doi: 10.17221/113/2015-SWR

Comparing Different Data Preprocessing Methods for Monitoring Soil Heavy Metals Based on Soil Spectral Features

Asa GHOLIZADEH[1], Luboš BORŮVKA[1], Mohammad Mehdi SABERIOON[2], Josef KOZÁK[1], Radim VAŠÁT[1] and Karel NĚMEČEK[1]

[1]Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Prague, Czech Republic; [2]Laboratory of Image and Signal Processing, Institute of Complex Systems, Faculty of Fisheries and Protection of Waters, University of South Bohemia in České Budějovice, Nové Hrady, Czech Republic

Abstract

Gholizadeh A., Borůvka L., Saberioon M.M., Kozák J., Vašát R., Němeček K. (2015): Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. Soil & Water Res., 10: 218–227.

The lands near mining industries in the Czech Republic are subjected to soil pollution with heavy metals. Excessive heavy metal concentrations in soils not only dramatically impact the soil quality, but also due to their persistent nature and indefinite biological half-lives, potentially toxic metals can accumulate in the food chain and can eventually endanger human health. Monitoring and spatial information of these elements require a large number of samples and cumbersome and time-consuming laboratory measurements. A faster method has been developed based on a multivariate calibration procedure using support vector machine regression (SVMR) with cross-validation, to establish a relationship between reflectance spectra in the visible-near infrared (Vis-NIR) region and concentration of Mn, Cu, Cd, Zn, and Pb in soil. Spectral preprocessing methods, first and second derivatives (FD and SD), standard normal variate (SNV), multiplicative scatter correction (MSC), and continuum removal (CR) were employed after smoothing with Savitzky-Golay to improve the robustness and performance of the calibration models. According to the criteria of maximal coefficient of determination (R^2_{cv}) and minimal root mean square error of prediction in cross-validation ($RMSEP_{cv}$), the SVMR algorithm with FD preprocessing was determined as the best method for predicting Cu, Mn, Pb, and Zn concentration, whereas the SVMR model with CR preprocessing was chosen as the final method for predicting Cd. Overall, this study indicated that the Vis-NIR reflectance spectroscopy technique combined with a continuously enriched soil spectral library as well as a suitable preprocessing method could be a nondestructive alternative for monitoring of the soil environment. The future possibilities of multivariate calibration and preprocessing with real-time remote sensing data have to be explored.

Keywords: heavy metals; preprocessing; support vector machine regression; visible-near infrared spectroscopy

Environmental problems caused by mining are considerable. Those affecting surface soils and vegetation are particularly substantial, because in any surface mining the land surface has to be removed to expose the mineral resource being mined, and in deep mining, any waste material usually has to be deposited at the surface. Consequently, careful programs of conservation and restoring of surface soils must be practiced (Bradshaw 2000).

Contamination of soil is by far one of the most significant effects of mining. Due to their persistent nature and long biological half-lives, elevated concentrations of heavy metals in soils can lead to their accumulation in the food chain, and can eventually influence human health (Wu *et al*. 2005; N'Guessan *et al*. 2009; Xie *et al*. 2012). Heavy metal concentrations in soils can be measured, but their determination is dependent on large-scale sampling

doi: 10.17221/113/2015-SWR

and physical or conventional analysis techniques. The negative side to this is that they are time-consuming, inefficient, and expensive when applied to large scale contaminated lands (Ren *et al*. 2009). Moreover, based on Xie *et al.* (2012), conventional methods for soil monitoring depend on the collection of numerous soil samples, followed by laboratory analyses, which involve complex processes such as separation and pre-concentration. In practice, sampling density and analytic diversity are frequently less than sufficient due to significant costs of analyses.

Visible-near infrared (Vis-NIR) reflectance spectroscopy is a rapid, non-destructive, reproducible, and cost-effective analytical method (Reeves 2010). It provides knowledge of the state of soil, giving results in real-time due to its portability. Furthermore, this method has been adjusted to analyze the spectrally active properties of sediment and soil samples. For example, many studies have shown that, under laboratory conditions, some soil constituents which have spectral features, such as Fe_2O_3 (Ji *et al*. 2002), carbonates (Ben-Dor & Banin 1990), organic matter (Dalal & Henry 1986; Reeves *et al*. 2002), and clay (Ben-Dor & Banin 1995; Kooistra *et al*. 2003), can be accurately determined by reflectance spectroscopy. Moreover, recent studies have shown that via the inter-correlation between spectrally featureless constituents and constituents with spectral features, even featureless soil constituents can be predicted by soil reflectance spectra. Although potentially harmful elements in soils at low concentrations do not have spectral features, the increased input of these elements from anthropogenic sources can often be absorbed or bound by these spectrally active constituents such as soil organic matter (SOM) and clay (Song *et al*. 2012). This makes it possible to study the characteristics of heavy metals in soils using Vis and NIR spectroscopy (Wu *et al*. 2005). Kooistra *et al*. (2001) found that there was a positive correlation between the SOM content and the contents of Zn and Cd in floodplains along the river Rhine in the Netherlands, and based on this correlation the Zn and Cd contamination levels were predicted. Kemper and Sommer (2002) successfully used reflectance spectroscopy to estimate Hg, Pb, and Sb contents in the Aznalcollar Mine area in Spain. Correlation analysis revealed that the most important wavelengths for prediction were attributed to absorption features of iron (Fe) and iron oxides (Fe_2O_3) in soils.

