Topsoil mapping using hyperspectral airborne data and multivariate regression modeling

Thomas Selige, Urs Schmidhalter
Chair of Plant Nutrition, Department of Plant Sciences, Technical University Munich,
Am Hochanger 2, 85350 Freising, Germany
selige@wzw.tum.de

Abstract

The spatial variability of topsoil texture and organic matter across fields was studied using airborne hyperspectral imagery to lead towards improved fine-scale soil mapping procedures. Two important topsoil features for precision farming applications, soil organic matter and texture, were correlated with spectral properties of the airborne HyMap scanner. Sand, clay, organic carbon and total nitrogen contents can be predicted quantitatively and simultaneously by a multivariate calibration approach using Partial Least Square Regression or Multiple Linear Regression. The suite of topsoil parameters can be determined simultaneously from a single spectral signature since the various features are represented by varying combinations of wavebands across the spectra.

Keywords: hyperspectral airborne data, topsoil mapping, multivariate regression, soil texture, soil organic matter

Introduction

Topsoils frequently show a fine tessellated pattern and heterogeneity across fields indicated by e.g. color, roughness, infiltration, erosion and surface sealing phenomena. Consequently topsoil heterogeneity causes differences in crop germination, nutrient and water uptake and thus markedly influence crop growth and plant coverage. This has implications for the pattern and the spatial extent of appropriate land use management practices and soil conservation strategies including site specific management in precision agriculture systems. For optimizing crop growth, soil tillage, seed bed preparation, fertilization and herbicide use in particular must be adapted to the local topsoil properties.

However, there is still no effective way to map fine scale soil heterogeneity so as to derive site specific data about topsoil physical/chemical characteristics. Several authors established relationships between soil spectral reflectance data and organic matter characteristics (Dalal & Henry 1986, Udelhoven *et al.* 2003) and soil texture (Al-Abbas *et al.*,1972, Ben-Dor & Banin 1995). Both groups of parameters play an interdependent and decisive role in assessing topsoil characteristics e.g. soil aggregation, aggregate stability and resistance to water and wind erosion (Neemann 1991) and as a consequence it would be an advantage to be able to map both sets of physical characteristics from the one set of image data.

The aim of the work reported here was to develop a method of mapping fine scale topsoil organic and texture parameters from a combination of field and hyperspectral image data. The work investigated the use of both Multiple Linear Regression and Partial Least Squares Regression to construct the models to estimate the soil physical/chemical variables from the image data. Field data was combined with the image data for the construction of the models. This innovative approach to digital soil mapping achieves the simultaneous estimation of a suite of topsoil parameters. The use of high resolution remotely sensed data avoids the need for interpolation with its attendant problems of accurate and reliable spatial prediction.

Material and methods

Study area

The East German study area *Wulfen* (11°55'E, 51°49'N) is characterised by a slightly undulated tertiary plain at 70 m altitude that is covered by a thin Loess layer up to 1.2 m and an alluvial plain (glacial valley) of the river Elbe at 50 m altitude that served as origin for the Aeolian deposit in the tertiary plain. The predominant soil type of the Loess covered Tertiary plain is Chernozem in conjunction with Cambisols and Luvisols. The alluvial plain is characterised by coarse sand to fine sand, loamy and clayey sediments. The predominant soil types are Mollic Gleysols, Fluvisols and Planosols. The fine-scale pattern of soil texture and organic matter within the fields of the landscapes results in highly diverse soil properties and virtually forces the application of site specific management.

Data sets

The remotely sensed data was acquired using the HyMap™ scanner (Integrated Spectronics Pty Ltd, Australia), installed on a Cessna Caravan aircraft. The scanner records spectra from 420 nm to 2480 nm wavelength, in 128 wavebands with full width half maximum (FWHM) bands of 15 and 20 nm for the (420 - 1803 nm range) and (1949 - 2480 nm range) ranges respectively. The imagery was acquired with 6m nadir pixels and a swath width of 30 degrees from 12:30 to 13:15 hours on 19th May 1999. The data was atmospherically and geometrically corrected using the ATCOR procedures (Richter & Schläpfer 2002) and a rectification procedure (Schläpfer & Richter 2002).

