MAPPING THE DISTRIBUTION OF CHEMICAL PROPERTIES IN SOIL PROFILES USING LABORATORY IMAGING SPECTROSCOPY, SVM AND PLS REGRESSION

Henning Buddenbaum[1], and Markus Steffens[2]

1. University of Trier, Remote Sensing Department, Trier, Germany; Buddenbaum(at)uni-trier.de
2. Department für Ökologie und Ökosystemmanagement, Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt, Technische Universität München, Freising-Weihenstephan, Germany; steffens(at)wzw.tum.de

ABSTRACT

Laboratory imaging spectroscopy as a tool for studying the three-dimensional properties of soils is introduced. Hyperspectral images of parallel slices of a soil core sampled in a custom-made steel box were made. We used image spectra from chemically analysed samples to train Support Vector Machine and Partial Least Squares regression models of chemical soil properties. Special attention has to be paid on the correct number of estimator variables. The models were applied on the images to create sub-millimetre scale maps of carbon, nitrogen, iron, aluminium and manganese content in the soil profiles.

INTRODUCTION

Although soil horizons are often seen as homogeneous they show clear patterns and large variations in chemical and physical properties on the sub-centimetre scale. Most analytical techniques for soil constituent analysis are destructive and cannot be repeated often enough for high spatial resolution examinations. These problems can be overcome by imaging proximal sensing techniques. Laboratory imaging spectroscopy offers a novel means of mapping various small-scale soil properties, either looking into the depth (as done in this study) or looking at the soil surface from above (as done by airborne or spaceborne remote sensors). The utility of imaging spectroscopy for the study of soil properties has been shown several times (e.g., 1,2,3).

We recorded hyperspectral images of soil profiles to map these small-scale variations without using destructive analytical techniques. Like the method introduced by Ben-Dor et al. (4) this technique allows looking into the depth. While our method until now was only applied in the laboratory and not in the field, it produces complete images of the profiles, not only point measurements.

Several chemical soil properties of small samples were analysed in a laboratory. Support Vector Machine (SVM) and Partial Least Squares (PLS) regression, two different statistical chemometric techniques (5), are compared to assess their ability to map chemical soil constituents in the soil profiles.

METHODS

Field soil sampling

A siltic stagnic luvisol (6) was sampled in a Norway spruce (*Picea abies*) monoculture near Freising (SE-Germany), approximately 35 km northeast of Munich.

A custom-made stainless steel box of 100 mm × 100 mm × 300 mm was used to sample 30 cm deep soil profiles. After removing the litter layer, the steel box was gently hammered vertically into the soil after litter removal and digged out so that an undisturbed profile was sampled. The soil core was oven-dried for 24 h.

Imaging Setup

The images were acquired using a hyperspectral scanner with 160 bands in the 400-1000 nm range (NEO HySpex VNIR-1600) mounted in a laboratory frame with a translation stage under the scanner. The translation stage moves the object in along-track direction while the push-broom scanner records lines across-track. Each line in the setup is about 10 cm wide and consists of 1600 pixels resulting in a sub-millimetre sampling distance. Light sources illuminate the currently scanned line from two directions in order to minimise shadows on the soil surface. A Spectralon® white reference panel was scanned with the sample so radiance values could be transformed to reflectance. More details on the imaging setup can be found in (7).

Soil sampling for laboratory analysis

Several samples of the soil profile were taken for chemical analyses, and then the top 15 mm of the soil profile were removed and the profile was scanned again. In total 66 samples were taken from seven layers and analysed for organic carbon, total nitrogen, iron, manganese and aluminium content using standard laboratory measurements. Repeating the scans for the seven layers allow for a three-dimensional characterization of the soil properties, albeit with a coarse resolution in one of the dimensions.

