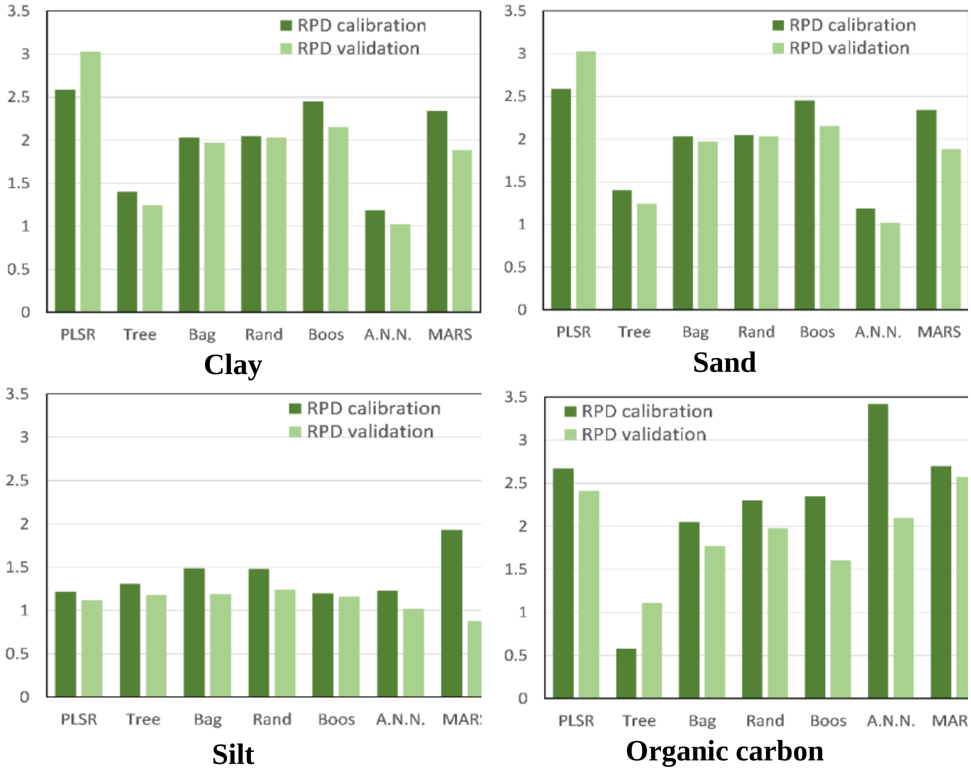


(a) **(b)**

4.2 *Multivariate and ensemble calibrations*

The capability of vis-NIR reflectance spectroscopy to predict the investigated soil properties among different statistical methods is summarised in Table 2, and shows that, as a general rule and considering both calibration and validation results, PLS gives the best results. The comparison among different models is immediately evident in Figure 3. However, the response of those methods in both Table 2 and Figure 3 gives slightly different results depending on the variable considered. In any case, clay and organic carbon are the best-predicted variables. Specifically, PLSR applied to two-thirds of the available sample set revealed good correlations between soil reflectance spectra and the considered soil properties, except for silt content. Based on the RPD values , the calibration models were excellent for log-OC (RPD = 2.67) and clay (RPD = 2.59) and good for sand (RPD = 1.95). For clay, models including 5 factors, based on the RMSE and AIC values, allowed to attain a cross-validation between predicted and measured data with R^2_{adj} of 0.855, 0.845 and 0.731, for log-OC, clay and sand, respectively. For silt, the only calibration possible performed a poor model, with an R^2_{adj} of 0.314 and RPD of 1.22.

Figure 3. Histograms of calibration and validation RPD of all statistical models for soil variables.



In some cases, increasing the number of factors gave slightly higher coefficients of regression (R^2), but increased RMSE, thus reducing the stability of the calibration models (i.e., leading to over-fitting) (Vågen et al., 2006; Wise et al., 2003). Therefore, we selected the most parsimonious model in terms of number of factors, based on the values of AIC and RMSE.