To ensure a most representative calibration soil data set, we did not focus on extended data sampling but rather designed the sample selection procedure in terms of soil forming geo-factors and factor combinations across the test site's landscapes. To represent a relevant spectrum of soil organic matter (SOM) and texture that is frequently found with arable soil, the study was based on 72 samples. From these samples, 46 were on 12 bare soil fields across the study area and used as a calibration test set for multivariate regression modeling purposes. For validation purposes (see also application and discussion paragraph), we additionally sampled two subsets of 16 and 12 samples respectively. All samples were passed through a 2 mm sieve and were air dried. The soils were analyzed for total amount of organic carbon (C_{org}) and the total amount of nitrogen (N_t) by

Table 1. Partial Least Square Regression (PLSR) and Multiple Linear Regression (MLR) models statistics for the different topsoil parameters.

Para-meter	Range (%)	n =	PLSR			MLR		
			Factors	R^2	RMSECV	Wavelength (nm)	R^2	RMSECV
C_{org}	0.7 - 3.85	46	7	0.90	0.29	800, 830, 1194, 1322	0.86	0.22
N_t	0.07 - 0.36	46	7	0.92	0.026	1194, 2115, 2185, 2220	0.87	0.019
Sand	16 - 88	46	9	0.95	9.7	2202, 2238, 2322, 2371	0.87	12.9
Clay	7 - 22	46	5	0.71	4.2	902, 950, 998, 1165	0.65	3.8

dry combustion using an elemental analyzer. The particle size distribution was analyzed using sieve analysis for the sand fractions and the coarse silt fraction and pipette analysis for the fine fractions of silt and clay. The min-max ranges of the soil data are shown in Table 1. Effects of soil surface moisture and roughness were excluded from this study by selecting bare soil fields after seed bed preparation and organizing the flight campaign after a period of soil surface drying.

Methodology

In this paper, we focus on the development of spectral model calibration towards large scale soil mapping of C_{org}, N_t and the sand and clay contents. Complexity of soil forming and therefore expected auto-correlations between soil parameters, as were found for example between C_{org} and N_t, has led to the use of multivariate calibration techniques. Two multivariate regression techniques were used to develop the models to estimate soil parameters from the hyperspectral image data. These techniques allow the simultaneous quantitative determination of several soil parameter from individual spectral signatures. Multivariate calibration was performed using Multiple Linear Regression (MLR) and Partial Least Square Regression (PLSR). Our goal was to test the applicability of multivariate regression techniques in hyperspectral remote sensing rather than to compare the statistical level of model fit. Thus a more simple (MLR) and a more sophisticated and data compression technique (PLSR) were chosen.

To optimise the derived PLSR model and ensure that it is robust against variability of natural factors, the whole spectra should be considered (Dardenne, 1996). Similarly, it would be usual to include all of the possibly occurring variations of natural factor combinations in the model construction so as to achieve a robust model algorithm (Schenk & Westerhaus, 1991). PLSR reduced the reflectance spectra to a few relevant factors and regressed them to the soil value of a given sample (Martens & Næs, 1989). The PLSR algorithm will automatically give high weights to the decisive wavelength regions and low or zero weights to uninformative wavelengths provided that the spectral and natural variability included in the calibration set is high enough. As distinct from PLSR, MLR performs a regression model selecting and combining the most significant wavebands from the spectrum. The methods are discussed in detail in Martens & Næs (1989) and Næs *et al.* (2002) respectively.

From these regression procedures, models were derived enabling prediction of the soil value from the spectra of samples with unknown soil value. Various wavelength regions and data pre-treatments were analysed using an optimization routine to find the best calibration algorithm. The algorithms with the lowest root mean square error of cross validation (RMSECV) was chosen as statistically the best. Since several data pre-treatments result in similar error, those were chosen that had the lowest number of factors included in the regression model (Næs *et al.*, 2002).

Finally, the optimal calibration model for each soil parameter was used to predict the particular parameter from the HyMap spectra for each HyMap image pixel of bare soils resulting in a map showing the distribution of the topsoil parameter. For independent prediction of the respective parameter, it was desirable similarly to find calibration algorithms for the different parameters that depend on different wavebands. The spectrum of each of the sample sites was extracted from the image data. For this, we applied a seeded region growing algorithm to identify the spectrally most similar neighboring pixels. From the spectra of these pixels, the mean spectrum of each site was calculated. The channel at 1949 nm was excluded from the spectra due to insufficient signal to noise ratio at this water induced absorption band.

