remote sensing

MDPI

Article

A Memory-Based Learning Approach as Compared to Other Data Mining Algorithms for the Prediction of Soil Texture Using Diffuse Reflectance Spectra

Asa Gholizadeh [1,*], Luboš Borůvka [1], Mohammadmehdi Saberioon [2] and Radim Vašát [1]

[1] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Prague 16521, Czech Republic; boruvka@af.czu.cz (L.B.); vasat@af.czu.cz (R.V.)

[2] Laboratory of Signal and Image Processing, Institute of Complex Systems, South Bohemia Research Centre of Aquaculture and Biodiversity of Hydrocenoses, Faculty of Fisheries and Protection of Waters, University of South Bohemia in České Budějovice, Nové Hrady 37333, Czech Republic; msaberioon@frov.jcu.cz

* Correspondence: gholizadeh@af.czu.cz; Tel.: +420-224-382-633

Academic Editors: José Alexandre Melo Demattê, Nicolas Baghdadi and Prasad S. Thenkabail
Received: 10 November 2015; Accepted: 12 April 2016; Published: 19 April 2016

Abstract: Successful determination of soil texture using reflectance spectroscopy across Visible and Near-Infrared (VNIR, 400–1200 nm) and Short-Wave-Infrared (SWIR, 1200–2500 nm) ranges depends largely on the selection of a suitable data mining algorithm. The objective of this research was to explore whether the new Memory-Based Learning (MBL) method performs better than the other methods, namely: Partial Least Squares Regression (PLSR), Support Vector Machine Regression (SVMR) and Boosted Regression Trees (BRT). For this purpose, we chose soil texture (contents of clay, silt and sand) as testing attributes. A selected set of soil samples, classified as Technosols, were collected from brown coal mining dumpsites in the Czech Republic (a total of 264 samples). Spectral readings were taken in the laboratory with a fiber optic ASD FieldSpec III Pro FR spectroradiometer. Leave-one-out cross-validation was used to optimize and validate the models. Comparisons were made in terms of the coefficient of determination (R^2_{cv}) and the Root Mean Square Error of Prediction of Cross-Validation ($RMSEP_{cv}$). Predictions of the three soil properties by MBL outperformed the accuracy of the remaining algorithms. We found that the MBL performs better than the other three methods by about 10% (largest R^2_{cv} and smallest $RMSEP_{cv}$), followed by the SVMR. It should be pointed out that the other methods (PLSR and BRT) still provided reliable results. The study concluded that in this examined dataset, reflectance spectroscopy combined with the MBL algorithm is rapid and accurate, offers major efficiency and cost-saving possibilities in other datasets and can lead to better targeting of management interventions.

Keywords: Technosols; model performance; VNIR/SWIR spectroscopy; PLSR; SVMR; BRT

1. Introduction

The accomplishment of sustainable agricultural and environmental management requires a better understanding of the soil at increasingly finer scales. Conventional soil sampling and laboratory analyses cannot effectively provide this information because they are slow and costly [1]. Visible and Near-Infrared (VNIR, 400–1200 nm) spectroscopy and Short-Wave-Infrared (SWIR, 1200–2500 nm) spectroscopy are non-destructive, rapid and low-cost methods that differentiate materials based on their reflectance in the wavelength range from 400–2500 nm. VNIR/SWIR spectroscopy was confirmed to be a superior substitute for conventional laboratory analysis of soil chemical properties, such as various forms of carbon [2], N, P, K contents, Cation Exchange Capacity (CEC), pH [3] and, to some

extent, physical parameters, including soil structure, bulk density and texture [4–6]. Actually, because the analysis of the clay fraction depends on features of the mineral content, VNIR/SWIR spectra can be of value for predicting clay content [7,8].

VNIR/SWIR spectroscopy allows for fast, cost-effective and intensive data collection, although problems related to instrumentation instability (and the differences in calibration between different devices used for the same purpose), environmental conditions and difficulties related to the scale of the experiment (global, regional, local, field) lead to variation in accuracy [9,10]. Under *in situ* measurement conditions with non-mobile or mobile instrumentation, additional challenges linked to diverse soil moisture content, color, dust, stones and excessive residues and surface roughness all affect the accuracy of the measurement [11,12]. To overcome one or more of these difficulties, some solutions were suggested and employed by researchers. These included the selection of proper instrumentation, improved spectra filtering and preprocessing [13], better control of ambient conditions [11] and the appropriate selection of multivariate statistical analysis [14,15].

Soil VNIR/SWIR spectra are non-specific; they include weak, wide and overlapping absorption bands. For this reason, information needs to be mathematically extracted from the spectra for correlating with soil parameters. Multivariate statistics are frequently used to calibrate soil prediction models. Quantitative spectral analysis of soil may therefore necessitate complicated techniques to detect the response of soil attributes from spectral characteristics [8]. Araújo *et al*. [8] stated that attention toward nonlinear data mining calibration techniques is escalating, as relationships between soil properties are not often linear in nature, mainly in libraries containing a broad variety of soils. When dealing with a heterogeneous sample set in which soil composition may vary considerably, the accuracy of linear regression methods decreases, because of the nonlinear nature of the relationship between spectral data and the dependent variable. Partial Least Squares Regression (PLSR) is the most common algorithm used to calibrate VNIR/SWIR spectra to soil properties [16–19]. Other approaches have also been used, for example Multiple Linear Regression (MLR) [20], Principle Component Regression (PCR) [21], Artificial Neural Networks (ANN) [22], Multivariate Adaptive Regression Splines (MARS) [23], PLSR with bootstrap aggregation (bagging-PLSR) [24] and Penalized Signal Regression (PSR) [25]. Brown [26] suggested the use of Boosted Regression Trees (BRT), and Kovačević *et al*. [27] and Gholizadeh *et al*. [28] recommended the use of Support Vector Machine Regression (SVMR) as the best solution for handling the calibration of sample populations. Memory-Based Learning (MBL) is a data-driven approach and can be defined as a lazy learning method. Despite other learning methods, the key aim in MBL is not to achieve a general or global target function. Instead, when an explanation for a new problem is required, experience in the form of a set of similar related samples is regained from memory, and then, those samples are merged to build the solution and explanation to the new problem [29]. Therefore, for each new problem, a new target function is obtained. A global target function may be very complex, while MBL can explain the target function as a set of less complex local (or locally stable) approximations [30]. In this case, nonlinear relationships can be simply determined. In contrast to complex learning techniques, such as ANN or SVMR, most of the MBL systems do not need a complex function fitting process [31], so it can be introduced as a supportive calibration algorithm that has been employed to analyze soil texture, in the spectral domain.

Evaluation and estimation of soil texture is essential for the mapping of regions at risk of soil erosion, driven by water and wind. Coarser-textured soils are more resistant to detachment and movement via raindrops and, so, are less influenced by water-assisted erosion [32]. Soils with silt content above 40% are believed to be extremely erodible, while clay particles can potentially bind with Soil Organic Matter (SOM) to shape aggregates, which help in their resistance to erosion [32]. Another incentive to determine a soil's texture is calculating a soil's capability to retain water or allow drainage; for example, clays can display swelling properties, absorbing and accumulating water within their layered lattice structure [33]. Such finer-textured clay-rich soils can retain more water for plant growth than sandy soils. However, under flood conditions, they have poor infiltration and drainage of overload water and, so, are prone to becoming saturated [34].

