

Article

Examining the Performance of PARACUDA-II Data-Mining Engine versus Selected Techniques to Model Soil Carbon from Reflectance Spectra

Asa Gholizadeh ^{1,2,*} , Mohammadmehdi Saberioon ³ , Nimrod Carmon ^{4,5}, Lubos Boruvka ¹ and Eyal Ben-Dor ^{4,*} 

¹ Department of Soil Science and Soil Protection, Faculty of Agrobiolgy, Food and Natural Resources, Czech University of Life Sciences Prague, Kamycka 129, 16500 Prague, Czech Republic; boruvka@af.czu.cz

² Czech Geological Survey, Klarov 3, 11800 Prague, Czech Republic

³ Laboratory of Signal and Image Processing, Institute of Complex Systems, South Bohemian Research Centre of Aquaculture and Biodiversity of Hydrocenoses, Faculty of Fisheries and Protection of Waters, University of South Bohemia in Ceske Budejovice, Zamek 136, 37333 Nove Hrad, Czech Republic; msaberioon@frov.jcu.cz

⁴ Remote Sensing Laboratory, Department of Geography and Human Environment, Porter School of Environment and Earth Science, Tel-Aviv University, Tel-Aviv 6997801, Israel; carmonmon@gmail.com

⁵ Porter School of Environment and Earth Science, Tel-Aviv University, Tel-Aviv 6997801, Israel

* Correspondence: gholizadeh@af.czu.cz (A.G.); bendor@post.tau.ac.il (E.B.-D.); Tel.: +420-22-438-2633 (A.G.); +972-3-640-7049 (E.B.-D.)

Received: 6 June 2018; Accepted: 20 July 2018; Published: 25 July 2018



Abstract: The monitoring and quantification of soil carbon provide a better understanding of soil and atmosphere dynamics. Visible-near-infrared-short-wave infrared (VIS-NIR-SWIR) reflectance spectroscopy can quantitatively estimate soil carbon content more rapidly and cost-effectively compared to traditional laboratory analysis. However, effective estimation of soil carbon using reflectance spectroscopy to a great extent depends on the selection of a suitable preprocessing sequence and data-mining algorithm. Many efforts have been dedicated to the comparison of conventional chemometric techniques and their optimization for soil properties prediction. Instead, the current study focuses on the potential of the new data-mining engine PARACUDA-II[®], recently developed at Tel-Aviv University (TAU), by comparing its performance in predicting soil oxidizable carbon (Cox) against common data-mining algorithms including partial least squares regression (PLSR), random forests (RF), boosted regression trees (BRT), support vector machine regression (SVMR), and memory based learning (MBL). To this end, 103 soil samples from the Pokrok dumpsite in the Czech Republic were scanned with an ASD FieldSpec III Pro FR spectroradiometer in the laboratory under a strict protocol. Spectra preprocessing for conventional data-mining techniques was conducted using Savitzky-Golay smoothing and the first derivative method. PARACUDA-II[®], on the other hand, operates based on the all possibilities approach (APA) concept, a conditional Latin hypercube sampling (cLHs) algorithm and parallel programming, to evaluate all of the potential combinations of eight different spectral preprocessing techniques against the original reflectance and chemical data prior to the model development. The comparison of results was made in terms of the coefficient of determination (R^2) and root-mean-square error of prediction ($RMSE_p$). Results showed that the PARACUDA-II[®] engine performed better than the other selected regular schemes with R^2 value of 0.80 and $RMSE_p$ of 0.12; the PLSR was less predictive compared to other techniques with $R^2 = 0.63$ and $RMSE_p = 0.29$. This can be attributed to its capability to assess all the available options in an automatic way, which enables the hidden models to rise up and yield the best available model.

Keywords: soil carbon; soil spectroscopy; preprocessing techniques; data-mining algorithms; PARACUDA-II[®]

1. Introduction

Soil carbon content is a valuable indicator of soil fertility and is a critical parameter in directing the soil and atmosphere dynamics of different agrotechnical processes. Concerns about the influence of soil-carbon-decline influences on soil quality have encouraged research on the expansion of accurate and effective methods of evaluating soil carbon [1]. Therefore, the development of more rapid, accurate, and cost-effective methodologies for soil analysis, and more specifically for carbon content estimation, is a major desire.

There is a widespread interest in using visible-near-infrared-short-wave infrared (VIS-NIR-SWIR) reflectance spectroscopy for carbon analysis due to its spectrally active nature [2]. The technique has become a well-recognized, rapid, non-destructive, and low-cost [3] method with minimal sample preparation requirements that can be applied in both the laboratory and the field using point and imaging spectral measurements [4–6]. Moreover, the method does not use any chemicals, and it has capability to measure several soil properties using a single scan and a large number of samples in a very short time [7].

In order to get the full advantages of VIS-NIR-SWIR reflectance spectroscopy and to reduce the negative effects and errors that arise during measurement, Ben-Dor et al. [8] suggested the development of standards and protocols using assurance processes. However, another effective solution is removing the information from the spectra mathematically so that they may be correlated with soil parameters using effective chemometric and multivariate calibration techniques [9,10]. Gholizadeh et al. [11] also stated that one way of minimizing the undesirable impacts is by adopting advanced preprocessing methods as well as the appropriate selection of multivariate statistical analysis algorithms. These approaches can significantly decrease the differences between spectral measurements of samples assessed by different operators and systems under different conditions [12] and improve the obtained prediction accuracy [13].

The mathematical reduction of spectra noise and extraction of suitable information from a great number of highly correlated spectral bands, as well as the selection of the appropriate technique, are effective tasks. For instance, Gholizadeh et al. [11] indicated that the first derivative preprocessing method gave the best result for removing spectra noise in the Czech Republic mining areas soils in comparison to second derivative, multiplicative scatter correction (MSC), standard normal variate (SNV), and continuum removal (CR). Regarding multivariate calibration techniques, partial least squares regression (PLSR) is the most commonly used multivariate calibration technique for soil spectral analysis [14,15]. Other approaches have also been used, for instance, stepwise multiple linear regression (SMLR) [16,17], principal components regression (PCR) [18], and multivariate adaptive regression splines (MARS) [19,20]. Likewise, some other data-mining techniques, such as artificial neural networks (ANN) [21,22], boosted regression trees (BRT) [13,23], random forests (RF) [10,24], support vector machine regression (SVMR) [25,26], and memory based learning (MBL) [13,27,28], have been reported to improve the accuracy of the calibration models.

Due to the fact that the target function's nature strongly affects the performance of the different prediction approaches, different studies provide different results. Moreover, preprocessing and calibration procedures represent a significant portion of the work and expenses for spectroscopy technique application. Therefore, introducing more efficient approaches is greatly needed. Accordingly, the new data-mining engine PARACUDA-II[®], recently developed at Tel-Aviv University (TAU) [29], has been designed to utilize parallel and automatic processing to build and process hundreds of diverse models in order to prevent errors or biases caused by a human operator in the loop when taking the model setting decision. The engine also enables one to check all possible preprocessing combinations along with different statistical methods automatically, which, in reality, is time demanding and difficult for a single operator to perform. Therefore, the current study aims to examine the PARACUDA-II[®] concept and performance against selected traditional manual techniques (PLSR, RF, BRT, SVMR, and MBL) to predict soil oxidizable carbon (Cox) content. This study was performed over bare soil sites within the Pokrok dumpsite in the Czech Republic.

2. Materials and Methods

2.1. Experimental Site, Soil Sampling and Analysis

The Pokrok dumpsite (50°60'N; 13°71'E) in the northeastern part of the Czech Republic was selected as the test site, where the soil samples for this study were collected (Figure 1).

