THE INTERNATIONAL SOCIETY OF
PRECISION AGRICULTURE PRESENTS THE
13th INTERNATIONAL CONFERENCE ON
PRECISION AGRICULTURE
July 31-August 4, 2016 • St. Louis, Missouri USA

Memory Based Learning: A New Data Mining Approach to Model and Interpret Clay Diffuse Reflectance Spectra

Gholizadeh A., Saberioon M.M., Borůvka L.

A paper from the Proceedings of the
13th International Conference on Precision Agriculture
July 31 – August 4, 2016
St. Louis, Missouri, USA

Abstract. Successful estimation of spectrally active clay in soil with Visible and Near-Infrared (VNIR, 400-1200 nm) and Short-Wave-Infrared (SWIR, 1200-2500 nm) spectroscopy depends mostly on the selection of an appropriate data mining algorithm. The aims of this paper were: to compare different data mining algorithms including Partial Least Squares Regression (PLSR), which is the most common technique in soil spectroscopy, Support Vector Machine Regression (SVMR), Boosted Regression Trees (BRT), and Memory Based Learning (MBL) as a very new promising approach for estimating the content of clay, to explore whether these methods show differences regarding their ability to predict clay from VNIR/SWIR data and to evaluate the interpretability of the results. The dataset consisted of 264 samples from large brown coal mining dumpsites in the Czech Republic. Spectral readings were taken in the laboratory with a fibreoptic ASD FieldSpec III Pro FR spectroradiometer. Leave-one-out cross validation was applied 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 clay by MBL outperformed the accuracy of the other algorithms. It produced the largest R^2_{cv} and smallest $RMSEP_{cv}$ values, followed by SVMR. Actually, the main goal of this work was to develop a suitable MBL approach for soil spectroscopy, it showed that MBL is a very promising approach to deal with complex clay VNIR/SWIR datasets. A systematic comparison like the one presented here is important as the nature of the target function has a strong influence on the performance of the different algorithms.

Keywords.

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 (Viscarra Rossel and McBratney 1998). 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 to 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 (Ji et al. 2015), N, P, K contents, Cation Exchange Capacity (CEC) and pH (Viscarra Rossel et al. 2006b) and, to some extent, physical parameters including soil structure, bulk density, and texture (Cécillon et al. 2009; Bellon-Maurel et al. 2010; Gholizadeh et al. 2014). Actually, because analysis of the clay fraction depends on features of the mineral content, VNIR/SWIR spectra can be of value for predicting clay content (Stenberg et al. 2010; Araújo et al. 2014).

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 (Mouazen et al. 2010; Gholizadeh et al. 2013). Under in situ measurement conditions with non-mobile or mobile instrumentation, additional challenges linked with diverse soil moisture content, colour, dust, stones, and excessive residues and surface roughness all affect the accuracy of the measurement (Mouazen et al. 2007; Waiser at al. 2007). 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 (Maleki et al. 2008), better control of ambient conditions (Mouazen et al. 2007), and the appropriate selection of multivariate statistical analysis (Gomez et al. 2008; Viscarra Rossel and Behrens 2010).

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 (Araújo et al. 2014). Araújo et al. (2014) 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 (Wold et al. 1983; Moros et al. 2009; Song et al. 2012; Saiano et al. 2013). Other approaches have also been used, for example, Multiple Linear Regression (MLR) (Dalal and Henry 1986), Principle Component Regression (PCR) (Pirie et al. 2005), Artificial Neural Networks (ANN) (Daniel et al. 2003), Multivariate Adaptive Regression Splines (MARS) (Shepherd and Walsh 2002), PLSR with Bootstrap Aggregation (bagging-PLSR) (Viscarra Rossel 2007), and Penalized Signal Regression (PSR) (Stevens et al. 2008). Brown (2007) suggested the use of Boosted

The authors are solely responsible for the content of this paper, which is not a refereed publication.. Citation of this work should state that it is from the Proceedings of the 13th International Conference on Precision Agriculture. EXAMPLE: Lastname, A. B. & Coauthor, C. D. (2016). Title of paper. In Proceedings of the 13th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

Regression Trees (BRT), and Kovačević *et al*. (2009) and Gholizadeh *et al*. (2015a) 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 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 (Ramirez-Lopez et al. 2013). 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 (Mitchell 1997). 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 (Kang and Cho 2008), 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 (Morgan 2005). 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 (Morgan 2005). 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 (Kosmas et al. 1999). 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 (Hewson et al. 2012).