Successful predictions of Soil Organic Carbon (SOC) using spectroscopy have been reported [35,36]. Soil water content has also been predicted under both laboratory and *in situ* conditions [37]. Clay content can be well estimated with VNIR [38–40], but much less attention has been given to the prediction of other soil textural classes, including silt and sand. The use of VNIR/SWIR reflectance spectroscopy offers a lower precision than for clay, especially with particular chemometric algorithms.

The question arises: why another study on different calibration approaches? As shown by Gholizadeh *et al*. [10,41], choosing the most robust calibration technique can help to achieve a more reliable and accurate prediction model. Moreover, different studies reveal different results, because the nature of the target function has a significant effect on the performance of the different prediction approaches. Therefore, in this context, the aim of this paper was to compare the performance of different state-of-the-art calibration methods, with special attention given to the MBL algorithm. The purpose was to provide the interpretation of the results for the prediction of soil texture using VNIR/SWIR diffuse reflectance spectra data by the best performing algorithm. This study was performed over bare soil sites within the Bílina and Tušimice area in the Czech Republic.

2. Materials and Methods

2.1. Study Area

Six dumpsites in the mines Bílina and Tušimice in the Czech Republic were selected (Figure 1): Pokrok (50°60′N; 13°71′E), Radovesice (50°54′N; 13°83′E), Březno (50°39′N; 13°36′E), Merkur (50°41′N; 13°30′E), Prunéřov (50°42′N; 13°28′E) and Tumerity (50°37′N; 13°31′E).

An amount of approximately 2500–3000 t per ha natural topsoil was extended as a cover one year before sampling on a part of each dumpsite. The topsoil material originated from humic horizons of natural soils of the region, mainly Vertisols and partially Chernozems (clayic and haplic). The topsoil was not mixed with the dumpsite material. Soil attributes differed somewhat between the six dumpsites. Some properties of the topsoils, including pH, SOM and texture, were determined using bulk control subsamples. The soil pH range for the entire area was 5.3–8.5. The SOM content range was 0.6%–3.8%.

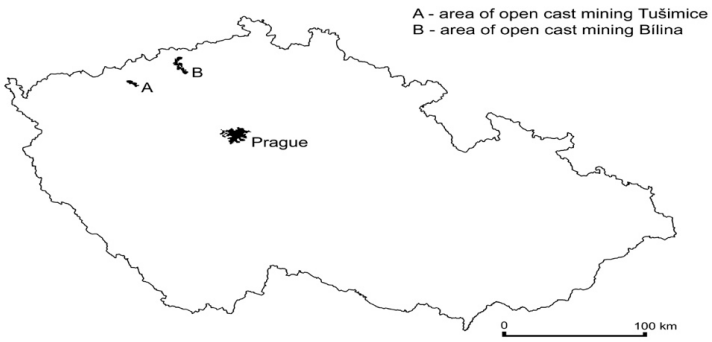

Figure 1. Map of the sampling locations in the Czech Republic.

2.2. Soil Sampling and Analysis

A total of 264 soil samples was collected: 103 samples on Pokrok, 40 samples on Radovesice, 25 samples on Březno, 38 samples on Merkur, 48 samples on Prunéřov and 10 samples on the Tumerity dumpsite. All soils were classified as Technosols, according to World Reference Base (WRB) for soil resources [42]. Roughly half of the sampling points were placed on the region with natural topsoil cover and half on the region without the cover. Sampling was made at a depth of 0–30 cm [18,43]. This depth corresponds to the common depth of a ploughed soil layer, as these soils will be mostly used as arable land in the future. The depth of the topsoil cover was also at least 30 cm.

The original samples were air-dried, crushed and sieved ($\leqslant$2 mm) and thoroughly mixed before analyzing. The particle size distribution (textural fractions of clay, silt and sand) was determined by the sedimentation hydrometer method [44]. Samples and standards were matrix matched, and all analyses were carried out in triplicate.

2.3. Spectral Data Measurements

Spectral reflectance was deliberated in the 350–2500-nm wavelength range using a fiber optic ASD FieldSpec III Pro FR spectroradiometer (ASD Inc., Denver, CO, USA) under laboratory conditions. The spectral resolution of the spectroradiometer was 3 nm for the region 350–1000 nm and 10 nm for the region 1000–2500 nm. Moreover, the radiometer bandwidth from 350–1000 nm was 1.4 nm, while it was 2 nm from 1000–2500 nm. Samples were illuminated using a stable direct current-powered 50-W tungsten-quartz-halogen lamp, which was installed on a tripod. The angle of incident illumination was 15°, and the distance between the illumination source and the sample was 30 cm. A fiber optic probe with an 8° field of view was used to collect reflected light from the sample. The probe was installed on a tripod and located approximately 10 cm vertically above the sample. Soil samples were placed in 9-cm diameter petri dishes, forming a 2-cm layer of soil to avoid beam reflectance from the bottom of the dish, due to downwelling solar and sky radiation penetrating into the soil approximately 1/2 wavelength [45], which could have the unwanted effect of modifying the soil spectra. Samples were levelled off using a blade to guarantee a flat surface flush with the top of the petri dishes, as a smooth soil surface ensures maximum light reflection and a large Signal-to-Noise Ratio (SNR) [46]. We measured all spectral readings in the center of the samples in a dark room to avoid interference from stray light. The final spectrum was an average based on 20 iterations from 4 directions, with 5 iterations per direction to improve the SNR. Each sample spectrum was corrected for background absorption before each single measurement to account for changes in temperature and air humidity; the spectral transmission of the fingertip was also corrected using a reference spectrum through a 1-mm layer of a white $BaSO_4$ panel standard [43,47].

2.4. Spectra Preprocessing

Murray [48] mentioned that removing outliers improves prediction accuracy; hence, the outliers were left out. Outliers were detected by using the principle of Mahalanobis distance (H) [49,50], applied on PCA-reduced data. The H statistic identified outliers whose spectra were different from other samples that made up the calibration set [51]. In the present study, an H value of 3 (based on the Mahalanobis distance) was chosen for the identification of outliers [52]. The detected spectral outliers were deleted from the calibration set. These samples should not belong to the population.

In order to calibrate a model that provides accurate predictive performance about the soil texture content in each soil sample, the captured soil spectra, jointly with laboratory data of the aforementioned parameters, were imported into R software (R Development Core Team, Vienna, Austria) to be processed. Spectra preprocessing algorithms entailed a range of mathematical techniques for refining light scattering in spectral reflectance measurements and data improvement before the data were used in calibration models. The first derivative transformation, which was utilized in this study, is very efficient for eliminating baseline offset and, according to some researchers, gives the best results and uppermost accuracy among other algorithms [28,41,53]. In this study, before all further spectra treatments, the noisy parts of the spectra, ranges 350–399 nm and 2450–2500 nm, were removed, and the spectra were subjected to Savitzky–Golay smoothing with a second-order polynomial fit and 11 smoothing points [18,54] for eliminating the artificial noise caused by the spectroradiometer device.

2.5. Comparison of Algorithms

Four different calibration techniques, PLSR, SVMR, BRT and MBL, were applied to calibrate spectral data with texture reference data and to describe the relationship between reflectance spectra and estimated soil texture. We present a brief summary of each algorithm in the following sections.