Results

Calibration procedure by PLSR

The PLSR algorithm will automatically give high weights to the decisive wavelength regions and low or zero weights to uninformative wavelengths provided that the spectral and natural variability included in the calibration set is high enough. PLSR reduces the whole reflectance spectra to a few relevant factors and regresses them to the measured parameter of a given sample. While doing so, it is recommended to use a calibration design which covers the whole range of possible values (Brereton 2000). It is also reported that a certain redundancy within the spectra is useful to stabilize PLSR models against noise (Martens & Næs 1989) and therefore have to be considered so as to be more robust than MLR calibration models. This might be relevant in particular for SOM parameters, since SOM has numerous broad overlapping absorption features located throughout the spectra and consequently influence the overall shape of the reflectance curve (Baumgartner *et al.* 1985). Thus, multivariate calibration was performed with PLSR first.
From the aforementioned regression, a model was derived enabling prediction of the C_{org} content from the spectra of samples with unknown C_{org}. The same calibration procedure was also employed to derive a prediction model for the N_t content as well as prediction models for the sand and clay contents. All wavelength regions of the spectra and different data pre-treatments were analyzed using an optimization routine to find the best calibration algorithm. Overall the min-max normalisation gave the best results for the different data pre-treatments. The PLSR models statistics are compiled in Table 1.

Calibration procedure by MLR

Since the PLSR procedure reduces the spectra to a few factors, it does not support the idea of identifying the most significant individual wavebands. Thus, we also applied multivariate calibration by MLR. Due to the dramatic increase of calculation time with the number of regression variables, the selection of algorithms was limited for the time being to regression models with a maximum of 4 spectral variables.
Thus, all possible subsets of regression models with at least 1 and up to any combination of 2, 3 and 4 spectral variables were calculated for each of the selected soil parameters. The algorithm with the lowest RMSECV was chosen as statistically the best for each soil variable. As before with the PLSR calibration, overall the min-max normalization gave the best results for the different data pre-treatments. The MLR models statistics are compiled in Table 1 as well.
The best MLR C_{org} model [1] was based on the spectral channels C_{26} (wavelength: 800 nm), C_{28} (830 nm), C_{52} (1194 nm) and C_{62} (1322 nm):

$$C_{ORG} = 3.5688 - 0.0318\ C_{26} + 0.0362\ C_{28} - 0.0173\ C_{52} + 0.0122\ C_{62} \qquad [\%] \qquad (1)$$

The best MLR Nt model [2] was based on the spectral channels C52 (wavelength: 1194 nm), C106 (2115 nm), C110 (2185 nm) and C112 (2220 nm):

$$N_T = 0.3691 - 0.0001529\ C_{52} - 0.0007059\ C_{106} + 0.002087\ C_{110} + 0.001208\ C_{112} \qquad [\%] \qquad (2)$$

The calibration model output is shown versus the measured reference values for C_{org} in Figure 1 and for N_t in Figure 2. Both models are in the same range of quality expressed by R^2 and RMSECV. Most channels are different in the two models, except for channel 52 at 1194 nm mean wavelength. This underlines the independency of the C_{org} and the N_t prediction from the spectral domain.

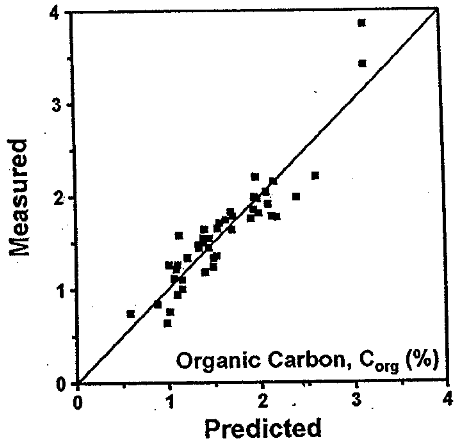

Figure 1. Predicted organic carbon (C_{org}) from calibrated model versus C_{org} values measured by dry combustion reference method.