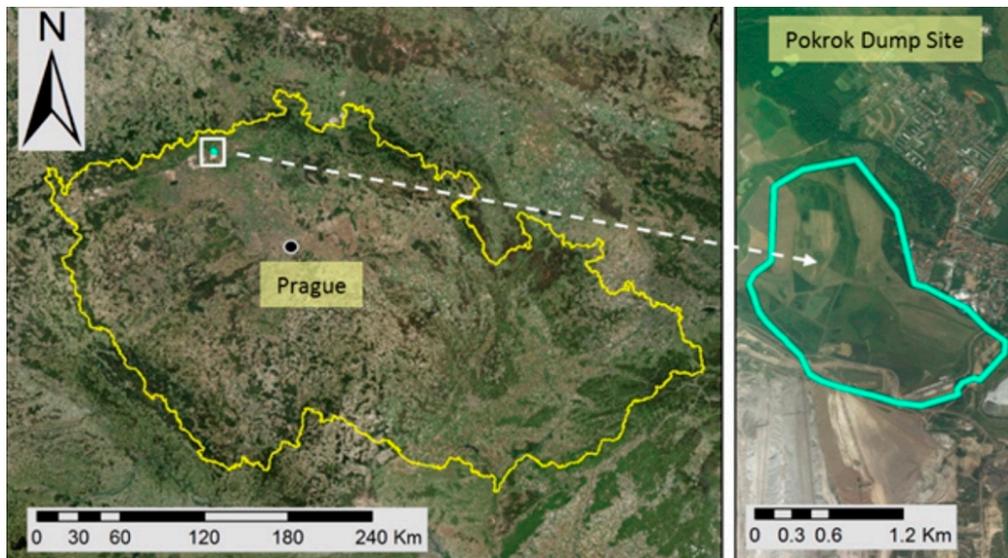


Figure 1. Location of Pokrok dumpsite in the northeastern part of the Czech Republic.

The dumpsite is formed by clay. One year before sampling, a cover of natural topsoil (≤ 25 cm) in an amount of 2500–3000 tons per ha was spread over a part of the area. Topsoil material originated from humic horizons of natural soils of the region, mainly Vertisols and partly Chernozems (clayic and haplic). Topsoil was not mixed with the dumpsite material. Some characteristics of the soils, including pH, soil organic matter, and texture were measured using bulk control subsamples due to their importance as environmental indicators. The soil pH range for the area was 5.3–8.5. The soil organic matter content range was 0.6–3.8%. Texture analysis, which was performed by the hydrometer method, showed that soil of the area had 37.30% clay, 33.10% sand and 29.60% silt. Disturbed and undisturbed soil samples were randomly collected at the dumpsite randomly. One-hundred and three (103) soil samples were collected and a GeoXM (Trimble Inc., Sunnyvale, California, USA) receiver recorded each sampling point's position with an accuracy of 1 m. Sampling was performed on a range of depths from 0 to 25 cm, which corresponds to the common depth of a ploughing soil layer, as these soils will be used as arable lands in the future. The soil samples were air-dried, ground, and sieved (≤ 2 mm) and were thoroughly mixed prior to the analysis and stored in plastic containers. The dichromate redox titration method was used to measure the soil Cox in three replications [30].

2.2. Soil Spectra Measurement

An ASD FieldSpec III Pro FR spectroradiometer (ASD Inc., Denver, CO, USA) with a high intensity contact probe was used to measure the spectral reflectance across the optical range (350–2500 nm). The spectral resolution of the spectroradiometer was 2 nm for 350–1050 nm regions and 10 nm for 1050–2500 nm regions. Furthermore, the radiometer's full width at half maximum (FWHM) from 350–1000 nm was 1.4 nm, whereas it was 2 nm from 1000–2500 nm. The measurement protocol started with 30 min of the instrument and light warming up. Air-dried, crushed, sieved and thoroughly mixed soil samples with 2 cm depth were placed in 9 cm diameter petri dishes to avoid beam reflectance from the bottom of the dish [31]. Samples were leveled off with a stainless-steel blade to make a flat surface flush with the top of the petri dish, as a smooth soil surface guarantees maximum light reflection and a

high signal-to-noise ratio (SNR) [32]. All spectral readings were measured in three replications in the center of the samples. Before the first scan and after every six measurements, a white Spectralon™ (Lab-sphere, North Sutton, NH, USA) was used to optimize the spectroradiometer [33]. For SNR improvement, 30 spectra were averaged for each soil measurement. It needs to be mentioned that in order to avoid each instrument's on/off problems (instability and uncertainty), all sample spectra measurements were done in a single day [8].

2.3. Spectra Preprocessing for Selected Data-Mining Techniques

After collecting the spectral measurements, first, the noisy portions between 350 and 399 nm and 2451 and 2500 nm were removed, followed by smoothing of the spectra using Savitzky-Golay with a second-order polynomial fit and 11 smoothing points [34,35] to eliminate the artificial noise caused by various conditions. Data from the laboratory were then preprocessed before the chemometric analysis with the selected data-mining techniques (PLSR, RF, BRT, SVMR, and MBL) as follows. The outliers of the spectra were left out using the principle of Mahalanobis distance (H) [36–38], which was applied on principle component analysis (PCA)-reduced data. In the present study, the number of removed outliers was four. Then, the first derivative calculation was used as a spectra preprocessing technique, the transformation of which is very effective for removing baseline offset [11,39].

2.4. Development of Calibration Models for Selected Data-Mining Techniques

Soil Cox was modelled using various data-mining algorithms to compare the prediction capability of the PLSR, RF, BRT, SVMR, and MBL to PARACUDA-II®[®], an all possibilities approach (APA) data-mining and machine-learning engine. It should be mentioned that the samples were divided into calibration-validation 75–25% groups. To maintain the independence of validation samples from calibration samples and cover variations in soil properties, the validation dataset was selected using random stratified selection.

2.4.1. Partial Least Square Regression (PLSR)

PLSR has been a popular technique in chemometric analysis and is used for reflectance spectra quantitative analysis. It reduces the data, calculation time, and noise with minimum loss of the information enclosed in the original variables [40,41]. It is closely related to principal component regression (PCR); however, the PLSR method links the compression and regression steps and chooses consecutive orthogonal factors that maximize the predictor and response variables' covariance [7,30,42–44]. By fitting a PLSR model, a few PLSR factors are determined that explain most of the variations in both predictors and responses [9]. Viscarra Rossel and Behrens [10] and Gholizadeh et al. [45] stated that PLSR decomposes X and Y variables and finds latent variables, which are both orthogonal and weighted linear combinations of X variables. These new X variables are then employed for prediction of Y variables, as follows:

$$X = Tp' + E, \quad (1)$$

$$Y = Tq + F, \quad (2)$$

where X is soil reflectance, Y is measured soil property, T is factor scores, p' and q are factor loadings and E and F are residuals.

The residual factors simulate noise and can be ignored. The resulting matrices and vectors usually have a significantly lower dimension than X and Y. Given a new reflectance X, the soil parameter Y can be predicted as a (bi) linear combination of the factor scores and factor loadings of X [10]. In PLSR, a crucial step is choosing the optimal number of latent variables in the calibration model, which will help to avoid underfitting and overfitting of data that generate poor prediction models [20,46]. The R package Caret was used for the PLSR model [47].

2.4.2. Random Forest (RF)

The RF is a collaborative data-mining technique developed by Breiman [48] for data classification and regression. It works by growing a group of regression trees based on binary recursive partitioning. The algorithm starts with a number of bootstrap samples (*ntree*) from the original data [49]. Then, with a modifying operation, in which some of the predictors (*mtry*) are randomly sampled, each *ntree* grows a regression tree and the algorithm selects the best split among the sampled variables rather than all variables [50]. According to Abdel Rahman et al. [51], the square root of the total number of variables is considered to be the default *mtry* value. Generally, the RF prediction for regression problems can then be written [52] as:

$$\frac{1}{M} \sum_{m=1}^M \hat{f}_m^*(x_0), \quad (3)$$

where M is the m th bootstrap resample tree ($m = 1, \dots, M$), x_0 is the covariate, and $\hat{f}_m^*(x_0)$ explains the prediction of an independent test case by the m th tree.

For predicting an independent test case C_0 with the covariate x_0 , the predicted value by the whole RF is gained by combining the results given by individual trees. RF hardly overfits when using more trees [48], although it does yield a limited generalization error [53,54]. This method does not require complex data pretreatment and is very fast compared to some data-mining algorithms such as ANN [55]. The R package Random Forest was used for prediction modelling [56].

2.4.3. Boosted Regression Trees (BRT)

The BRT method has been suggested by Brown [57] as a reliable data-mining technique for VIS-NIR-SWIR spectroscopy of soil attributes. Analysis of BRT principally carries out a binary recursive partitioning of the dataset [58,59]. A predicted value is obtained as the average of all the measurements at each grouped terminal node. Multiple predictions are generated based on resampling and weighting, which belong to the group of collective methods [60]. Boosted models can be created in the following general form:

$$F(x; \{\beta_m, a_m\}_0^M) = \sum_{m=0}^M \beta_m h(x; a_m), \quad (4)$$

where $h(x; a)$ is simple classification function or base learner with parameters “ a ” and input variables “ x ”, m is the model step, and β_m is the weighting coefficient.