2.5.1. Partial Least Square Regression

The PLSR has turned into a popular method used in chemometrics that is applied for quantitative analysis of diffuse reflectance spectra. It decreases the data, noise and calculation time, with minor loss of the information contained in the original variables [55], and its arithmetic can be referred to Wold *et al.* [16]. It is strongly related to PCR in that both use statistical rotations to defeat the problem

of high dimensionality and multicollinearity [39,56]. They both compress the data before completing the regression. The difference is that the PLSR algorithm combines the compression and regression steps, and it selects successive orthogonal factors that maximize the covariance between predictor and response variables [3,15,56,57]. By fitting a PLSR model, one expects to discover a few PLSR factors that clarify most of the variation in both predictors and responses [58]. As stated by Gholizadeh *et al.* [10], Viscarra Rossel and Behrens [15] and Bilgili *et al.* [59], PLSR decomposes X and Y variables and finds new factors, called latent variables, which are both orthogonal and weighted linear combinations of X variables. These new X variables are then used for prediction of Y variables, according to:

$$X = Tp' + E \tag{1}$$

$$Y = Tq + F \tag{2}$$

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

Variables X and Y are mean-centered by subtracting column averages from each observation in the column prior to decomposition. The decomposition is performed simultaneously and in such a way that the first few factors describe most of the variation in X and Y. The residual factors resemble noise and can be ignored, hence the addition of residuals E and F. Generally, the resulting matrices and vectors have a much lower dimension than X and Y. Given a new reflectance X, thus, the soil attribute Y can be assessed as a (bi)linear combination of the factor scores and factor loadings of X [15]. It can be said that in PLSR, an essential step is the selection of the optimal number of latent variables in the calibration model to avoid under-fitting and over-fitting of data that would generate models with poor prediction potential [43,59].

2.5.2. Support Vector Machine Regression

The SVMR approach is a supervised, nonparametric and statistical learning method [60]. It has been identified to strike the correct balance between the accuracy gained from a given limited amount of training patterns and the generalization capability to handle unseen data. The algorithm is nonlinear and is employed in classification and multivariate calibration issues [27]. In this method, model complication is finite by the learning algorithm itself, which avoids over-fitting. Based on Vapnik [60], SVMR is a kernel-based learning method from statistical learning theory. The kernel-based learning method uses an implicit mapping of the input data into a high dimensional feature space described by a kernel function [61]. Using this so-called kernel-trick [62], it is possible to obtain a linear hyperplane as a decision function for nonlinear problems and then apply a back-transformation in the nonlinear space [15]. The ε-SVMR employs training data to obtain a model represented as a so-called ε-insensitive loss function (tube, band), which maps independent data with maximum ε deviation from dependent training data [56]. Error within the predetermined distance ε from the true value is ignored, and error greater than ε is penalized by the soil property. Finally, the model diminishes the complexity of the training data to a significant subset of so-called support vectors. Therefore, consider a given training set of N data points, $\{x_k, y_k\}_{k=1}^{N}$, with input data, which is an n-dimensional data vector ($x \in R^N$), and output, which is the one-dimensional vector space ($y \in r$). The subsequent equation for prediction has been described by Vapnik [63]:

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \tag{3}$$

where:
b, scalar threshold

$K(x, x_k)$, kernel function
α, Lagrange multiplier
N, number of data
x_k, input data
y, output

The Radial Basis Function (RBF) has been used in this study because RBF tends to give good efficiency under general smoothness assumption and can be estimated as below:

$$K(x, x_k) = exp\left\{-\frac{(x - x_k)^T (x - x_k)}{2\sigma^2}\right\},\ k = 1, \ldots, N \tag{4}$$

where:
σ, width of the radial basis function
T, transpose

2.5.3. Boosted Regression Trees

Based on Brown [26], BRT have been suggested as an ideal data-mining or pattern-recognition tool for VNIR/SWIR spectroscopy of soil properties. Boosted Regression Tree (BRT) analysis basically performs a binary recursive partitioning of the dataset [64,65]. At each terminal node, a predicted value is gained as the average of all of the measurements that were grouped in that node. The method makes multiple predictions that are based on resampling and weighting and belongs to the group of ensemble techniques [66]. It has the ability to take in a large number of weak relationships in a predictive model, and it is not sensitive to outliers in the calibration dataset [8]. Following Friedman [66], boosted models can be stated in the general form:

$$F(x; \{\beta_m, a_m\}_0^M) = \sum_{m=0}^{M} \beta_m h(x; a_m) \tag{5}$$

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

The base learner is applied in order to reweigh calibration datasets, such that observations with larger residuals receive proportionally higher weights in subsequent iterations [67]. The final classification is calculated with a weighted vote, as shown in Equation (5) [26]. The primary advantages of BRT include: (i) the ability to include a large number of weak relationships in a predictive model; (ii) insensitivity to outliers in the calibration dataset; (iii) no necessity for uniform data transformations; and (iv) relative immunity to over-fitting [68,69].

2.5.4. Memory-Based Learning

Memory-Based Learning (MBL) is a data-driven approach, which is closely related to Case-Based Reasoning (CBR). Like CBR, MBL resembles the human reasoning process [29,70]: remember earlier situations; reconcile them for solving the existing problem; study the possibility to solve the problem with the new solution; and memorize the skill for knowledge development. Actually, MBL is based on the idea that intelligent behavior can be achieved by analogical analysis, rather than by the use of abstract mental and rule-based processing [30]. Based on Daelemans [71], MBL is a family of learning algorithms that, in preference to performing clear and precise generalization, compares new problem cases to cases seen in training, which have been stored in memory, and it is a sort of lazy learning. It builds hypotheses directly from the training cases themselves [72]. This means that the hypothesis complexity can grow with the data. In contrast to other learning methods, the main goal in MBL is not

to obtain a general or global target function. In MBL, when a solution for a new problem is essential, the experience in the form of a set of analogous related samples is recovered from memory, and then, those samples are merged to create the solution to the new problem. Consequently, for each new problem, a new target function is developed. Actually, MBL carries out interpolation locally, which is based on a local reference set or spectral library. This means that nonlinear relationships can be simply resolved. There are two sets of data needed in the MBL calibration method. A set of n reference samples (e.g., spectral library), $(X_r, Y_r) = \{X_{ri}, Y_{ri}\}_{i=1}^{n}$, and a set of m samples as the prediction set, $(X_u, Y_u) = \{X_{uj}, Y_{uj}\}_{j=1}^{m}$, where Y_u is unknown. Prior to modeling, it is necessary to seek and find out k-nearest neighbors of each data in the prediction set, and then, a local model is calibrated with these referenced neighbors for predicting the corresponding value in Y_u from X_u. Correlation dissimilarity was used in this study for Nearest Neighbor (NN) selection. The NN of each sample specifies its most similar sample in terms of its VNIR/SWIR principal components. The local models are then fitted by applying weighted average PLS, which is the weighted mean of all of the predicted values created by the multiple PLS models between a maximum and minimum number of PLS components. The weight for each component can be evaluated as below:

$$w_j = \frac{1}{s_{1:j} \times g_j} \tag{6}$$

where:

$s_{1:j}$, root mean square of the spectral residuals of the unknown sample when a total of j-th PLS components is used

g_j, root mean square of the regression coefficient corresponding to the j-th PLS components

Further details on MBL regression can be found in Ramirez-Lopez *et al.* [29].

2.6. Assessment of VNIR/SWIR Predictions Performances

Proper fitting was achieved using leave-one-out cross-validation in which the models were constructed each time by leaving one sample out of the calibration dataset in order to use in the validation process until all samples were left out once.