Regression models for chemometric mapping

The image spectra of the sampled regions and their respective chemical contents were used for training SVM and PLS regression models. Because of the limited number of reference areas, model parameters were optimised using five-fold cross validation (8). Each regression model was evaluated by calculating cross-validated variance explained (R^2), root mean square error (*RMSE*) and relative *RMSE* (%*RMSE*) according to

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_{p,i} - y_{o,i}\right)^2}$$

$$\%RMSE = 100 * \frac{RMSE}{\overline{y_o}}$$

where

n – number of observations

y_p – predicted values

y_o – observed values

$\overline{y_o}$ – mean of the observed values.

SVMs are based on machine-learning theory. They can be used for both classification and regression problems and provide excellent generalization capabilities, are robust to high input space dimension and a low number of samples (9). The basic principle of SVM regression (10) is that an input vector is mapped from the input domain into a higher dimensional feature space via a kernel function, where data are spread out in a way that facilitates the finding of an interpolation function (11). In our case, a Gaussian radial basis function kernel was chosen. SVM regression was carried out using the software ImageSVM (Version 2.1, Humboldt-Universität zu Berlin, Germany; Software available at www.hu-geomatics.de) inside the EnMAP-Box (Version 1.1, Humboldt-Universität zu Berlin, Germany). ImageSVM has been used successfully in classification tasks, e.g. by van der Linden & Hostert (12) and by Waske et al. (13). The kernel and loss function parameters *g* and *C* are optimized automatically by the software using a grid search method: The software trains SVM regression models with combinations of *g* and *C* and selects the model with the best performance (14). In our case minimum and maximum values of 0.1 and 1000, respectively, and a multiplier of 1.33 were set for the two parameters. *RMSE* was chosen as performance measure. Starting with the minimum value, each parameter is multiplied by the multiplier until the maximum value is reached, and every possible parameter combination is tested. In forward selection mode a feature selection was carried out to use only the minimum number of input features required for a satisfactory estimation of the chemical properties.

PLS regression (15) is a widely used approach in chemometrics and hyperspectral remote sensing that generalizes and combines features from principal component analysis and multiple regression (16,17). The method is particularly well suited for calibration on a small number of samples with experimental noise in both dependent and independent variables. In addition, the method can be used even if the number of observations is smaller than the number of wavelengths (4). The PLS method consists of a projection of the data into a low-dimensional space formed by a set of orthogonal latent variables by a simultaneous decomposition of *X* (spectral matrix) and *Y* (chemical concentration matrix) that maximizes the covariance between *X* and *Y* (18). The PLS regression analysis outputs an offset and a regression coefficient for each spectral band. The estimated chemical concentration for a spectrum is calculated by linear combination of the spectrum and the regression coefficients and addition of the offset. Two sets of regression coefficients are shown exemplarily in Figure 2. The weight of certain wavelengths for the estimation of a variable can be deduced from the plots. PLS regression was carried out using the Matlab Statistics Toolbox (Version 2008b, The Mathworks). The number of components required for a good estimation without overfitting was determined by visual inspection of the resulting chemometric maps and by calculating cross-validated *RMSE*, the Akaike Information Criterion (AIC), and the Schwarz-Bayes Criterion (*SBC*):

$$AIC = ln(\sigma_{Res}^2) + \frac{M}{n} \cdot 2$$

$$SBC = ln(\sigma_{Res}^2) + \frac{M}{n} \cdot \ln n$$

with

σ_{Res}^2 – variance of residuals

M – number of estimated parameters

n – number of observations.

Learning curves that were used to determine the number of components used in the chemometric mapping are depicted in Figure 1. The *RMSE* values of the SVR estimations (Figure 1 left) show a more or less steep decline with growing number of considered bands. In the case of Al, Fe, and C a constant level is reached at about 25 bands, while the Mn and N curves do not saturate as clearly. While Fe and C, which are known to be optically active, reaching nearly zero *RMSE* in the SVR estimations, the other constituents that can probably only be indirectly detected, stay well above that mark. We always chose the lowest possible number of bands that still resulted in visually reasonable looking maps.