The main benefits of BRT are the potential to include a large number of weak relationships in a predictive model, insensitivity to outliers in the calibration dataset, relative immunity to overfitting, and no requirement for uniform data transformations [61–63]. The R package GBM was used for the BRT modelling [64].

2.4.4. Support Vector Machine Regression (SVMR)

The SVMR technique is a supervised, nonparametric, statistical learning, and kernel-based approach [65]. It provides balance between the accuracy obtained from a given limited amount of training patterns and the simplification ability to manage unseen data. The technique is nonlinear and is applied in classification and multivariate calibration [66]. Model complication in SVMR is limited by the learning algorithm itself, which prevents overfitting. The ε -SVMR uses training data to create a model that maps independent data with maximum ε deviation from dependent training dataset [30]. Error within the prearranged distance ε from the true value is avoided and error greater than ε is disciplined by the soil attribute. The model decreases the training data complication to a subset that is called support vectors. Vapnik [67] has described the subsequent equation for prediction as below:

$$y(x) = \sum_{k=1}^N \alpha_k K(x, x_k) + b, \quad (5)$$

where b is the scalar threshold, $K(x, x_k)$ is the kernel function, α is the Lagrange multiplier, N is the number of data, x_k is the input data, and y is the output data.

For this study, SVM with radial basis function as one of the most popular kernels was applied. A radial basis function can be calculated using the below equation:

$$\psi(x, x_k) = \exp \left\{ - \frac{\| (x - x_k)^2 \|}{2\sigma^2} \right\}, k = 1, \dots, N, \quad (6)$$

where σ is the width of the radial basis function, which here was determined by a grid search method using repeated cross validation approach. Additionally, the grid search method was used for choosing the best parameters for the model. The R package Caret was used for the SVMR model [47].

2.4.5. Memory Based Learning (MBL)

The MBL is a data-driven technique that recalls former situations, adapts them for solving the remaining issues, studies the option to solve the problem with the new explanation, and memorizes the skill for knowledge development [27,68]. The algorithm can be obtained more reliable by analogical analysis compared to the use of abstract mental and rule-based processing [69]. Daelemans and Van den Bosch [70] stated that MBL is a kind of lazy-learning approach that compares new problems with cases realized in training and stored in memory. In order to solve a new problem, the experience is retrieved from memory in the form of a set of analogous related samples which are merged and the solution to the new problem is built [71]. In fact, for each new problem, a new target function is established. In MBL, two sets of data are required—a set of n reference samples (e.g., spectral library) and a set of m samples as the prediction set. It should be noted that it is essential to find out the k -nearest neighbors of each data in the prediction set before modelling [13,27].

Correlation dissimilarity was applied in the current study for nearest-neighbor selection, which defines each sample's most comparable sample in terms of its VIS-NIR-SWIR principal components. Afterwards, the local models are close fitted using weighted average PLS of all the predicted values generated by the multiple PLS models between a maximum and minimum number of PLS components. The weight of each component is calculated as follows:

$$w_j = \frac{1}{s_{1:j} \times g_j} \quad (7)$$

where $s_{1:j}$ is the root-mean-square of the spectral residuals of the unknown sample when a total of j th PLS components is applied and g_j is the root-mean-square of the regression coefficient corresponding to the j th PLS components. The R package resemble was used for the MBL modelling [72].

2.4.6. PARACUDA-II®

PARACUDA-II® is a new machine-learning and data-mining engine developed at the remote sensing laboratory of TAU by Carmon and Ben-Dor [29]. It is a program based on the APA concept, a conditioned Latin hypercube sampling (cLHS) and parallel programming technique, which offers the automatic assessment of all possible combinations of manipulations (preprocessing) to the original reflectance and chemical data before the modelling procedure. PARACUDA-II® has four key steps that each carries a particular purpose during the modelling process: (i) outlier detection and elimination; (ii) preprocessing and transformations; (iii) model development and validation; and (iv) population analysis and selection of the best model (Figure 2).

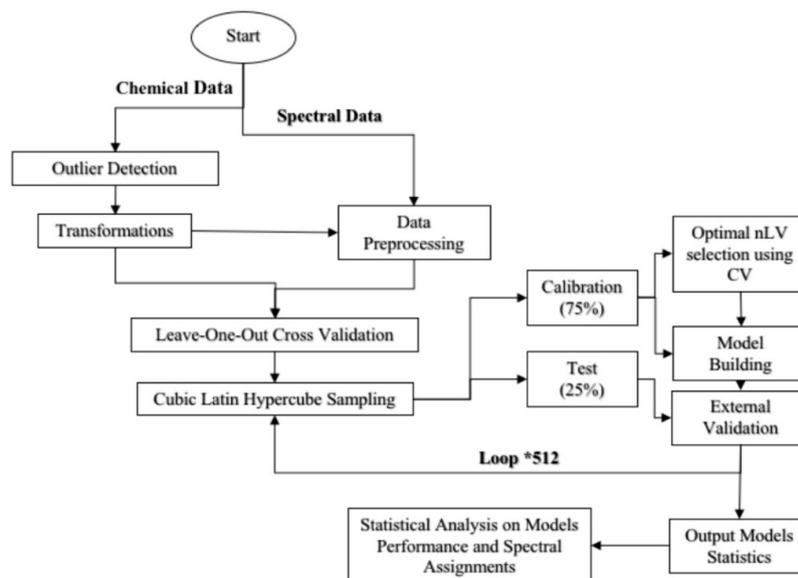


Figure 2. PARACUDA-II[®] processing framework.

The first step in the PARACUDA-II[®] arrangement is an outlier detection and removing module for the spectral and the chemical datasets. The z-score outlier test is used in which samples with chemical values above or below ± 2 are eliminated from the population. For spectral outlier detection, a PCA calculation is applied in order to extract the first two factors. Samples beyond a 95% confidence ellipse are identified and removed. Next, the samples are divided into calibration-validation 75–25% sets using a cLHs technique in which the grouping with the most co-variability is found based on the chemical values, which confirms the relatively similar value distribution between the two calibration and validation sets. On the calibration set samples, preprocessing techniques are performed both for the chemical and spectral values. The next step is preprocessing and transformations. During this step, the chemical values are transformed using a Box-Cox algorithm to obtain a more normal distribution [73,74]. The spectral data are preprocessed using eight different spectral preprocessing algorithms, namely, moving average, the first and the second derivatives, absorbance transformation, CR, SNV, MSC, and final smoothing, in all possible combinations (up to 120 preprocessing sequences). The correlation between each wavelength in each preprocessing combination and the chemical values is assessed. For each wavelength then, the preprocessing combination with the highest correlation to the chemical values is selected and all manipulations for the spectral dataset are extracted. A new dataset comprising the values of various and optimal preprocessing methods for every wavelength separately is yielded as the final product of this step. The best combination preprocessing of spectral data is then modeled with the Box-Cox transformed chemical data using a PLSR sequence limited to 15 factors. The number of factors is selected by finding the minimum root mean square error (RMSE) for each factor using cross-validation PLSR models. Then, a PLSR model is created in step three with the corresponding number of factors on the calibration group samples. A PLSR model is developed on transformed and preprocessed data without overfitting during the third step. The sequence begins with the sampling routine and ends when the prediction model evaluation is repeated 512 times. The validation set samples are preprocessed with the same process as the calibration samples to examine the developed model. The model is then applied on the samples and the predicted values are transformed back from Box-Cox values to original chemical values. For evaluating the model's performance at the fourth step, the coefficient of determination (R^2) between the measured and the predicted chemical values is calculated and saved. The sequence here begins with the sampling routine and ends when the prediction model assessment is repeated 512 iterations. The model resulting in the highest R^2 and lowest RMSE, among all created models, is chosen as the best available model. For providing spectral assignments, two calculations are performed: (1) a R^2 per wavelength for

the preprocessed data and (2) the weighted average beta coefficients of the best model. These spectra are beneficial for understanding the significant spectral ranges of specific chemical parameters and preparing further consideration of the results. Outputs of PARACUDA-II[®] consist of two files. One summarizes the report of the calibration group, validation group, cross-validation, and the two spectral assignment spectra in Excel format, which provides measured and predicted values for each parameter. The other file is an applicable model for applying to new spectral data in MATLAB format, which is helpful for further validation or practical purposes and is practical on either point spectral data or hyperspectral images directly from the PARACUDA-II[®] interface.

