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

Asa Gholizadeh
Mohammadmehdi Saberioon
Luboš Borůvka

**Department of Soil Science and Soil Protection, Faculty of Agrobilogy, Food and Natural Resources, Czech University of Life Sciences Prague**

# Introduction

- Evaluation and estimation of soil texture

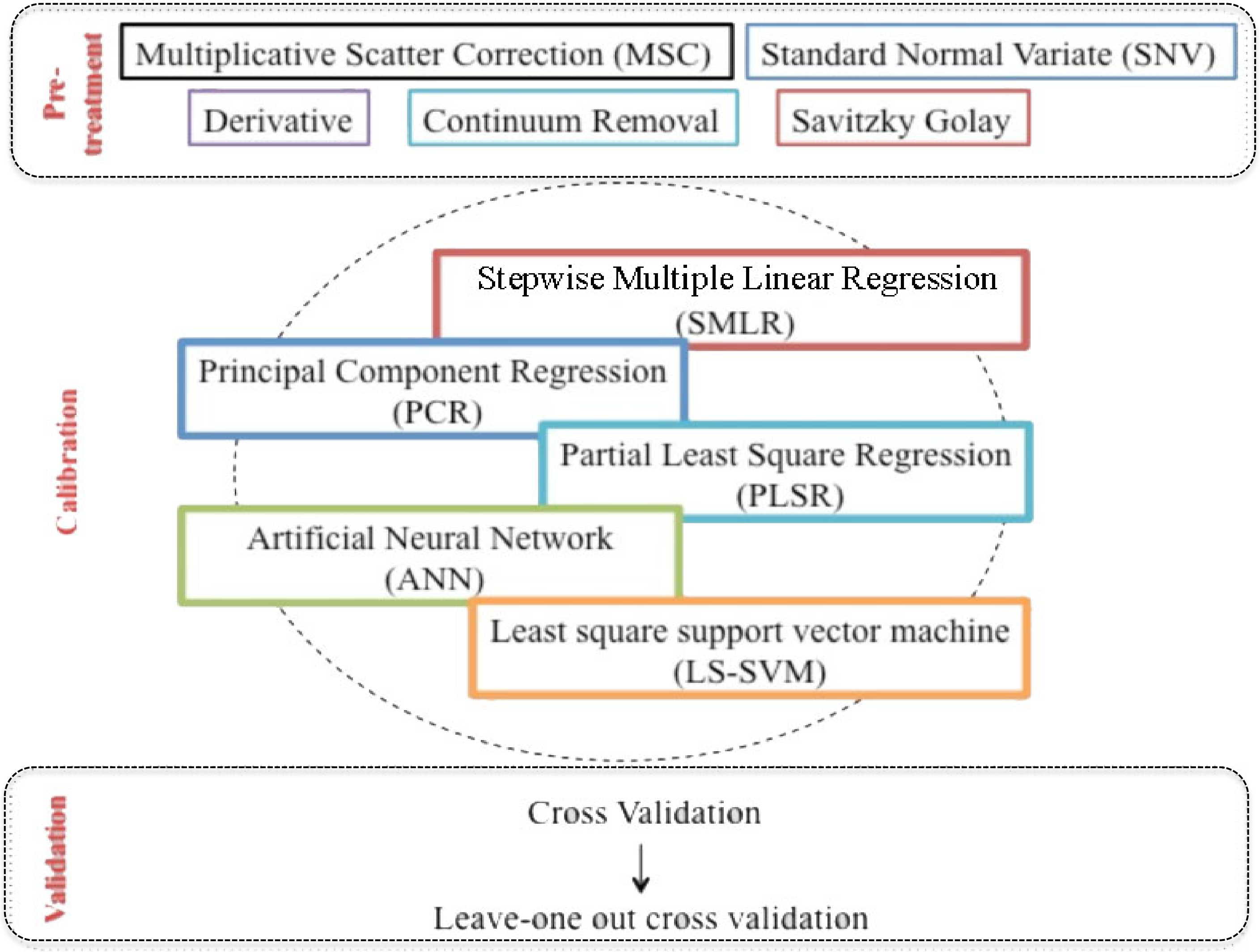
- Some parameters affect VNIR/SWIR spectroscopy accuracy

