Quantitative Assessment of Soil Parameters in Western Tajikistan

using a Soil Spectral Library Approach

Diploma Thesis

Submitted to the Department of Geography
University of Zurich

by Bruno Seiler, 2006
Diploma Thesis
Submitted to the Department of Geography
University of Zurich
by Bruno Seiler, 2006

Supervisors:

Prof. Dr. K. I. Itten
Dr. M. Kneubühler
Remote Sensing Laboratories
Department of Geography
University of Zurich

B. Wolfgramm
Centre for Development and Environment
University of Berne

Contact:

Bruno Seiler
Schlosshofstrasse 28
CH - 8400 Winterthur
brunoseiler@hotmail.com
Acknowledgements

I would like to speak out my thanks to all the people who assisted and supported me during the research and writing of my diploma thesis. Special acknowledgements deserve

- Prof. Dr. K. I. Itten for the perusal of my manuscript,
- Dr. M. Kneubühler for the scientific support,
- B. Wolfgramm for the scientific assistance and her Russian skills,
- NCCR north-south for the financial support,
- P. Sosin for the delicious dinner and beverages,
- G. Nekushoeva for the welcoming dinner and the translating,
- the Soil Science Research Institute in Dushanbe for the work place,
- Roman for the layout,
- Muriel for the corrections
- and Madeleine for her patience.

Winterthur, October 2006

B. Seiler
Summary

The diploma thesis at hand deals with modelling chemical and physical soil properties using a spectral library approach. Thereby statistical methods are used to predict soil properties quantitatively based on spectral information. The thesis is integrated in the framework of the Work Package 4: Natural resources in sustainable development of the NCCR North-South. Additionally this thesis is a collaboration of the Remote Sensing Laboratories with B. Wolfram, Ph.D. student at the Centre for Development and Environment at the University of Berne.

The investigation area is situated in Tajikistan around the capital city Dushanbe. Three test areas of 10 x 10 km size were determined and sampling plots were chosen using a random sampling scheme. In total 1465 soil samples were taken. The samples were air-dried, grinded and sieved to 2 mm. Spectral measurements of the prepared soil samples were taken with a Field Spec Pro FR in the spectral range from 350 to 2500 nm at a 1 nm resolution. Each soil sample was measured twice and averaged. The spectral data was resampled to a 10 nm resolution to reduce the amount of data. 260 soil samples were selected as calibration and validation samples by principal component analysis. They were analysed in the laboratory for the physical and chemical properties total C, organic C, total N, pH, CaCO$_3$, extractable P, exchangeable Ca, Mg and K, and the fractions clay, silt and sand. The cation exchange capacity was additionally approximated by calculation. As auxiliary variable soil color was quantified from the spectral data using the CIE color system.

Three statistical methods for building calibration models were tested against each other: multiple linear regression with continuum removed data, principal component regression and regression tree with first derivative data. One third of the chemically analysed samples were used for random hold out validation. Multiple linear regression turned out to be the best performing statistical method for the used data set and was therefore used to calibrate the prediction models. In order to improve the prediction accuracy the occurring soils were grouped using a classification tree and only the soil group of main interest was used for modelling.

No adequate models could be built for the soil properties extractable P, exchangeable Ca, Mg, and K, the fractions clay, silt and sand, and the cation exchange capacity. For the remaining soil properties good results were obtained: total C ($R^2 = 0.76$, RMSEP = 4.36 g kg$^{-1}$), total N ($R^2 = 0.83$, RMSEP = 0.30 g kg$^{-1}$), SOC ($R^2 = 0.81$, RMSEP = 3.30 g kg$^{-1}$), pH ($R^2 = 0.61$, RMSEP = 0.157) and CaCO$_3$ ($R^2 = 0.72$, RMSEP = 4.63 %).

The spectral library approach for prediction of soil properties has a high potential to substitute standard laboratory methods where rapid and inexpensive analysis is required. Further research is needed especially in the field of statistical methods and in the application on satellite images which would make monitoring of soil conditions with little effort possible.
Contents

Part I  Introduction  

1. Context of research  
  1.1 Scientific framework  
  1.2 Research aims  

2. Investigation area  

Part II  Theory  

3. NIR-SWIR reflectance spectroscopy  
  3.1 Molecular vibrations  

4. Soil color  
  4.1 Munsell Color Code vs. CIE Color Variables  

5. Multivariate statistical analysis  

Part III  Methods  

6. Soil sampling and preparation  
  6.1 Sampling design  
  6.2 Preparation of soil samples  

7. Spectral measurements of soil samples  
  7.1 Measurement procedure  

8. Chemical analysis of a selection of soil samples  
  8.1 Sample selection  
  8.2 Selection of soil laboratory  
  8.3 Methods for soil chemical analysis  
  8.4 Accuracy assessment and repeat measurements  
  9.1 Correction of steps
9.2 Omitting wave bands with low signal to noise ratio 50
9.3 Data compression 52
9.4 Scatter Correction 52

