QUANTITATIVE ASSESSMENT OF SOIL PARAMETERS IN WESTERN TAJIKISTAN USING A SOIL SPECTRAL LIBRARY APPROACH

Bruno Seiler[a*], Mathias Kneubühler[a], Bettina Wolfgramm[b], Klaus I. Itten[a]

[a] Remote Sensing Laboratories (RSL), Department of Geography, University of Zürich
Winterthurerstrasse 190, 8057 Zürich, Switzerland, E-mail: bseiler@geo.unizh.ch
[b] Centre for Development and Environment, Institute of Geography,
University of Berne, Steigerhubelstrasse 3, 3008 Berne, Switzerland

KEY WORDS: Soil, Sampling, Hyper spectral, Statistics, Correlation, Modelling, Prediction

ABSTRACT:

Soil degradation is a major problem in the agriculturally dominated country of Tajikistan, which makes it necessary to determine and monitor the state of soils. For this purpose a soil spectral library was established as it enables the determination of soil properties with relatively low costs and effort. A total of 1465 soil samples were collected from three 10x10 km test sites in western Tajikistan. The diffuse reflectance of the samples was measured with a FieldSpec PRO FR from ASD in the spectral range from 380 to 2500 nm in laboratory. 166 samples were finally selected based on their spectral information and analysed on total C and N, organic C, pH, $CaCO_3$, extractable P, exchangeable Ca, Mg and K, and the fractions clay, silt and sand. Multiple linear regression was used to set up the models. Two third of the chemically analysed samples were used to calibrate the models, one third was used for hold-out validation. Very good prediction accuracy was obtained for *total C* ($R^2 = 0.76$, RMSEP = 4.36 g kg^{-1}), *total N* ($R^2 = 0.83$, RMSEP = 0.30 g kg^{-1}) and *organic C* ($R^2 = 0.81$, RMSEP = 3.30 g kg^{-1}), good accuracy for *pH* ($R^2 = 0.61$, RMSEP = 0.157) and $CaCO_3$ ($R^2 = 0.72$, RMSEP = 4.63 %). No models could be developed for *extractable P*, *exchangeable Ca*, *Mg* and *K*, and the fractions *clay*, *silt* and *sand*. It can be concluded that the spectral library approach has a high potential to substitute standard laboratory methods where rapid and inexpensive analysis is required.

1. INTRODUCTION

Soil degradation is a major problem in the agriculturally dominated country of Tajikistan. In the 1990s, increasing poverty triggered by the civil war and the transformation of the economy led to widespread cultivation of steep slopes, formerly used as grazing land. In these areas water erosion is considered to be the fastest and most widespread soil degradation process, also having a highly negative impact on soil fertility (Wolfgramm, 2007). It is therefore necessary to determine and monitor the state of soils. The soil spectral library approach is highly suitable for this task as it enables the determination of soil properties for a big number of soil samples with relatively low costs and effort. The objectives of the present work were:

- to model several soil properties based on the spectral information of the soil samples with satisfying accuracy.
- to evaluate the predictive ability of three approaches *multiple linear regression with continuum removed spectra*, *principal component regression* and *regression tree with first derivatives of spectra*.

2. MATERIALS AND METHODS

A total of 1465 soil samples were collected from three 10 to 10 km test areas in the vicinity of the Tajik capital city Dushanbe using a random sampling scheme (Wolfgramm, 2007).

The soil samples were air-dried. 50 g of each soil sample were weighed to be further prepared for the spectrometer measurements and chemical analysis. Further preparation included grinding and sieving through a 2 mm sieve to minimize differences in grain size.