Successful predictions of Soil Organic Carbon (SOC) using spectroscopy have been reported (Cozzolino and Morón 2003; Sørensen and Dalsgaard 2005). Soil water content has also been predicted under both laboratory and in situ conditions (Mouazen et al. 2006). Clay content can be well estimated with VNIR (Ben-Dor and Banin 1995; Brown et al. 2006; Wetterlind and Stenberg 2010), but the use of VNIR/SWIR reflectance spectroscopy offers a lower precision for clay with particular chemometric algorithms.

As shown by Gholizadeh et al. (2013; 2015b), 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 MBL algorithm. The purpose was to provide interpretation of the results for the prediction of clay 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 (Fig. 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 to 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.

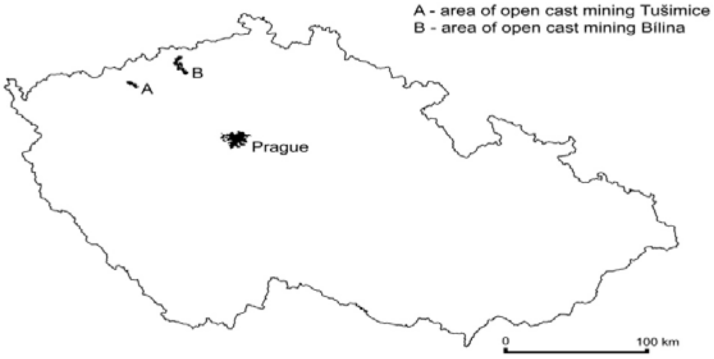

A - area of open cast mining Tušimice
B - area of open cast mining Bílina
A
B
Prague
0
100 km

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

2.2. Soil Sampling and Analysis

A total of 264 soil samples were collected. 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 in the depth of 0-30 cm (Song et al. 2012; Xie et al. 2012). This depth corresponds to the common depth of a ploughed soil layer, as these soils will be mostly used as arable land in future. The depth of the topsoil cover was also at least 30 cm. The original samples were air-dried, crushed and sieved (≤ 2 mm), and thoroughly mixed before analyzing. Clay content was determined by the sedimentation hydrometer method (Gee and Bauder 1986). Samples and standards were matrix matched and all analyses were carried out in triplicates.

2.3. Spectral Data Measurements

Spectral reflectance was deliberated in the 350-2500 nm wavelength range using a fiberoptic ASD FieldSpec III Pro FR spectroradiometer (ASD Inc., Denver, Colorado, 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 fiberoptic probe with 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. 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 (Mouazen et al. 2005). All spectral readings were measured in the centre 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 signal-to-noise ratio. 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 finger tip was also corrected using a reference spectrum through a 1 mm layer of white $BaSO_4$ panel standard (Xie et al. 2012; Workman 1999).

2.4. Spectra Preprocessing

Murray (1988) mentioned that removing outliers improves prediction accuracy; hence, the outliers were left out. Outliers were detected by using the principle of Mahalanobis distance (H) (Mark and Tunnell 1985; Shenk and Westerhaus 1991), applied on PCA-reduced data. In the present study, an H value of 3 (based on the Mahalanobis distance) was chosen for identification of outliers (Gomez et al. 2012). The detected spectral outliers were deleted from the calibration set.

In order to calibrate a model that provides accurate predictive performance about the clay content in each soil sample, the captured soil spectra, jointly with laboratory data of the parameter, were imported into R software (R Development Core Team, Vienna, Austria) to be processed. 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 (Duckworth 2004; Gholizadeh et al. 2015a; Gholizadeh et al. 2015b). 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 (Ren et al. 2009; Song et al. 2012) 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 clay reference data and to describe the relationship between reflectance spectra and estimated clay. A brief summary of each algorithm is presented in the following sections.

2.5.1. Partial Least Square Regression (PLSR)

The PLSR decreases the data, noise, and calculation time, with minor loss of the information contained in the original variables (Vasques et al. 2008) and its arithmetic can be referred to Wold et al. (1983). It is strongly related to PCR in that both use statistical rotations to defeat the problem of high dimensionality and multicollinearity (Brown et al. 2006; Vohland et al. 2011). They both compress the data before completing the regression. The difference is that PLSR algorithm combines the compression and regression steps and it selects successive orthogonal factors that maximize the covariance between predictor and response variables (Wold et al. 2001; Viscarra Rossel et al. 2006b; Viscarra Rossel and Behrens 2010; Vohland et al. 2011). By fitting a PLSR model, one expects to discover a few PLSR factors that clarify most of the variation in both predictors and responses (Martens and Næs 1989). As stated by Gholizadeh et al. (2013), Viscarra Rossel and Behrens (2010) and Bilgili et al. (2010), 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.