The main challenge limiting the application of spectroscopy for heavy metals assessment is finding suitable data preprocessing and calibration strategies. Choosing the most robust technique can help to achieve a more reliable prediction model. Spectral preprocessing methods are employed to remove any inappropriate information, which cannot be handled correctly by the modelling techniques (Gholizadeh *et al.* 2013). Actually, these methods aim to decrease the noise and enhance possible spectral features connected with the property studied. Some frequently used preprocessing methods, such as multiple scatter correction (MSC), standard normal variate (SNV), Savitzky-Golay, continuum removal (CR) and derivatives, which are mostly used in the multivariate calibration techniques for spectroscopy, can be carried out to determine the best data.

Calibration refers to relating a set of spectral parameters that are derived from the spectral information to the materials in question. Several common methods have been adopted to use multivariate calibration methods to extract the relevant part of the information for a very large dataset in soil applications. These methods include stepwise multiple linear regression (SMLR) (Vasques *et al*. 2008), principle component regression (PCR) (Nocita *et al.* 2013), partial least squares regression (PLSR) (Stevens *et al*. 2010), random forests (RF) (Viscarra Rossel & Behrens 2010; Ji *et al*. 2012), artificial neural networks (ANN) (Hidaka *et al*. 2011), and support vector machine regression (SVMR) (Stevens *et al*. 2010; Chen *et al*. 2012). According to some researchers, using SVMR can overcome the problems of other calibration methods, as the above-mentioned calibration methods require the creation of robust and generalized models due to their potential tendency to over-fit the data (Vapnik 1995; Gholizadeh *et al*. 2013). Based on work by Vapnik (1995) and Gholizadeh *et al*. (2015), SVMR is a supervised non-parametric statistical learning technique; thus, it represents a different method class compared with the previous techniques.

To the best of our knowledge, comparison of different preprocessing methods has not yet been commonly used to analyze heavy metals, in the spectra domain. Therefore, the objective of this study was to evaluate the feasibility of Vis-NIR spectroscopy in the rapid concentration prediction of selected heavy metals, including Mn, Cu, Cd, Zn, and Pb, and to compare the performance of different preprocessing algorithms and the SVMR method for multivariate calibrations, in the vicinity of anthropogenic soils on brown coal mining dumpsites of the Czech Republic. It is envisaged that this rapid and inexpensive

doi: 10.17221/113/2015-SWR

method for obtaining accurate information on heavy metals concentrations would be valued for providing reference data for monitoring the soil environment by proximal and remote sensing.

MATERIAL AND METHODS

Study area and soil sampling. Six dumpsites located in mines Bílina and Tušimice (Figure 1), the Czech Republic, were selected: 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).

All the dumpsites are formed by clays. On a part of each dumpsite, a cover with natural topsoil was spread in an amount of approximately 2500 to 3000 t per ha one year before sampling (Borůvka *et al*. 2012). The topsoil material originated from humic horizons of natural soils of the region, particularly Vertisols, and partly also Chernozems (clayic and haplic). The topsoil was not mixed with the dumpsite material. Individual soil properties differed slightly between the six dumpsites. Some characteristics of the soils, including pH, SOM, and texture were measured using the bulk control subsamples, since they are important environmental indicators. Specifically, the soil pH range for the whole area was 5.3–8.5. The SOM content range was 0.4–3.8%. Texture analysis, which was performed by the pipette method, showed that soil of the area had 37.30% clay, 33.10% sand, and 29.60% silt on average.

Disturbed and undisturbed soil samples were collected at all dumpsites randomly, as follows: Pokrok (103), Radovesice (40), Březno (25), Merkur (38), Prunéřov (48), Tumerity (10). Samples were taken from the depth of 0–30 cm (Song *et al*. 2012; Xie *et al*. 2012) corresponding to the common depth of a ploughing soil layer, as these soils will be used as arable land in the future. The depth of the topsoil cover, where it was applied, was also at least 30 cm.

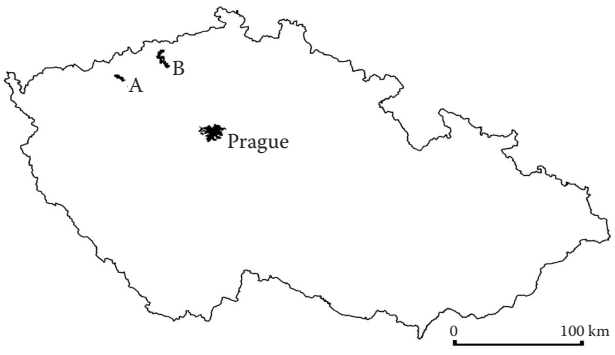

Figure 1. Map of the sampling locations in the Czech Republic; A – area of open cast mining Tušimice, B – area of open cast mining Bílina