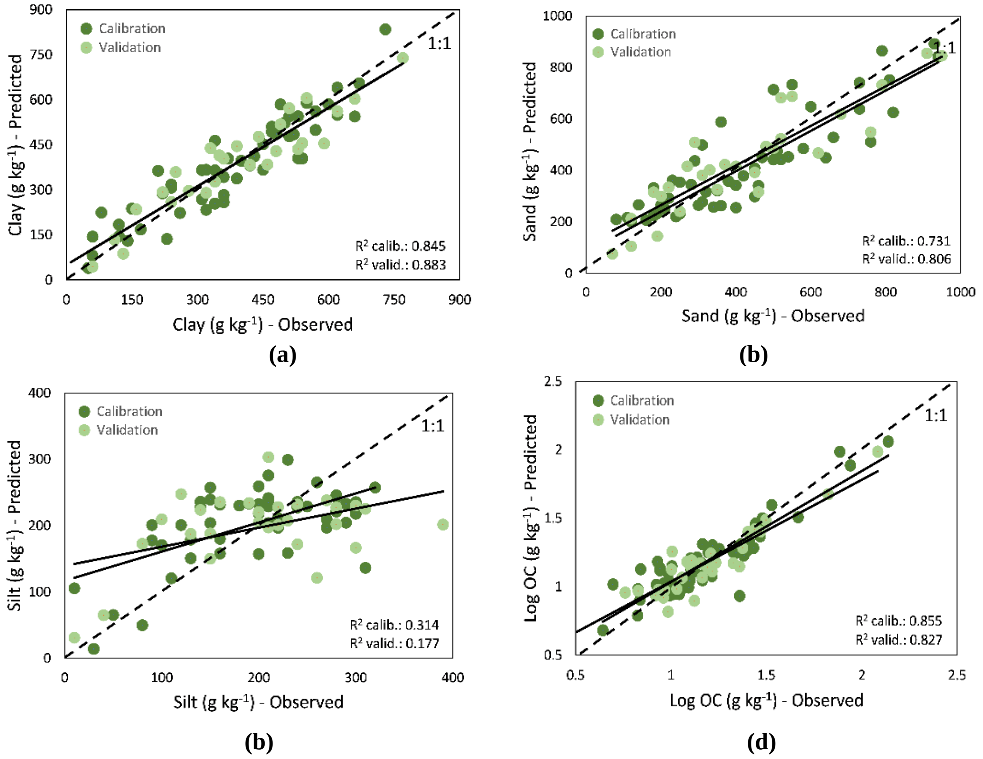
Leave-one-out calibration models constructed through vis-NIR reflectance spectroscopy and PLSR are empirical; therefore, validations of these models are better performed using a data set that is independent of the one used for calibration (Volkan Bilgili et al., 2010). Validation using the remaining one third of the available samples indicated excellent models for the prediction of clay (R^2_{adj} = 0.883 and RPD = 3.03), very good models for log-OC (R^2_{adj} = 0.827 and RPD = 2.41) and for sand (R^2_{adj} = 0.806 and RPD = 2.10), and a poor model for silt (R^2_{adj} = 0.177 and RPD = 1.11).

Table 2. RMSE and RPD for different methods applied on calibration and validation sample for Clay, Sand, Silt and OC contents in soil samples.