Variables X and Y are mean-centred 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. 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 (Viscarra Rossel and Behrens 2010). 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 (Bilgili et al. 2010; Xie et al. 2012).

2.5.2. Support Vector Machine Regression (SVMR)

The SVMR approach is a supervised, nonparametric and statistical learning method (Vapnik 1995). 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 (Kovačević et al. 2009). In this method, model complication is finite by the learning algorithm itself, which avoids over-fitting. Based on Vapnik (1995), 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 (Karatzoglou et al. 2015). Using this so called kernel-trick (Boser at al. 1992), 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 (Viscarra Rossel and Behrens 2010). 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 (Vohland et al. 2011). 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.

2.5.3. Boosted Regression Trees (BRT)

Based on Brown (2007), BRT have been suggested as an ideal data-mining or pattern-recognition tool for VNIR/SWIR spectroscopy of soil properties. Boosted Regression Trees (BRT) analysis basically performs a binary recursive partitioning of the dataset (Breiman et al. 1984; Steinberg and Colla 1997). At each terminal node, a predicted value is gained as the average of all 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 (Friedman 2001). 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 (Araújo et al. 2014).

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 (Friedman and Hastie 2000; Friedman and Meulman 2003).

2.5.4. Memory Based Learning (MBL)

MBL resembles the human reasoning process (An 2005; Ramirez-Lopez et al. 2013): 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 (Mitchell 1997). Based on Daelemans (2005), MBL is a family of learning algorithms that, in preference to performing clear and precise generalization, compares new problem cases with 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 (Russell and Norvig 2003). 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. It means that nonlinear relationships can be simply resolved.

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 once left out.

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:

$$R_{cv}^2 = \left(1 - \frac{\sum_{i=1}^{N}(y_i' - y_i)^2}{\sum_{i=1}^{N}(y_i' - \bar{y}_i)^2}\right) \qquad \textbf{(1)}$$

$$RMSEP_{cv} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i' - y_i)^2} \qquad \textbf{(2)}$$

Where:

y'_i – predicted value, y_i – observed value, $\bar{y}_i$– mean of y value and N − number of samples.

3. Results and Discussion

3.1. Descriptive Analysis of Clay

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 wide range of clay content, especially in Prunéřov (ranged from 6.1 to 60.7%). Ramirez-Lopez et al. (2013) 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.

Table 1. Descriptive statistics of clay content in the studied sample set according to location

	n	Min	Max	Mean	SD	CV (%)
Pokrok	103	7.5	53.3	36.7	8.7	23.6
Radovesice	40	18.1	52.9	41.9	7.8	18.5
Březno	25	28.9	61.4	39.9	5.9	14.9
Merkur	38	17.7	59.9	47.5	6.5	13.8
Prunéřov	48	6.1	60.7	40.5	12.6	31.1
Tumerity	10	31.6	68.4	50.7	11.5	22.7

3.2. Soil Spectral Properties

A visual assessment of the spectra permitted to remove parts of spectra which are known as the noisiest parts at the edges of the spectrum, and the final spectral library considered the spectral range from 400 to 2450 nm. Sets of spectra were defined qualitatively by identifying the positive and negative peaks (Fig. 2), which appear at particular wavelengths (Viscarra Rossel et al. 2006b).