The spectral readings were conducted with an ASD spectrometer, a FieldSpec Pro FR. The instrument has the ability to detect light in the spectral range from 350 to 2500 nm. This includes the spectral regions VIS (350-700 nm), NIR (700-1400nm) and a large part of SWIR (1400-3000 nm) For the purpose of standardization a special lamp was used to illuminate the samples, a Muglight from ASD (ASD, 2007b). Its design minimizes measurement errors associated with stray light and specular reflected components as the sample is illuminated and the reflected radiance is measured from below through a small window. For the white reference measurement a Spectralon from Labsphere was used. The fore-optic field of view of the spectrometer was set to 8 degrees, the scans to be averaged for white reference to 10 and for dark current to 25. A 1 nm sampling resolution was used for the measurements. The soil samples were measured twice with an approximate 90 degree turning in between and the two measurements were averaged. Spectral bands with a signal to noise ratio lower than 90 were removed from the data set. This affected the wavelengths from 350 to 430 and 2440 to 2500 nm. The spectral data was resampled to a 10 nm resolution by keeping every tenth nanometer value to reduce the amount of data. Scattering differences due to different grain size distributions were removed using two methods: *continuum removal* (ENVI, 2007) and *first*

* Corresponding author

derivative with a Savitzky-Golay filter with 20 nm window. Visual observation showed that continuum removal was more effective in removing the scattering effect. Figure 1 shows a typical continuum removed spectrum of a Tajik soil sample. The spectrum shows strong absorption bands of O-H (water) at 1414 and 1914 nm and a band related with CH absorption at 2208 nm (Cozzolino, 2003).

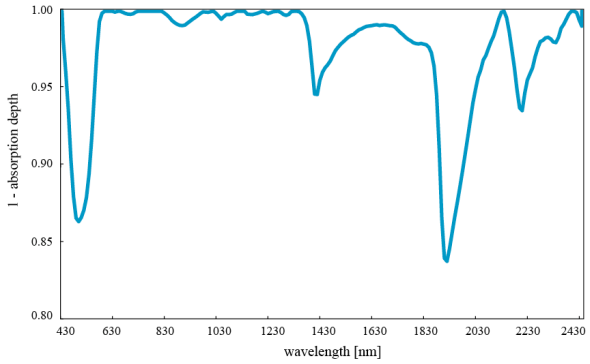

Figure 1: Continuum removed spectrum of Tajik soil sample

Soil color is strongly correlated to certain soil properties as for example $CaCO_3$ (Ben-Dor & Banin, 1994; Jarmer et al., 2000). Therefore, the soil color was extracted from the spectra using the CIE color system (Wiszecky & Stiles, 2001) resulting in the three color variables X (red), Y (green), Z (blue). They were used as additional input variables.

As at this point of work no information on the chemical composition was available the selection of the calibration and validation samples had to be made based on the spectra of the soil samples. Principal Component Analysis was used to select evenly distributed samples regarding the first and second Principal Components, all in all 252 samples were selected.

The chemical analysis of the selected soil samples was carried out in the ICRAF laboratory in Kenya due to reasons of costs and accuracy. The soil samples were analysed for *total C*, *organic C*, *total N*, *pH*, *$CaCO_3$*, *extractable phosphorus*, *exchangeable calcium*, *magnesium* and *potassium*, and the *fractions clay*, *silt* and *sand*. Table 1 shows the applied methods of soil analysis.

Table 1: Methods of soil analysis.

Soil Property	**Method**
Total C	Dry combustion method
Soil organic C	Dry combustion method on decarbonised samples
Total N	Dry combustion method
pH	pH-meter
$CaCO_3$	Weight loss method, $CaCO_3$ was removed using 10% HCl
Extractable phosphorus	Modified Olsen Extractant
Exchangeable calcium	KCl extractant
Exchangeable magnesium	KCl extractant
Exchangeable potassium	Modified Olsen Extractant
Clay (<0.002 mm) Silt (0.002 to 0.05 mm) Sand (>0.05 mm)	Hydrometer method, after penetration with H_2O_2 to remove organic matter and 10 % HCl to remove the soluble salts.