10. Calculation of CIE color variables 56

11. Separation of soil types 58
   11.1 Occurring soil types 58
   11.2 Approach of separation 60

12. Calibration methods 63
   12.1 Selection of calibration methods 63
   12.2 Multiple linear regression 64
   12.3 Principal component regression 66
   12.4 Regression tree 66

13. Development of calibration program in IDL 70
   13.1 Variable selection 70
   13.2 Verification of model assumptions 71
   13.3 F-tests 74
   13.4 Model verification 75

14. Model validation 76
   14.1 Selection of calibration/validation samples 76
   14.2 Accuracy measures 76
   14.3 Validation based on repeat measurements 80

15. Prediction of new samples 80

Part IV Results & Discussion 83

16. Results of chemical analysis 84
   16.1 Distributions of chemical soil properties 84
   16.2 Correlations between chemical soil properties 86

17. Results of soil color quantification 88
   17.1 Correlation between CIE color variables and chemical soil properties 88

18. Result of soil type separation 90
   18.1 Influence on the distribution of the chemical soil properties 90
   18.2 Impact on model accuracy 90

19. Comparison of applied statistical methods 92
Part I

Introduction
1. Context of research

1.1 Scientific framework

The diploma thesis at hand is integrated in the framework of the Work Package 4: Natural resources in sustainable development (WP4) of the NCCR North-South. The NCCR North-South started in 2002 and is one of 14 long-term research programs implemented by the Swiss National Science Foundation. The main goal of this program is to focus on ways of mitigating the syndromes of global change in urban, peri-urban, semi-arid and highland-lowland areas [25].

Additionally this diploma thesis is a collaboration of the Remote Sensing Laboratories (RSL) with B. Wolfgramm, Ph.D. student at the centre for development and environment (CDE), a department of the Institute of Geography at the University of Berne. The overall goal of her study is to gain a thorough understanding of the impacts of land use on land resources in the hill zone of western Tajikistan, which will allow to identify opportunities for sustainable land resource management for the specific land use systems considered. The specific objectives of the study are as follows:

- To provide an overview of the types, the extent and the dynamics of land use patterns including underlying causes and trends

- To identify and analyse the relevant degradation processes at different spatial and temporal scales

- To identify and analyse the relevant land conservation approaches and measures which have been or are applied within the specific land use systems

- To reveal the links and dependencies between land use and land degradation / conservation

- To identify land management options for sustainable natural resource management within specific land use systems

- To evaluate, adopt, adapt and develop methodologies for the assessment and analysis of land degradation and conservation and the impact of land use on land resources.

For spatial and temporal analysis of spectral information the use of Remote Sensing methods was planned. On this basis the collaboration between B. Wolfgramm and the Remote Sensing Laboratories (RSL), a department of the Institute of Geography at the University of Zurich was established. At RSL two diploma thesis were carried out:
1.2 Research aims

At the outset of this study two overall objectives were determined: (i) to develop models for predicting chemical soil properties based on spectral information, (ii) to map chemical soil properties using LANDSAT satellite imagery. As the second objective could not be followed up due to technical (see chapter 22) and time constraints the focus was set on the first objective. The overall goal is the following:

What chemical soil properties can be predicted from spectral information for the loess areas in western Tajikistan? What are major constraints?

The specific objectives of this study are:

- to build a soil spectral library of a sample set including around 1500 sub- and topsoil samples from three similar test areas,
- to select around 250 representative samples for chemical analysis and for calibration and validation of the models,
- to test the ability of different statistical methods for building models to predict chemical soil properties from spectral information,
- to evaluate the integration of soil color in the model building process,
- to apply the models to new data/a new area,
- and to compare the results to other studies.
Tajikistan’s territory is spreading over a total area of 142’500 km². The study concentrates on the western part of the country. Landscape in western Tajikistan is characterised by a high bio-physical variation of land units, from steep highland areas with peaks reaching over 5000 m a.s.l. to loess plateaus and to lowland plains at only 300 m a.s.l. In 1999 the population of western Tajikistan reached nearly 6 million. The calculated average population density in western Tajikistan is 62.5 inhabitants per km². However, most population and economic activities are concentrated in a few broad valleys and towns. Tajikistan has a high population growth rate and a growing percentage among the rural population: the share rose from 68% in 1990 to 74% in 2000. The share of agriculture in employment has increased from 45% in 1990 to 65% in 2000 (ADB, 2003). On the other hand flat arable land is very much restricted in Tajikistan and has to bear the increasing pressure of a growing population. Under these circumstances sustainable land resource use is a crucial issue for the country’s economic development and key to Tajikistan’s future [62].