2.5. Performance of Models

In order to evaluate the model performance for the prediction of Cox, the statistical parameters of R^2 and $RMSE_p$ were used. R^2 is a measure of how well the variation of one variable explains the variation of the other and shows the percentage of the variation explained by a best-fit regression line, which is calculated for the data, and $RMSE_p$ indicates the prediction error. Generally, the largest R^2 and the smallest $RMSE_p$ give the best prediction model [75].

3. Results

3.1. Soil Cox Statistics

The descriptive statistics of the Cox, determined by the conventional wet chemistry analysis, are shown in Figure 3 and Table 1. It can be observed that Cox concentration was relatively low with mean and maximum values of 1.62% and 3.80%, respectively. It is also obvious that there was a large variation in soil carbon (ranging from 0.40 to 3.80%), underlining the varied and diverse origin of the samples. Coefficients of variation (CV) highlighted that the Cox had relatively high CV (29%), which shows that it varied rather highly and its distribution was heterogeneous. The rather wide range of variability indicates that this site is a reasonably optimal case study as according to Kuang and Mouazen [76], as less promising results of prediction capability of soil calibration models are expected in cases of low soil variability. The Cox was left-skewed, with a higher mean than median (1.62% versus 1.57%, respectively).

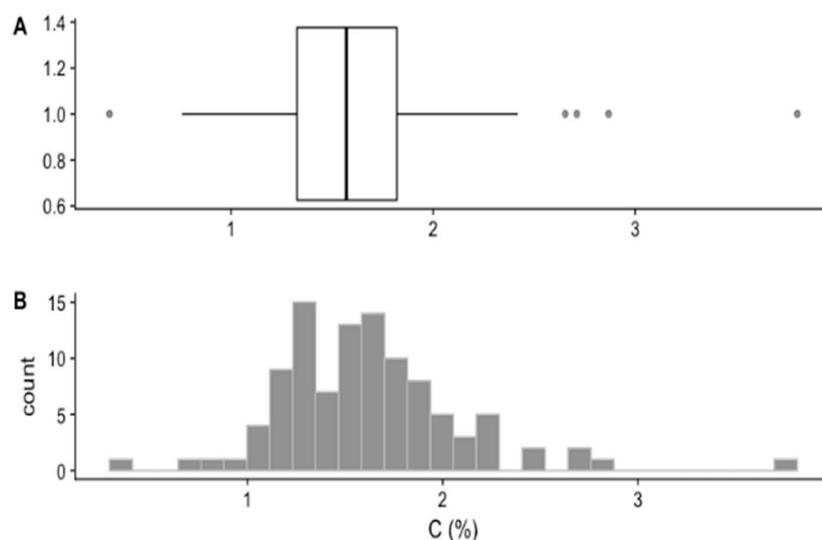


Figure 3. (A) Box-and-whiskers plot with outliers and (B) Histogram.

Table 1. Descriptive statistics of Cox.

Characteristic	Cox (%)
Mean	1.62
Median	1.57
Min.	0.40
Max.	3.80
Std. Dev.	0.47
CV	29

3.2. Soil Spectral Response

Spectral curves of analyzed soils are presented in Figure 4. An average raw spectra form is typical for soil reflectance, with a gradual increase through VIS wave range (400–700 nm), almost flat segment in NIR (700–1000 nm), and somewhat lower reflectance values in SWIR-II (1900–2500 nm) [77]. Few observed absorption features can be attributed to the presence of water (at 1400 nm and 1900 nm) and clay minerals (at ~2200 nm) [2,57,78]. The spectra shape was a key for differentiation of average spectra after the first derivative preprocessing through visual inspection (Figure 4b). These spectra were quite different from the raw spectra. There were more features of high variability around 460–550 nm, 1400 nm, 1900–2000 nm, and 2200 nm, which is typical for the noise-removed and preprocessed spectra of the first derivative technique [11,26].

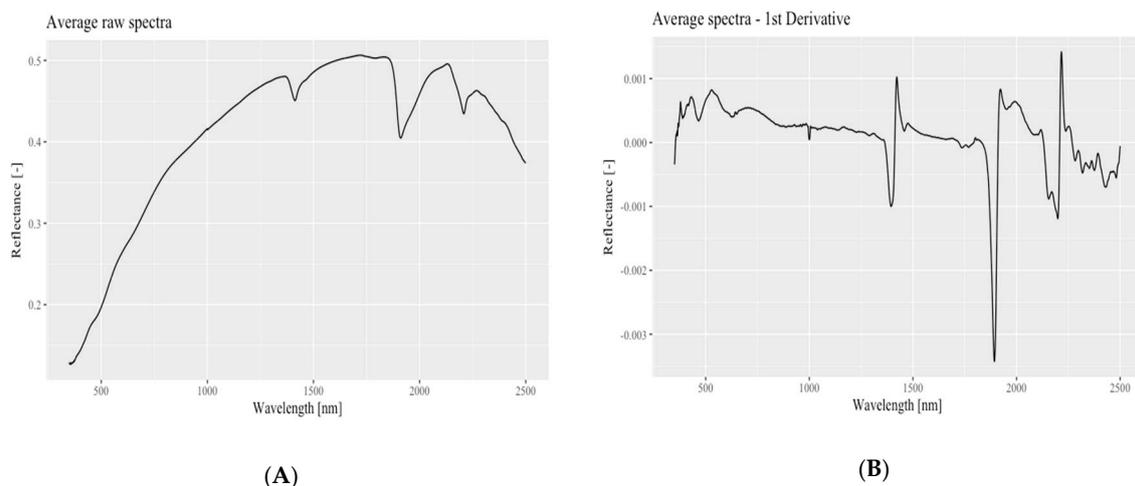


Figure 4. The spectra of soil samples derived from (A) raw spectra and (B) 1st derivative preprocessing technique.

Within the PARACUDA-II[®] procedure, different cumulative preprocessing sequences for each band in the data were applied. In fact, each band was preprocessed with a different preprocessing sequence, resulting in a nonlinear spectral dataset. Therefore, displaying an average spectrum of the data after this routine was not adequate. Consequently, to visualize the implication of the applied preprocessing approaches, correlograms for raw spectra, first derivative, and PARACUDA-II[®] were built (Figure 5).