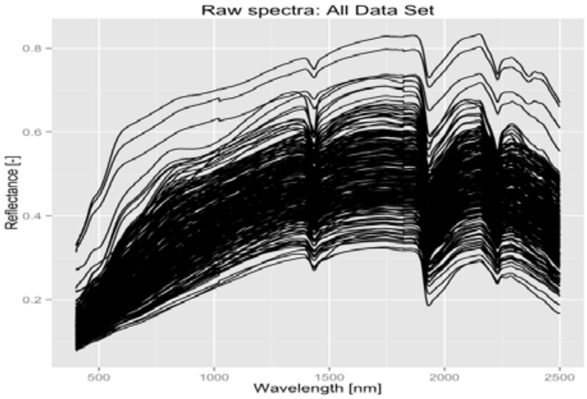
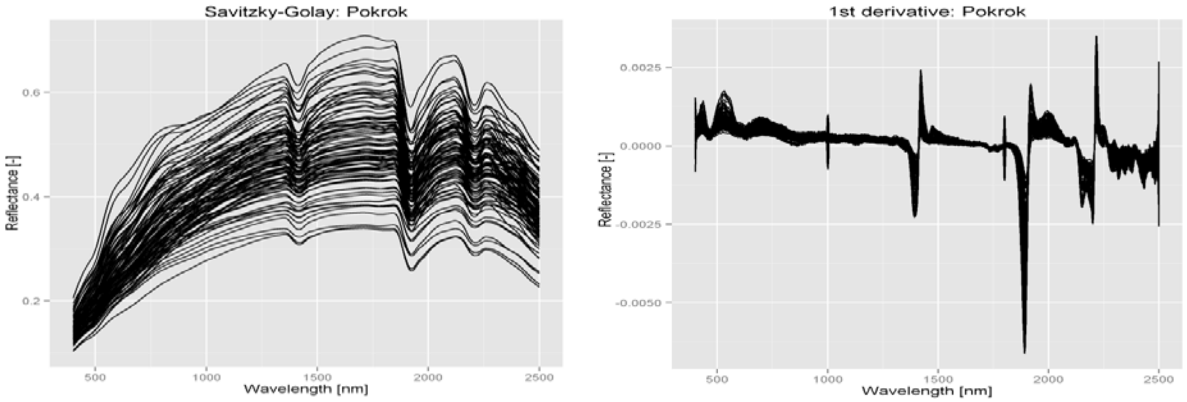


Fig. 2. Representative VNIR/SWIR spectra of soil samples

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 (Sherman and Waite 1985; Ji et al. 2002; Wu et al. 2005; Viscarra Rossel et al. 2006a). Based on Sherman and Waite (1985), 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) (Viscarra Rossel and Behrens 2010). In the NIR region, O-H bonds in clay minerals would have a general influence on the reflectance spectra (Kooistra et al. 2003; Ren et al. 2009; Song et al. 2012). 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 (Sherman and Waite 1985). 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 (Araújo et al. 2014; Bishop et al. 1994). 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 (Viscarra Rossel et al. 2006a). The traits around 2000-2500 nm are linked to the characteristics of SOM and clay minerals (Kooistra et al. 2003; Ren et al. 2009). Based on Viscarra Rossel et al. (2006a), the prominent peaks near 2200 nm, 2300 nm and 2400 nm may be attributed to 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 (Viscarra Rossel and Behrens 2010). 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 (Post and Noble 1993; Ben-Dor et al. 1997). It should be noted that band positions and wavelength peaks may vary with composition (Hunt and Salisbury 1970).

3.3. Spectra Preprocessing and Model Calibration

In order to create a robust prediction model and to discover the impression of 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 (Awiti et al. 2008; Kuang and Mouazen 2012; Gholizadeh et al. 2015b). Smoothed only spectra, by Savitzky-Golay filter, as well as smoothed and preprocessed spectra, using Savitzky-Golay plus first derivative, of all selected soil samples were illustrated below (Fig. 3). The comparison between Figs. 2 and 3 revealed that the main difference between the spectra is a baseline shift. It also showed that the Savitzky-Golay and first derivatives preprocessing techniques can remove additive baseline effects and thus are more useful for the prediction of soil texture than the original spectra. The first derivative spectra generally amplify the absorption features indicative of the contents of the soil materials, and also reduce variation among samples (Martens and Næs 1989).



(a) (b)

Fig. 3. Smoothed only (a) and smoothed and 1st derivative preprocessed (b) soil spectra

The capability of spectral reflectance spectra to predict clay content using PLSR, SVMR, BRT, and MBL techniques was studied. The comparison of prediction accuracy and model performance from the different algorithms were presented in this part (Fig. 4) which showed the validation results of predicted and measured values of clay content using PLSR, SVMR, BRT and MBL algorithms.

In the multivariate calibration, based on R^2_{cv} and $RMSEP_{cv}$, which have been reported as standard methods for validation of the prediction models, clay fraction gained consistent estimates. Compared with PLSR; SVMR, BRT, and MBL gave smaller $RMSEP_{cv}$ and larger R^2_{cv}.

The MBL technique gave better prediction of clay compared to other methods, giving the smallest error.