Variables	**Models**	**Calibration sample**		**Validation sample**	
		RMSE	**RPD**	**RMSE**	**RPD**
Clay	PLSR	67.61	2.59	62.46	3.03
	Tree	125.18	1.40	152.51	1.24
	Bagging	86.31	2.03	95.76	1.97
	Random Forest	85.47	2.05	93.07	2.03
	Boosting	71.28	2.45	88.05	2.15
	A.N.N.	146.72	1.19	185.56	1.02
	M.A.R.S	74.70	2.34	100.54	1.88
Sand	PLSR	114.61	1.95	102.40	2.32
	Tree	120.02	1.86	134.17	1.77
	Bagging	116.83	1.91	117.16	2.03
	Random Forest	120.48	1.86	113.18	2.10
	Boosting	118.17	1.89	121.67	1.96
	A.N.N.	186.15	1.20	233.66	1.02
	M.A.R.S	117.48	1.90	135.04	1.76
Silt	PLSR	64.99	1.22	77.19	1.12
	Tree	60.20	1.31	73.19	1.18
	Bagging	52.99	1.49	73.03	1.19
	Random Forest	53.45	1.48	69.60	1.24
	Boosting	65.80	1.20	74.53	1.16
	A.N.N.	64.44	1.23	84.99	1.02
	M.A.R.S	41.09	1.93	97.90	0.88
OC	PLSR	0.12	2.67	0.13	2.41
	Tree	0.20	0.58	0.28	1.11
	Bagging	0.15	2.05	0.17	1.77
	Random Forest	0.14	2.30	0.15	1.98
	Boosting	0.13	2.35	0.19	1.61
	A.N.N.	0.09	3.42	0.15	2.10
	M.A.R.S	0.12	2.70	0.12	2.57

To make the reader more comfortable with the results of PLSR prediction, scatterplots of the predicted vs measured values for these properties are shown in Figure 4. In this plot, the values of the $\mathrm{R}^2_{\mathrm{adj}}$ and regression's straight lines of both the datasets are also highlighted. In order to make exhaustive the discussion about the results of this work, the outcomes of the other models cannot be overlooked.

Figure 4. Scatterplots of observed vs predicted soil properties for calibration and validation data sets in PLSR.



(a) (b)

(b) (d)

In particular, considering singularly all different models examined through the statistical-computational environment R (R Core Team, 2020), there are some models with behaviour similar to PLSR. One of these models is MARS, that has performed very good/excellent values in terms of RPD in Clay and OC ($2.3 \leq \text{RPD}_{\text{cal}} \leq 2.7$) and good ones in Sand and Silt. It is remarkable that MARS is better than PLSR to predict OC, as evidenced by the excellent values of RMSE and RPD in both calibration and validation datasets. In the validation phase, this model returns good outputs ($1.8 \leq \text{RPD}_{\text{val}} \leq 2.6$), excluding Silt, in which predictions are not recommended ($\text{RPD}_{\text{val}} \leq 1$). The second model in order of goodness of predictions is Boosting, thanks to its good values of RPD in both calibration and validation terms ($1.6 \leq \text{RPD}_{\text{cal,val}} \leq 2.5$). Also in this case, RPD for Silt variable have unacceptable values ($\text{RPD}_{\text{cal,val}} \leq 1.2$).

Another good model in prediction of content of soils are RF, in fact in this case study, it has performed very good values of RPD both in calibration and validation sets for Clay, Sand and OC ($2 \leq \text{RPD}_{\text{cal,val}} \leq 2.3$), while fair and poor ones for Silt

($1.2 \leq \text{RPD}_{\text{cal,val}} \leq 1.5$). It is also very useful to highlight that RF has the best compromise, in RPD outcomes, to predict Silt among the various models performed.

Bagging is an alternative model that has carried out good RPD values between 1.5 and 2 for Clay, Sand, and OC, while between 1 and 1.5 for Silt. ANN instead, has performed lower values of RPD in Clay Sand and Silt, denoting itself as a poor model to predict soil texture but an excellent model in order to predict OC with values extremely good ($\text{RPD}_{\text{cal}} = 3.4$ and $\text{RPD}_{\text{val}} = 2.1$). The worst model performed is RT, which in all variables considered, carried out poor/slightly fair values of RPD in both calibration and validation datasets, but its usage is still not recommended.

5. Conclusions

This paper aims to evaluate the goodness of different multivariate and statistical ensemble methods in order to predict some soil properties.

Analogies and differences with our results appear in other papers, where authors applied similar techniques in different geographic areas, performing statistical calibration. For all properties we analyzed, PLSR is the technique that gived best results. This technique is the most complete and it is useful to predict many soil properties. Anyway, other alternative methods give good results. It would be worth if these techniques would be applied to deepen studies in soil properties predictions. Please refer to future studies in order to develop and broaden these issues, which are the subject of numerous papers.