The main land use systems in western Tajikistan and their respective share of the agricultural land are the following [62]:

- Summer and winter pastures, rangelands and forests found on marginal lands in the hill and mountainous zones across all western Tajikistan, covering 78% of the agricultural land.

- Rainfed agriculture (mainly cereal crops) located in the hill zones, covering 10% of the agricultural land.

- Irrigated agriculture (mainly cotton crops) concentrated in the semi-arid lowlands, covering 12% of the agricultural land.

The hill zone of western Tajikistan consist mainly of easily erodable loess deposits. The increasing pressure of the growing population mentioned above led to wide-spread 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 [46], also having a highly negative impact on soil fertility [61]. Figure 2.4 shows the impact of erosion: From erosion rills to the formation of deep gullies.
Figure 2.1  Map of Tajikistan and neighbouring countries. The red square is indicating the position of the clipped LANDSAT image in figure 2.2.

Figure 2.2  Clipping of LANDSAT 7 image from 22.08.2000 showing the three test areas Yavan (I), Faizabad (II), Varzob (III).
Within the hill zone three test areas 10 x 10 km size representing the loess deposits and the different climatic regions were selected. Rainfall characteristics vary from the south with 400 mm per year to the northeast with up to 900 mm. The yearly distribution of precipitation is given in figure 2.3. It shows monthly precipitation averaged over the years 1988 to 2002 for the two test areas Yavan and Faizabad. The humid season is spring whereas summer is very dry. Soils are defined as brown carbonate by the local Tajik definition system. Soils dominated by granodiorit mother rock are also situated in the study area. Typical values for SOC are 1-2 %. CaCO3 contents vary between 2-30 %, depending on the mother rock, but also on the state of erosion [54].

The three test areas are marked in the clipping of the LANDSAT image in figure 2.2, the location of the clipping in Tajikistan is given as red square in figure 2.1. Yavan test area is located in the southern, Faizabad in the northeastern and Varzob in the northwestern part of the image. Figure 2.4 shows the impact of erosion: From erosion rills to the formation of deep gullies.
Figure 2.3  Monthly precipitation in the two test areas Faizabad (black) and Yavan (grey) averaged over the years 1988 to 2002. Source: B. Wolfgramm
Figure 2.4  Erosion in the loess deposits of western Tajikistan. From erosion rills (above) to deep gullies (below).
Part II

Theory
3. NIR-SWIR reflectance spectroscopy

The aim of NIR-SWIR reflectance spectroscopy is to measure and interpret reflected radiation in the NIR (0.7-1.4µm) and SWIR (1.4-3µm) region of the electromagnetic radiation. In literature this is often referred to in short as NIR spectroscopy.

When electromagnetic radiation incidents on a material it is partially reflected, partially transmitted and partially absorbed, depending on the material and on the wavelength of the radiation. If the material is thick enough such that the transmitted radiance is equal to zero, the relation can be simplified to $\rho + \alpha = 1$, where $\rho$ is the rate of reflection and $\alpha$ the rate of absorption. A spectrometer is able to measure the rate of the reflected radiance as percentage of the incident radiance. In remote sensing this is often transformed to absorption via continuum removal as this makes information on the material’s composition easier accessible.

The interpretation of the spectral information is based on the Beer’s Law. It states that the absorbance of a species at a particular wavelength of electromagnetic radiation is proportional to the concentration of the absorbing species and to the length of the path of the electromagnetic radiation through the sample containing the absorbing species [25]:

$$ A(\lambda) = e(\lambda) \cdot l \cdot c $$

where

- $A$ : absorbance
- $e$ : absorptivity of the species at a specific
- $l$ : length of path
- $c$ : concentration
- $\lambda$ : wavelength

Multivariate statistical methods can make use of this relation and determine the concentration of specific molecules in a sample based on spectral reflectance/absorbance information.