For the PLSR models the *AIC* and *SBC* were calculated in addition to the *RMSE* (Figure 1, right). These information criteria have the advantage over *RMSE* that they usually have a minimum that identifies the optimal number of variables in the model. *SBC* is more strict in the way additional parameters are punished, so that the *SBC* curve is always above the *AIC* curve after the one-parameter model. The more latent variables are considered in a PLSR model, the more complex it becomes. This is illustrated exemplarily by Figure 2 which shows the regression coefficients for all spectral channels in the PLSR models for Al estimation with 4 and with 10 latent variables.

Two samples for the organic-rich top horizons contained very high concentrations of carbon and nitrogen. Because no reliable robust models could be established with these outliers, they were excluded from model calibration. In consequence, the results should not be considered valid for regions of very high carbon concentration.

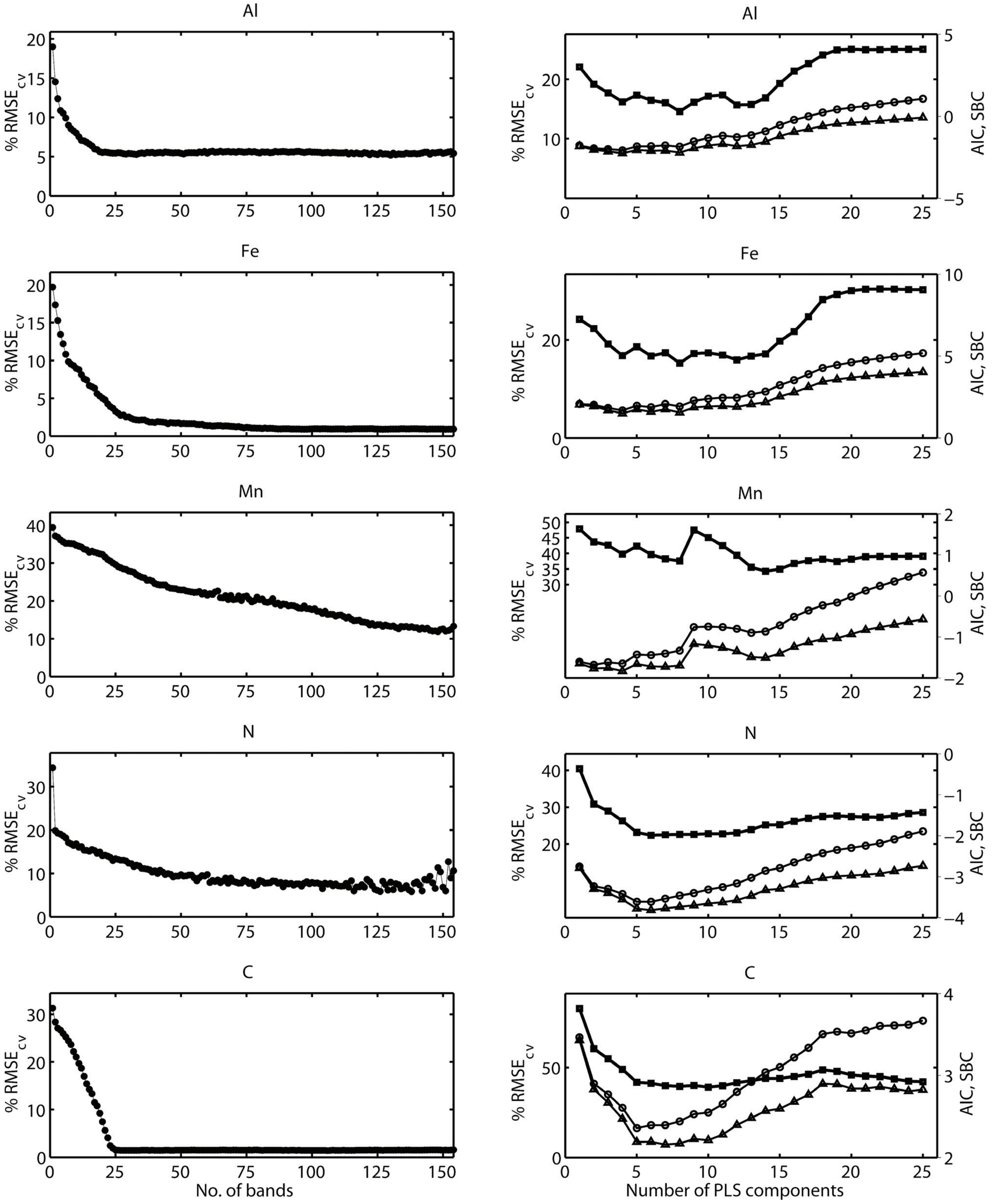

Figure 1: Learning curves for SVM regression (left) and PLS regression (right). PLSR accuracy is given in terms of %RMSE (squares), AIC (triangles), and SBC (circles).

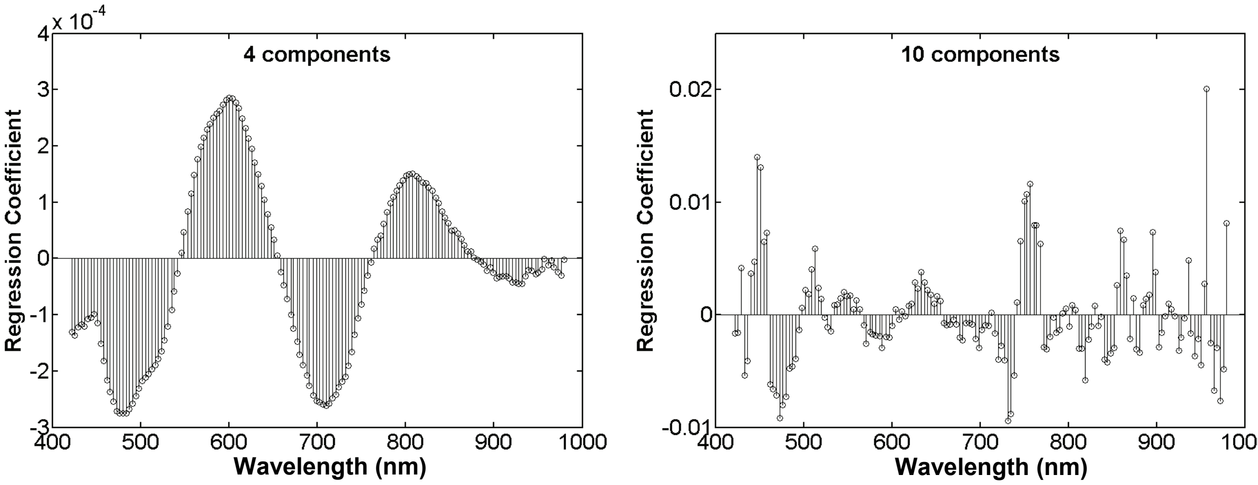

Figure 2: Regression coefficients in the PLSR models for Al estimation with 4 components (left) and 10 components (right).

RESULTS

Applying the statistical models described in the previous section to the images resulted in detailed maps of the chemical properties. The maps clearly showed the heterogeneity of the chemical properties and the different structures in the soil horizons. Optically active constituents like organic carbon showed less noisy maps than constituents that are only indirectly linked to absorption features.