Soil analysis. All samples were placed into plastic bags, stored in cool dark containers, and brought to the laboratory for conventional analyses of heavy metals (including Mn, Cu, Cd, Zn, and Pb) and reflectance measurements. Small quantities of dried soil samples were ground to 2 mm mesh. Total concentrations of heavy metals were then determined by digesting soil samples (< 0.149 mm fraction) with a mixture of concentrated hydrochloric and nitric acids (4:1, v:v) (McGrath & Cunliffe 1985; Xie *et al*. 2012), and then analyzed by inductively coupled plasma optical emission spectrometry (ICP-OES). Samples and standards were matrix matched and all analyses were performed in triplicates. Moreover, organic carbon (OC) was measured by combusting samples from which carbonates were removed using 1 N HCl in a NCS Analyzer Flash 2000 (Thermo Scientific, Massachusetts, USA) and the clay content was determined with the hydrometer method (Buurman *et al*. 1997; Kooistra *et al*. 2001).

Reflectance spectroscopy measurement. Reflectance was measured in the 350–2500 nm wavelength range by a FieldSpec 3 spectroradiometer (Analytical Spectral Devices Inc., Colorado, USA) with a contact probe. The spectral resolution of the spectroradiometer was 3 nm for the region 350–1000 nm and 10 nm for the region 1000–2500 nm. Furthermore, the radiometer bandwidth from 350–1000 nm was 1.4 nm, while from 1000–2500 nm it was 2 nm. Samples were illuminated using a stable direct current powered 50 W tungsten-quartz-halogen lamp, which was mounted on a tripod. The angle of incident illumination was 15°, and the distance between the illumination source and the sample was 30 cm. A fibre-optic probe with 8° field of view was used to collect reflected light from the sample. The probe was mounted on a tripod and positioned about 10 cm vertically above the sample. The sample dish was over-filled with soil sample and then levelled off using a blade to ensure a flat surface flush with the top of the dish. The final spectrum was an average based on 20 iterations from 4 directions with 5 iterations per direction to increase the signal-to-noise ratio. Each sample spectrum was corrected for background absorption by division of the reference spectrum of a standardized white $BaSO_4$ panel.

Preprocessing methods and validation. Outliers are commonly defined as observations that are not consistent with the majority of the data (Pearson 2002;

doi: 10.17221/113/2015-SWR

Chiang *et al.* 2003), such as observations that deviate significantly from normal values. An outlier can be defined as (*i*) a spectral outlier when the sample is spectrally different from the rest of the samples, and (*ii*) a concentration outlier when the predicted value has a residual difference significantly greater than the mean of the predicted values (> 2.5 times). For all samples, outliers were detected and eliminated before establishing the regression model (Wu *et al.* 2005). Murray (1988) noted that removing outliers may increase prediction accuracy; hence the outliers were left out.

In order to calibrate a model that provides accurate predictive performance about the quantity of heavy metals contained in each soil sample, the captured soil spectra together with laboratory data of heavy metals were imported into R software (R Development Core Team 2011) to be processed. From a total of 264 samples taken for laboratory analysis, mostly subsets were used to determine the content of heavy metals. The number of samples subjected to individual analysis was then as follows: the entire data were tested for Mn; 148 samples were tested for Pb; 115 for Cu and Zn, and 104 samples for Cd.

The first step in spectroscopy analysis often consists of preprocessing to assess and possibly improve data quality. This step may take more time than the analysis itself. Spectral preprocessing techniques consist of a variety of mathematical methods for correcting light scattering in reflectance measurements and data enhancement before the data is used in calibration models. Usually, preprocessing techniques can be divided into four categories, namely smoothing, baseline removal, scaling, and normalization (Xie *et al.* 2012). The first category is smoothing, which is used for noise (random measurement error) reduction. Savitzky-Golay smoothing as one of the popular choices is an averaging algorithm that fits a least squares polynomial to the data points, and then the value to be averaged is predicted from the polynomial (Savitzky & Golay 1964). In some forms of spectroscopy, one can encounter a baseline, or "background signal" that is far away from the zero level. Since this influences measurements like peak height and peak area, it is of utmost importance to correct for such phenomena; thus, the second category of preprocessing is the baseline removal. There are several baseline removal methods available such as spectral derivative transformation, which is one of the best methods for removing baseline effects (Duckworth 2004): the first derivative (FD) is very effective for removing baseline offset; the second derivative (SD) is very effective for both the baseline offset and linear trend from a spectrum (Duckworth 2004; Rinnan *et al.* 2009). Another way to remove scatter effects in spectroscopy is MSC, which is a transformation method used to compensate for multiplicative and/or additive scatter effects in the data (Chu *et al.* 2004). Another preprocess is range scaling. This method is applicable when the total intensity in the spectra is sample-dependent, and samples need to be scaled in such a way that intensities can be compared. Among different range scaling methods, SNV is a widely used technique, especially in NIR applications, which corrects the multiplicative interferences of light scatter and particle size by centering and scaling each individual spectrum (Duckworth 2004). In other words, after scaling, every spectrum will have a mean of zero and a standard derivation of 1. Another method for preprocessing is normalization which is normally used for absorption feature enhancement and identification. This study used continuum removal (CR) as a normalization method, which generates new spectral data by dividing the envelope curve of a continuum on raw reflectance spectra (Clark & Roush 1984). CR is effective at isolating specific absorption features, and removing the effects of changing slopes and overall reflectance levels (Kokaly *et al.* 2003).