BRT showed lower $RMSEP_{cv}$ than PLSR, but PLSR still showed relatively good prediction of clay content. The results were comparable to R^2 values introduced in the literature by other authors (Ben-Dor et al. 1997; Chang et al. 2001; Minasny and McBratney 2008; Bilgili et al. 2010) who utilized wavelength UV, VIS, NIR, and MIR wavebands. Regarding BRT, Brown et al. (2006) also found that this method can outperform PLSR and recognized this to its capacity to contain interactions and nonlinear relationships. Authors also compared the prediction with the SVMR, which was found to give better prediction than PLSR and even BRT. However, SVMR was less accurate when compared to MBL. SVMR had a small $RMSEP_{cv}$ for predicting clay, 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 (Kang and Cho 2008).

For PLSR, SVMR, and BRT, these findings coincided with the results of some other studies. Viscarra Rossel and Behrens (2010) 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 ($R^2_{cv} = 0.84$, $RMSEP_{cv} = 7.63$). Araújo et al. (2014) 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. Their study agreed with Brown (2007), 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 (Jalabert et al. 2010). On the other hand, Vasques et al. (2008), 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. Ramirez-Lopez et al. (2013) introduced 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.

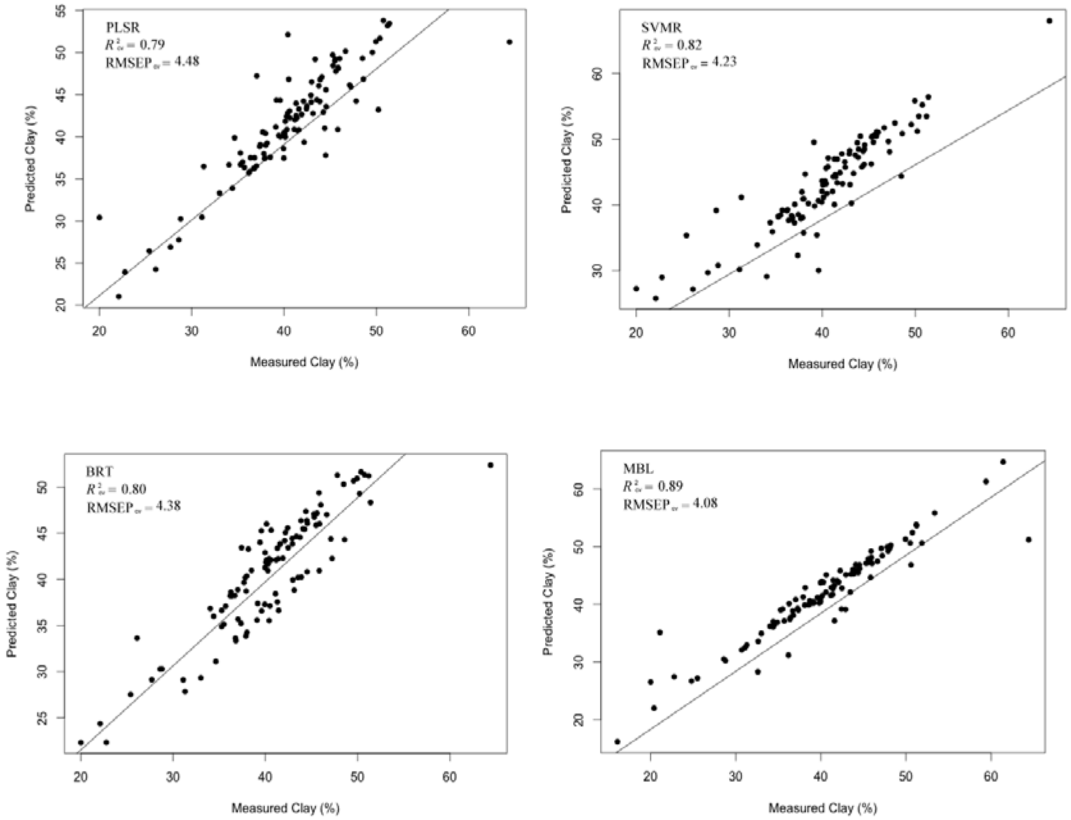


Fig. 4. Scatterplots of measured versus predicted clay obtained by PLSR, SVMR, BRT, and MBL

It could be concluded from this study that the prediction accuracy of data mining techniques, MBL particularly, will be high in fine-textured fields, which reflects the influence of the direct spectral responses of clay, especially in the NIR range. Therefore, the Merkur dumpsite soil, which contained more clay than other brown coal mining dumpsites, can be predicted more accurately with higher R^2_{cv} and lower $RMSEP_{cv}$. These results supported those studies that found SVMR as a very promising method for the determination of clay content (Viscarra Rossel and Behrens 2010; Araújo et al. 2014). 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. 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 which have the best adaptability to the structure of the data to be analyzed.