The ability of the techniques to predict soil texture classes was evaluated by calculating the corresponding coefficient of determination of cross-validation (R^2_{cv}) and Root Mean Square Error of Prediction of Cross-Validation ($RMSEP_{cv}$). The R^2_{cv} and $RMSEP_{cv}$ were calculated based on the following equations:

$$R^2_{cv} = \left(1 - \frac{\sum_{i=1}^{N} \left(y'_i - y_i\right)^2}{\sum_{i=1}^{N} \left(y'_i - \overline{y_i}\right)^2}\right) \tag{7}$$

$$RMSEP_{cv} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y'_i - y_i\right)^2} \tag{8}$$

where:

y'_i, predicted value

y_i, observed value

$\overline{y_i}$, mean of y value

N, number of samples

Actually, R^2 shows the percentage of the variance in the y variable that is calculated by the x variables. An R^2 value between 0.50 and 0.65 demonstrates that more than 50% of the variance in y is calculated by variable x, so that differentiation between high and low condensation is possible. An R^2 value between 0.66 and 0.81 displays estimated quantitative predictions, while an R^2 between 0.82 and 0.90 manifests good prediction. Calibration models having an R^2 value higher than 0.91 are assumed excellent [73].

3. Results and Discussion

3.1. Soil Textural Properties

Summary statistics for the soil samples from the six dumpsites, including minimum, maximum, mean, Standard Deviation (SD) and Coefficient of Variation (CV), are shown below (Table 1). The samples under study represented a narrow range of silt, especially in Tumerity (ranging from 22.9%–30.6%); however, they varied widely in the case of clay and sand content. Ramirez-Lopez *et al.* [29] also observed a wide range of clay in their study, which was reportedly due to the high variability of the region in terms of parent material. The data also showed that the Tumerity area was more clayey than other dumpsites, followed by Merkur and Radovesice. Pokrok and Prunéřov were considerably coarser-textured than the other dumpsites, as the sand content was generally higher there.

The comparison of properties' CV revealed that among all properties, sand had the highest CV, particularly in the Radovesice and Tumerity areas (51.9% and 51.7%, respectively); therefore, sand varied the most as compared to other considered attributes. Silt showed the lowest CV, especially in Tumerity (11.3%), which means that it is more homogeneous than the other attributes.

Table 1. Descriptive statistics of soil texture in the studied sample set according to location.

Item	Clay	Silt	Sand
		(%)	
		Pokrok (n = 103)	
Min	7.5	23.8	11.3
Max	53.3	44.9	63.6
Mean	36.7	33.9	29.3
SD	8.7	4.4	9.2
CV (%)	23.6	13.0	31.3
		Radovesice (n = 40)	
Min	18.1	28.2	11.1
Max	52.9	48.0	53.5
Mean	41.9	38.2	19.8
SD	7.8	5.7	10.3
CV (%)	18.5	14.9	51.9
		Březno (n = 25)	
Min	28.9	26.0	9.1
Max	61.4	44.6	34.8
Mean	39.9	32.9	22.1
SD	5.9	4.7	6.1
CV (%)	14.9	14.3	27.6
		Merkur (n = 38)	
Min	17.7	24.3	14.7
Max	59.9	37.6	54.9
Mean	47.5	30.2	22.4
SD	6.5	3.8	5.3
CV (%)	13.8	12.7	23.6
		Prunéřov (n = 48)	
Min	6.1	12.6	14.3
Max	60. 7	48.9	74.3
Mean	40.5	31.2	28.3
SD	12.6	7.6	12.7
CV (%)	31.1	24.4	45.0
		Tumerity (n = 10)	
Min	31.6	22.9	2.7
Max	68.4	30.6	37.8
Mean	50.7	22.9	21.3
SD	11.5	2.6	11.0
CV (%)	22.7	11.3	51.7

3.2. Soil Spectral Properties

A visual assessment of the spectra permitted to remove parts of spectra that are known as the noisiest parts at the edges of the spectrum, and the final spectral library considered the spectral range from 400–2450 nm. Sets of spectra were defined qualitatively by identifying the positive and negative peaks (Figure 2), which appear at particular wavelengths [3].

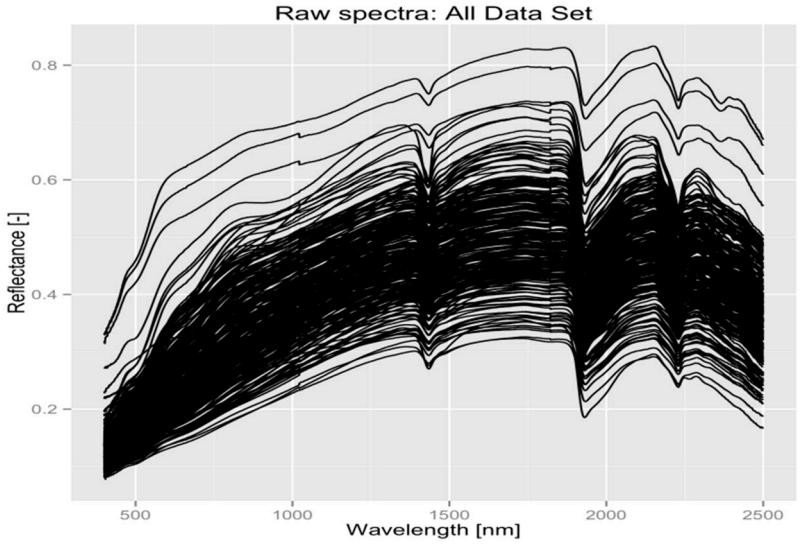

Figure 2. Representative VNIR/SWIR spectra of soil samples.

The spectra have absorption peaks overlapping near 430 nm and 530 nm in the Visible (VIS) region, which state the presence of iron oxides, and are caused by paired and single Fe^{3+} electron transitions to a higher energy state [74–77]. Based on Sherman and Waite [74], the 530-nm band is also credited to absorption limits of extreme charge transfer that occur in the Ultraviolet (UV). The 650-nm shoulder in the spectrum of the soil samples may exhibit the entity of small amounts of hematite (Fe_2O_3) [15]. In the NIR region, O-H bonds in clay minerals would have a general influence on the reflectance spectra [18,54,78]. It can be said that the group of positive peaks near 900 nm may represent absorption caused by electronic transitions in goethite due to Laporte forbidden transitions [74]. The small absorption bands occurring near 1200 nm, 1400 nm and 1900 nm may be due to the vibrational combinations and overtones of molecular water contained in various locations in minerals [8,79]. To be more accurate, the group of peaks near 1400 nm may be attributed to the first overtone of the O-H stretch; the peaks near 1700 nm, due to the first overtone of the C-H stretch; the prominent group of peaks near 1900 nm may be related to H-O-H bend with the O-H stretches [77]. The traits around 2000–2500 nm are linked to the characteristics of SOM and clay minerals [54,78]. Based on Viscarra Rossel *et al.* [77], the prominent peaks near 2200 nm, 2300 nm and 2400 nm may be attributed to the metal-OH bend plus O-H stretch combinations. For example, the absorption near 2204 nm occurs due to the absorption of Al-OH, and the small absorption near 2280 nm may be related to Fe-OH, as Fe is replaced in the octahedral sheet [15]. In the spectrum of soil samples, the absorption near 2380 nm, the minor shoulder near 2350 nm, plus that near 2345 nm may correspond to the presence of illite or mixtures of smectite and illite due to additional Al-OH features [80,81]. It should be noted that band positions and wavelength peaks may vary with composition [82].