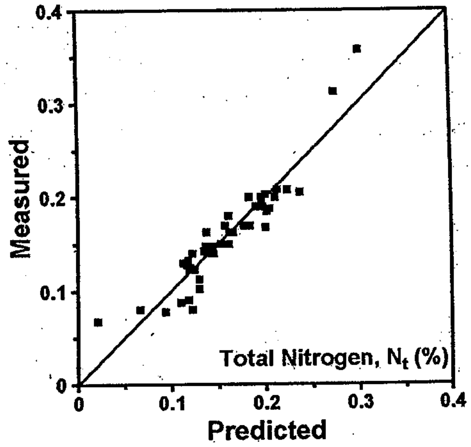

Figure 2. Predicted total nitrogen (N_t) from calibrated model versus N_t values measured by dry combustion reference method.

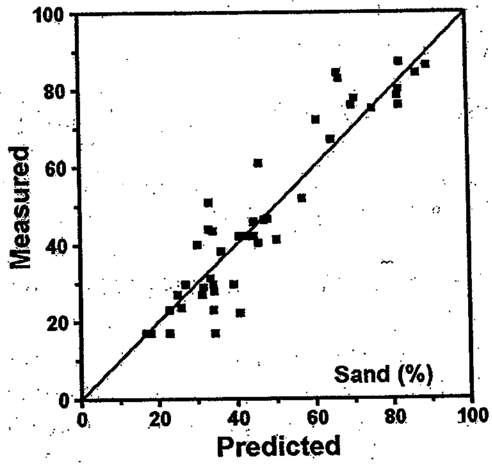

Figure 3. Predicted sand content from calibrated model versus sand values measured by sieve analysis.

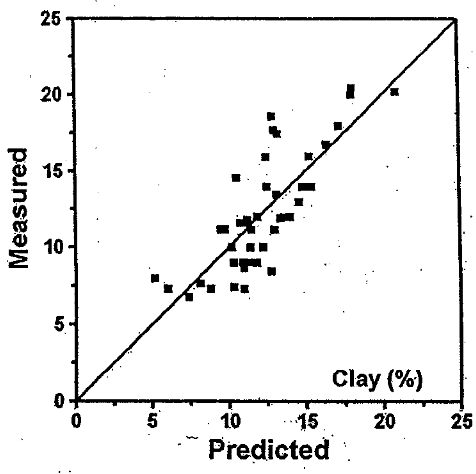

Figure 4. Predicted clay content from calibrated model versus clay values measured by pipette analysis.

The best MLR sand model [3] was based on the spectral channels C_{111} (wavelength: 2202 nm), C_{113} (2238 nm), C_{118} (2322 nm) and C_{121} (2371 nm):

$$SAND = 3.3 - 0.68\ C_{111} + 1.15\ C_{113} - 0.76\ C_{118} + 0.3\ C_{121} \qquad [\%] \qquad (3)$$

The best MLR clay model [4] was based on the spectral channels C34 (wavelength: 902 nm), C37 (950 nm), C40 (998 nm) and C51 (1165 nm):

$$CLAY = 19.48 - 0.19\ C_{34} + 0.14\ C_{37} + 0.14\ C_{40} - 0.09\ C_{51} \qquad [\%] \qquad (4)$$

The calibration model output is shown versus the measured reference values for sand in Figure 3 and for clay in Figure 4. The sand model fits well to the calibration data as expressed by R^2 and RMSECV, whereas the clay model is characterized by a much weaker regression fit. Both texture models employ different wavebands likewise as against the SOM parameters models. All four models are independent in their parameter prediction from the spectral domain.

Application

The optimal calibration models were used to calculate the C_{org}, N_t, sand and clay contents from the HyMap spectra of each HyMap image pixel. Both the PLSR and the MLR calibration procedures led to models which are characterized by comparable results in terms of R^2 and RMSECV. As the MLR models were more suitable for simple grid operations, they were used to calculate C_{org}, N_t, sand and clay maps of reference fields. Resulting topsoil maps (Figures 5 to 8) show the distributions of C_{org}, N_t, sand and clay contents across an 88 ha-sized reference field as an example. The C_{org} values (Figure 5) range from 1.2 % (light color) to 2.5 % (dark color). The N_t values (Figure 6) range from 0.13 % (light color) to 0.24 % (dark color). The sand values (Figure 7) range from 10 % (light color) to 50 % (dark color). The clay values (Figure 8) range from 5 % (light color) to 20 % (dark color). As silt content supplements sand and clay to 100 % (silt = 100-sand-clay), a silt map was calculated from the sand and clay map (Figure 9). A silt model was attempted so far.