4. Conclusion or Summary

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 clay content, using soil samples taken from six brown coal mining dumpsites of the Czech Republic. To validate the results, comparison with three other commonly used methods (PLSR, SVMR and BRT) was made. The results revealed that in the full spectral domain, MBL provided better predictions (lower $RMSEP_{cv}$) than the SVMR. The other two methods, PLSR and BRT, although significant, they still lay back from the MBL performance. 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 condition, providing key increments in effectiveness and cost-saving compared with traditional soil analytical techniques. It increases the model accuracy, reduces the number of samples to be analyzed for precision management applications in fields, and can be applied as supplementary information in combination with spatial statistical methods to monitor soil conditions. Based on 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.

Acknowledgements

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

References

An, A. (2005). Classification methods. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 144-149). New York: Idea Group Inc.

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B. (2014). Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *European Journal of Soil Science*, 65(5), 718-729.

Awiti, A.O., Walsh, M.G., Shepherd, K.D., Kinyamario, J. (2008). Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence. *Geoderma,* 143, 73-84.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends in Analytical Chemistry,* 29(9), 1073-1081.

Ben-Dor, E., & Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal*, 59**,** 364-372.

Ben-Dor, E., Inbar, Y., Chen, Y. (1997). 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 Sensing of Environment*, 61(1), 1-15.

Bilgili, A.V., Van Es, H.M., Akbas, F., Durak, A., Hively, W.D. (2010). Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *Journal of Arid Environments*, 74, 229-238.

Bishop, J.L., Lane, M.D., Dyar, M.D., Brown, A.J. (1994). Reflectance and emission spectroscopy study of four groups of phyllosilicates: smectites, kaolinite-serpentines, chlorites and micas. *Clays and Clay Minerals*, 43, 35-54.

Boser, B.E., Guyon, I.M., Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *5th Annual ACM Workshop on COLT* (pp. 144-152). Pittsburgh: ACM Press.

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and regression trees. The Wadsworth*

Statistics/Probability Series. Belmont, CA: Wadsworth International Group.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma,* 132, 273-290.

Brown, D.J. (2007). Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma*, 140, 444-453.

Cécillon, L., Barthes, B.G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., Brun, J.J. (2009). Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). *European Journal of Soil Science,* 60(5), 770-784.

Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh Jr, C.R. (2001). Near infrared reflectance spectroscopy-principal components regression analysis of soil properties. *Soil Science Society of America Journal*, 65, 480-490.

Cozzolino, D., & Morón, A. (2003). The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics. *Journal of Agricultural Science,* 140, 65-71.

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

Dalal, R.C., & Henry, R.J. (1986). Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. *Soil Science Society of America Journal,* 50, 120-123.

Daniel, K.W., Tripathi, N.K., Honda, K. (2003). Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). *Australian Journal of Soil Research*, 41, 47-59.

Duckworth, J. (2004). Mathematical data preprocessing. In C.A. Roberts, J. Workman Jr., J.B. III Reeves (Eds.), *Near-Infrared Spectroscopy in Agriculture* (pp. 115-132). Madison, WI: ASA-CSSA-SSSA.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189-1232.

Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: a statistical view of boosting. *Ann. Stat.* **2000**, 28(2), 337-374.

Friedman, J.H., & Meulman, J.J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine,* 22(9), 1365-1381.

Gee, G.W., & Bauder, J.W. (1986). Particle-size analysis. In A. Klute (Ed.), *Methods of Soil Analysis, Part 1* (pp. 383-411). Madison, WI: ASA and SSSA.

Gholizadeh, A., Borůvka, L., Saberioon, M.M., Vašát, R. (2013). 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. *Applied Spectroscopy,* 67, 1349-1362.

Gholizadeh, A., Amin, M.S.M., Borůvka, L., Saberioon, M.M. (2014). Models for estimating the physical properties of paddy soil using visible and near infrared reflectance spectroscopy. *Journal of Applied Spectroscopy,* 81(3), 534-540.

Gholizadeh, A., Borůvka, L., Vašát, R., Saberioon, M.M., Klement, A., Kratina, J., Tejnecký, V., Drábek, O. (2015a). Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: A case study. *PLoS One,* 10(2), e0117457.