3.3. Spectra Preprocessing and Model Calibration

In order to create a robust prediction model and to discover the impression of the spectral sampling interval on the prediction accuracy, Savitzky–Golay smoothing with second-order polynomial fit and 11 smoothing points with subsequent first derivative preprocessing technique were applied prior to

model calibration [41,83,84]. Smoothed spectra only, by Savitzky–Golay filter, as well as smoothed and preprocessed spectra, using Savitzky–Golay plus the first derivative, of all selected soil samples are illustrated below (Figure 3). The comparison between Figures 2 and 3 reveals that the main difference between the spectra is a baseline shift. It also shows that the Savitzky–Golay and first derivative preprocessing techniques can remove additive baseline effects and minimize variation among samples caused by variation in grinding and optical setup. Moreover, they increase the resolution of superposed peaks, decrease noise and enhance possible spectral features connected to the property studied and, thus, are more useful for the prediction of soil texture than the original spectra [58]. The first derivative spectra generally amplify the absorption features indicative of the contents of the soil materials and also reduce variation among samples [58].

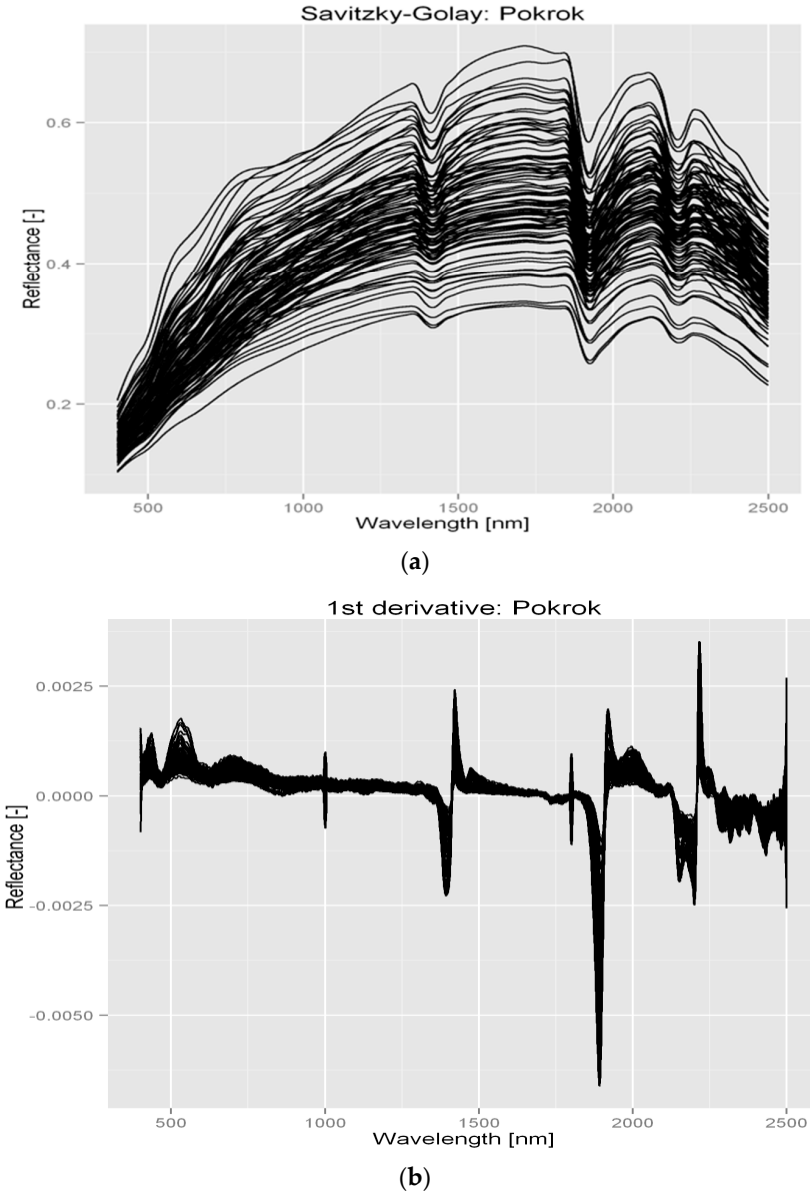

Figure 3. Smoothed only (**a**) and smoothed and first derivative preprocessed (**b**) soil spectra.

The capability of spectral reflectance spectra to predict soil attributes using PLSR, SVMR, BRT and MBL techniques was studied. The comparison of prediction accuracy and model performance from the different algorithms is presented in this part (Table 2). Figure 4 also shows the validation results of predicted and measured values of parameters using the PLSR, SVMR, BRT and MBL algorithms.

Table 2. Statistics of the leave-one-out cross-validation for four different calibration techniques: PLSR, SVMR, BRT and Memory-Based Learning (MBL).

Data Mining Algorithms	Clay		Silt		Sand	
	(%)					
	R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$
PLSR	79	4.48	71	5.79	68	6.63
SVMR	82	4.23	76	5.19	69	6.57
BRT	80	4.38	74	5.31	69	6.59
MBL	89	4.08	81	4.90	76	6.04