- Solutions to overcome the abovementioned difficulties

- Attention toward 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

- Choosing the most robust calibration algorithm

# Common Preprocessing, Calibration and Validation Methods



Pre-treatment

Multiplicative Scatter Correction (MSC)

Standard Normal Variate (SNV)

Derivative

Continuum Removal

Savitzky Golay

Calibration

Stepwise Multiple Linear Regression (SMLR)

Principal Component Regression (PCR)

Partial Least Square Regression (PLSR)

Artificial Neural Network (ANN)

Least square support vector machine (LS-SVM)

Validation

Cross Validation

↓

Leave-one out cross validation

# Methodology

## Study Area

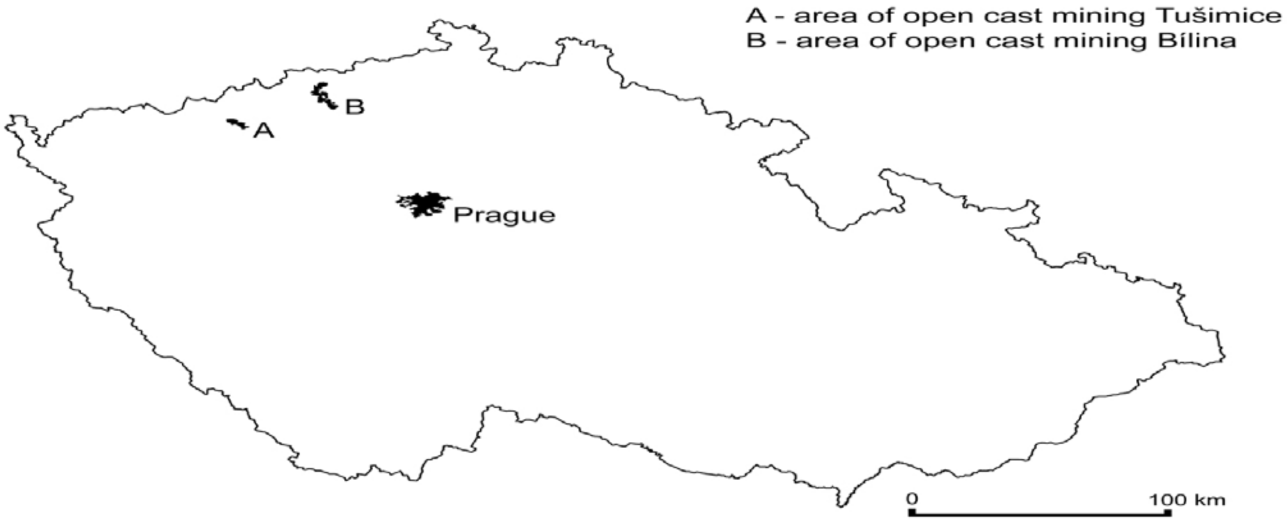
- 6 dumpsites in mines Bílina and Tušimice, the Czech Republic: Pokrok, Radovesice, Březno, Merkur, Tumerity and Prunéřov.

- All formed by clay.

- The range soil pH for the whole area was 5.3-8.5.

- The SOM content range was 0.6-3.8%.

- 37.30% clay, 33.10% sand and 29.60% silt.