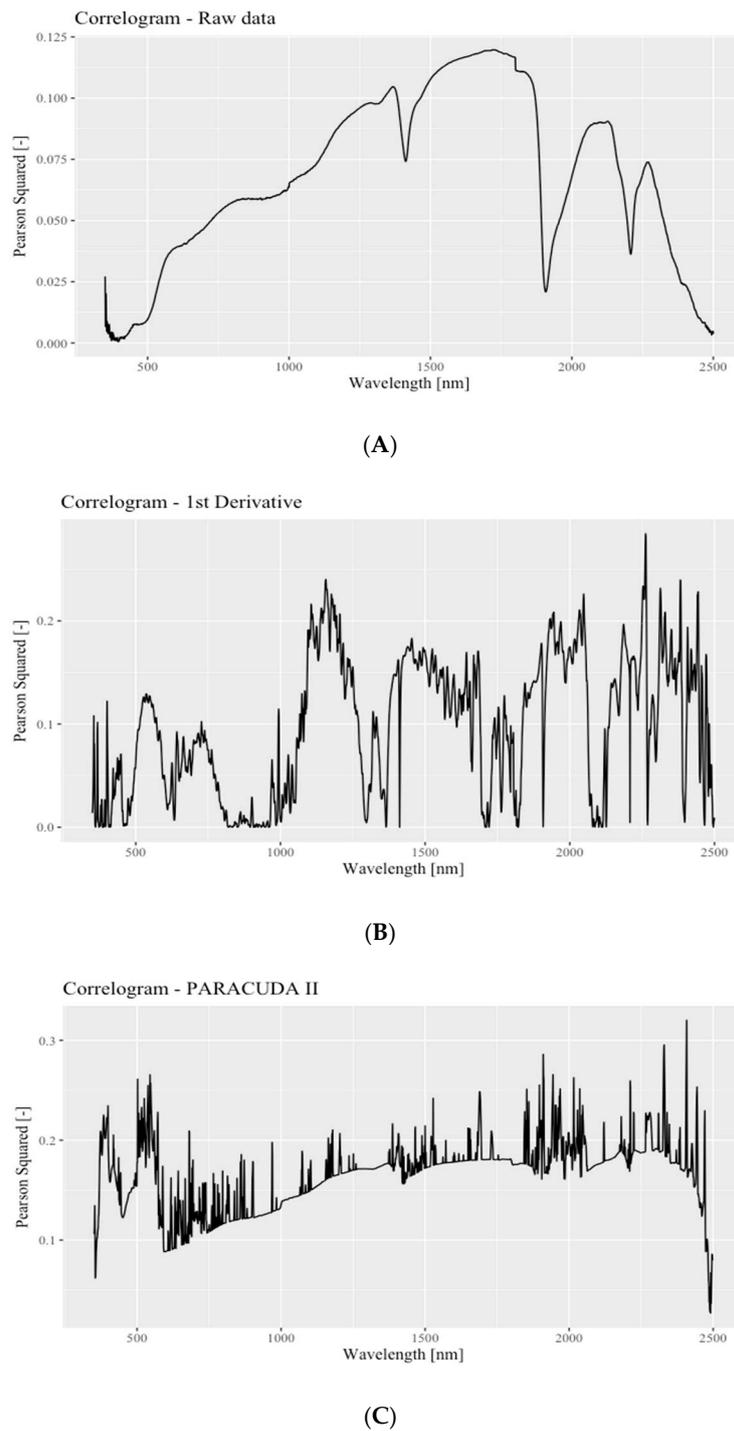


Figure 5. The correlogram of the spectra at (A) raw spectra, (B) after the 1st derivative, and (C) after the PARACUDA-II[®] preprocessing stage.

Figure 5 highlights that, for obtaining a more superior correlogram, the learning ability of the raw or preprocessed spectra was noticeably sensitive to the selection of a proper preprocessing technique. This means that the differences in correlation between each wavelength (in raw data as well as in each preprocessing approach) and the chemical values (Figure 5) contributed to the successful role of the PARACUDA-II[®].

3.3. Calibration Model Performance

Scatterplots in Figure 6 show the results of predicted versus measured Cox using six applied data-mining techniques on validation datasets. The difference in predicting carbon among PLSR, RF, and BRT was not that obvious according to the scatterplots. Visually, a rather nonsignificant different pattern can also be seen in Cox prediction using SVMR and MBL based on Figure 6. All techniques for Cox showed overall acceptable patterns. However, a difference is noticeable between PARACUDA-II[®] and other algorithms, especially with PLSR, RF, and BRT. There were significantly less scatters in prediction of Cox when the PARACUDA-II[®] engine was used.

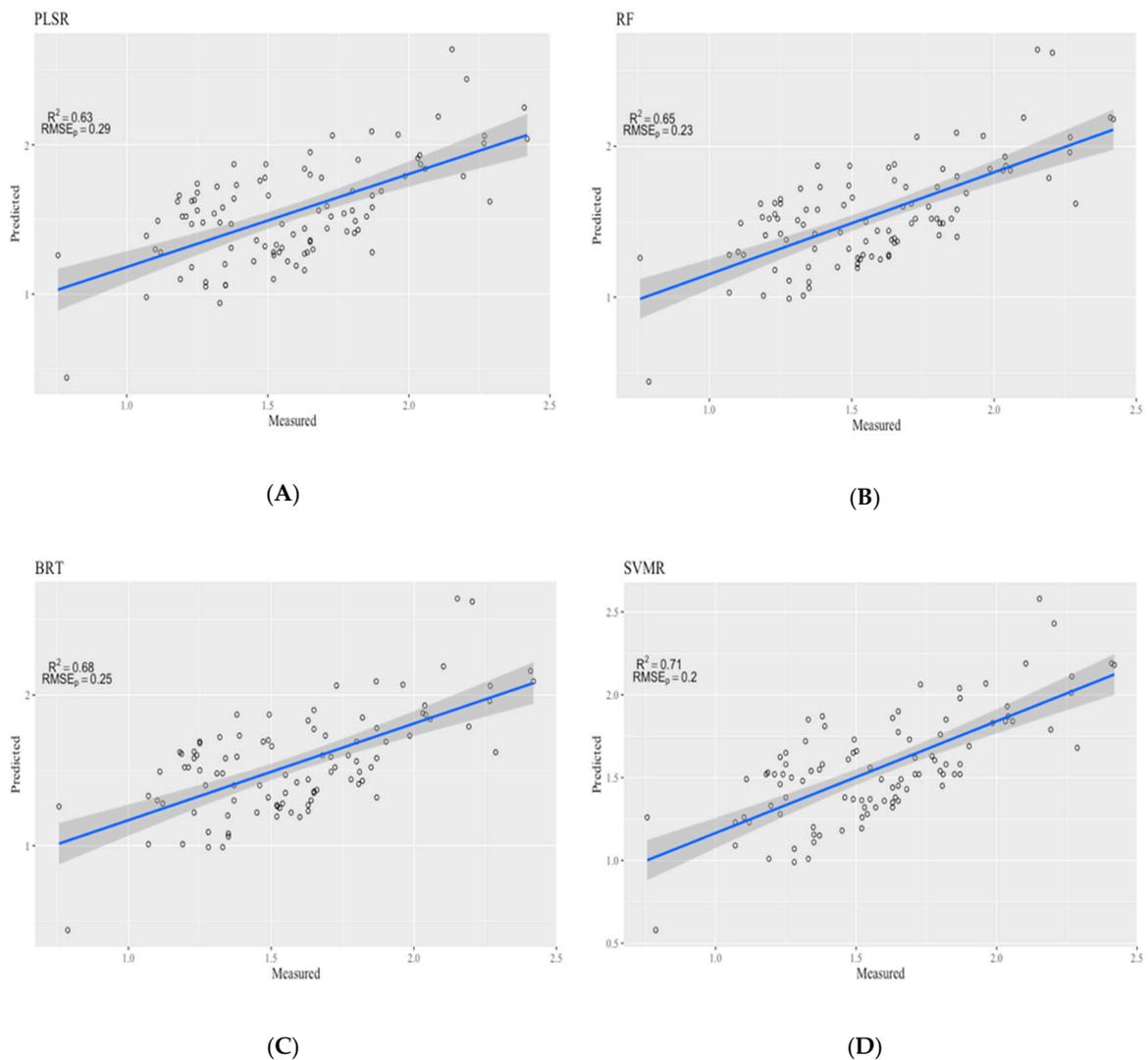


Figure 6. Cont.

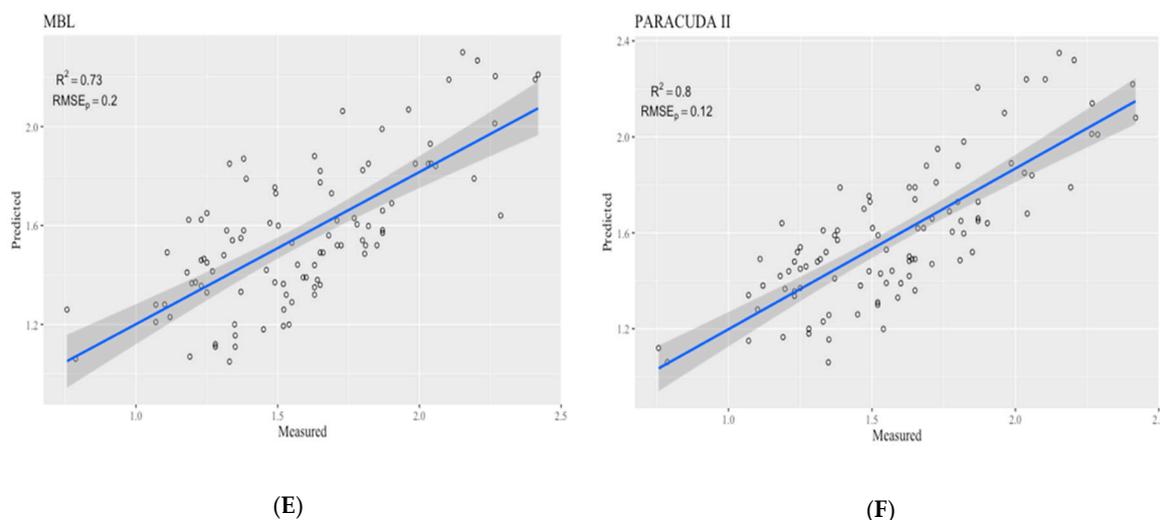


Figure 6. Scatterplots of predicted versus measured Cox using (A) PLSR, (B) RF, (C) BRT, (D) SVMR, (E) MBL, and (F) PARACUDA-II[®] techniques on validation dataset.