Statement

All authors reviewed and revised the manuscript, approved the final version, and agreed to submit the revised manuscript for publication. The authors state that they have no disclosure to declare.

References

AA.VV., (1995), ‘*S-Plus, Guide to Statistics*’, Seattle-WA: MathSoft.

Akaike, H., (1973), ‘Second International Symposium on Information Theory’, in *Information theory and an extension of the maximum likelihood principle, B. Petrov & F. Csaki, eds. Budapest: Akademiai Kiado,* pp267-281.

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., (2014), 'Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques', *Eur. J. Soil Sci.* 65 (5), pp718–729.

Bellman, R., (1954) 'The theory of dynamic programming'. *Bull. Amer. Math. Soc.*, 60(6), pp. 503-515.

Breiman, L., Friedman, J., Olshen, R. & Stone, C., (1984), '*Classification and Regression Trees'*, New York - London: Chapman & Hall.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., (2006), 'Global soil characterization with VNIR diffuse reflectance spectroscopy', *Geoderma* 132 (3–4), pp273–290.

Burger, J.; Geladi, P., (2007), 'Spectral pre-treatments of hyperspectral near infrared images: Analysis of diffuse reflectance scattering', *J. Near Infrared Spec* 15, pp29–37.

Clark, R.N.; Roush, T.L., (1984), 'Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications', *J. Geophys. Res.* 89, pp6329–6340.

Cozzolino, D. and Moròn A., (2003), 'The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics', *Journal of Agricultural Science*, 140, pp65–71.

Daubechies, I., (1992), 'Ten Lectures on Wavelets', *Society for Industrial and Applied Mathematics: Philadelphia*, PA, USA, p341.

De Jong, S., (1992), 'The analysis of spectroscopical data to map soil types and soil crusts of Mediterranean eroded soils', *Soil Technol.* 5, pp199–211.

Di Gennaro, A., (2002), 'I sistemi di terra della Campania', *SELCA: Florence, Italy*, p63.

Ding, J.; Yang, A.; Wang, J.; Sagan, V.; Yu, D., (2018), 'Machine-learning-based quantitative estimation of soil organic carbon content by VIS/NIR spectroscopy', *Peer J.* 5714, pp1–14.

Drury, S.A.; (1993)*, 'Image interpretation in geology'*, Chapman & Hall: London, 1993.

Dunn, B.W.; Beecher, H.G.; Batten, G.D.; Ciavarella, S., (2002), 'The potential of near-infrared reflectance spectroscopy for soil analysis - A case study from the Riverine Plain of South-Eastern Australia'. *Aust. J. Exp. Agric*., 42, pp607–614.

Efron, B.; Tibshirani, R.J., (1993), '*An Introduction of the Bootstrap'*, 1st ed., Chapman and Hall: New York, NY, USA, p436.

Ergon, R., (2009), 'Re-interpretation of NIPALS results solves PLSR inconsistency problem', *J. Chemom.* 23/1, pp72-75.

Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wilström, C.; Wold, S., (2006), '*Multi-and Megavariate Data Analysis; Part I-Basic principles and applications*', Umetrics Academy: Umeå, Sweden, p425.

FAO, Food and Agriculture Organization of United Nations, (2015), 'Status of the World's soil resources', Rome, pp607.

Friedman, J. H., (1991), '*Multivariate adaptive regression splines*', Annals of Statistics 19, pp1-67.

Fystro, G., (2002), 'The prediction of C and N content and their potential mineralisation in heterogeneous soil samples using Vis-NIR spectroscopy and comparative methods', *Plant Soil* 246, pp139–149.

Gene, G., Heath, M. & Wahba, G., (1979), 'Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter', *Technometrics* 21, 2, pp215–230.