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

# Soil Sampling and Analysis

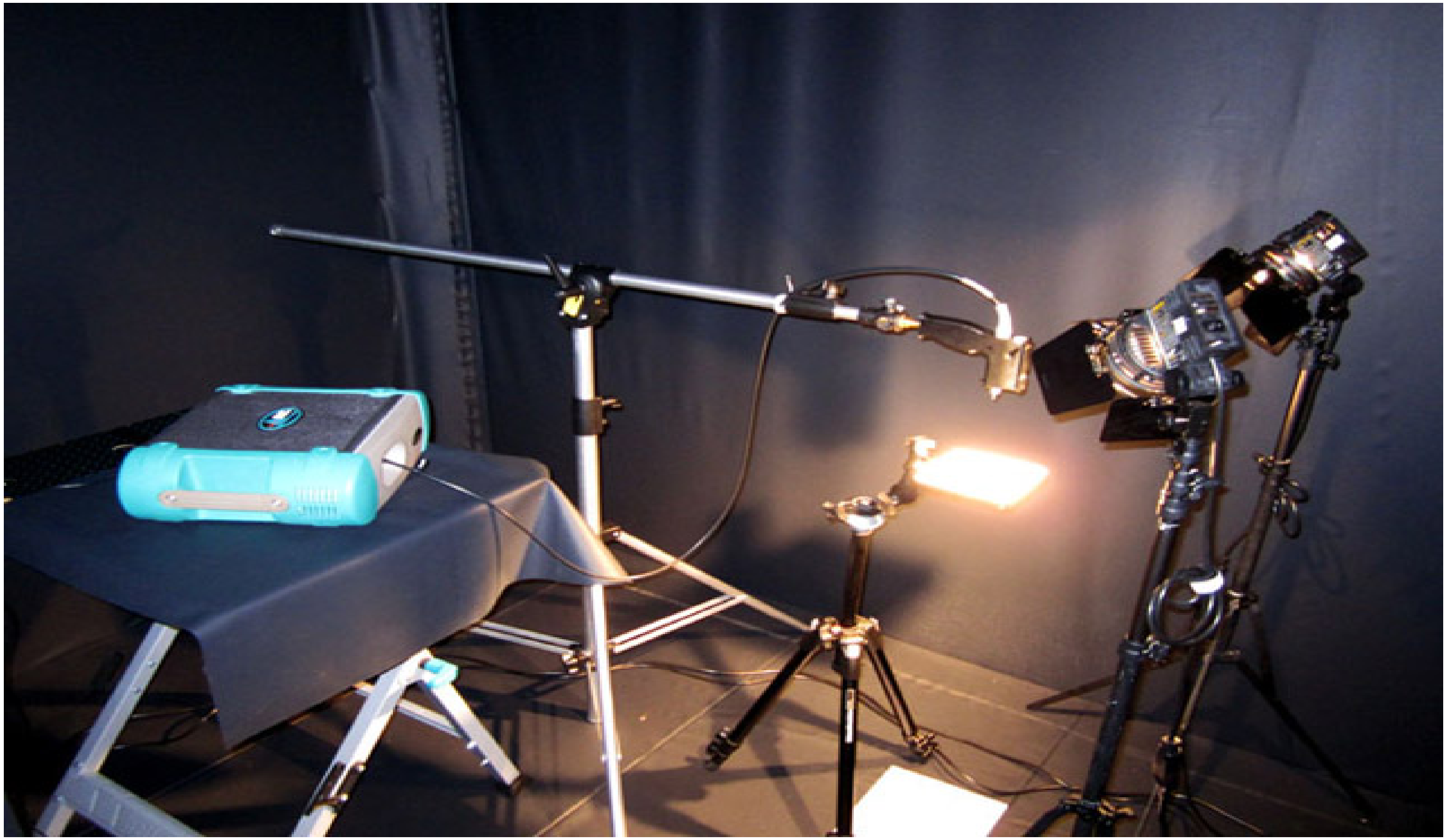
- 264 soil samples

Sampling was made in the depth of 0 to 30 cm

- The samples were air-dried and sieved through a 2 mm mesh.
- All samples were then saved for analyses clay.
- Clay content determined using sedimentation hydrometer method.
- Samples and standards were matrix matched and all analyses were carried out in triplicates.

# Reflectance Spectroscopy Measurement

- Reflectance was measured in the 350-2500 nm wavelength range by a fiberoptic ASD FieldSpec III Pro FR spectroradiometer (ASD Inc., Denver, Colorado, USA) with a contact probe under laboratory condition.