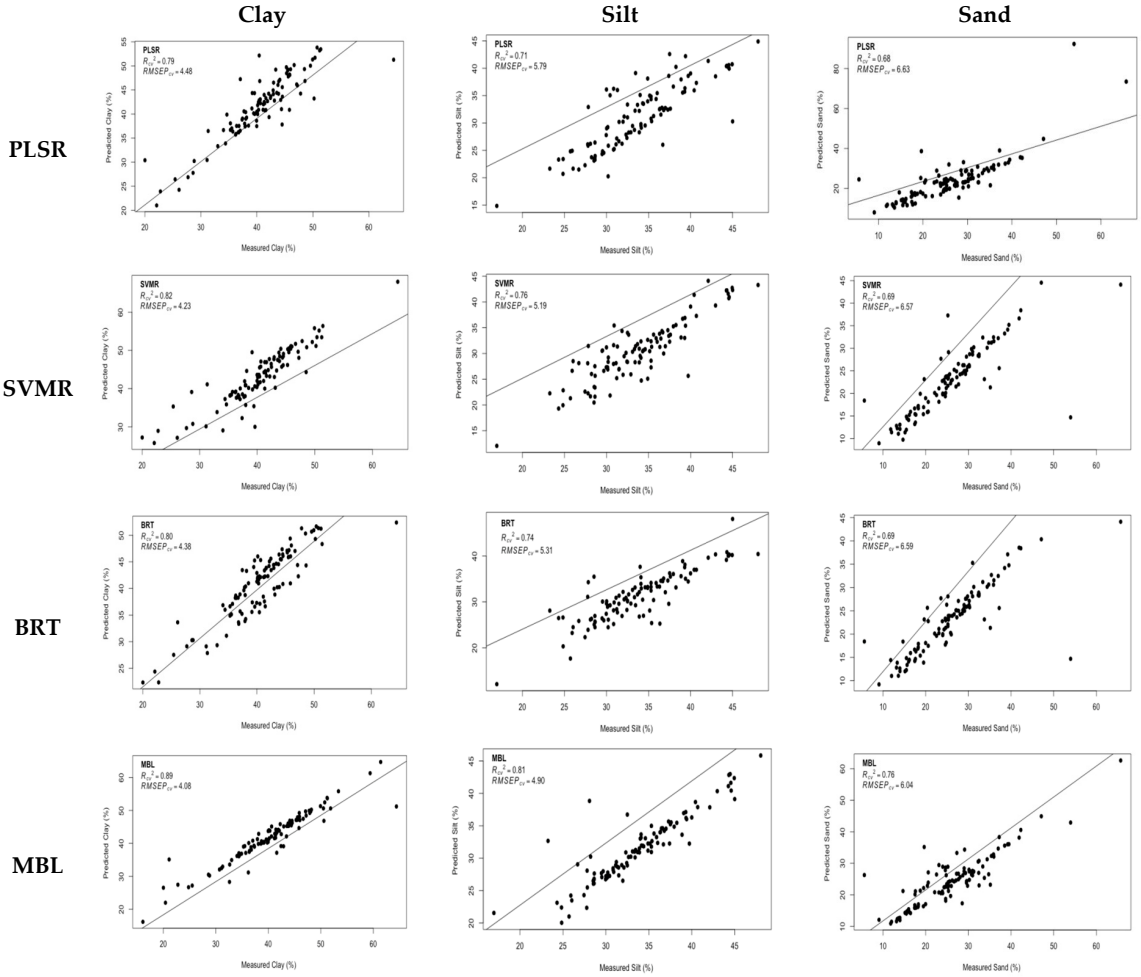

Figure 4. Scatterplots of measured *versus* predicted obtained by PLSR, SVMR, BRT and MBL.

In the multivariate calibration, based on R^2_{cv} and $RMSEP_{cv}$, which have been reported as standard methods for the validation of the prediction models, the most consistent estimates were commonly gained for the clay fraction. This may be related to a different rate of uncertainty in the determination of each particular textural class by the hydrometer method [85]. Compared to PLSR, SVMR, BRT and MBL gave smaller $RMSEP_{cv}$ and larger R^2_{cv} for the prediction of clay, silt and sand (Table 2).

The MBL technique gives generally better prediction of soil texture compared to the other methods, giving the smallest error. Boosted Regression Trees (BRT) showed lower $RMSEP_{cv}$ for all parameters than PLSR, but PLSR still showed relatively good prediction of soil texture. Actually, for the PLSR model, R^2_{cv} values ranged between 0.68 and 0.79, for clay (0.79), silt (0.71) and sand (0.68). These results

are comparable to R^2 values introduced in the literature by other authors [59,81,86,87] who utilized wavelength UV, VIS, NIR and Mid-Infrared (MIR) wavebands. Regarding BRT, Brown *et al*. [39] also found that this method can outperform PLSR and recognized its capacity to contain interactions and nonlinear relationships. They mentioned that the notably improved results of BRT are not surprising as BRT can incorporate complex nonlinear relationships and interactions, whereas PLSR is built on the linear relationship between predictors and the target variable of interest. More accurate outputs of BRT in comparison to PLSR are also a result of some superiorities of this method, such as 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 [26,67,69]. The authors also compared the prediction to the SVMR, which was found to give better prediction than PLSR and even BRT. However, SVMR was less accurate when compared to MBL. The premier performance of SVMR can also be explained by the inclusion of nonlinear and interaction effects, as well as linear combinations of variables. It is able to approximate nonlinear functions between multidimensional spaces [59]. This algorithm can derive a linear hyperplane as a decision function for nonlinear problems, which can be considered as another reason for the method's excellence [62]. Interestingly, for the prediction of sand, SVMR and BRT had the same R^2_{cv} value (0.69). Support Vector Machine Regression (SVMR) had a small $RMSEP_{cv}$ for predicting soil texture, while MBL had even lower $RMSEP_{cv}$ and higher R^2_{cv}. This is most likely because nonlinear relationships can be merely determined by MBL [31]. The superior consequences of MBL generally can be related to the selection of a more appropriate neighbor to calibrate local models, as well as the inclusion in each local model as a source of additional predictor variables [29].

For PLSR, SVMR and BRT, these findings coincided with the results of some other studies. Viscarra Rossel and Behrens [15] applied these methods, amongst others, for the prediction of clay, based on VNIR/SWIR spectra using a large spectral library with 1104 soil samples. Without feature selection, SVMR showed the most successful prediction model (R^2_{cv} = 0.84, $RMSEP_{cv}$ = 7.63) due to its ability to solve the multivariate calibration problems. Araújo *et al*. [8] compared PLSR, SVMR and BRT for their ability to determine clay from 7172 samples of seven different soil types collected from several areas of Brazil. Their goal was to explore the chance of increasing the performance of VNIR/SWIR data in the assessment of clay content in this library. They found that SVMR outperformed BRT and PLSR for clay prediction. Araújo *et al*. [8] mentioned that SVMR superiority relates to the capability of this technique to reduce problems with heterogeneity and nonlinearity. Their study agreed with Brown [26], who compared BRT and PLSR techniques for analyzing soil characteristics with VNIR/SWIR and found BRT to be the superior approach. These authors used 4184 diverse, well-characterized and mostly independent soil samples. Actually, the BRT method tends to be insensitive to the impacts of outliers and can handle omitted values and correlated variables. It also permits the embodiment of a potentially large number of irrelevant predictors [88]. On the other hand, Vasques *et al*. [55], using 554 samples collected in profiles to a depth of 180 cm in north-central Florida, discovered that the BRT model provided the worst results among many multivariate techniques, including PLSR, when tested for total carbon, SOC and clay. Based on their results, one explanation of why BRT was not as good as the other multivariate techniques is the fact that it produces discrete outputs predicting a single value at each terminal node [55]. Ramirez-Lopez *et al.* [29] introduced the Spectrum-Based Learner (SBL) technique, which is a kind of MBL and combines local distance matrices and the spectral features as predictor variables. They used this method for model calibration of clay content, SOC and exchangeable Ca (Ca^{++}) and found that SBL produced more accurate results than the other calibration methods (PLSR and SVMR) for all measured parameters. They found that the SBL approach derives additional predictive information (a characteristic that is not explored by any of the other algorithms) from the spectra. In addition, it carries a more suitable neighbor selection by using the distance matrix [29].

In this study, clay was predicted reliably, whereas the prediction of silt and sand was fair and moderately successful. It could be concluded that the prediction accuracy of data mining techniques,

MBL particularly, will be higher in fine-textured fields than coarse-textured ones, which reflects the influence of the direct spectral responses of clay, especially in the NIR range. Therefore, the Merkur dumpsite soil, which contains more clay than other brown coal mining dumpsites, can be predicted more accurately with higher R^2_{cv} and lower $RMSEP_{cv}$. These results support those studies that found SVMR as a very promising method for the determination of clay content [8,15]. Some researchers claim that spectral predictive mechanisms may differ from one population of soil samples to another. This difference may be caused by the decomposition stage of SOM, the nature of existing compounds and the influence of other relevant factors, such as texture, soil moisture or iron oxides [11,15,38,56]. To the best of our knowledge, the MBL algorithm has not yet been commonly used to analyze and predict soil properties, including soil texture.

Differences between the multivariate methods were more remarkable for clay, but results from the different multivariate approaches were very similar for all properties. For all parameters, MBL provided the best calibration results; followed by SVMR, BRT and PLSR. We believe that the successful performance of MBL results from the combination of two important characteristics of this technique: (i) the storage of earlier situations in memory to reconcile them for solving the existing problem; and (ii) seeking and finding out k-nearest neighbors of each data to calibrate local models with these referenced neighbors. Actually, those statistical methods with the highest efficiency are the ones that have the best adaptability to the structure of the data to be analyzed.