Geoportale Regione Campania, (accessed on 10 February 2019), *Sistema Informativo Territoriale della Regione Campania. Available online*: https://sit2.regione.campania.it/content/carta-utilizzazione-agricola-dei-suoli.

Gholizadeh, A., Borůvka, L., Saberioon, M., Vašát, R., (2016), 'A memory-based learning approach as compared to other data mining algorithms for the prediction of soil texture using diffuse reflectance spectra', *Remote Sens*. 8 (4), 341.

Graupe, D. and Lynn, J.W. (1969) 'Some aspects regarding mechanistic modelling of recognition and memory', Cybernetics 3, pp119-141.

Hecht-Nielsen, R., (1987) 'Counter propagation networks', *Applied Optics* 26, pp4979-4984

Helland, I., (1998), 'On the structure of partial least square regression', *Commu. Stat* 17, pp311-388.

Hopfield, J.J., (1982) 'Neural networks and physical systems with emergent collective computational abilities', *Proceedings of the National Academy of Sciences* 79, pp2554-2558

Irons, J.R.; Weismiller, R.A.; Petersen, G.W., (1989), '*Theory and Applications of optical remote sensing'*; G. Asrar, Ed.; Wiley: New York, pp66-106.

James, G., Witten, D., Hastie, T. & Tibishirani, R., (2013), 'An Introduction to Statistical Learning with Application in R', *G. Casella, S. Fienberg & I. Olkin, eds. Springer Texts in Statistics. New York: Springer Science+Business Media*, pp1-426.

Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., Van der Putten, W.H., Bardgett, R.D., Moolenaar, S., Mol G., Jansen, B., Fresco, L.O., (2016) 'The significance of soils and soil science towards realization of the United Nations Sustainable: Development Goals', *SOIL An interactive open-access journal of the European Geosciences Union, Copernicus Publications*, Vol. 2 N°2, pp.111-128.

Kuang, B.; Tekin, Y. and Mouazen, A.M., (2015), 'Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content', *Soil & Tillage Research Science Direct* 146, pp243-252.

Lee, R.J. (1959) 'Generalization of learning in a machine'. *Proc. 14th ACM National Meeting*.

Lee, K.S., Lee, D.H., Sudduth, K.A., Chung, S.O., Kitchen, N.R., Drummond, S.T., (2009), 'Wavelength identification and diffuse reflectance estimation for surface and profile soil properties', *Trans. ASABE* 52 (3), pp683–695.

Leone, A.P., (2000), 'Spettrometria e valutazione della riflettanza spettrale dei suoli nel dominio ottico 400–2500 nm', *Riv. Ital. Telerilevamento* 9, pp3–28.

Leone, A.P.; Viscarra-Rossel, A.R.; Amenta, P.; Buondonno, A., (2012), 'Prediction of soil properties with PLSR and vis-NIR spectroscopy: Application to Mediterranean soils from Southern Italy'. *Curr. Anal. Chem*., 8, pp283–299.

Leone, A.P.; Leone, G.; Leone, N.; Galeone, C.; Grilli, E.; Orefice, N.; Ancona, V., (2019), ‘Capability of diffuse Reflectance Spectroscopy to predict soil water retention and related soil properties in an irrigated lowland district of southern Italy’, *Water Science and Technology* 11, 1712.

Li, B.; Morris, J.; Martin, E.B., (2002), ‘Model selection for partial least squares regression’, *Chemom. Intell. Lab. Syst.* 1, pp79–89.

Lucadamo, A.; Amenta, P.; Leone, N. (2020), ‘Soil texture prediction via reduced k-means principal component multinomial regression’, *Socio-Economic Planning Sciences* (in press).

Lucadamo, A.; Leone, A., (2015), ‘Principal Component Multinomial Regression and spectrometry to predict soil texture’, *Journal of chemometrics* 29 (9), pp 514–520.

Martens, H., (1985), ‘Multivariate Calibration’, Dr. techn. Thesis, Technical University of Norway.

Martens, H.; Naes, T., (1989), ‘*Multivariate Calibration’*, John Wiley & S.: Chichester, UK.