# Preprocessing Strategies

- Captured soil spectra

- Laboratory data of clay (264 samples)

- The noisy parts of the spectra range (350-399 and 2450-2500 nm) were cut out

- The outliers were detected → **Mahalanobis distance (H) applied on PCA-reduced data**

- Baseline offset was removed → **First derivative**

- The artificial noised were eliminated → **Savitzky-Golay**

# Comparison of Algorithms

**PLSR**

PLSR decomposes *X* and *Y* variables and finds new factors (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*.

$$X = Tp' + E$$

$$Y = Tq + F$$

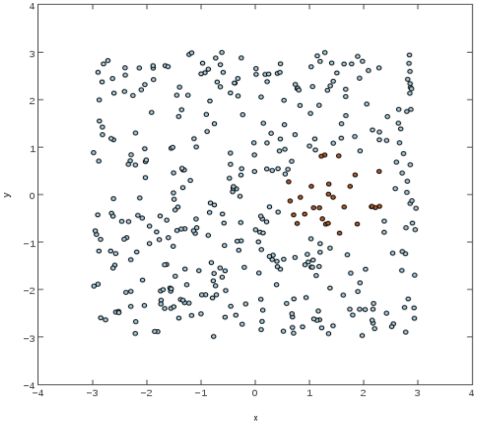
Where:
*X* – soil reflectance
*Y* – measured soil property
*T* – factor scores
*p'* and *q* – factor loadings
*E* and *F* – residuals

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.

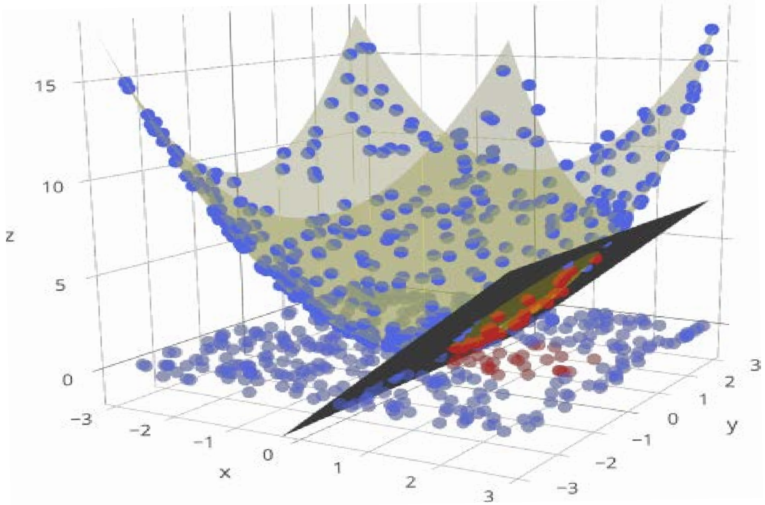
# Comparison of Algorithms (cont.)

## SVMR

- SVMR is a supervised, nonparametric method.
- To avoid over-fitting, SVMR uses learning algorithm itself.



SVMR



- The subsequent equation for prediction has been described below:

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

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

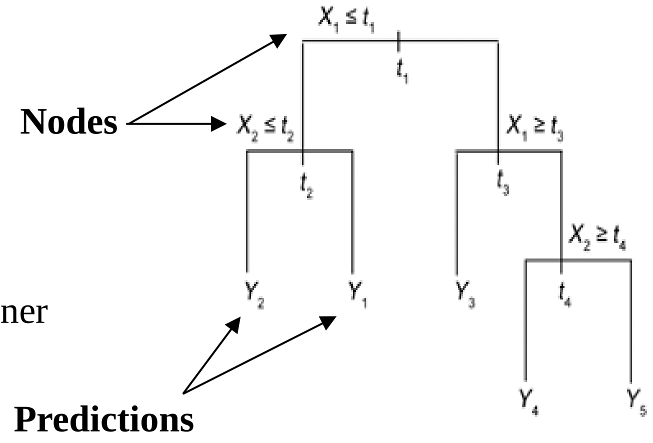
Where:
*b* – scalar threshold
*K (x, x$_k$)* – kernel function
*α* – Lagrange multiplier
*N* – number of data
*x$_k$* – input data
*y* – output

# Comparison of Algorithms (cont.)

## BRT

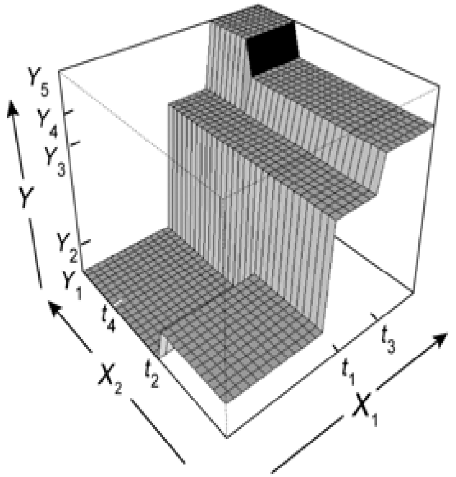
- Boosted models can be stated in the general form:

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


Nodes

Predictions

Where:
*h (x; a)* – simple classification function or base learner with parameters *a* and input variables *x*
*m* – model step
$\beta_m$ – weighting coefficient



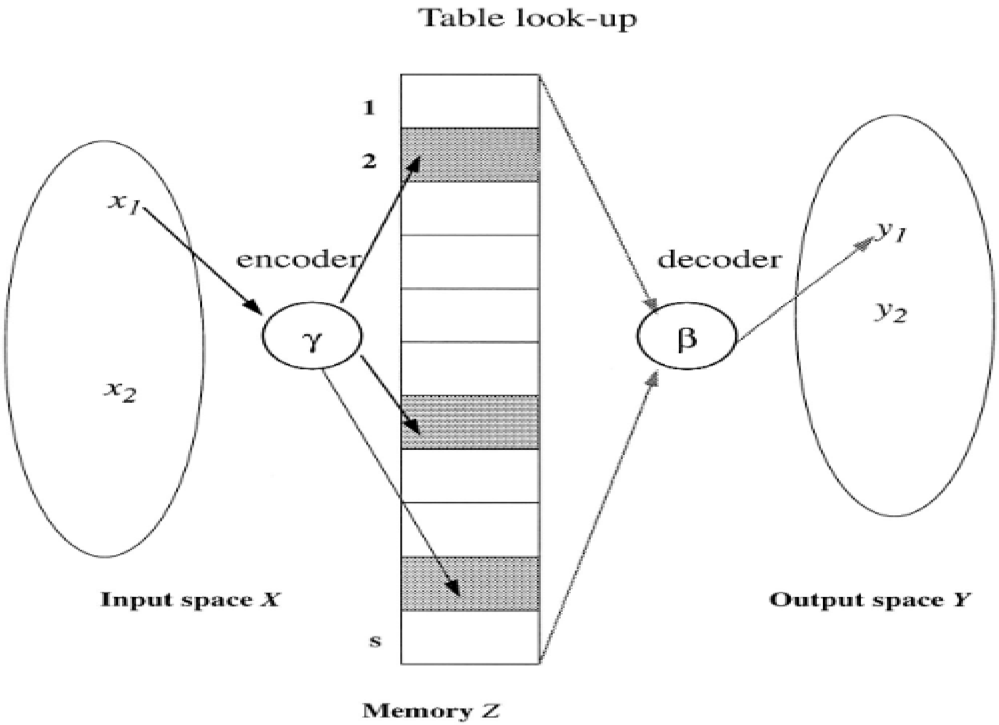
- The primary advantages of BRT include:

(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
(iv) relative immunity to over-fitting

# Comparison of Algorithms (cont.)

## MBL

- MBL resembles the human reasoning process: 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.
- MBL carries out interpolation locally which is based on a local reference set or spectral library.



$s_{1:j}$

# Comparison of Algorithms (cont.)

**MBL**

- ➢ 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 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:

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

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

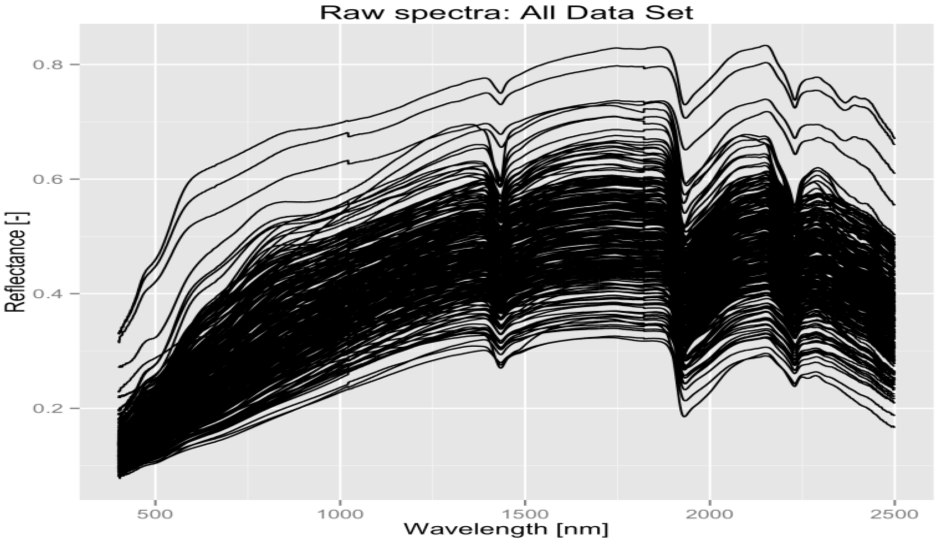
$g_j$ – root mean square of the regression coefficient corresponding to the $j^{th}$ PLS components