In this study, prior to all further spectra treatments, the noisy part of the spectra range (350–399 and 2450–2500 nm) was cut out in order to remove the artificial noise caused by the spectroradiometer instrument. Then, 7 forms of spectra preprocessing methods were used to remove the non-constituent-related effects in spectra data and to develop optimal models by SVMR for estimating soil properties. The six forms were raw reflectance (R), SNV, MSC, Savitzky-Golay smoothing with a second-order polynomial fit and 11 smoothing points, the FD and the SD transformation, and CR. For SVMR prediction we used radial basis function kernel contained in *e1071* R package (Meyer *et al.* 2012). All spectral preprocessing transformations were also carried out using R package.

Accuracy assessment of techniques. Assessment of the methods accuracy was carried out using a leave-one-out cross-validation approach (R^2_{cv} and $RMSEP_{cv}$). The $RMSEP_{cv}$ was computed as follows:

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

doi: 10.17221/113/2015-SWR

where:
y_i' – predicted value
y_i – observed value
N – number of samples

The smallest $RMSEP_{cv}$ value was related to the optimal calibration model.

In fact, R^2 indicates the percentage of the variance in the Y variable that is accounted for by the X variables. An R^2 value between 0.50 and 0.65 indicates that more than 50% of the variance in Y is accounted for by variable X, so that discrimination between high and low concentrations can be made. An R^2 value in the range of 0.66–0.81 indicates approximate quantitative predictions, whereas R^2 of 0.82– 0.90 reveals a good prediction. Calibration models having an R^2 value above 0.91 are considered excellent (Williams 2003).

RESULTS AND DISCUSSION

Soil samples descriptive statistics and correlations. Descriptive statistical results of soil parameters in the six dumpsites are shown in Table 1. The comparison of coefficients of variation (CV) of different

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

Item	Cu	Mn	Cd	Pb	Zn	Clay	SOM
			(mg/kg)			(%)	
Pokrok (n = 103)							
Min	5.50	198.3	0.01	7.60	8.30	7.50	0.41
Max	35.70	869.1	0.73	42.40	127.10	53.26	3.83
Mean	13.76	599.4	0.27	18.43	25.26	36.72	1.62
SD	3.58	118.6	0.11	5.32	15.77	8.65	0.47
CV (%)	26	20	40	29	62	24	29
Radovesice (n = 40)							
Min	6.42	254.1	0.03	4.70	9.38	18.12	0.89
Max	22.10	844.1	0.30	49.60	66.85	52.92	2.05
Mean	14.20	541.3	0.17	13.70	21.98	41.91	1.35
SD	3.45	125.1	0.05	6.40	11.15	7.75	0.24
CV (%)	24	23	30	47	51	19	18
Březno (n = 25)							
Min	9.01	473.3	0.00	10.90	11.49	28.93	1.18
Max	38.81	885.8	0.37	21.60	200.27	61.41	1.78
Mean	14.37	680.9	0.16	14.17	41.50	39.98	1.53
SD	5.95	105.9	0.11	2.97	41.62	5.97	0.18
CV (%)	41	16	64	21	100	15	12
Merkur (n = 38)							
Min	7.29	318.0	0.04	9.30	6.95	17.69	0.98
Max	16.76	787.3	0.27	55.90	32.22	59.87	2.16
Mean	12.22	590.0	0.16	17.53	13.56	47.45	1.59
SD	1.77	100.7	0.06	7.23	4.19	6.54	0.26
CV (%)	14	17	39	41	31	14	16
Prunéřov (n = 48)							
Min	8.40	41.6	0.00	0.90	6.60	6.09	0.95
Max	92.24	984.0	0.24	24.80	213.11	60.67	3.58
Mean	15.81	552.6	0.11	14.38	26.83	40.49	1.84
SD	14.36	224.4	0.06	4.82	39.32	12.58	0.49
CV (%)	91	41	55	34	147	31	27
Tumerity (n = 10)							
Min	12.29	496.8	0.00	9.50	15.50	31.63	0.76
Max	20.34	1027.6	0.20	14.50	48.56	68.40	1.68
Mean	15.03	753.1	0.12	12.25	25.61	50.74	1.28
SD	2.40	192.3	0.05	1.38	10.32	11.53	0.36
CV (%)	16	26	44	11	40	23	28

SD – standard deviation; CV – coefficient of variation

doi: 10.17221/113/2015-SWR

Table 2. Pearson's correlation coefficients between measured soil properties

	Cu	Mn	Cd	Pb	Zn	Clay	SOM
Cu	1						
Mn	0.15*	1					
Cd	0.24*	0.56	1				
Pb	0.43**	0.27**	0.40**	1			
Zn	0.85**	0.23**	0.23*	0.44**	1		
Clay	–0.30	0.23	0.43*	–0.26	0.46*	1	
SOM	0.31**	0.14	0.45**	0.20**	0.23**	0.71**	1

SOM – soil organic matter; significant at *P = 0.05 and **P = 0.01, respectively (two-tailed)

contaminants showed that among all parameters Zn had the highest CV, especially in the Prunéřov area (147%). However, in Merkur and Tumerity, the highest CV that shows the most variety belongs to Cd, as compared to other measured parameters. In contrast, Pb in Tumerity had the lowest CV (11%), which shows it is more homogeneous than the other properties. CV of clay and SOM were lower than CV of all contaminants (12% < CV < 31%) except those in Pokrok and Tumerity, which had rather low CV.