Soil Cox was predicted with R^2 values between 0.63–0.80 (Table 2), which are classified as fair to good models [18]. The statistical accuracy obtained using PLSR, RF, and BRT indicated that for the Cox content, the methods of prediction could give a reasonable indicator based on spectra from soil samples. While predictions by them were close, BRT outperformed RF, followed by PLSR, which performed the least well. While SVMR and MBL yielded almost similar results, they were more highly predictive than the PLSR, RF, and BRT approaches. However, according to the criteria of maximal R^2 and minimal $RMSE_p$, PARACUDA-II[®], with $R^2 = 0.80$ and $RMSE_p = 0.12$, was considered to be the best technique among the others.

Table 2. Performance of Cox prediction on validation dataset using different data-mining techniques.

Data-Mining Technique	Cox (%)		
	R^2	$RMSE_p$	Bias
PLSR	0.63	0.29	−0.021
RF	0.65	0.23	−0.020
BRT	0.68	0.25	−0.017
SVMR	0.71	0.20	−0.014
MBL	0.73	0.20	−0.013
PARACUDA-II [®]	0.80	0.12	0.003

4. Discussion

Considering the correlograms and preprocessing performance of PARACUDA-II[®] (the highest correlation −0.32 compared to the first derivative and raw spectra, 0.28 and 0.124, respectively) in Figure 5, PARACUDA-II[®] appeared as a robust engine for spectral denoising and preprocessing purposes. This is because of its capability to check all possible preprocessing combinations along with different statistical methods automatically using eight different spectral preprocessing algorithms in all possible combinations (totally 120 sequences).

Regarding the assessment of soil carbon using various data-mining algorithms, Viscarra Rossel and Behrens [10] showed that for the prediction of soil carbon, the RF and BT models produced the largest $RMSE_p$ values and were thus the least accurate. They mentioned that, generally, tree ensemble approaches (RF and BT) perform weakly. However, Brown [42], Brown et al. [57], and Gholizadeh et al. [79] proved the advantage of BRT over PLSR for analyzing soil characteristics with VIS-NIR-SWIR reflectance data. More accurate outputs of BRT in comparison to PLSR are due to

some of its superiorities, including insensitivity to outliers in the calibration dataset as well as the capability to utilize a large number of weak classifiers and thereby make maximum use of the entire spectrum [57,61,63,80]. Conversely, in an experiment by Vasques et al. [40] to predict soil carbon, the BRT model provided the worst results among many multivariate techniques, including PLSR. Based on their results, one explanation could be the fact that BRT produces discrete outputs predicting a single value at each terminal node [40]. This alteration in spectral predictive mechanisms may be originated by the carbon situation in soil, the nature of available compounds, and the effect of other relevant factors, such as soil moisture, texture, or iron oxides [2,5,10,30]. Moreover, depending on the geographic region and its condition, one method may outperform several others [41].

By comparing the results of the SVMR and other techniques, it can be seen that the SVMR model produced the highest R^2 and the smallest $RMSE_p$ for Cox prediction rather than PLSR, RF, and BRT. These results are supported by the results obtained by Viscarra Rossel and Behrens [10], Gholizadeh et al. [13], Araujo et al. [23], Sorensen et al. [24], and Morellos et al. [25]. The more exceptional performance of SVMR in comparison to PLSR, RF, and BRT can be explained by its high ability to deal with the nonlinear patterns as well as its ability to approximate nonlinear functions between multidimensional spaces [20,81,82]. Viscarra Rossel and Behrens [10] stated that SVMR is a nonlinear and flexible method, capable of modelling complex, nonlinear, and linear relationships between variables. It can develop a linear hyperplane as a decision function for non-linear issues, which reduces problems with heterogeneity and nonlinearity and can be considered as an additional reason for the method's merit [23,83]. Nevertheless, SVMR was less accurate when compared to MBL, which is in agreement with Gholizadeh et al. [13] on soil texture prediction. The better outputs of MBL can be related to the potential of the technique for choosing a more proper neighbor to calibrate local models and for being involved in each local model as a source of further predictor variables [27,28,84].

The results gained in our study proved the superior potential of the PARACUDA-II[®] for soil Cox estimation in the considered region. In fact, among all examined techniques, it is interesting to note that the PARACUDA-II[®] provided the best calibration results. Table 2 shows that the spectroscopic model developed from the PARACUDA-II[®] had a R^2 value of 0.80 for an $RMSE_p$ of 0.12, and both were more improved than models using other algorithms. These findings confirmed the results of another study by Gholizadeh et al. [12] that demonstrated the capability of the PARACUDA-II[®] to perform as an effective machine for providing the most appropriate model within a given population. They conducted an experiment in which PARACUDA-II[®] and PLSR were used to analyze two spectral datasets, acquired from different protocols at different laboratories, for estimation of some soil attributes namely Cox, pH-H₂O, pH-KCl, and selected forms of Fe and Mn in agricultural soils. Their results indicated that under both measurement protocols, PARACUDA-II[®] performed noticeably more effectively compared to regular PLSR. This superiority can be attributed to the capability of the PARACUDA-II[®] to apply preprocessing algorithms on the spectral data using an APA automatic approach. It has the efficiency to explore in parallel several data manipulations and to generate many calibration-validation groups' partitions [85]. In addition, this system also applies a dual outlier detection module to recognize problematic samples both in the spectral and chemical/physical domains. Accordingly, as it is capable of checking all the existing options, it can generate hidden models that are not accessible by running regular schemes such as PLSR, RF, BRT, SVMR, and MBL.

The current study strengthened the superiority of the PARACUDA-II[®] engine performance. It concluded that the best model could not be found by a random selection of a given chemometric method or a given preprocessing technique, although the APA should also be taken into account. As APA requires a skilled person and considerable time, it cannot be achieved if no automatic approach such as PARACUDA-II[®] is used.

5. Summary and Conclusions

In this study, the performance of five data-mining techniques (PLSR, RF, BRT, SVMR, and MBL) was compared against the PARACUDA-II[®] engine in order to predict Cox in the Pokrok dumpsite

located in the northeastern part of the Czech Republic. PARACUDA-II[®] is an engine recently developed at TAU which has been designed based on APA and cLHs techniques as well as parallel programming to assess all possible combinations of manipulations (preprocessing) of the original reflectance and chemical data before a modelling procedure. The results of correlograms derived from raw spectra, the first derivative, and PARACUDA-II[®] showed that the learning capability of the raw or preprocessed spectra were apparently sensitive to the selection of an appropriate preprocessing method. Therefore, regarding the potential of PARACUDA-II[®] in all possible combinations of eight different spectral preprocessing techniques (120 sequences in total), it appeared as a robust engine for spectral denoising and preprocessing purposes. The results of the calibration models in comparison indicated that all algorithms outperformed the PLSR, the most commonly used multivariate technique for soil spectral analysis, in modelling. However, PLSR still provided acceptable accuracy for the prediction of Cox. The most considerable finding was that the PARACUDA-II[®] engine, with highest prediction results, was the best option for soil carbon prediction compared to the selected regular schemes (PLSR, RF, BRT, SVMR, and MBL). This is essentially because of its potential to assess all the available options and extract the hidden models. It also surpasses other methods in the automatic procedure it offers, which permits searching for the best available model. Moreover, the automatic process option of the PARACUDA-II[®] promises to be an effective way of reducing the need for both analytical time and a skilled person. It can be concluded that the PARACUDA-II[®] data-mining approach is a powerful tool for obtaining more significant outputs that cannot be achieved using other techniques. In addition, taking into account that PARACUDA-II[®] can be run automatically with no man in the loop and with promising efficiency, it can pave the road for many more applications and analysis that could not be executed before. It should be mentioned that PARCUDAII[®] now has the IP of the TAU and is under pending patent process. It is currently used for scientific collaboration with institutions that are part of joint projects with Remote Sensing Laboratory of TAU (TAU-RSL); however, the idea is to make this engine commercially available to scientists in the near future.

Author Contributions: A.G. and M.S. conceived and designed the experiment; N.C. and M.S. performed and analyzed the data; E.B.-D. and A.G. interpreted the results; A.G. wrote the paper; and A.G., E.B.-D., and L.B. reviewed the paper.