In the study area high heterogeneity of soils was observed. Thus a homogeneous dataset of brown carbonate soils was determined (166 samples) using a CART classification tree. This resulted in higher prediction accuracy of MLR models. Three approaches were evaluated for their ability to build models for predicting soil properties based on spectral information:

- Multiple linear regression with a stepwise procedure (MLR) with continuum removed data
- Principal component regression (PCR) with first derivative data
- Regression tree (CART, 2007) with first derivative data

2/3 of the brown carbonate soil samples were used to calibrate the models, 1/3 randomly chosen samples were used for validation. Table 2 shows the R^2 and RMSEP (Root Mean Squared Error of Prediction) for the different methods. Despite of the good results PCR was considered to be not suitable for this application, as new data can not be predicted with the developed models. This is connected to the calculation of the principal components where the calculation of one sample is influenced by all other samples in the data set. As the combination of continuum removed spectra and MLR was performing better than the combination of spectra transformed to first derivatives and CART, it was chosen as method for this work.

Table 2: Comparison of MLR, PCR and CART models at the example of predicting total C and total N. (*cr*: continuum removed spectra. *1.deriv*: 1. derivative of spectra).

	Approach	**R^2**	**RMSEP**
Total C	MLR (cr)	0.72	0.486
	PCR (1.deriv)	0.76	0.436
	CART (1.deriv)	0.63	0.559
Total N	MLR (cr)	0.83	0.030
	PCR (1.deriv)	0.78	0.033
	CART (1.deriv)	0.80	0.035

3. RESULTS AND DISCUSSION

The following correlations could be observed between the soil properties:

- Correlation between total N and SOC: $R^2 = 0.76$. Total N is for the largest part bounded in organic matter, which explains the high correlation (Gisi, 1997).
- Correlation between total C and $CaCO_3$: $R^2 = 0.53$. Total C is made up of inorganic C (mainly $CaCO_3$ in these soils) and organic C content. The analysed Tajik soils show high concentrations of $CaCO_3$ and low SOC contents, which explains the high correlation between total C and $CaCO_3$.

It was not possible to set up models for all measured soil properties. No models with satisfying accuracy could be built for extractable P, exchangeable Ca, Mg and K, and for the fractions clay, silt and sand.

Poor calibrations for extractable soil properties have also been reported by Udelhoven et al. (2003), whereas the findings of Shepherd and Walsh (2002) are contradictory: they modelled exchangeable Ca and Mg with good accuracy. Successful modelling of the fractions clay, silt and sand was obtained in other studies (Chang, 2001; Cozzolino, 2003). It can be assumed that non-linear methods are more suitable

for the determination of the particle size fractions and the extractable and exchangeable soil properties.

Table 3 shows descriptive statistics for the soil properties for which adequate models could be built. The accuracy measures for the MLR models are given in Table 4. All models use between 3 and 6 components. The R^2 values lie between 0.72 and 0.83 for all but the pH model. As the range of the pH values is very low, i.e. between 7.01 and 8.66, modelling of pH in these soils is difficult. The range of $CaCO_3$ is very high with 3.7 to 414 g per kg soil.

Table 3: Descriptive statistics for the soil properties

	Unit	Min	Max	Mean	Standard Deviation
Total C	g kg^{-1}	2.5	57.7	29.4	10.8
Organic C	g kg^{-1}	0.2	46	10.9	7.1
Total N	g kg^{-1}	0.2	3.8	1.3	0.6
pH	log H	7.01	8.66	8.05	0.25
$CaCO_3$	g kg^{-1}	3.7	414	194	93.7

Table 4: Accuracy measures for the calibration models. (R^2 *C*: R^2 Calibration, R^2 *P*: R^2 Prediction).