In the study area, the estimated mean concentration of Cd (0.27 mg/kg) and Pb (18.43 mg/kg) in Pokrok was higher than at the other locations. Nevertheless, most values are under current limit values (1 and 70 mg/kg for Cd and Pb, respectively) given by Czech legislation for agricultural soils (Ministry of Environment 1993). An exception can be values of Zn in some soil samples that exceed the limit values (100 mg/kg) significantly, suggesting strong anthropogenic contamination from the mining industries nearby. The long history of metal manufacturing and processing resulted in both point and diffuse pollution of heavy metals, especially Zn in the study area.

Table 2 highlights the linear correlation coefficients between all the examined elements for 264 samples.

The relationships between clay, SOM, and heavy metals exhibited different features, which reflected different metal-affinity in different circumstances (Xie *et al*. 2012). Based on all samples, the correlation between clay and SOM for the total data set was moderate (r = 0.71). However, SOM were positively correlated with all heavy metals except Mn concentration (0.20 < r < 0.71), for the clay content, moderate to low correlations were found with Zn (r = 0.46) and Cd (r = 0.43). These results give evidence to earlier observations (Kooistra *et al*. 2001) that the amounts of metals contained in river floodplains was determined by the contents of clay and SOM. Actually, Zn and Cd had significant correlations with both clay and SOM, whereas Cu and Pb only showed a significant correlation with SOM. Interestingly, Mn showed no dependence to clay and SOM.

Spectral response of soil samples. Selected representative soil sample spectra are illustrated in Figure 2, and it can be seen that the soil samples had similar Vis-NIR reflectance spectra. The Vis-NIR reflectance spectra of soil had weak absorption peaks overlapping at around 430, 500, 530, and 650 nm in the Vis region. Fe_2O_3 can absorb or bond with other metal cations or hydroxyl groups, and this has a spectral activity in the Vis region (Wu *et al*. 2005). Therefore, the peaks in the Vis region were mainly related to electronic transitions of the Fe^{3+} in the goethite or hematite component of the Fe_2O_3 (Ji *et al*. 2002; Wu *et al*. 2005).

In the NIR region, O-H bonds in the SOM and clay minerals would have a general impact on the reflectance spectra (Kooistra *et al*. 2003; Ren *et al*. 2009; Song *et al*. 2012). The 1400 nm peak is related to O-H bonds in the hydroxyl or clay minerals, such as smectite and illite (White 1971; Xie *et al*. 2012) and the 1900 nm peak is mainly related to the absorption peak of the O-H bond in water (Clark *et al*. 1990). The 2200 nm peak is a combined result of the O-H bonds found in clay minerals such as kaolinite, illite, and smectite