Funding: This research was funded by the Czech Science Foundation (project No. 18-28126Y) and partially funded by the Ministry of Education, Youth and Sport of the Czech Republic projects CENAKVA (project No. CZ.1.05/2.1.00/01.0024) and CENAKVA II (project No. LO1205 under the NPU I program).

Acknowledgments: Authors wish to thank Vit Penizek and Karel Nemecek for their help and support.

Conflicts of Interest: The authors declare no conflict of interest. Furthermore, the funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Smith, P. Monitoring and verification of soil carbon changes under Article 3.4 of the Kyoto Protocol. *Soil Use Manag.* **2004**, *20*, 264–270. [[CrossRef](#)]
2. Ben-Dor, E.; Banin, A. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. *Soil Sci. Soc. Am. J.* **1995**, *59*, 364–372. [[CrossRef](#)]
3. Reeves, J.B., III. Near-versus Mid-Infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma* **2010**, *158*, 3–14. [[CrossRef](#)]
4. Ben-Dor, E.; Patkin, K.; Banin, A.; Karnieli, A. Mapping of several soil properties using DAIS-7915 hyperspectral scanner data—A case study over clayey soils in Israel. *Int. J. Remote Sens.* **2002**, *23*, 1043–1062. [[CrossRef](#)]
5. Mouazen, A.M.; Maleki, M.R.; De Baerdemaeker, J.; Ramon, H. On-line measurement of some selected soil properties using a VIS-NIR sensor. *Soil Till. Res.* **2007**, *93*, 13–27. [[CrossRef](#)]
6. Viscarra Rossel, R.A.; Cattle, S.R.; Ortega, A.; Fouad, Y. In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy. *Geoderma* **2009**, *150*, 253–266. [[CrossRef](#)]

7. Viscarra Rossel, R.A.; Walvoort, D.J.J.; McBratney, A.B.; Janik, L.J.; Skjemstad, J.O. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75. [[CrossRef](#)]
8. Ben-Dor, E.; Ong, C.; Lau, I.C. Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma* **2015**, *245–246*, 112–124. [[CrossRef](#)]
9. Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley and Sons: Chichester, UK, 1989; p. 419.
10. Viscarra Rossel, R.A.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [[CrossRef](#)]
11. Gholizadeh, A.; Boruvka, L.; Vasat, R.; Saberioon, M.M. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil Water Res.* **2015**, *10*, 218–227. [[CrossRef](#)]
12. Gholizadeh, A.; Carmon, N.; Ben-Dor, E.; Boruvka, L. Agricultural soil spectral response and properties assessment: Effects of measurement protocol and data mining technique. *Remote Sens.* **2017**, *9*, 1078. [[CrossRef](#)]
13. Gholizadeh, A.; Saberioon, M.M.; Boruvka, L.; Vasat, R. A memory-based learning approach as compared to other data mining algorithms for the prediction of soil texture using diffuse reflectance spectra. *Remote Sens.* **2016**, *8*, 341. [[CrossRef](#)]
14. Wold, S.; Martens, H.; Wold, H. The multivariate calibration method in chemistry solved by the PLS method. In *Matrix Pencils, Lecture Notes in Mathematics*; Ruhe, A., Kagstrom, B., Eds.; Springer: Heidelberg, Germany, 1983; Volume 973, pp. 286–293.
15. Conforti, M.; Castrignanò, A.; Robustelli, G.; Scarciglia, F.; Stelluti, M.; Buttafuoco, G. Laboratory-based Vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. *Catena* **2015**, *124*, 60–67. [[CrossRef](#)]
16. Shibusawa, S.; Imade Anom, S.W.; Sato, S.; Sasao, A.; Hirako, S. Soil mapping using the real-time soil spectrophotometer. In Proceedings of the 3rd European Conference on Precision Agriculture, Agro Montpellier, France, 18–20 June 2001; pp. 497–508.
17. Gholizadeh, A.; Amin, M.S.M.; Saberioon, M.M.; Boruvka, L. Visible and near infrared reflectance spectroscopy to determine chemical properties of paddy soils. *J. Food Agric. Environ.* **2013**, *11*, 859–866.
18. Chang, C.-W.; Laird, D.A.; Mausbach, M.J.; Hurburgh, C.R., Jr. Near-infrared reflectance spectroscopy–principal components regression analysis of soil properties. *Soil Sci. Soc. Am. J.* **2001**, *65*, 480–490. [[CrossRef](#)]
19. Shepherd, K.D.; Walsh, M.G. Development of reflectance spectral libraries for characterization of soil properties. *Soil Sci. Soc. Am. J.* **2002**, *66*, 988–998. [[CrossRef](#)]
20. Bilgili, A.V.; Van Es, H.M.; Akbas, F.; Durak, A.; Hively, W.D. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *J. Arid Environ.* **2010**, *74*, 229–238. [[CrossRef](#)]
21. Mouazen, A.M.; Kuang, B.; De Baerdemaeker, J.; Ramon, H. Comparison 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* **2010**, *158*, 23–31. [[CrossRef](#)]
22. Kuang, B.; Tekin, Y.; Mouazen, A.M. 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 Till. Res.* **2015**, *146*, 243–252. [[CrossRef](#)]
23. Araujo, S.R.; Wetterlind, J.; Dematte, J.A.M.; Stenberg, B. 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.* **2014**, *65*, 718–729. [[CrossRef](#)]
24. Sorenson, P.T.; Small, C.; Tappert, M.C.; Quideau, S.A.; Drozdowski, B.; Underwood, A.; Janz, A. Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy. *Can. J. Soil Sci.* **2017**, *97*, 241–248. [[CrossRef](#)]
25. Morellos, A.; Pantazi, X.E.; Moshou, D.; Alexandridis, T.; Whetton, R.; Tziotzios, G.; Wiebensohn, J.; Bill, R.; Mouazen, A.M. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosyst. Eng.* **2016**, *152*, 104–116. [[CrossRef](#)]
26. Nawar, S.; Mouazen, A.M. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *Catena* **2017**, *151*, 118–129. [[CrossRef](#)]

27. Ramirez-Lopez, L.; Behrens, T.; Schmidt, K.; Stevens, A.; Dematte, J.A.M.; Scholten, T. The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets. *Geoderma* **2013**, *195*–196, 268–279. [[CrossRef](#)]
28. Clairotte, M.; Grinand, C.; Kouakoua, E.; Thebault, A.; Saby, N.P.A.; Bernoux, M.; Barthes, B.G. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma* **2016**, *276*, 41–52. [[CrossRef](#)]
29. Carmon, N.; Ben-Dor, E. An advanced analytical approach for spectral-based modelling of soil properties. *Int. J. Emerg. Technol. Adv. Eng.* **2017**, *7*, 90–97.
30. Vohland, M.; Besold, J.; Hill, J.; Frund, H.C. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma* **2011**, *166*, 198–205. [[CrossRef](#)]
31. Jensen, J.R. *Remote Sensing of the Environment: An Earth Resource Perspective*; Prentice Hall: Upper Saddle River, NJ, USA, 2007; p. 525.
32. Mouazen, A.M.; De Baerdemaeker, J.; Ramon, H. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil Till. Res.* **2005**, *80*, 171–183. [[CrossRef](#)]
33. Shi, T.; Wang, J.; Chen, W.; Wu, G. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. *Intl. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 95–103. [[CrossRef](#)]
34. Ren, H.Y.; Zhuang, D.F.; Singh, A.N.; Pan, J.J.; Qid, D.S.; Shi, R.H. Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: A case study. *Pedosphere* **2009**, *19*, 719–726. [[CrossRef](#)]
35. Song, Y.; Li, F.; Yang, Z.; Ayoko, G.A.; Frost, R.L.; Ji, J. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang river delta, China. *Appl. Clay Sci.* **2012**, *64*, 75–83. [[CrossRef](#)]
36. Gomez, C.; Lagacherie, P.; Coulouma, G. Regional predictions of eight common soil properties and their spatial structures from hyperspectral Vis-NIR data. *Geoderma* **2012**, *189–190*, 176–185. [[CrossRef](#)]
37. Mark, H.L.; Tunnell, D. Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Anal. Chem.* **1985**, *57*, 1449–1456. [[CrossRef](#)]
38. Shenk, J.S.; Westerhaus, M.O. Population definition, sample selection, and calibration procedure for near infrared reflectance spectroscopy. *Crop Sci.* **1991**, *31*, 469–474. [[CrossRef](#)]
39. Duckworth, J. Mathematical data preprocessing. In *Near-Infrared Spectroscopy in Agriculture*; Roberts, C.A., Workman, J., Jr., Reeves, J.B., III, Eds.; ASA-CSSA-SSSA: Madison, WI, USA, 2004; pp. 115–132.
40. Vasques, G.M.; Grunwald, S.; Sickman, J.O. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma* **2008**, *146*, 14–25. [[CrossRef](#)]
41. Yu, X.; Liu, Q.; Wang, Y.; Liu, X.; Liu, X. Evaluation of MLSR and PLSR for estimating soil element contents using visible/near-infrared spectroscopy in apple orchards on the Jiaodong peninsula. *Catena* **2016**, *137*, 340–349. [[CrossRef](#)]
42. Brown, D.J.; Shepherd, K.D.; Walsh, M.G.; Mays, M.D.; Reinsch, T.G. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *132*, 273–290. [[CrossRef](#)]
43. Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [[CrossRef](#)]
44. Maleki, M.R.; Mouazen, A.M.; De Keterlaere, B.; Ramon, H.; De Baerdemaeker, J. On-the-go variable-rate phosphorus fertilisation based on a visible and near infrared soil sensor. *Biosyst. Eng.* **2008**, *99*, 35–46. [[CrossRef](#)]
45. Gholizadeh, A.; Boruvka, L.; Saberioon, M.M.; Vasat, R. Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. *Appl. Spectrosc.* **2013**, *67*, 1349–1362. [[CrossRef](#)] [[PubMed](#)]
46. Xie, X.; Pan, X.Z.; Sun, B. Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a Copper smelter. *Pedosphere* **2012**, *22*, 351–366. [[CrossRef](#)]
47. Kuhn, M. Building predictive models in R using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [[CrossRef](#)]
48. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
49. Cutler, A.; Cutler, D.R.; Stevens, J.R. *Random Forests*; Springer: Boston, MA, USA, 2012; pp. 157–175.