For a better understanding of which molecules are absorbing in which spectral region and why, the theory of molecular vibrations has to be consulted. It deals with the process of absorption on the molecular level and will be introduced in the following subchapter.
3.1 Molecular vibrations

The interactions between electromagnetic radiation in the NIR and SWIR region and molecules are described through the theory of molecular vibrations. To simplify certain calculations the wavenumber is partly used in the explanations instead of the wavelength. The wavenumber is inversely related to the wavelength and has the unit cm\(^{-1}\), it describes the number of waves per centimeter:

\[ \kappa = \frac{1}{\lambda} \] where \( \kappa \) = wavenumber, \( \lambda \) = wavelength

At ambient temperature most of the molecules are in their fundamental vibrational energy levels. The group of atoms that form the molecules displace one atom in relation to the other in a frequency that is defined by the strength of the chemical bond and the mass of the individual bonded atoms or their groups. The amplitudes of these vibrations are of a few nanometers and will increase if energy is transferred to the molecule by a photon of a distinct wavelength [44].

Molecules with \( N \) atoms have \( 3N \) degrees of freedom, three of which represent translational motion in mutually perpendicular directions (the \( x-, y-, \) and \( z- \) axes) and three represent rotational motion about the \( x-, y-\) and \( z- \) axes. The remaining \( 3N-6 \) degrees of freedom give the number of ways that the atoms in the molecule can vibrate, i.e. the number of vibrational modes. For each mode all the atoms vibrate at a certain characteristic frequency. The energy difference for transitions between the ground state of vibration and the first excited state of most vibrational modes corresponds to the energy of radiation in the MIR spectrum above \( 2.5 \) µm [22]. Nonetheless effects of these transitions are visible in the NIR region of electromagnetic radiation through overtones or combinations of fundamental vibrations:

- Overtones are approximately multiples of the fundamental vibrations. A fundamental vibration, “\( f \)”, will give rise to a series of absorptions \( 2f, 3f, 4f, \ldots \) which are known as the first, second, third overtone [43].

- In a combination of fundamental vibrations, absorption of a photon of NIR energy is shared between two or more vibrations, which would be individually observed as fundamentals in the MIR region above \( 2.5 \) µm. For instance the fundamental absorptions \( f_{1} \) and \( f_{2} \) occurring at \( 3000 \) and \( 1600 \) cm\(^{-1}\) would give rise to a combination band at approximately \( 3000 + 1600 = 4600 \) cm\(^{-1}\). This translates to \( 2174 \) nm, which is a NIR absorption [43].

Overtones and combination bands are usually much weaker than the fundamental modes from which they are derived. The only exception is when these bands are enhanced by Fermi resonance, which occurs when an overtone or combination band absorbs at approximately the same frequency of a fundamental mode involving the same atoms. Although many overtone and combination bands still absorb in the MIR region, the first and second overtones of C-H, O-H and N-H stretching vibrations are found above \( 4000 \) cm\(^{-1}\), i.e. in the NIR region. In table 3.1 an extract of absorption features of these functional groups between 2300 and 2500 nm is given. The question remains how molecules
without the functional groups C-H, O-H, N-H can be detected in the NIR region [22/41]. In this case, the most important reasons are the presence of electronic transitions and the occurrence of pseudo-correlation:

- Electronic transitions are caused by the movement of electrons from one orbit to a higher-energy orbit again connected with the absorption of photons of a distinct energy. These processes are normally observed in the VIS or UV regions of the electromagnetic spectrum, but they can also occur in the NIR region, especially in the 780-1100 nm region [43].

- Pseudo-correlations result from the interactions of different molecules. If the concentration of a molecule A depends on the concentration of a molecule B, they are correlated. Assuming that the molecule A shows distinct absorption features that can be used for modelling and molecule B does not, molecule B will adopt the features of molecule A as they are correlated. Although no direct physical relation exists between molecule B and the absorption features of molecule A this pseudo-correlation can be used for modelling. This results in calibration equations for the two molecules that use the same wavelengths and only show differences regarding the regression coefficients.

A further influence comes from the physical properties of the analysed material. The strength of absorption is certainly a function of physical properties such as density and particle size. Baumgardner et al. [2] specify several investigations in this field of research. Generally one can say that the reflectance increases with decreasing particle size.

The consequence of these different mechanisms makes an NIR spectrum very complex, but often this complexity is hidden in a few rather broad areas of absorption. The fact that the same atoms are involved in many different absorptions means that these absorptions can be used via statistical analysis [43].

Although the absorption features of molecules are much stronger in the MIR region and the spectra are easier to interpret, NIR has several advantages over MIR-spectroscopy; the sample preparation is less complicated, field-measurement possible, the detectors are more sensitive, it is less cost-intensive and the spectra are very amenable to multivariate statistical analysis [22].