# Results

## Clay and Spectral Properties

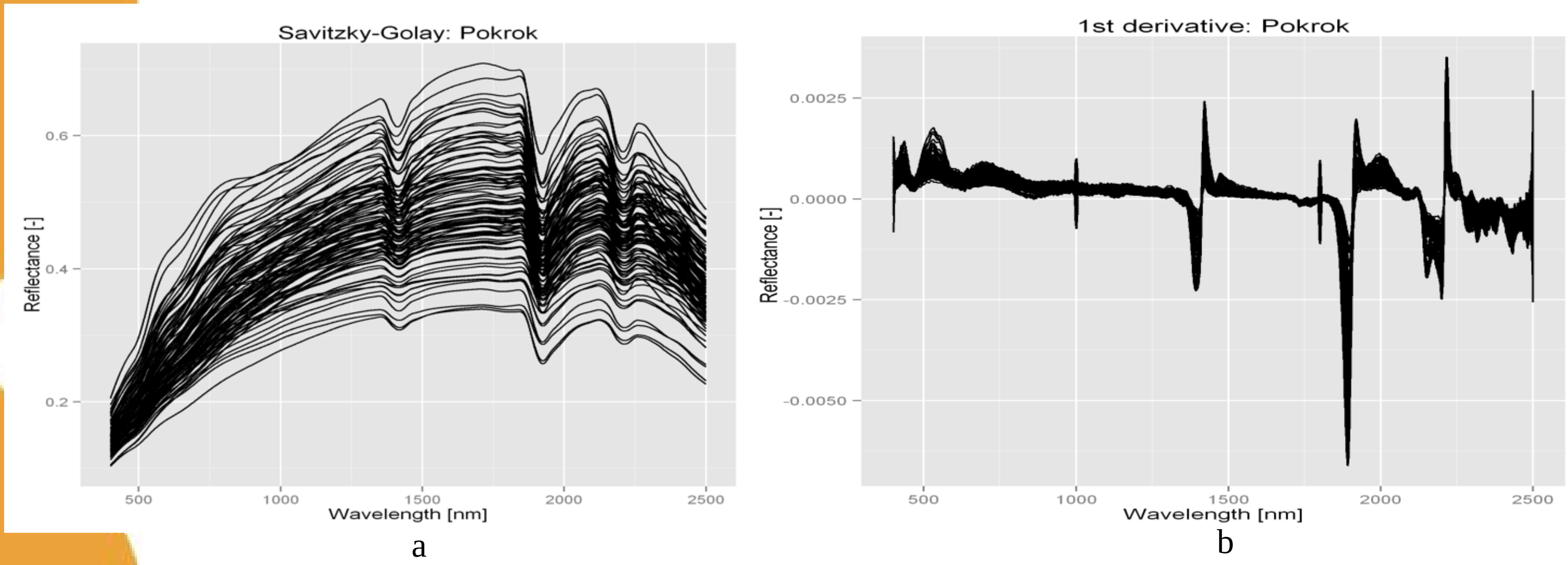
Descriptive statistics of clay content in the studied sample set according to location