A validation test using 12 independent samples from this field (not used in the calibration procedures) indicated a close relation of $R^2 = 0.89^{***}$ (*** = significant level at P <0,001) and

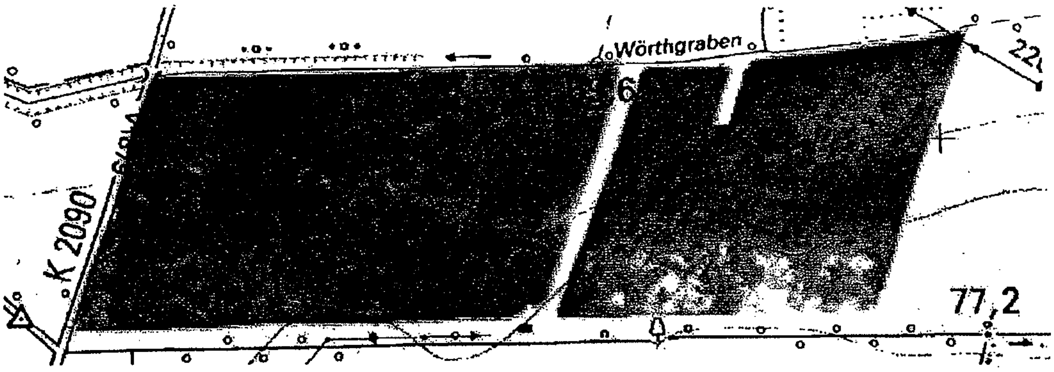

Figure 5. Spatial distribution of the organic carbon content (Corg) across the 88 ha reference field "Pfingstbreite" (range from 1.2 % = light color to 2.5 % = dark color)

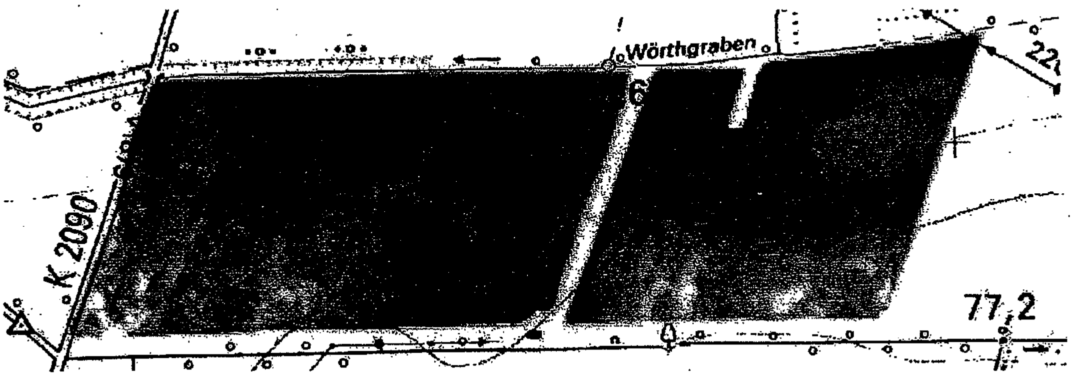

Figure 6. Spatial distribution of the total nitrogen content (Nt) across the 88 ha reference field "Pfingstbreite" (range from 0.13 % = light color to 0.24 % = dark color)