The VIS region of the electromagnetic radiation was additionally included in the analysis as it contains information about soil color which is correlated to distinct chemical soil properties (See chapter 4 for details).
4. Soil color

Soil color is a differentiating characteristic for many soil classes in all modern classification systems and permits estimations on chemical soil properties [30]. The most important relations between soil properties and color are the following:

- Soil organic matter:
  Soil organic matter is known to have a strong influence on soil reflectance. A general observation has been that, as organic matter content increases, soil reflectance decreases throughout the 0.4 to 2.5 µm wavelength range and therewith the lightness of color [2].

- Moisture:
  It is a common observation that most soils appear darker when wet than when dry. This results from decreased reflectance of incident radiation in the visible region of the spectrum [2].

- Iron oxides:
  The type and relative amount of iron oxides are known to influence soil color substantially. An increase of iron oxides leads to a decrease of overall reflection in the VIS and NIR region [32]. Yellowish, orange and reddish hues of soils are caused by Iron bearing minerals including Fe oxides, hydroxides, oxyhydroxides and hydroxysulfates [48].

- CaCO$_3$:
  The concentration of CaCO$_3$ influences the lightness of soils substantially. An increasing CaCO$_3$ concentration implicates an increasing lightness of color [3].

Including soil color in calibration modelling of soil properties seems to have a high potential, as the work of Jarmer et al [31] shows for CaCO$_3$. They found a high correlation ($R^2 = 0.90$) between the CIE chromaticity value $x$ and the amount of CaCO$_3$.

4.1 Munsell Color Code vs. CIE Color Variables

The most frequently used system to determine color of soils is the Munsell Color System. It is based on three color parameters: the Munsell hue, value and chroma. The system was established in the beginning of the 20th century by the American painter Albert Henry Munsell [17]. Today it is the most common of the color systems which are based on the principle of uniform color space, that is a color space where the Euclidean differences between colors are equivalent to the human perception of those differences [42]. The composition of the color system is given in figure 4.1. It is made up of three variables that describe the value, the chroma and the hue of a color. In 1942 the organization American Standards declared that the Munsell color system should be used to specify the color of
surfaces. A factor that made this color system even more interesting for soil scientists was the fact that the Munsell Color system was available as color chips which is very useful for the use in field work and makes the matching more precise and simple [17]. There are however some limitations of this Color System compared to newer Systems like those created by the Comission Internationale de l’Eclairage (CIE). Those limitations include [48]:

- The resolution of color assignment is limited by the resolution of the chips. Interpolation between chips is possible in principle, but difficult in practice as it has to be done in a three-dimensional space.

- Due to metamerism, the assignment of a color may depend on the light source that is often not standardized, and on the physiological properties of the observer’s color receptors.

- Differences in gloss and other surface properties between the chips and the soil may influence the assignment.

These three factors cause a substantial scattering of color coordinates when colors are determined repeatedly by one or several observers [45]. These limitations can be eliminated by calculating the color from diffuse reflectance spectra [19]. On this account it was decided on calculating the color from the diffuse reflectance spectra using the CIE XYZ System. The detailed description of the calculation process is given in chapter 10.
<table>
<thead>
<tr>
<th>Functional group</th>
<th>Wavelength [nm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-H bend second overtone</td>
<td>2300/2310</td>
</tr>
<tr>
<td>C-H stretch/CH₄ deformation</td>
<td>2322/2330</td>
</tr>
<tr>
<td>C-H stretch/C-H deformation</td>
<td>2335</td>
</tr>
<tr>
<td>CH₄ bend second overtone</td>
<td>2352</td>
</tr>
<tr>
<td>C-H stretch/C-C stretch combination</td>
<td>2380</td>
</tr>
<tr>
<td>C-H combination</td>
<td>2470</td>
</tr>
<tr>
<td>Symmetr. C-N-C stretch overtone</td>
<td>2470</td>
</tr>
<tr>
<td>C-H stretch/C-C stretch combination</td>
<td>2488</td>
</tr>
<tr>
<td>C-H stretch/C-C and C-O-C stretch</td>
<td>2500</td>
</tr>
</tbody>
</table>

Table 3.1 Extract of approximate wavelengths of some common functional groups between 2300 and 2500 nm [44].

Figure 4.1 Munsell Color System, simplified. Circle: The 10 main hues for maximal chroma and value 5. Five chroma steps for hue 5YR(Yellow-Red).
5. **Multivariate statistical analysis**

This study attempts to explain a dependent variable, for example soil organic carbon content, using a number of independent variables, in this case spectral reflectance/absorbance information of different wavelengths. Multivariate statistics are used for this purpose. The goal is to substitute the costly and time-consuming method of the chemical analysis of soil samples through the less expensive and much faster method of reflectance spectroscopy. It can be assumed that the used region VIS-NIR-SWIR of the electromagnetic radiation is containing sufficient information on the chemical composition of soils, as other studies in this field of research sufficiently demonstrated [8/35/40/50]. For the development of a model that can predict chemical soil properties from spectral information calibration is needed. For this purpose a number of soil samples are selected (chapter 8.1), chemically analysed (see chapter 8.2-8.4) and then used to calibrate and validate the model (see chapters 12/14).