McCarty, G.W.; Reeves, J.B., III; Reeves, V.B.; Follett, R.F.; Kimble, J.M., (2002), ‘Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement’, *Soil Sci. Soc. Am. J.*, 66, pp640–646.

Milton; E.J. (1987), ‘Principles of field spectroscopy’, *Int. J. Remote Sens*., 12, pp1807-1827.

MIPAF, Ministero delle Politiche Agricole e Forestali, (2000), ‘*Metodi di Analisi Chimica del Suolo*’; Franco Angeli: Milan, Italy, p536.

Mouazen, A.M., Kuang, B., De Baerdemaeker, J., Ramon, H., (2010), 1Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy’, *Geoderma* 158 (1), pp23–31.

Mouazen, A.M.; Maleki, M.R.; Cockx, L.; Van Meirvenne, M.; Van Holm, L.H.J.; Merckx, R.; De Baerdemaeker, J.; Ramon, H., (2009), ‘Optimum three-point linkage set up for improving the quality of soil spectra and the accuracy of soil phosphorus measured using an on-line visible and near infrared sensor’, *Soil and Tillage Research*, Volume 103, Issue 1, pp144-152.

Mouazen, A.M.; Maleki, M.R.; De Baerdemaeker, J.; Ramon, H., (2007), ‘On-line measurement of some selected soil properties using a VIS-NIR sensor’, *ScienceDirect Soil & Tillage Research* 93, pp13-27.

Mouazen, A.M.; Maleki, M.R.; De Ketelaere, D.; Ramon, H.; De Baerdemaeker, J.; (2008), ‘On-the-go variable-rate phosphorus fertilisation based on a visible and near-infrared soil sensor’, *Biosystems engineering Science direct* 99, pp35-46.

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., Mouazen, A.M., (2016), ‘Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy’, *Soil Tillage Res.* 155, pp510–522.

Pell, R., Ramos, L. & Manne, R., (2007), ‘The model space in partial least squares regression’, *J. Chemom.* 21, pp165-172.

Phatak, A. & De Jong, S., (1997), ‘The geometry of partial least squares’, *J. Chemo*. 11, pp311-338.

R Core Team, (2020), '*R: A language and environment for statistical computing. R Foundation for Statistical Computing*', Vienna, Austria. URL https://www.R-project.org/.

Ripley, B., (1996), '*Pattern Recognition and Neural Networks'*. Cambridge: Univ Press.

Rosenblatt, F., (1958) The perceptron, a probabilistic model for information storage and organization in the brain, *Psychol. Rev.* 65, pp386-408.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) 'Learning internal representations by error propagation', in *Parallel Distributed Processing. Explorations in the Microstructures of Cognition*, eds. Rumelhart, D.E. and McClelland, J.L. MIT press, Cambridge, MA.

Shi, Z., Ji, W., Viscarra-Rossel, R.A., Chena, S., Zhou, Y., (2015), 'Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis–NIR spectral library', *European Journal of Soil Science*.

Stenberg, B.; Viscarra-Rossel, R.A; Mouazen, A.M.; Wetterlind, J. (2010), 'Visible and near infrared spectroscopy in soils science'. *Advances in Agronomy*, 107, pp163-215.

Stevens, A., Udelhoven T., Denis A., Tychon, B., Lioy R., Hoffmann L., Van Wesemael B., (2010), 'Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy', *Geoderma* 158, pp32-45.

Stevens, A.; Nocita, M.; Tóth, G.; Montanarella, L.;Van Wesemael, B.; (2013), 'Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy', *PLSoS One* 8(6), e66409.

Vågen, T.G.; Shepherd, K.D.; Walsh, M.G., (2006), 'Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy', *Geoderma* 133, pp281–294.

Van der Meer, F., (1999), 'Can we map swelling clays with remote sensing?', *Int. J. Appl. Earth Obs. Geoinf.* 1, pp27–35.