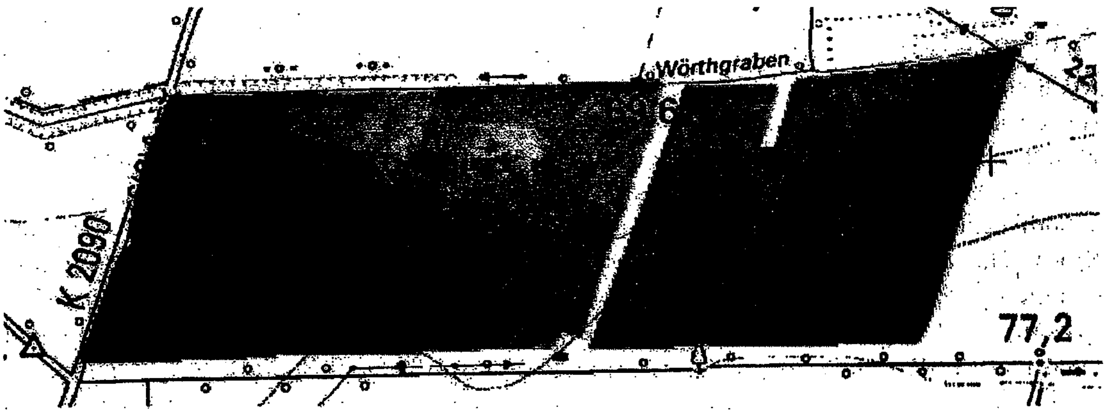

Figure 7. Spatial distribution of the sand content across the 88 ha reference field "Pfingstbreite" (range from 10 % = light color to 50 % = dark color)

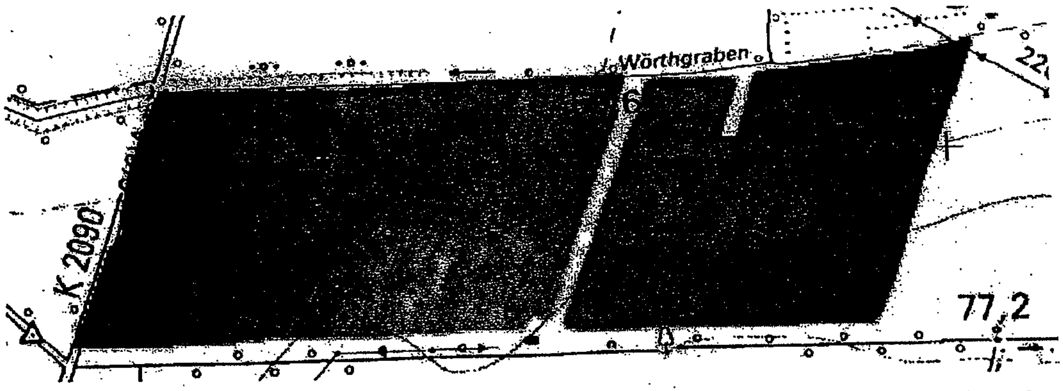

Figure 8. Spatial distribution of the clay content across the 88 ha reference field "Pfingstbreite" (range from 5 % = light color to 20 % = dark color)

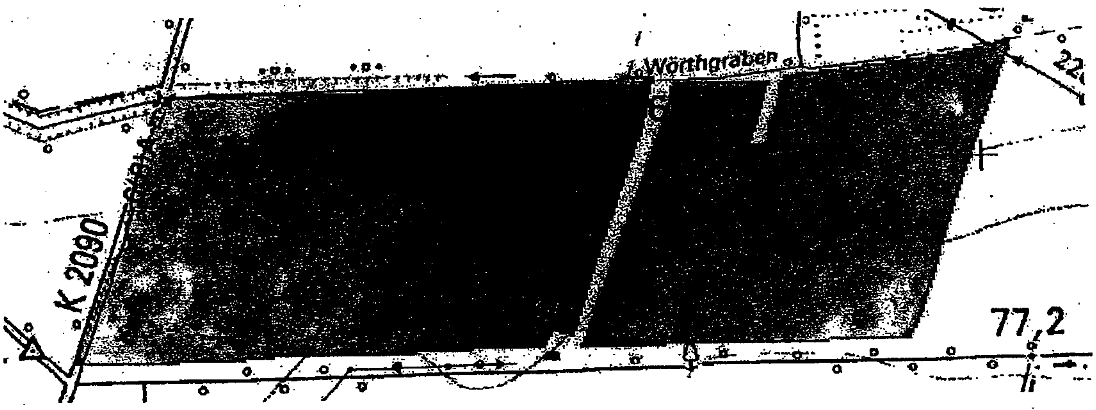

Figure 9. Spatial distribution of the silt content across the 88 ha reference field "Pfingstbreite" (range from 30 % = light color to 75 % = dark color)

$R^2 = 0.91$***, respectively between the C_{org} and N_t content predicted from the Hymap data using the MLR models and the C_{org} and N_t content measured by dry combustion method as standard reference value. The validation test gave relations of $R^2 = 0.94$*** and $R^2 = 0.64$***, respectively between the sand and clay content predicted from the Hymap data using the MLR models and the sand and clay) content measured by sieve and pipette analysis as standard reference value.