	R^2 C	RMSEC	R^2 P	RMSEP
Total C	0.85	3.48	0.76	4.36
Org. C	0.74	3.54	0.81	3.30
Total N	0.72	0.30	0.83	0.30
PH	0.61	0.13	0.61	0.16
$CaCO_3$	0.62	58.7	0.72	46.3

The resulting calibration equations for the different soil properties are given in Table 5. Some of the spectral bands that were selected by the MLR models in this study are consistent with other studies:

- 1860 nm and 2180 nm are absorption bands of carbonates (Jarmer et al., 2000; Hunt&Salisbury, 1971)
- 2320 nm, 2340 nm, 2380 nm are strong absorption bands of carbonates (Jarmer et al., 2000). Figure 2 shows the absorption band at 2340 nm at the example of four soil samples with 15 to 359 g kg^{-1} $CaCO_3$.
- 1530 nm is an absorption of N-H stretch first overtone (ASD, 2007a).

Table 5: Resulting MLR calibration equations (*X* denotes red color in the CIE color system, selected wavelengths [nm] are in italic).

Total C [g kg^{-1}] =	-4161.8+86.2(*X*)+1851.2(*2180*)-1097.7(*1490*)-1319.5(*2320*)+940.8(*1860*)+3831.7(*1300*)
Org. C [g kg^{-1}] =	-1426.6+1681.5(*1870*)-1050.7(*1530*)+879.6(*2180*)
Total N [g kg^{-1}] =	-62.94+127(*1870*)-88.3(*1530*)+29.1(*900*)
pH =	17.8-52.3(*1860*)-24.4(*X*)+36.1(*1700*)+18.0(*720*)
$CaCO_3$ [g kg^{-1}] =	-7956.5+1121.9(*X*)-5440.7(*2340*)+13230.9(*2380*)

The appearance of the color variable 'X' is not surprising, as the brightness of soil color is strongly correlated to the $CaCO_3$ content, and as total C is strongly correlated to $CaCO_3$. The models for organic C and total N both use the wavelengths 1870 nm and 1530 nm. This is due to the high correlation between the two soil properties.

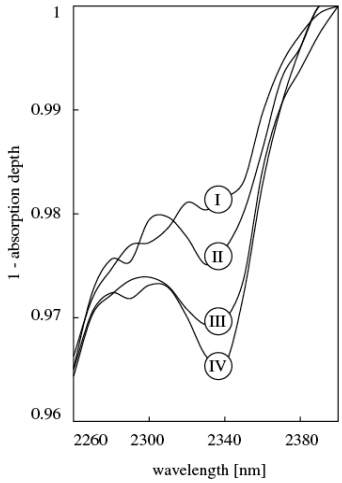

Figure 2: Absorption band of $CaCO_3$ at 2340 nm, shown at the example of four soil samples with 15 (I), 110 (II), 288 (III) and 359 g kg^{-1} (IV) of $CaCO_3$.

Figures 3, 4, 5, 6, 7 illustrate the accuracy of the model predictions for *total C*, *organic C*, *total N*, *pH* and *$CaCO_3$* using plots of the chemical reference data versus the predicted chemical values for the validation data set. The more closely the points approximate the 1:1 line, the more accurate is the predictive model. The models for *total C*, *organic C* and *total N* are very accurate, whereas the models for *pH* and *$CaCO_3$* show rather low accuracies.

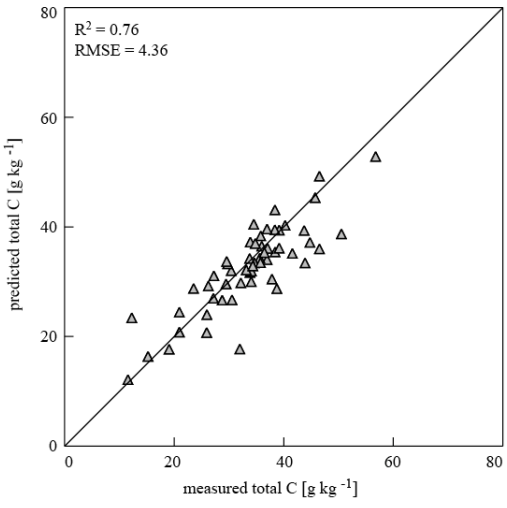

Figure 3: Chemical reference data versus predicted values for the validation data set of the total C model.