Vasques, G.M.; Grunwald, S.; Sickman, J.O., (2008), 'Comparison of multivariate methods for inferential modelling of soil carbon using visible/near-infrared spectra'. *Geoderma*, 146, pp14-25.

Venables, W. & Ripley, B., (1997), '*Modern Applied Statistics with S-Plus'*, Springer.

Vibhute, A.D; Kale, K.V.; Mehrotra, S.C.; Dhumal, R.K.; Nagne, A.D., (2018), 'Determination of soil physicochemical in farming sites through visible, near-infrared diffuse reflectance spectroscopy and PLSR modeling', *Ecological Processes, Springer Open*, pp7-26.

Viscarra-Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O.; (2006), 'Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties', *Geoderma*, 131 (1), pp59–75.

Viscarra-Rossel, R.A., (2007), 'Robust modelling of soil diffuse reflectance spectra by bagging-partial least square regression', *J. Near Infrared Spec* 15, pp39–47.

Viscarra-Rossel, R.A., (2008), 'ParLeS: Software for chemometric analysis of spectroscopic data', *Chemom. Intell. Lab. Syst*. 90, pp72–83.

Viscarra-Rossel, R.A., Lark, R.M., (2009), 'Improved analysis and modelling of soil diffuse reflectance spectra using wavelets', *Eur. J. Soil Sci*. 60 (3), pp453–464.

Viscarra-Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthes, B.G., Baratholomeus, H.M., Bayer, A.D., Bernoux, M., Bottcher, K., Brodsky, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morras, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Rufasto Campos, E.M., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlns, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., (2016), 'A global spectral library to characterize the world's soil', *Earth Sci. Rev.* 155, pp198–230.

Vohland, M., Besold J., Hill J., Fründ H.C., (2018), 'Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy', *Geoderma* 166, pp198-205.

Wang, S.Q., Li, W.D., Li, J., Liu, X.S., (2013), 'Prediction of soil texture using FT-NIR spectroscopy and PXRF spectrometry with data fusion', *Soil Sci.* 178 (11), pp626–638.

Were K., Tien Bui D., Dick Ø.B., Singh B.R., (2015), 'A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape', *Ecological Indicators* 52, pp394-403.

Wetterlind, J., Stenberg, B. and Jonsson, A., (2008), 'Near infrared reflectance spectroscopy compared with soil clay and organic matter content for estimating within-field variation in N uptake in cereals', *Plant and soil*. 302, 1-2, pp317-327.

Widrow, B. and Hoff, M.E., (1960) 'Adaptive switching circuits', *Proc. IRE WESCON conf.*, New York, pp. 96-104

Widrow, B. and Winter, R. (1988) 'Neural nets for adaptive filtering and adaptive pattern recognition', *Computer*, 21, pp. 25-39.

Williams, P.C., (1987), 'Variables affecting near-infrared reflectance spectroscopic analysis', *Near-Infrared Technology in the Agricultural and Food Industries*, American Association of Cereal Chemists Inc, pp143–167.

Wise, B.M.; Gallagher, N.B.; Bro, R.; Shaver, J.M., (2003), '*PLS Toolbox Version 3.0 for Use with Matlab*', Eigenvector Research Inc., p171.

Wold, H., (1973), '*Multivariate Analysis*', vol III, pp383-407.

Wold, S., Martens, H. & Wold, H., (1983), 'Lecture Notes in Mathematics'. In: A. Ruhe & B. Kagstrom, eds. *Proceedings of the Conference on Matrix Pencils.* Heidelberg, Germany: Springer-Verlag, pp286-290.

Wold, S., Johannson, E. & Cocchi, M., (1993) '3D QSAR in Drug Design, Theory, Methods, and Applications', *ESCOM Science Publishers: The Netherlands*, pp523-550.

Wold, S., Eriksson, L., Trygg, J. & Kettaneh, N., (2004), '*Proceedings of COMPOSTAT 2004',*. Prague, Physica Verlag: Germany, pp522-529.

Xu S., Zhao Y., Wang M., Shi X., (2018), 'Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy', *Geoderma* 310, pp29-43.