With both calibration techniques, PLSR and MLR, we achieved similar results for R^2 and RMSECV. This is an unexpected result as there is generally multi-collinearity and strong auto-correlation between soil and spectral data. These characteristics have been found to cause a problem when using MLR but not when using PLSR (Næs *et al.*, 2002) and as a consequence it was expected that PLSR would give a statistically better result than MLR. Further investigation is required of this issue, for which an extended data set over that used in this study would be necessary.

The relatively weak regression model for clay content might be attributed to the relatively narrow range of data values. The effect of soil surface moisture was excluded from this study by organizing flight campaigns after crop seeding and a period of bare soil surface drying.

Conclusion

This remote sensing approach shows the potential benefits of using image data with carefully located in-situ field data in digital soil mapping. The results also indicate that soil mapping procedures must be adapted to the soil parameter of interest and that multivariate calibration techniques allow calibration modelling as a generic procedure. With the HyMap™ spectrometer, the wavebands that are relevant to the mapping of the different soil parameters can be recorded and used in adapted calibration models to simultaneously predict a suite of different soil parameters. The method proposed provides a means of simultaneously estimating topsoil SOM and texture in an extensive, rapid and non-destructive manner, whilst avoiding the spatial accuracy problems associated with interpolation. The use of remotely sensed data in the manner proposed in this paper can also be used to monitor and better understand the influence of management and land use practices on SOM composition and content. In precision agriculture, it can be used to establish the precise spatial locations of specific management practices, as a pre-requisite to much of the modeling and estimation that needs to be conducted for variable rate applications. How this modeling and estimation is best integrated with site specific management has not yet been completely resolved.

Acknowledgment

This study was partly funded by the German Federal Ministry of Education and Research (BMBF)

References

Al-Abbas, A. H., Swain, P. H., and Baumgartner, M. F. 1972. Relating organic matter and clay content to the multispectral radiance of soils. Soil Science 114 477-485.

Baumgartner, M.F., Silva, L.F., Biehl, L.L., Stoner, E., 1985. Reflectance properties of soils. Advances in Agronomy 38 1-44.

Ben-Dor, E. and 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.

Brereton,R.G., 2000. Introduction to multivariate calibration in analytical chemistry. Analyst 125 2125-2154.

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.

Dardenne, P. 1996. Stability of NIR spectroscopy equations. NIR news 7 (5) 8-9.

Martens, H. and Næs, T. 1989. Multivariate calibration. John Wiley & Sons, Chichester, UK.

Næs, T., Isaksson, T. Fearn, T. and Davies, T. 2002. A User-Friendly Guide to Multivariate Calibration and Classification. NIR-Publication, Chichester, UK.

Neemann, W. 1991. Bestimmung des Bodenerodierbarkeitsfaktors für winderosionsgefährdete Böden Norddeutschlands (Determination of soil erodibility factors for wind-erosion endangered soils in Northern Germany). Geologisches Jahrbuch, Reihe F 25, 131 pp.

Richter, R., D. Schläpfer (2002). Geo-atmospheric processing of airborne imaging spectrometry data. Part 2: atmospheric/topographic correction. International Journal of Remote Sensing 23 2631-2649.

Schenk, J. S. and Westerhaus, M. O. 1991. Population definition, sample selection, and calibration procedures for near infrared spectroscopy. Crop Science 31 469-474.

Schläpfer, D., R. Richter (2002). Geo-atmospheric processing of airborne imaging spectrometry data. Part 1: parametric orthorectification. International Journal of Remote Sensing. 23 2609-2630.

Udelhoven, T., Emmerling, C., Jarmer, T., 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.