50. Nawar, S.; Mouazen, A.M. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. *Sensors* **2017**, *17*, 2428. [[CrossRef](#)] [[PubMed](#)]
51. Abdel Rahman, A.M.; Pawling, J.; Ryczko, M.; Caudy, A.A.; Dennis, J.W. Targeted metabolomics in cultured cells and tissues by mass spectrometry: Method development and validation. *Anal. Chim. Acta* **2014**, *845*, 53–61. [[CrossRef](#)] [[PubMed](#)]
52. Segal, M.; Xiao, Y. Multivariate random forests. *WIREs Data Min. Knowl. Discov.* **2011**, *1*, 80–87. [[CrossRef](#)]
53. Prasad, A.M.; Iverson, L.R.; Liaw, A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* **2006**, *9*, 181–199. [[CrossRef](#)]
54. Peters, J.; De Baets, B.; Verhoest, N.E.C.; Samson, R.; Degroeve, S.; De Becker, P.; Huybrechts, W. Random forests as a tool for ecohydrological distribution modelling. *Ecol. Modell.* **2007**, *207*, 304–318. [[CrossRef](#)]
55. Caruana, R.; Niculescu-Mizil, A. An Empirical Comparison of Supervised Learning Algorithms. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 161–168.
56. Liaw, A.; Wiener, M. Classification and Regression by Random Forest. *R News* **2002**, *2*, 18–22.
57. Brown, D.J. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma* **2007**, *140*, 444–453. [[CrossRef](#)]
58. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, USA, 1984; p. 358.
59. Steinberg, D.; Colla, P. *CART: Tree-Structured Non-Parametric Data Analysis*; Salford Systems: San Diego, CA, USA, 1997.
60. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
61. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [[CrossRef](#)]
62. Friedman, J.H.; Meulman, J.J. Multiple additive regression trees with application in epidemiology. *Stat. Med.* **2003**, *22*, 1365–1381. [[CrossRef](#)] [[PubMed](#)]
63. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **2000**, *28*, 337–374. [[CrossRef](#)]
64. Ridgeway, G. Gbm: Generalized Boosted Regression Models. Available online: <https://CRAN.R-project.org/package=gbm> (accessed on 12 May 2018).
65. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
66. Kovacevic, M.; Bajat, B.; Trivic, B.; Pavlovic, R. Geological units classification of multispectral images by using support vector machines. In Proceedings of the International Conference on Intelligent Networking and Collaborative Systems, New York, NY, USA, 4–6 November 2009; pp. 267–272.
67. Vapnik, V. *Statistical Learning Theory*; Wiley-Interscience: New York, NY, USA, 1998.
68. An, A. Classification methods. In *Encyclopedia of Data Warehousing and Mining*; Wang, J., Ed.; Idea Group Inc.: New York, NY, USA, 2005; pp. 144–149.
69. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997; pp. 24–51.
70. Daelemans, W.; Van den Bosch, A. *Memory-Based Language Processing*; Cambridge University Press: Cambridge, UK, 2005; p. 189.
71. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice Hall, Pearson Education Inc.: Upper Saddle River, NJ, USA, 2003; p. 733.
72. Ramirez-Lopez, L.; Stevens, A. Resemble: Regression and Similarity Evaluation for Memory-Based Learning in Spectral Chemometrics R Package Version 1.2.2. 2016. Available online: <https://cran.r-project.org/web/packages/resemble/resemble.pdf> (accessed on 1 June 2018).
73. Box, G.E.P.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. B (Methodol.)* **1964**, *1964*, 211–252.
74. Sarathjith, M.C.; Das, B.S.; Wani, S.P.; Sahrawat, K.L. Dependency measures for assessing the covariation of spectrally active and inactive soil properties in diffuse reflectance spectroscopy. *Soil Sci. Soc. Am. J.* **2014**, *78*, 1522–1530. [[CrossRef](#)]
75. Kusumo, B.H.; Hedley, M.J.; Hedley, C.B.; Tuohy, M.P.; Arnold, C.G. The use of diffuse reflectance spectroscopy for in situ carbon and nitrogen analysis of pastoral soils. *Aust. J. Soil Res.* **2008**, *46*, 623–635. [[CrossRef](#)]

76. Kuang, B.; Mouazen, A.M. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. *Eur. J. Soil Sci.* **2011**, *62*, 629–636. [[CrossRef](#)]
77. Ben-Dor, E.; Irons, J.R.; Epema, G.F. Soil reflectance. In *Manual of Remote Sensing, Remote Sensing for the Earth Sciences*; Rencz, A.N., Ed.; John Wiley & Sons: New York, NY, USA, 1999; pp. 111–188.
78. Brunet, D.; Barthes, B.G.; Chotte, J.L.; Feller, C. Determination of carbon and nitrogen contents in Alfisols, Oxisols and Ultisols from Africa and Brazil using NIRS analysis: Effects of sample grinding and set heterogeneity. *Geoderma* **2007**, *139*, 106–117. [[CrossRef](#)]
79. Gholizadeh, A.; Boruvka, L.; Vasat, R.; Saberioon, M.M.; Klement, A.; Kratina, J.; Tejnecky, V.; Drabek, O. Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: A case study. *PLoS ONE* **2015**. [[CrossRef](#)] [[PubMed](#)]
80. Jalabert, S.S.M.; Martin, M.P.; Renaud, J.P.; Boulonne, L.; Jolivet, C.; Montanarella, L. Estimating forest soil bulk density using boosted regression modeling. *Soil Use Manag.* **2010**, *26*, 516–528. [[CrossRef](#)]
81. Stevens, A.; Udelhoven, T.; Denis, A.; Tychon, B.; Liroy, R.; Van Wesemeal, B. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* **2010**, *158*, 32–45. [[CrossRef](#)]
82. Zornoza, R.; Guerrero, C.; Mataix-Solera, J.; Scow, K.M.; Arcenegui, V.; Mataix-Beneyto, J. Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biol. Biochem.* **2008**, *40*, 1923–1930. [[CrossRef](#)] [[PubMed](#)]
83. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In *5th Annual ACM Workshop on COLT*; Haussler, D., Ed.; ACM Press: Pittsburgh, PA, USA, 1992; pp. 144–152.
84. Gupta, A.; Vasava, H.B.; Das, B.S. Choubey, K. Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region. *Geoderma* **2018**, *325*, 59–71. [[CrossRef](#)]
85. Carmon, N.; Ben-Dor, E. Mapping Asphaltic Roads' Skid Resistance Using Imaging Spectroscopy. *Remote Sens.* **2018**, *10*, 430. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).