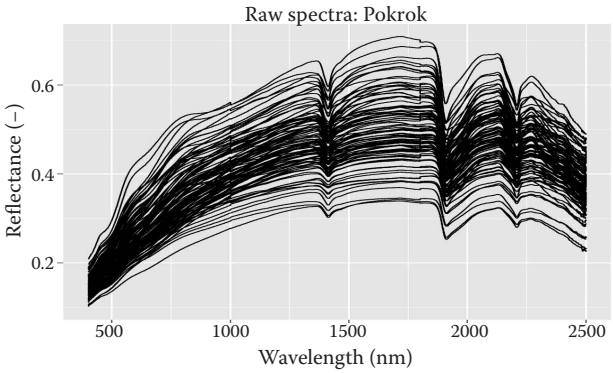

Figure 2. Representative Vis-NIR spectra of soil samples

doi: 10.17221/113/2015-SWR

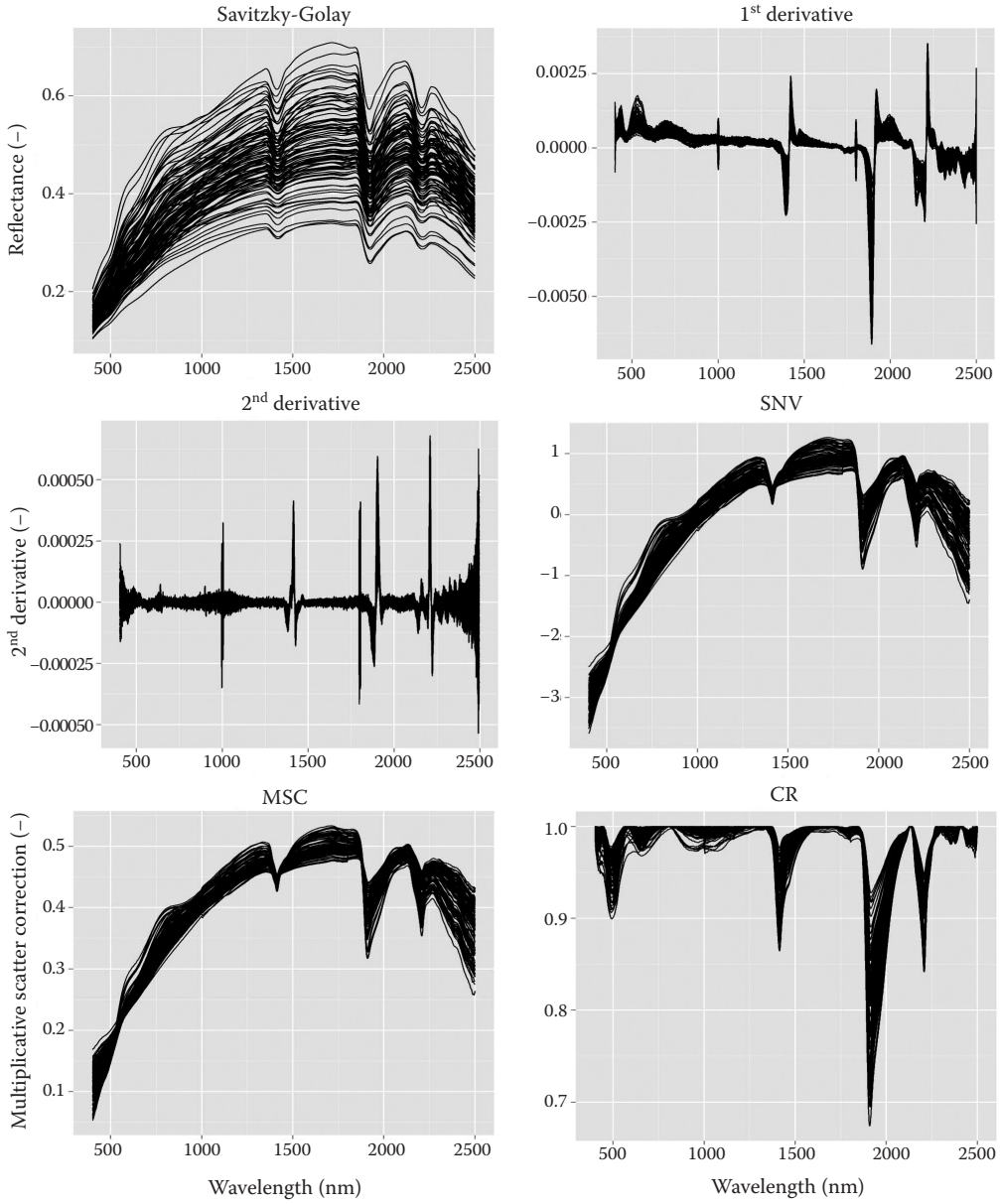

Figure 3. Spectra of soil samples from Pokrok dumpsite preprocessed using different methods
SNV – standard normal variate; MSC – multiplicative scatter correction; CR – continuum removal

(Madejova & Komadel 2001; Nayak & Singh 2007). Additionally, there were some weak peaks between 2250 and 2450 nm, which were associated with the C-H bonds in SOM, including lignin and humic acid (Ben-Dor *et al.* 1997; Song *et al.* 2012), as well as carbonates (White 1971; Ben-Dor & Banin 1990).

Data preprocessing and optimal SVMR algorithm. In order to establish a robust prediction model and explore the influence of spectral sampling interval on the prediction accuracy, different spectral preprocessing techniques were performed. A visual inspection of the spectra allowed detection of some spectral readings possibly affected by measurement errors. These were removed, and the final spectral library had a total of 264 soil spectra. Smoothed spectra by Savitzky-Golay, and all preprocessed spectra of all selected soil samples with different preprocessing techniques in the location that had the most samples (Pokrok) are shown in Figure 3. Other locations also showed the same pattern.

Support vector machine regression (SVMR) was applied for constructing optimal models. Accordingly, the SVMR method with the FD preprocessing was determined as the final technique for predicting Cu, Mn, Pb, and Zn concentration, whereas the

doi: 10.17221/113/2015-SWR

Table 3. Prediction results of the preprocessing models and support vector machine regression (SVMR) for the heavy metals concentration (in mg/kg)

Preprocessing	Cu		Mn		Cd		Pb		Zn	
	R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$	R^2_{cv}	$RMSEP_{cv}$
No preprocessing	55	6.37	30	122.43	72	0.05	28	4.00	40	17.87
FD	78	4.02	60	97.18	80	0.04	68	2.97	77	13.67
SD	72	6.70	59	98.10	79	0.04	65	3.01	60	15.82
SNV	23	9.14	19	131.76	32	0.09	29	3.99	26	20.63
MSC	21	9.58	19	131.76	25	0.09	23	4.22	24	21.32
CR	58	6.16	47	108.73	82	0.04	60	3.30	68	14.33

FD – first derivative; SD – second derivative; SNV – standard normal variate; MSC – multiplicative scatter correction; CR – continuum removal; R^2_{cv} – coefficient of determination; $RMSEP_{cv}$ – root mean square error of prediction in cross-validation

SVMR with the CR preprocessing was chosen as the best algorithm for predicting Cd. Table 3 shows cross-validation results of the SVMR for the heavy metals concentration using different preprocessing methods on the spectra.