Table 1: Regression accuracy of the chosen PLS and SVM regression models

Element	SVM-R				PLSR			
	R²	%RMSE	RMSE	# comp.	R²	%RMSE	RMSE	# comp.
C	0.97	17.34	1.14	30	0.81	39.57	2.59	8
N	0.98	9.07	0.04	23	0.74	31.14	0.013	6
Fe	0.96	5.09	0.55	8	0.53	16.80	1.83	4
Al	0.95	6.84	0.12	18	0.45	16.17	0.28	4
Mn	0.58	34.48	0.30	30	0.50	38.22	0.33	7

Regression accuracies of the models are stated in Table 1. All models are unbiased and at least of acceptable quality. PLS regression models generally have lower accuracy values than SVM regression models. The chosen number of components in the SVM-R models is higher than in the PLSR models, but the influence of the number of components is lower. Figure 3 shows a true-colour depiction of the hyperspectral image of the soil profile’s first layer with centimetre scales at the left. The other two images show the spatial estimations of organic carbon content as calculated using SVM-R (center) and PLSR (right). The most striking difference between these carbon maps is that in the lower part of the SVM-R image shadows seem to have been identified as high carbon concentrations, while this is not the case for the PLSR image. Carbon concentrations in the top 2 centimetres may be invalid because training areas from the carbon rich horizon were excluded from model calibration.

Figure 4 shows the cross-validated estimations for C concentrations of the sampled areas by both regression models. Most SVM-R estimated values lie within a narrow tube around the 1:1 line, as can be observed in other similar studies (9). Only values further than a small ε from the 1:1 line become support vectors for the SVR (14). PLSR estimations exhibit a larger scattering around the 1:1 line, but are also unbiased. The same general behaviour can be observed in the estimations of the other chemicals.