The basic problems in multivariate calibration are the following [43]:

- **collinearity problem**
  - Strong collinearity particularly exists between spectral wavelengths. This means that a near or total linear relationship can be observed between two or more variables. There are two solutions for this problem: On the one hand the variables for the model can be selected carefully using a stepwise procedure (chapter 13.1), on the other hand new variables can be calculated that dont show collinearity any more using principal component regression (PCR) (chapter 12.3) or partial least square regression (PLSR).

- **calibration data selection**
  - The selected calibration samples should cover the variability of soil properties in the study area and the number has to be sufficiently high. (chapter 8.1)

- **outlier problem**
  - Outliers are samples of a different population or samples afflicted with measuring errors. They need to be detected and removed from the dataset as they can cause strong falsifications of the model. (chapter 13.2)

A number of different multivariate statistical methods have been applied for modeling soil properties from soil spectral reflectance. The most commonly used calibration techniques are partial least square regression (PLSR)[8/14/35/40/57], principal component regression (PCR) [11] and multiple linear regression (MLR)[31/32/35]. Alternatively new techniques are regression tree, multivariate adaptive regression splines and neural network [9/50]. Three methods were finally selected for this study: MLR, PCR and regression tree. The selection process was partly based upon a test data set from Kenyan soil data received from C.Hett [25] and is given in chapter 13.
Part III

Methods
6. Soil sampling and preparation

6.1 Sampling design

An example for the sampling design is given in figure 6.1. For each of the three test areas Faizabad, Yavan and Varzob a raster with 100 points was applied from which 15 points were selected randomly. Areas with villages and bare rock were excluded from this selection. 15 clusters with a 460 m radius were centered on the selected points, each consisting of 13 sampling plots systematically located. The plots are lined up at 58, 115, 230 and 460 m, which is a 2, 4, 8, 16 x plot diameter distance along three radial lines placed at 120° angle to one another. This structure allows the analysis of the spatial characteristics of land use/land cover and soil parameters most efficiently. With a diameter of 30 meters the size of the sample plots corresponds with the ground resolution of the Landsat images. Two sampling pits were placed within one sample plot. If the plot showed a uniform land use the two sampling pits were located in such way that the variability within a plot was covered best. For example one sampling pit representing dense cover of wheat and one sampling pit representing sparse cover of wheat within one sample plot. If more than one land use type was present in a plot the land use type taking up more space was sampled. [58]

One top-, 0 to 20 cm, and one subsoil sample, 20 to 50 cm, was taken per sampling pit. Two approaches were applied: Partly the four soil samples of a plot were packed separately, partly the two topsoil and the two subsoil samples were mixed and packed as two composite samples. The first method is more detailed and permits to determine the within plot variability whereas the second method is time- and cost-saving.

The soil samples were labelled with an abbreviation for the test Area, FA for Faizabad, YA for Yavan and VZ for Varzob, followed by cluster number, plot number, and a number indicating the plot’s soil samples, uneven number for topsoil, even number for subsoil samples. Example: FA880802 is located in the test area Faizabad, cluster 88, plot 08, and is a subsoil sample.
Figure 6.1  Sampling design for Faizabad test area. Source: B. Wolfgramm
The soil samples were collected by B. Wolfgramm together with students from the Tajik Agrarian University, the test areas Faizabad and Yavan in Summer 2004 and Varzob in Summer 2005. In total 1465 samples were taken:

- Testarea Faizabad 666 soil samples (no composite samples)
- Testarea Varzob 412 soil samples (12 clusters as composite, 3 as separate samples)
- Testarea Yavan 387 soil samples (12 clusters as composite, 3 as separate samples)

A Garmin eTrex was used for the measurement of the position with an average accuracy of about 15 meters. The color of the soil samples was determined in the field using Munsell color chips. For further information on soil color see chapter 4.

Further ground truth was collected like visible indicators for land degradation and conservation and land cover classes according to the FAO land cover classification system. For details see Wolfgramm (Ph.D., forthcoming).