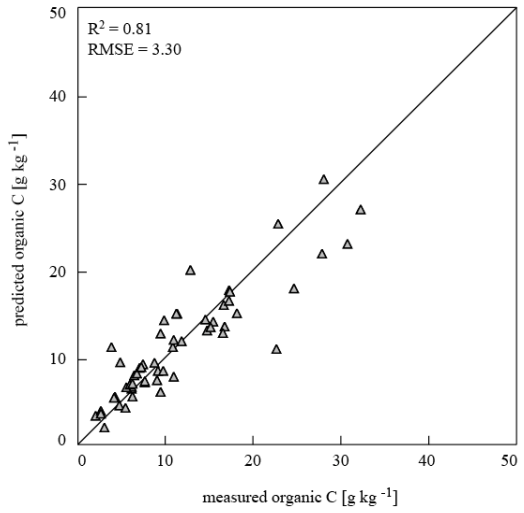

Figure 4: Chemical reference data versus predicted values for the validation data set of the organic C model.

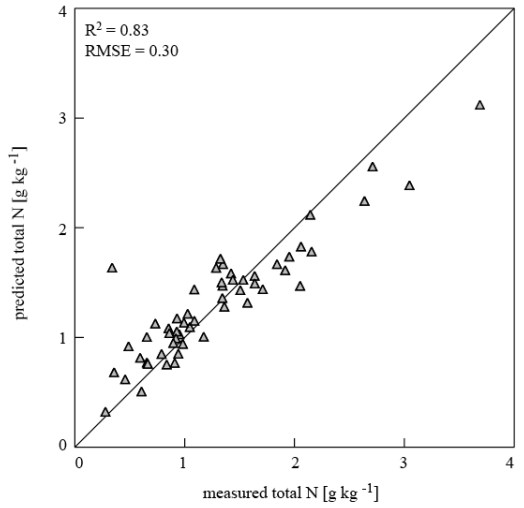

Figure 5: Chemical reference data versus predicted values for the validation data set of the total N model.

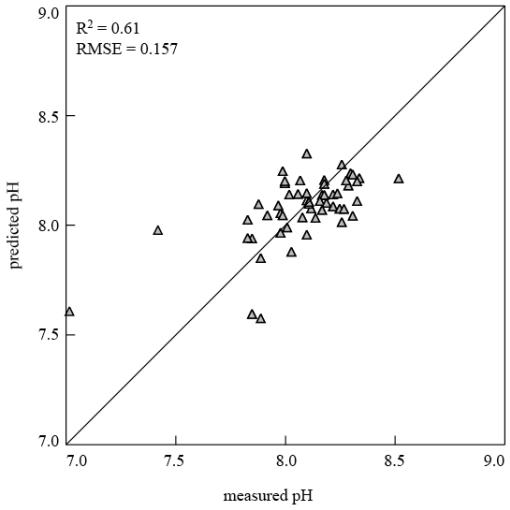

Figure 6: Chemical reference data versus predicted values for the validation data set of the pH model.

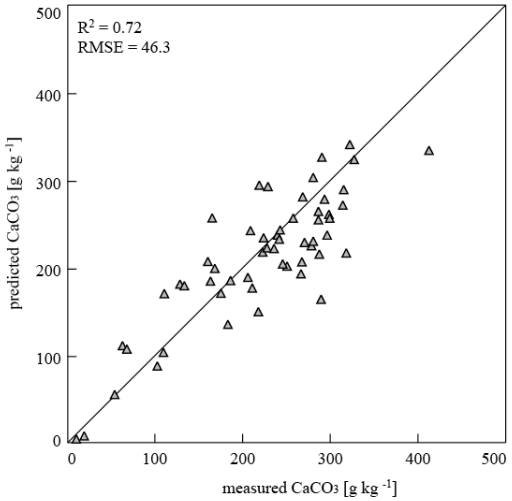

Figure 7: Chemical reference data versus predicted values for the validation data set of the $CaCO_3$ model.