4. Summary and Conclusions

This study focused on the performance of the new MBL method for soil spectroscopy analysis across the VNIR/SWIR spectral region for the prediction of soil texture, using soil samples taken from six brown coal mining dumpsites of the Czech Republic. To validate the results, a comparison with three other commonly-used methods (PLSR, SVMR and BRT) was made. To the best of our knowledge, this is the first time that MBL has been used for soil texture.

The results revealed that in the full spectral domain, MBL provided better predictions (lower $RMSEP_{cv}$) for all tested soil properties than the SVMR. The other two methods, PLSR and BRT, although significant, still have poorer performance than the MBL.

Our results (using PLSR, SVMR and BRT) were usually in line with those of other studies using the same methods, even though they were conducted at different scales and in other geographic regions.

Considering the high spatial variability and the expensive and time-consuming measurements of soil properties, VNIR/SWIR reflectance spectroscopy coupled with MBL can offer a rapid monitoring test for screening conditions, providing key increments in effectiveness and cost savings compared to traditional soil analytical techniques. It increases the model accuracy, reduces the number of samples to be analyzed for precision management applications in the field and can be applied as supplementary information in combination with spatial statistical methods to monitor soil conditions. Based on the very promising results of the MBL method's performance, implementation of further studies with other soil datasets over different geographic scales is highly recommended in order to check the MBL robustness and stability.

Acknowledgments: The authors acknowledge the scientific suggestions of Eyal Ben-Dor for improving the quality of the manuscript. They also like to thank the assistance of Aleš Klement for the measurement of spectra and soil texture. Moreover, the authors are grateful to Christopher Ash for English editing. The mining company Severoceske Doly is also appreciated for enabling this research.

Author Contributions: Asa Gholizadeh conceived of the experiment, analyzed the results and is the main author of the article. Mohammadmehdi Saberioon contributed to the writing of the article in his area of expertise and also in analyzing the results and their interpretation. Radim Vašát helped with modeling algorithm coding. Luboš Borůvka was the project supervisor and participated in all stages of the work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Viscarra Rossel, R.A.; McBratney, A.B. Soil chemical analytical accuracy and costs: Implications from precision agriculture. *Aust. J. Exp. Agric.* **1998**, *38*, 765–775. [CrossRef]
2. Ji, W.; Viscarra Rossel, R.A.; Shi, Z. Improved estimates of organic carbon using proximally sensed vis–NIR spectra corrected by piecewise direct standardization. *Eur. J Soil Sci.* **2015**, *66*, 670–678. [CrossRef]
3. 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]
4. Cécillon, L.; Barthes, B.G.; Gomez, C.; Ertlen, D.; Genot, V.; Hedde, M.; Stevens, A.; Brun, J.J. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). *Eur. J. Soil Sci.* **2009**, *60*, 770–784. [CrossRef]
5. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.M.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends Analyt. Chem.* **2010**, *29*, 1073–1081. [CrossRef]
6. Gholizadeh, A.; Amin, M.S.M.; Borůvka, L.; Saberioon, M.M. Models for estimating the physical properties of paddy soil using visible and near infrared reflectance spectroscopy. *J. Appl. Spectrosc.* **2014**, *81*, 534–540. [CrossRef]
7. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* **2010**, *107*, 163–215.
8. Araújo, S.R.; Wetterlind, J.; Demattê, 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]
9. 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]
10. Gholizadeh, A.; Borůvka, L.; Saberioon, M.M.; Vašát, 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]
11. 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]
12. Waiser, T.H.; Morgan, C.L.S.; Brown, D.J.; Hallmark, C.T. *In situ* characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. *Soil Sci. Soc. Am. J.* **2007**, *71*, 389–396. [CrossRef]
13. 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. *Biosys. Eng.* **2008**, *99*, 35–46. [CrossRef]
14. Gomez, C.; Lagacherie, P.; Coulouma, G. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma* **2008**, *148*, 141–148. [CrossRef]
15. Viscarra Rossel, R.A.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [CrossRef]
16. 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-Verlag: Heidelberg, Germany, 1983; Volume 973, pp. 286–293.
17. Moros, J.; De Vallejuelo, S.F.O.; Gredilla, A.; De Diego, A.; Madariaga, J.M.; Garrigues, S.; De La Guardia, M. Use of reflectance infrared spectroscopy for monitoring the metal content of the estuarine sediments of the Nerbioi-Ibaizabal River (Metropolitan Bilbao, Bay of Biscay, Basque Country). *Environ. Sci. Technol.* **2009**, *43*, 9314–9320. [CrossRef] [PubMed]
18. 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]
19. Saiano, F.; Oddo, G.; Scalenghe, R.; La Mantia, T.; Ajmone-Marsan, F. DRIFTS sensor: Soil carbon validation at large scale (Pantelleria, Italy). *Sensors* **2013**, *13*, 5603–5613. [CrossRef] [PubMed]

20. Dalal, R.C.; Henry, R.J. Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. *Soil Sci. Soc. Am. J.* **1986**, *50*, 120–123. [CrossRef]
21. Pirie, A.; Singh, B.; Islam, K. Ultra-violet, visible, near-infrared, and mid infrared diffuse reflectance spectroscopic techniques to predict several soil properties. *Aus. J. Soil Res.* **2005**, *43*, 713–721. [CrossRef]
22. Daniel, K.W.; Tripathi, N.K.; Honda, K. Artificial neural network analysis of laboratory and *in situ* spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). *Aus. J. Soil Res.* **2003**, *41*, 47–59. [CrossRef]
23. 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]
24. Viscarra Rossel, R.A. Robust modelling of soil diffuse reflectance spectra by bagging-partial least squares regression. *J. Near Infrared Spec.* **2007**, *15*, 39–47. [CrossRef]
25. Stevens, A.; Van Wesemael, B.; Bartholomeus, H.; Rosillon, D.; Tychon, B.; Ben-Dor, E. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma* **2008**, *144*, 395–404. [CrossRef]
26. 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]
27. Kovačević, M.; Bajat, B.; Trivic, B.; Pavlovic, R. Geological units classification of multispectral images by using support vector machines. In *International Conference on Intelligent Networking and Collaborative Systems*; Badr, Y.K., Caballe, S., Xhafa, F., Abraham, A., Gros, B., Eds.; IEEE: New York, NY, USA, 2009; pp. 267–272.
28. Gholizadeh, A.; Borůvka, L.; Vašát, R.; Saberioon, M.M.; Klement, A.; Kratina, J.; Tejnecký, V.; Drábek, 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**, *10*, e0117457. [CrossRef] [PubMed]
29. Ramirez-Lopez, L.; Behrens, T.; Schmidt, K.; Stevens, A.; Demattê, 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]
30. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
31. Kang, P.; Cho, S. Locally linear reconstruction for instance-based learning. *Pattern Recognit.* **2008**, *41*, 3507–3518. [CrossRef]
32. Morgan, R.P.C. *Soil Erosion and Conservation*, 3rd ed.; Blackwell: Malden, MA, USA, 2005.
33. Kosmas, C.; Kirby, M.; Geeson, N. *The Medalus Project Mediterranean Desertification and Land Use. Manual on Key Indicators of Desertification and Mapping Environmentally Sensitive Areas to Desertification*; EUR 18882; European Commission, Energy, Environment and Sustainable Development: Brussels, Belgium, 1999.
34. Hewson, R.D.; Cudahy, T.J.; Jones, M.; Thomas, M. Investigations into soil composition and texture using infrared spectroscopy (2–14 μm). *Appl. Environ. Soil Sci.* **2012**, *2012*, 1–12. [CrossRef]
35. Cozzolino, D.; Morón, A. The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics. *J. Agric. Sci.* **2003**, *140*, 65–71. [CrossRef]
36. Sørensen, L.K.; Dalsgaard, S. Determination of clay and other soil properties by near infrared spectroscopy. *Soil Sci. Soc. Am. J.* **2005**, *69*, 159–167. [CrossRef]
37. Mouazen, A.M.; Karoui, R.; De Baerdemaeker, J.; Ramon, H. Characterization of soil water content using measured visible and near infrared spectra. *Soil Sci. Soc. Am. J.* **2006**, *70*, 1295–1302. [CrossRef]
38. 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]
39. 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]
40. Wetterlind, J.; Stenberg, B. Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. *Eur. J. Soil Sci.* **2010**, *61*, 823–843. [CrossRef]
41. Gholizadeh, A.; Borůvka, L.; Vašát, 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]
42. IUSS Working Group WRB. World Reference Base for Soil Resources 2014. *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*; World Soil Resources Reports No. 106. FAO: Rome, Italy. Available online: http://www.fao.org/3/a-i3794e.pdf (accessed on 1 October 2015).
43. 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]