Gholizadeh, A., Borůvka, L., Vašát, R., Saberioon, M.M. (2015b). Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil and Water Research,* 10(4), 218-227.

Gomez, C., Lagacherie, P., Coulouma, G. (2008). Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma*, 148(2), 141-148.

Gomez, C., Lagacherie, P., Coulouma, G. (2012). Regional predictions of eight common soil properties and their spatial structures from hyperspectral Vis–NIR data. *Geoderma*, 189-190, 176-185.

Hewson, R.D., Cudahy, T.J., Jones, M., Thomas, M. (2012). Investigations into soil composition and texture using infrared spectroscopy (2-14 µm). *Applied and Environmental Soil Science*, 1-12.

Hunt, G.R., & Salisbury, J.W. (1970). Visible and near-infrared spectra of minerals and rocks. I. Silicate Minerals. *Modern Geol*ogy, 4, 283-300.

Jalabert, S.S.M., Martin, M.P., Renaud, J.P., Boulonne, L., Jolivet, C., Montanarella, L. (2010). Estimating forest soil bulk density using boosted regression modeling. *Soil Use and Management*, 26, 516-528.

Ji, J.F., Balsam, W., Chen, J., Liu, L.W. (2002). Rapid and quantitative measurement of hematite and goethite in the Chinese loess-paleosol sequence by diffuse reflectance spectroscopy. *Clays and Clay Minerals.* 50(2), 208-216.

Ji, W., Viscarra Rossel, R.A., Shi, Z. (2015). Improved estimates of organic carbon using proximally sensed vis–NIR spectra corrected by piecewise direct standardization. *European Journal of Soil Science*, 66(4), 670-678.

Kang, P., & Cho, S. (2008). Locally linear reconstruction for instance-based learning. *Pattern Recognition*, 41, 3507-3518.

Karatzoglou, A., Smola, A., Hornik, K. (2015). Kernlab: Kernel-based machine learning lab. Available online: http://cran.r-project.org/web/packages/kernlab/index.html. Accessed 30 September 2015.

Kooistra, L., Wanders, J., Epema, G.F., Leuven, R., Wehrens, R., Buydens, L.M.C. (2003). The potential of field spectroscopy for the assessment of sediment properties in river floodplains. *Analytica Chimica Acta*, 484(2), 189-200.

Kosmas, C., Kirby, M., Geeson, N. (1999). *Manual on: Key indicators of desertification and mapping environmentally sensitive areas to desertification.* EUR 18882, European Commission, Energy, Environment and Sustainable Development.

Kovačević, M., Bajat, B., Trivic, B., Pavlovic, R. (2009). Geological units classification of multispectral images by using support vector machines. In Y.K. Badr, S. Caballe, F. Xhafa, A. Abraham, B. Gros (Eds.), *International Conference on Intelligent Networking and Collaborative Systems* (pp. 267-272). New York: IEEE.

Kuang, B., & Mouazen, A.M. (2012). Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. *European Journal of Soil Science*, 63, 421-429.

Maleki, M.R., Mouazen, A.M., De Keterlaere, B., Ramon, H., De Baerdemaeker, J. (2008). On-the-go variable-rate phosphorus fertilisation based on a visible and near infrared soil sensor. *Biosystem Engineering*, 99(1), 35-46.

Mark, H.L., & Tunnell, D. (1985). Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Analytical Chemistry*, 57, 1449-1456.

Martens, H., & Næs, T. (1989). *Multivariate calibration*. New York: John Wiley and Sons.

Minasny, B., & McBratney, A.B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 94, 72-79.

Mitchell, T.M. (1997). *Machine learning*. New York: McGraw-Hill.

Morgan, R.P.C. (2005). *Soil erosion and conservation*. Malden, Mass: Blackwell.

Moros, J., De Vallejuelo, S.F.O., Gredilla, A., De Diego, A., Madariaga, J.M., Garrigues, S., De La Guardia, M. (2009). 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). *Environmental Science & Technology*, 43, 9314-9320.

Mouazen, A.M., De Baerdemaeker, J., Ramon, H. (2005). Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil & Tillage Research,* 80, 171-183.

Mouazen, A.M., Karoui, R., De Baerdemaeker, J., Ramon, H. (2006). Characterization of soil water content using measured visible and near infrared spectra. *Soil Science Society of America Journal,* 70, 1295-1302.

Mouazen, A.M., Maleki, M.R., De Baerdemaeker, J., Ramon, H. (2007). On-line measurement of some selected soil properties using a VIS-NIR sensor. *Soil & Tillage Research,* 93, 13-27.