6.2 Preparation of soil samples

The soil samples were dried in the sun as moisture strongly affects the spectral response of soils [6/37]. 50 g of each soil sample were weighed to be further prepared for the spectrometer measurements, the remaining material was stored in plastic bags in the basement of the SSRI in Dushanbe. Further preparation included grinding and sieving through a 2 mm sieve to minimize differences in grain size. Finally the prepared samples were packed into labelled paper bags. Figures 6.2 to 6.4 show photos of the sampling, preprocessing and storing of the soil samples.
Figure 6.2  Soil sample preparation. Top to bottom: sampling pit, drying of samples. Photo: B. Wolfgramm
Figure 6.3  Soil sample preparation. Top to bottom: grinding, sieving. Photo: B. Wolfgramm
Figure 6.4  Soil sample preparation. Top to bottom: storage of the prepared samples in paperbags, storage of the remaining soil material in plastic bags.
7. Spectral measurements of soil samples

The spectral measurements of the soil samples were taken at the Soil Science Research Institute in Dushanbe, Tajikistan. The readings were conducted with an ASD spectrometer, a FieldSpec Pro FR owned by the Department of Geography of the University of Zurich. 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. 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. The Muglight utilizes an internal Tungsten Quartz Halogen light source and views a 12 mm spot size [1]. For the spectrum of the light source see appendix (figure A.1). A photograph of the measurement setup is given in figure 7.1.

In Tajikistan the electricity is subject to certain fluctuations. As the instability of the DC power supply can affect the operation of as well the spectrometer as the muglight a voltage regulator was switched in.

For the white reference measurement a Spectralon from Labsphere was used. To measure the white reference under the same conditions as the soil samples in Petri dishes, the bottom of a Petri dish was cut and placed under the spectralon.

The soil samples were filled in Petri dishes for the measurement process. As the Petri dishes affect the radiance they had to be chosen with caution. It was decided to use Petri dishes from Schott made of DURAN®, a borosilicate glass. In the spectral range from 350 to 2200 nm the absorption of DURAN® is negligibly low (see appendix, figure A.2). From 2200 to 2500 the absorption increases and lowers the reflectance signal of the soil sample resulting in a lower signal to noise ratio [49]. It should be kept in mind that the quality of the spectral measurements in this region may be affected.

7.1 Measurement procedure

The fore-optic field of view was set to 8 degrees, the scans to be averaged for white reference to 10 and for dark current to 25. The samples were measured using a 1 nm resolution. Optimization was renewed two times a day, the white reference after the measurement of each soil sample. Soil samples were measured twice with an approximate 90 degree turning in between and the two measurements were averaged.

A problem of the indoor measurement was the dust originating from the handling of the soil samples. The window of the muglight had to be cleaned with care after every measurement to avoid falsifications. In a protocol the number of the soil samples with the corresponding two spectrum numbers, irregularities and particular observations were noted. In addition the soil samples were
Figure 7.1  Measurement setup for spectral analysis. Left side: muglight with soil sample in Petri dish. Right side: spectrometer with interface.
photographed with a digital camera, an example is given in the figure 7.2.

Beside the spectral measurements of each soil sample two additional measurements were carried out:

- A soil sample was repeatedly measured without changing its position. This allowed an estimation of the signal to noise ratio for the given measuring setup. It is specified in chapter 9.2.
- Two whole plots, FA64 and YA24, were measured twice to determine the repeatability of the different statistical methods. Details on this topic are given in chapter 19.3.
Figure 7.2  Lab photo of soil sample FA550302 with paperbag. Photo: B. Wolfram
8. Chemical analysis of a selection of soil samples

8.1 Sample selection

The selection of soil samples to be chemically analyzed is very sensitive. It should cover the whole range of the different chemical properties’ concentration in the sample set to allow representative calibration models. As at this point of work no information on the chemical composition was available the selection had to be made based on the spectra of the soil samples.

Principal Component Analysis was used to facilitate analyzing the big quantity of data. This is a method of data reduction, which makes it possible to analyse data using a few newly derived variables only, the principal components (PCs) (see chapter 12.3 for details). To remove the scattering effect (see chapter 9.4) the spectra were transformed using a Savitzky-Golay first derivative [47] with a three point filter window first.

Only samples from the Faizabad and Yavan test areas were selected as the samples of Varzob area were not taken at the time of the sample selection.

For each cluster (for example FA88) evenly distributed samples regarding the PCs were selected and the samples of the center points (plot number 05) were added if they had not already been selected before. The final selection contained 252 samples. For further information see B. Wolfgramm [Ph. D., forthcoming].