44. Gee, G.W.; Bauder, J.W. Particle-size analysis. In *Methods of Soil Analysis, Part 1*; Klute, A., Ed.; ASA and SSSA: Madison, WI, USA, 1986; pp. 383–411.
45. Jensen, J.R. *Remote Sensing of the Environment: An Earth Resource Perspective*; Prentice Hall: Upper Saddle River, NJ, USA, 2000.
46. Mouazen, A.M.; De Baerdemaeker, J.; Ramon, H. Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil Tillage Res.* **2005**, *80*, 171–183. [CrossRef]
47. Workman, J.J., Jr. Review of process and non-invasive near-infrared and infrared spectroscopy: 1993–1999. *Appl. Spectrosc. Rev.* **1999**, *34*, 1–89. [CrossRef]
48. Murray, I. Aspects of interpretation of NIR spectra. In *Analytical Application of Spectroscopy*; Creaser, C.S., Davies, A.M.C., Eds.; Royal Society of Chemistry: London, UK, 1988; pp. 9–21.
49. Mark, H.L.; Tunnell, D. Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Anal. Chem.* **1985**, *57*, 1449–1456. [CrossRef]
50. 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]
51. Cozzolino, D.; Morón, A. Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions. *Soil Till. Res.* **2006**, *85*, 78–85. [CrossRef]
52. 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]
53. 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.
54. 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]
55. 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]
56. Vohland, M.; Besold, J.; Hill, J.; Fruend, 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]
57. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
58. Martens, H.; Næs, T. *Multivariate Calibration*; John Wiley and Sons: New York, NY, USA, 1989.
59. 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]
60. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, NY, USA, 1995.
61. Karatzoglou, A.; Smola, A.; Hornik, K. Kernlab: Kernel-Based Machine Learning Lab. Available online: http://cran.r-project.org/web/packages/kernlab/index.html (accessed on 30 September 2015).
62. 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.
63. Vapnik, V. *Statistical Learning Theory*; Wiley-Interscience: New York, NY, USA, 1998.
64. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and regression trees. In *The Wadsworth Statistics/Probability Series*; Wadsworth International Group: Belmont, CA, USA, 1984; p. 358.
65. Steinberg, D.; Colla, P. *CART: Tree-Structured Non-Parametric Data Analysis*; Salford Systems: San Diego, CA, USA, 1997; p. 342.
66. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
67. 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]
68. Friedman, J.H.; Meulman, J.J. Multiple additive regression trees with application in epidemiology. *Stat. Med.* **2003**, *22*, 1365–1381. [CrossRef] [PubMed]
69. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **2000**, *28*, 337–374. [CrossRef]
70. 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.

71. Daelemans, W.; van den Bosch, A. *Memory-Based Language Processing*; Cambridge University Press: Cambridge, UK, 2005.
72. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice Hall, Pearson Education Inc.: Upper Saddle River, NJ, USA, 2003; p. 733.
73. Williams, P. *Near-Infrared Technology-Getting the Best out of Light*; PDK Projects: Nanaimo, BC, Canada, 2003.
74. Sherman, D.M.; Waite, T.D. Electronic spectra of Fe^{3+} oxides and oxyhydroxides in the near infrared to ultraviolet. *Am. Mineral.* **1985**, *70*, 1262–1269.
75. Ji, J.F.; Balsam, W.; Chen, J.; Liu, L.W. Rapid and quantitative measurement of hematite and goethite in the Chinese loess-paleosol sequence by diffuse reflectance spectroscopy. *Clay Clay Min.* **2002**, *50*, 208–216. [CrossRef]
76. Wu, Y.; Chen, J.; Wu, X.; Tian, Q.; Ji, J.; Qin, Z. Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. *Appl. Geochem.* **2005**, *20*, 1051–1059. [CrossRef]
77. Viscarra Rossel, R.A.; McGlynn, R.N.; McBratney, A.B. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *137*, 70–82. [CrossRef]
78. Kooistra, L.; Wanders, J.; Epema, G.F.; Leuven, R.; Wehrens, R.; Buydens, L.M.C. The potential of field spectroscopy for the assessment of sediment properties in river floodplains. *Anal. Chim. Acta* **2003**, *484*, 189–200. [CrossRef]
79. Bishop, J.L.; Lane, M.D.; Dyar, M.D.; Brown, A.J. Reflectance and emission spectroscopy study of four groups of phyllosilicates: Smectites, kaolinite-serpentines, chlorites and micas. *Clay Clay Min.* **1994**, *43*, 35–54. [CrossRef]
80. Post, J.L.; Noble, P.N. The near-infrared combination band frequencies of dioctahedral smectites, micas, and illites. *Clay Clay Min.* **1993**, *41*, 639–644. [CrossRef]
81. Ben-Dor, E.; Inbar, Y.; Chen, Y. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sens. Environ.* **1997**, *61*, 1–15. [CrossRef]
82. Hunt, G.R.; Salisbury, J.W. Visible and near-infrared spectra of minerals and rocks. I. Silicate Minerals. *Mod. Geol.* **1970**, *4*, 283–300.
83. Awiti, A.O.; Walsh, M.G.; Shepherd, K.D.; Kinyamario, J. Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence. *Geoderma* **2008**, *143*, 73–84. [CrossRef]
84. Kuang, B.; Mouazen, A.M. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. *Eur. J. Soil Sci.* **2012**, *63*, 421–429. [CrossRef]
85. Eshel, G.; Levy, G.J.; Mingelgrin, U.; Singer, M.J. Critical evaluation of the use of laser diffraction for particle-size distribution analysis. *Soil Sci. Soc. Am. J.* **2004**, *68*, 736–743. [CrossRef]
86. 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]
87. Minasny, B.; McBratney, A.B. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemom. Intell. Lab. Syst.* **2008**, *94*, 72–79. [CrossRef]
88. 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]