According to the criteria of minimal $RMSEP_{cv}$ and maximal R^2_{cv}, the SVMR method with the FD pretreatment ($RMSEP_{cv}$ = 4.02, 97.18, 2.97, 13.67 and R^2_{cv} = 0.78, 0.60, 0.68, 0.77) were considered as the foremost techniques for predicting Cu, Mn, Pb, and Zn, respectively. However, the CR preprocessing method was chosen as the best algorithm for Cd ($RMSEP_{cv}$ = 0.24, R^2_{cv} = 0.04).

Furthermore, MSC and SNV spectral preprocessing gave the lowest values for all heavy metals, while compared to that of the methods with no preprocessing, the prediction ability with the preprocessing of FD, SD, and CR was improved. In fact, the application of most of the preprocessing methods for these heavy metals increased the accuracy of prediction, which was similar to the conclusion drawn by Ren *et al.* (2009). These results show that for all heavy metals several techniques can give a robust prediction on the basis of spectra from soil samples. Earlier research has shown that calibration models, in which spectra are not preprocessed, are more sensitive to changes in operating or environmental conditions compared to models for which preprocessing is applied (Moros *et al.* 2009). Although Kooistra *et al.* (2001) obtained good results in a study where no preprocessing was used for all parameters; one could still decide to use a preprocessing method to avoid this problem, as the differences in cross-validation error are relatively small.

In this study, spectra measured in the laboratory were used for predicting metal concentration levels using different data preprocessing methods. The research information provided an alternative tool for the investigation of contaminated soils by remote sensing. Reflectance spectra collected in the laboratory after preprocessing were suitable for developing future prediction models of the measured heavy metals, especially using the FD preprocessing technique. However, the results should be validated and explored by further investigations on different geographical scales, because conditions in the laboratory might be more convenient than those in the field, and reflectance spectra collected under field conditions are affected by natural soil surface conditions (e.g. roughness, humidity, vegetation cover, etc.), atmosphere, and illumination (Leone & Sommer 2000). For the applications where we have to deal with relatively noisy spectra, the application of wavelength selection could be a promising preprocessing method. These techniques will be further studied on the basis of spectral data collected directly in the field. Moreover, soil type is also a major factor in constructing models, especially for the applications of remote sensing in a large-scale soil contamination survey.

CONCLUSION

The present research provided an alternative tool for the prediction of soil heavy metal contaminations and contaminated soil ecosystems using laboratory spectral reflectance and remote sensing. The study showed that high resolution spectra of soil samples taken from the Czech dumpsites can be used for predicting Cu, Mn, Cd, Pb, and Zn contamination levels. Cu and Pb only displayed significant correlation with SOM, but Zn and Cd had significant correlations with clay and SOM. The inter-correlation between heavy metals

doi: 10.17221/113/2015-SWR

and the spectrally active constituents of soils is the major mechanism by which the spectrally featureless toxic metals can be predicted. Soil spectroscopy in the Vis-NIR region with SVMR technique was shown to be a very promising method for the determination of metal concentrations in anthropogenic soils on brown coal mining dumpsites of the Czech Republic. The FD preprocessing method gave the best validation results for Cu, Mn, Pb, and Zn. However, MSC and SNV spectral preprocessing showed weak prediction for all the measured heavy metals. The future possibilities of these methods in remote sensing have to be explored.

Acknowledgements. The authors acknowledge the financial support of the EC Operational Program (Project No. ESF/MEYS CZ.1.07/2.3.00/30.0040) and of the Czech University of Life Sciences Prague. The assistance of Mr. Ch. Ash for English revision is also acknowledged.

References

Ben-Dor E., Banin A. (1990): Near-infrared reflectance analysis of carbonate concentration in soils. Applied Spectroscopy, 44: 1064–1069.

Ben-Dor E., Banin A. (1995): Near infrared analysis (NIRA) as a rapid method to simultaneously evaluate several soil properties. Soil Science Society of American 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: 1–15.

Borůvka L., Kozák J., Mühlhanselová M., Donátová H., Nikodem A., Němeček K., Drábek O. (2012): Effect of covering with natural topsoil as a reclamation measure on brown-coal mining dumpsites. Journal of Geochemical Exploration, 113: 118–123.

Bradshaw A. (2000): The use of natural processes in reclamation – Advantages and difficulties. Landscape Urban Planning, 51: 89–100.

Buurman P., Pape Th., Muggler C.C. (1997): Laser grain-size determination in soil genetic studies. Soil Science, 162: 211–218.

Chen Q., Guo Z., Zhao J., Ouyang Q. (2012): Comparisons of different regressions tools in measurement of antioxidant activity in green tea using near infrared spectroscopy. Journal of Pharmaceutical and Biomedical Analysis, 60: 92–97.

Chiang L.H., Pell R.J., Seasholtz M.B. (2003): Exploring process data with the use of robust outlier detection algorithms. Journal of Process Control, 13: 437–449.