| | *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 |



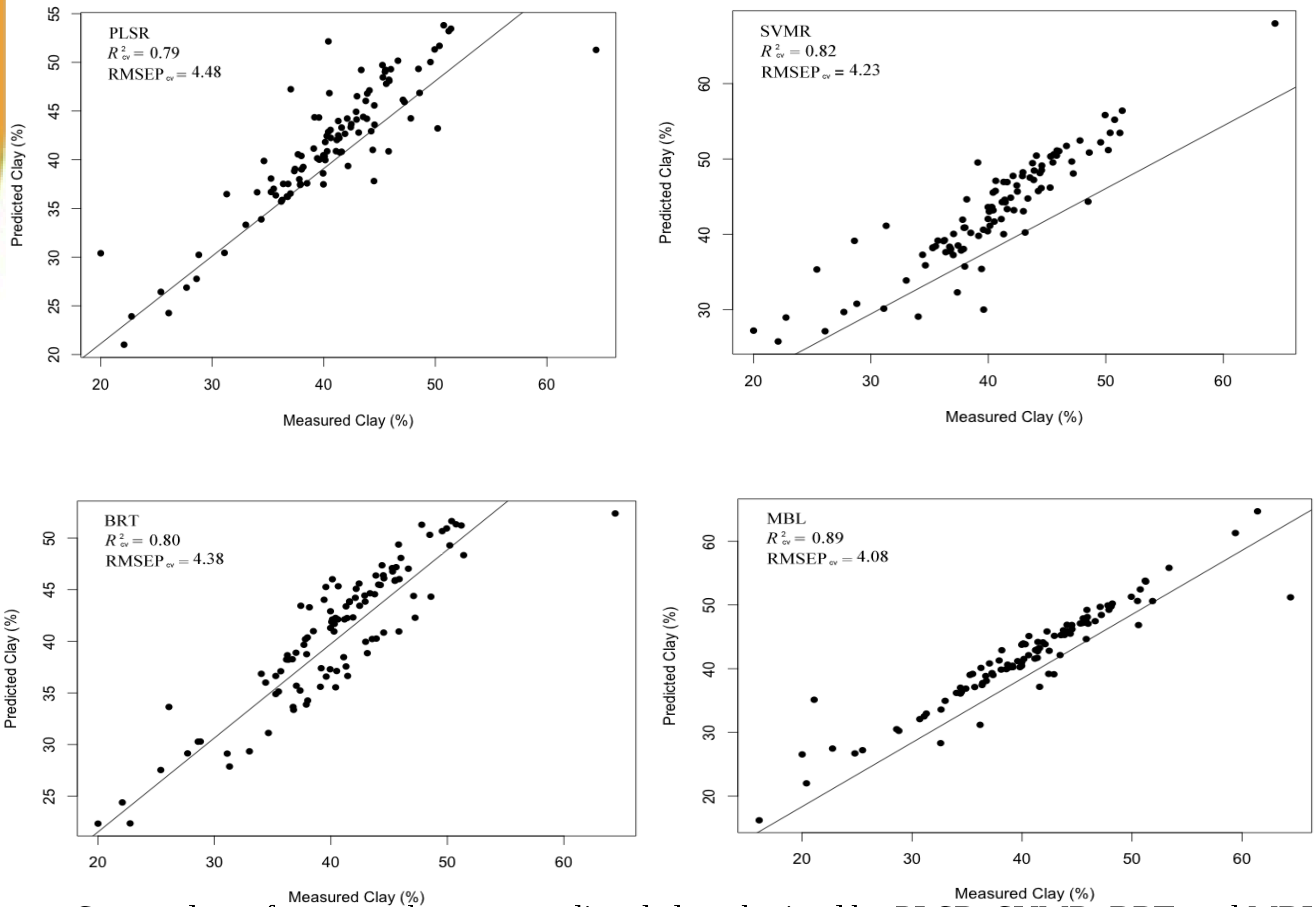
Representative VNIR/SWIR spectra of soil samples

# Spectra Preprocessing



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

# Model Calibration



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

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, the most common method, still showed relatively good prediction of clay content.

# Conclusions

- MBL provided better predictions (lower $\text{RMSEP}_{cv}$) than the SVMR. The other two methods, PLSR and BRT, although significant, they still lay back from the MBL performance.

- MBL increases the model accuracy and is a very promising approach to deal with complex clay VNIR/SWIR datasets.

- 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.

# Thank you