4. CONCLUSIONS

Modelling soil properties using VIS-NIR-SWIR spectroscopy is a powerful tool to determine chemical and physical soil properties more rapidly and inexpensively than standard soil characterization techniques. Further research is necessary, especially in the field of statistical methods. The results indicate that the pre-processing method is crucial. Further, the potential of the many different statistical methods used in this and other studies should be evaluated further, especially with regard to sample sets with highly heterogeneous soils. This work shows that the classical method of MLR with continuum removed data has the advantage to simplify comparisons among different studies as it uses physically defined absorption bands, whereas some more complex methods - for example multivariate adaptive regression splines (MARS, 2007) - seem to have the ability to model more soil properties. For modelling of key properties in soil fertility (e.g. organic and inorganic C) for homogeneous soils, the MLR method seems appropriate. For heterogeneous soils and other soil properties such as extractable and exchangeable soil parameters or soil fractions more complex methods must be used. The developed models bear the potential to be applied to future hyperspectral imaging data for regional and rapid assessment of soil state in the area.

5. REFERENCES

ASD, 2007a. Analytical Spectral Devices. Introduction to NIR Technology. http://www.asdi.com/ASD-600510_NIR-Introduction.pdf (accessed March 20, 2007).

ASD, 2007b. Analytical Spectral Devices. Muglight. http://www.asdi.com/products-accessories-hisp.asp (accessed March 20, 2007).

Ben-Dor, E., Banin, A., 1994. Visible and Near-Infrared (0.4-1.1 μm) Analysis of Arid and Semiarid Soils. *Remote Sensing of Environment*, 48, pp. 261-274.

CART, 2007. Classification and Regression Trees. Software by Salford Systems. http://www. salford-systems.com (accessed March 20, 2007).

Chang, C.-W., et al., 2001. Near-Infrared Reflectance Spectroscopy - Principal Component Regression Analysis of Soil Properties. *Soil Science Society of America Journal*, 65, pp. 480-490.

Cozzolino, D., Morón, A., 2003. The Potential of Near-Infrared Reflectance Spectroscopy to Analyse Soil Chemical and Physical Characteristics. *Journal of Agricultural Science*, 140, pp. 65-71.

ENVI, 2007. Software by ITT Visual Information Solutions. http://www.ittvis.com (accessed March 20, 2007).

Gisi, U., 1997. *Bodenökologie*. 2nd edition. Georg Thieme Verlag, Stuttgart, New York.

Hunt, G.R., Salisbury, J.W., 1971. Visible and Near-Infrared Spectra of Minerals and Rocks: II. In: *Modern Geology*, Nr. 2, pp. 23-30.

Jarmer, T., et al., 2000. Spectral Detection of Inorganic Carbon Content along a Semi-Arid to Hyper-Arid Climatic Gradient in the Judean Desert (Israel). *Second EARSeL Workshop on Imaging Spectroscopy*, Enschede, 2000, CD-ROM.

MARS, 2006. Multivariate Adaptive Regression Splines. Software by Salford Systems. http:// www.salford-systems.com (accessed March 20, 2007).

Shepherd, K.D., Walsh, M.G., 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*, 66, pp. 988-998.

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, pp. 319-329.

Wiszecky, G., Stiles, W.S., 2001. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, New York.

Wolfgramm, B., Seiler, B., Kneubühler, M., Liniger, H.-P., 2007. Spatial assessment of erosion and its impact on soil fertility in the Tajik foothills. *EARSeL eProceedings*, 6 (1), pp. 12-25.