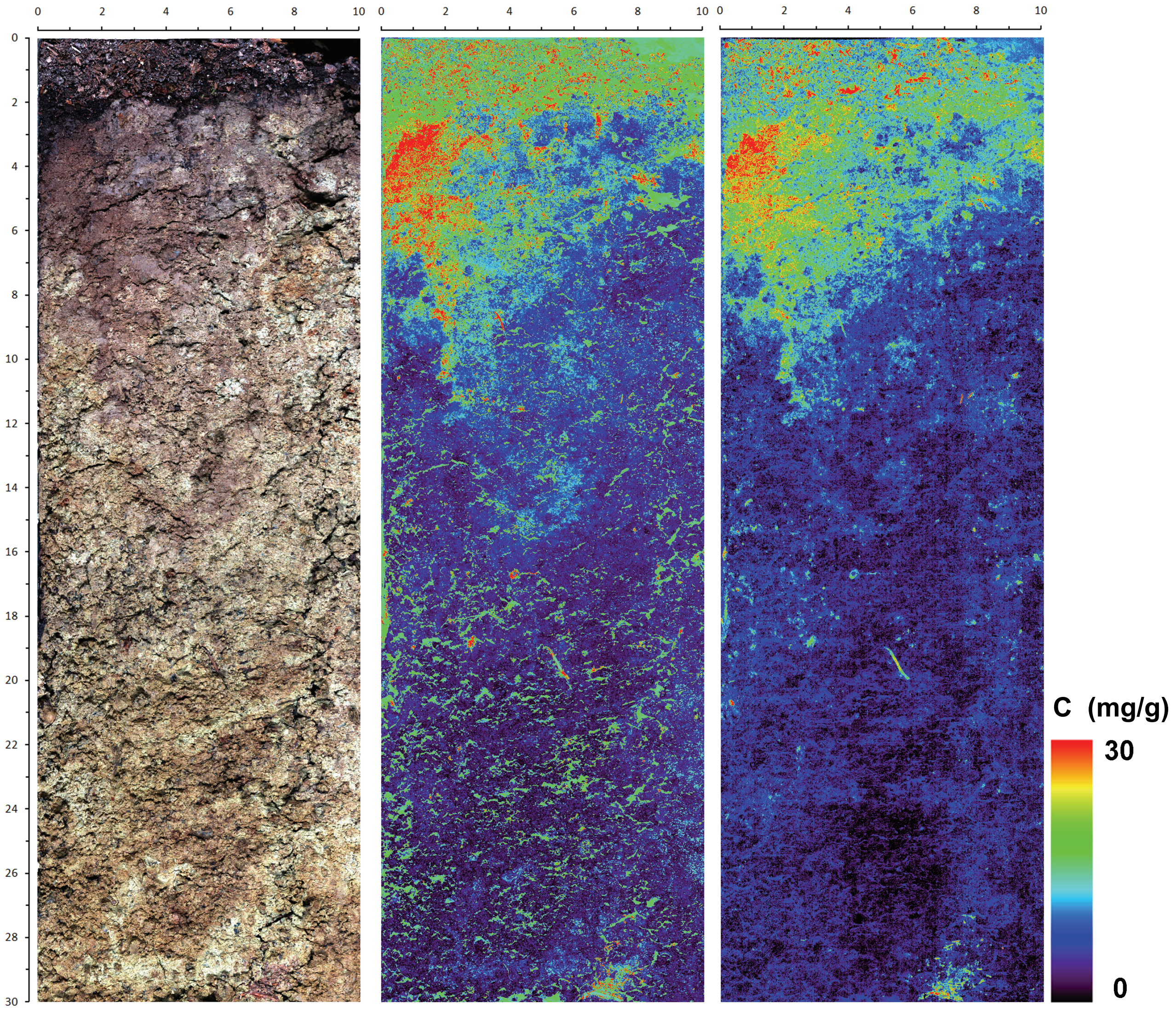

Figure 3: Left: True-colour image of soil profile with centimetre scales, middle: SVM regression of C content, right: PLS regression of C content.

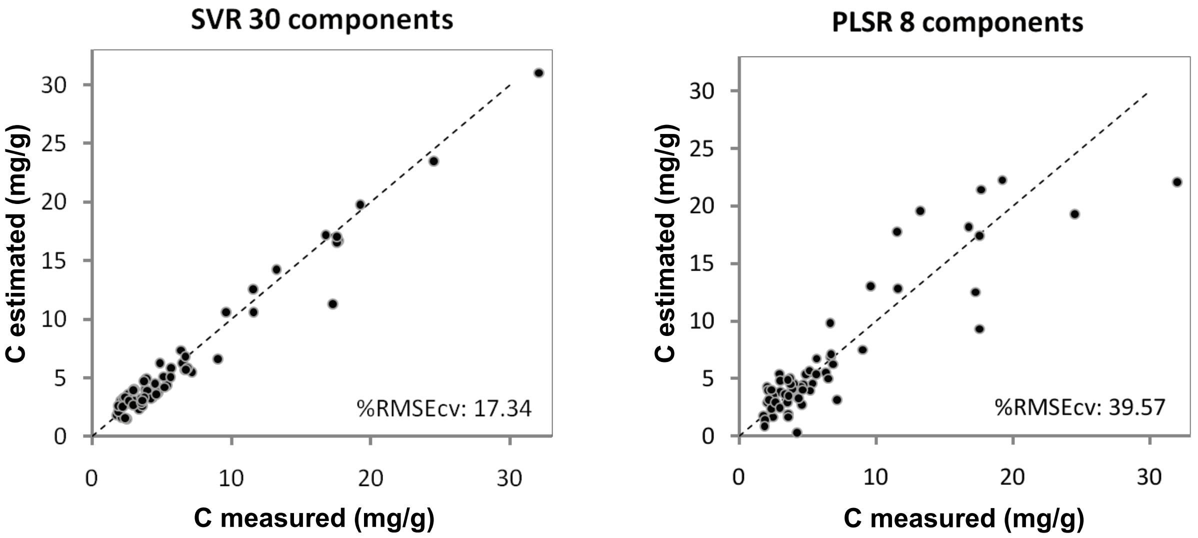

Figure 4: Measured and estimated carbon concentrations, modelled with SVR (left) and PLSR (right), with 1:1 lines.