Chu X.L., Yuan H.F., Lu W.Z. (2004): Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique. Progress in Chemistry, 16: 528–542.

Clark R.N., Roush T.L. (1984): Reflectance spectroscopy: quantitative analysis techniques for remote sensing application. Journal of Geophysical Research, 89: 6329–6340.

Clark R.N., King T.V.V., Klejwa M., Swayze G.A., Vergo N. (1990): High spectral resolution reflectance spectroscopy of minerals. Journal of Geophysical Research, 95 (B8): 12653.

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 American Journal, 50: 120–123.

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

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

Hidaka Y., Kurihara E., Hayashi K. (2011): Near infrared spectrometer for a head feeding combine for measuring rice protein content. Japan Agricultural Research Quarterly, 45: 63–68.

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: 208–216.

Ji W.J., Li X., Li C.X., Zhou Y., Shi Z. (2012): Using different data mining algorithms to predict soil organic matter based on visible near infrared spectroscopy. Spectroscopy and Spectral Analysis, 32: 2393–2398.

Kemper T., Sommer S. (2002): Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environmental Science and Technology, 36: 2742–2747.

Kokaly R.F., Despain D.G., Clark R.N., Livo K.E. (2003): Mapping vegetation in Yellowstone National Park using spectral feature analysis of AVIRIS data. Remote Sensing of Environment, 84: 437–456.

Kooistra L., Wehren R., Leuven R.S.E., Buydens L.M.C. (2001): Possibilities of visible-near-infrared spectroscopy for the assessment of soil contamination in river flood plains, Analytica Chimica Acta, 446: 97–105.

Kooistra L., Wanders J., Epema G.F., Leuven R., Wehrens R., Buydens L.M.C. (2003): The potential of field spectros-

doi: 10.17221/113/2015-SWR

copy for the assessment of sediment properties in river floodplains. Analytica Chimica Acta, 484: 189–200.

Leone P.L., Sommer S. (2000): Multivariate analysis of laboratory spectra for the assessment of soil development and soil degradation in the Southern Apennines (Italy). Remote Sensing of Environment, 72: 346–359.

Madejova J., Komadel P. (2001): Baseline studies of the clay minerals society source clays: infrared methods. Clays and Clay Minerals, 49: 410.

McGrath S.P., Cunliffe C.H. (1985): A simplified method for the extraction of metals Fe, Zn, Cu, Ni, Cd, Pb, Cr, Co and Mn from soils and sewage sludges. Journal of the Science of Food and Agriculture, 36: 794–798.

Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F. (2012): e1071: Misc Functions of the Department of Statistics (e1071), R Package Version 1.6-1. Wien, TU Wien.

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 and Technology 43: 9314–9320.

Murray I. (1988): Aspects of interpretation of NIR spectra. In: Creaser C.S., Davies A.M.C. (eds): Analytical Application of Spectroscopy. London, Royal Society of Chemistry: 9–21.

Nayak P., Singh B. (2007): Instrumental characterization of clay by XRF, XRD and FTIR. Bulletin of Materials Science, 30: 235–238.

N'Guessan Y.M., Probst J.L., Bur T., Probst A. (2009): Trace elements in stream bed sediments from agricultural catchments (Gascogne region, S-W France): where do they come from? Science of the Total Environment, 407: 2939–2952.

Nocita M., Stevens A., Noon C., Van Wesemael B. (2013): Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. Geoderma, 199: 37–42.

Pearson R.K. (2002): Outliers in process modeling and identification. IEEE Transactions on Control Systems Technology, 10: 55–63.

R Development Core Team. (2011): R: A language and environment for statistical computing. R foundation for Statistical Computing. Available at http://www.R-project.org

Reeves J.B. III (2010): Near versus mid infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma, 158: 3–14.

Reeves III J.B., McCarty G.W., Mimmo T.V., Reeves V.B., Follet R.F., Kimble J.M., Galletti G.C. (2002): Spectroscopic calibrations for the determination of C in soils. Transactions of the 17th World Congress of Soil Science, 10: 1–9.

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.

Rinnan A., van den Berg F., Engelsen S.B. (2009): Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry, 28: 1201–1222.

Savitzky A., Golay M.J.E. (1964): Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36: 1627–1639.

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.

Stevens A., Udelhoven T., Denis A., Tychon B., Lioy R., Hoffmann L., Van Wesemael B. (2010): Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma, 158: 32–45.

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., Behrens T. (2010): Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158: 46–54.

White W. (1971): Infrared characterization of water and hydroxyl ion in the basic magnesium carbonate minerals. American Mineralogist, 56: 46–53.

Williams P. (2003): Near-infrared Technology – Getting the Best out of Light. Nanaimo, PDK Projects.

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.

Received for publication June 12, 2015
Accepted after corrections September 2, 2015

Corresponding author:

Dr. Asa Gholizadeh, Česká zemědělská univerzita v Praze, Fakulta agrobiologie, potravinových a přírodních zdrojů, katedra pedologie a ochrany půd, Kamýcká 129, 165 21 Praha 6-Suchdol, Česká republika; e-mail: gholizadeh@af.czu.cz