Mouazen, A.M., Kuang, B., De Baerdemaeker, J., Ramon, H. (2010). 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,* 158, 23-31.

Murray, I. (1988). Aspects of interpretation of NIR spectra. In C.S. Creaser, & A.M.C. Davies (Eds.), *Analytical application of spectroscopy* (pp. 9-21). London, UK: Royal Society of Chemistry.

Pirie, A., Singh, B., Islam, K. (2005). Ultra-violet, visible, near-infrared, and mid infrared diffuse reflectance spectroscopic techniques to predict several soil properties. *Australian Journal of Soil Research,* 43, 713-721.

Post, J.L., & Noble, P.N. (1993). The near-infrared combination band frequencies of dioctahedral smectites, micas, and illites. *Clays and Clay Minerals,* 41, 639-644.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets. *Geoderma*, 195-196, 268-279.

Ren, H.Y., Zhuang, D.F., Singh, A.N., Pan, J.J., Qid, D.S., Shi, R.H. (2009). Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: A case study. *Pedosphere*, 19, 719-726.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach.* New Jersey: Prentice Hall, Pearson Education Inc.

Saiano, F., Oddo, G., Scalenghe, R., La Mantia, T., Ajmone-Marsan, F. (2013). DRIFTS sensor: soil carbon validation at large scale (Pantelleria, Italy). *Sensors, 13,* 5603-5613.

Shenk, J.S., & Westerhaus, M.O. (1991). Population definition, sample selection, and calibration procedure for near infrared reflectance spectroscopy. *Crop Science,* 31, 469-474.

Shepherd, K.D., & Walsh, M.G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal,* 66, 988-998.

Sherman D.M., & Waite T.D. (1985). Electronic spectra of Fe^{3+} oxides and oxyhydroxides in the near infrared to ultraviolet. *American Mineralogist,* 70, 1262-1269.

Song, Y., Li, F., Yang, Z., Ayoko, G.A., Frost, R.L., Ji, J. (2012). Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. *Applied Clay Science,* 64, 75-83.

Sørensen, L.K., & Dalsgaard, S. (2005). Determination of clay and other soil properties by near infrared spectroscopy. *Soil Science Society of America Journal,* 69, 159-167.

Steinberg, D., & Colla, P. (1997). *CART: tree-structured non-parametric data analysis*. San Diego, CA: Salford Systems.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy,* 107:163-215.

Stevens, A., Van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., Ben-Dor, E. (2008). Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma*, 144, 395-404.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vasques, G.M., Grunwald, S., Sickman, J.O. (2008). Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma*, 146, 14-25.

Viscarra Rossel, R.A. (2007). Robust modelling of soil diffuse reflectance spectra by bagging-partial least squares regression. *Journal of Near Infrared Spectroscopy*, 15, 39-47.

Viscarra Rossel, R.A., & McBratney, A.B. (1998). Soil chemical analytical accuracy and costs: implications from precision agriculture. *Australian Journal of Experimental Agriculture,* 38, 765-775.

Viscarra Rossel, R.A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma***,** 158, 46-54.

Viscarra Rossel, R.A., McGlynn, R.N., McBratney, A.B. (2006a). Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma*, 137, 70-82.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O. (2006b). Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 131, 59-75.

Vohland, M., Besold, J., Hill, J., Fruend, H.C. (2011). Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma*, 166**,** 198-205.

Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T. (2007). In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. *Soil Science Society of America Journal*, 71(2), 389-396.

Wetterlind, J., & Stenberg, B. (2010). Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science,* 61, 823-843.

Wold, S., Martens, H., Wold, H. (1983). The multivariate calibration method in chemistry solved by the PLS method. In A. Ruhe, & B. Kagstrom (Eds.), *Proceeding of Conf Matrix Pencils, Lecture Notes in Mathematics* (pp. 286-293). Heidelberg, Germany: Springer-Verlag.

Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 58, 109-130.

Workman, JR. (1999). Review of process and non-invasive near-infrared and infrared spectroscopy: 1993-1999. *Applied Spectroscopy Reviews,* 34(1&2), 1-89.

Wu, Y., Chen, J., Wu, X., Tian, Q., Ji, J., Qin, Z. (2005). Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. *Applied Geochemistry,* 20, 1051-1059.

Xie, X., Pan, X.Z., Sun, B. (2012). Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a Copper smelter. *Pedosphere,* 22, 351-366.