CONCLUSIONS AND OUTLOOK

A novel technique for analysing soils in up to 30 cm depth by laboratory imaging spectroscopy is introduced. This paper deals with analyses of the chemical properties of the soil profiles, but further analyses of the laboratory imaging spectroscopy data include horizon and inclusion classifications, analyses of the small-scale structure and high-resolution mapping of soil colour. The technique of laboratory imaging spectroscopy is extendable to further objects of geo- and soil-scientific interest like drill cores, stone samples and leaves.

Both PLSR and SVR are appropriate chemometric techniques for the applications presented. Although the cross-validated *R²* and *RMSE* values of SVR are better than those of PLSR, in most cases the maps produced by PLSR seem more realistic. Further examinations of the chemometric techniques and of spectral pre-processing steps like continuum removal, spectral derivatives and standard normal variate with de-trending will be explored in the near future with this data set and with other soil cores that have also been hyperspectrally imaged the same way.

ACKNOWLEDGEMENTS

Hans and Florian Steffens are gratefully acknowledged for technical assistance (welding the steel box). Joachim Hill is acknowledged for providing the imaging spectrometer. We thank the anonymous reviewers for helpful comments on the manuscript.

REFERENCES

1 Stoner E R & M F Baumgardner, 1981. Characteristic variations in reflectance of surface soils. Soil Science Society of America Journal, 45: 1161-1165

2 Ben-Dor E, S Chabrillat, J A M Demattê, G R Taylor, J Hill, M L Whiting & S Sommer, 2009. Using imaging spectroscopy to study soil properties. Remote Sensing of Environment, 113: S38-S55

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

4 Ben-Dor E, D Heller & A Chudnovsky, 2008. A novel method of classifying soil profiles in the field using optical means. Soil Science Society of America Journal, 72: 1113-1123

5 Atzberger C, M Guérif, F Baret & W Werner, 2010. Comparative analysis of three chemometric techniques for the spectroradiometric assessment of canopy chlorophyll content in winter wheat. Computers and Electronics in Agriculture, 73: 165-173

6 IUSS Working Group WRB, 2006. World Reference Base for Soil Resources 2006 (FAO, Rome) 145 pp.

7 Buddenbaum H & M Steffens, 2011. Laboratory imaging spectroscopy of soil profiles. Journal of Spectral Imaging, 2(a2), 1-5

8 Schlerf M, C Atzberger, J Hill, H Buddenbaum, W Werner & G Schüler. 2010. Retrieval of chlorophyll and nitrogen in Norway spruce (*Picea abies* L. Karst.) using imaging spectroscopy. International Journal of Applied Earth Observation and Geoinformation, 12: 17-26

9 Durbha S S, R L King & N H Younan, 2007. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Remote Sensing of Environment, 107: 348-361

10 Smola A J & B Schölkopf, 2004. A tutorial on support vector regression. Statistics and Computing, 14: 199-222

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

12 van der Linden S & P Hostert, 2009. The influence of urban structures on impervious surface maps from airborne hyperspectral data. Remote Sensing of Environment, 113: 2298-2305

13 Waske B, S van der Linden, JA Benediktsson, A Rabe & P Hostert, 2010. Sensitivity of Support Vector Machines to random feature selection in classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 48: 2880-2889

14 Rabe A, S van der Linden & P Hostert, 2009. Simplifying support vector machines for regression analysis of hyperspectral imagery. Proceedings of the 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers 2009, Grenoble)

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

16 Udelhoven T, C Emmerling & T Jarmer, 2003. Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study. Plant and Soil, 251: 319-329

17 Farifteh J, F D Van der Meer, C Atzberger & E J M Carranza, 2007. Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). Remote Sensing of Environment, 110: 59-78

18 Vohland M & C Emmerling, 2011. Determination of total soil organic C and hot water-extractable C from VIS-NIR soil reflectance with partial least squares regression and spectral feature selection techniques. European Journal of Soil Science, 62: 598-606