8.2 Selection of soil laboratory

For the selection of the soil laboratory, ten soil samples were partly selected from the existing sample set and partly newly collected because of the big quantity that was needed. Every sample was split up into 6 subsamples and two of this subsamples were given to a laboratory at a time as double blind samples. In addition the VIS/NIR/SWIR spectra of the ten samples were measured.

In a first step three soil laboratories were tested against each other:

- Soil Science Research Institute SSRI, Dushanbe, Tajikistan
- State Institute of Land Planning GIPROZEM, Dushanbe, Tajikistan
- Geographic Institute of University of Berne GIUB, Switzerland

It was planned to do the chemical analysis in a soil laboratory in Tajikistan. On the one hand to enforce the collaboration with local scientists, to profit from their knowledge and to allow for continuous enlargement of the soil spectral library at later stages, on the other hand to avoid the complicated shipping out of the country and to save costs.
Figure 8.1  Soil organic matter (%) for the 10 laboratory test samples. Two subsamples measured by each laboratory. Bottom: Subsamples A and B for each soil laboratory plotted against each other. SSRI (square), GIUB (circle), GIPROZEM (triangle). Source: B. Wolfgramm
The idea was to test the Tajik laboratories against a laboratory in Switzerland disposing of the newest equipment to gain information on the laboratories accuracy.

In figure 8.1 the soil organic matter of the two subsamples for each soil laboratory are plotted against each other. The more a point approximates the \((x=y)\) line, the less is the measured difference in organic matter content between the two subsamples. Particularly the GIPROZEM results are not very trustworthy as they show high absolute differences of up to 1.5 % organic matter which corresponds to a relative error of 80 %.

Since the Tajik laboratories could not provide all the requested analysis and results showed rather high within lab variation it was decided to conduct the analysis in the ICRAF laboratory in Kenya. This was a cost-saving alternative to a Swiss laboratory and the collaboration between B. Wolfgramm and K. Shepherd who is working for the ICRAF and has published several papers about modelling soil properties based on spectral information \([50/57]\) made it even more interesting.

For further information see B. Wolfgramm [Ph. D., forthcoming]..

**8.3 Methods for soil chemical analysis**

The chemical analysis of the soil samples included the total of carbon, soil organic carbon (SOC), the total of nitrogen, pH, CaCO3, extractable P, exchangeable Ca, Mg and K, the fractions clay, silt, sand. One additional chemical property was calculated from the analysed data: The cation exchange capacity (CEC) was simply approximated by summing the exchangeable calcium, magnesium and potassium. The methods for soil chemical analysis are shown in table 8.1.

**8.4 Accuracy assessment and repeat measurements**

To determine the accuracy of the ICRAF soil laboratory, the ten laboratory test samples (see chapter 8.2) were added to the sample set as double blind samples. Coefficients of variation (CVs) were calculated for the results of these double blind samples, averaged and compared to the ICRAF lab-own CVs and to international standards.

\[
CV = \sqrt{\frac{\sum_{i=1}^{q} \left( x_i - \frac{1}{q} \sum_{i=1}^{q} x_i \right)^2}{\frac{1}{q} \sum_{i=1}^{q} x_i}}
\]

Apart from this, samples which could not be properly predicted in a first modelling attempt were listed and a repeat analysis was carried out at the ICRAF laboratory. If the CV of the first and the second measurement of a sample did not strongly excess the values of the labown CVs the average of the two samples was used for modelling. If this was not the case the second measurement was used.
<table>
<thead>
<tr>
<th>Soil property</th>
<th>Mean CV (%) of LT samples</th>
<th>CV (%)</th>
<th>CV (%)</th>
<th>Repeat analysis: Number of samples</th>
<th>Repeat analysis: mean CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total C</td>
<td>0.9</td>
<td>2.5</td>
<td>4-5</td>
<td>35</td>
<td>6.0</td>
</tr>
<tr>
<td>SOC</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total N</td>
<td>2.4</td>
<td>2.7</td>
<td>7-8</td>
<td>35</td>
<td>13.2</td>
</tr>
<tr>
<td>pH</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Extr. P</td>
<td>13.2</td>
<td>-</td>
<td>8-9</td>
<td>147</td>
<td>16.8</td>
</tr>
<tr>
<td>Exch. Ca</td>
<td>4.0</td>
<td>5.1*</td>
<td>5-7</td>
<td>31</td>
<td>8.8</td>
</tr>
<tr>
<td>Exch. Mg</td>
<td>3.7</td>
<td>3.1*</td>
<td>8-11</td>
<td>21</td>
<td>3.2</td>
</tr>
<tr>
<td>Exch. K</td>
<td>3.7</td>
<td>3.1*</td>
<td>9-10</td>
<td>21</td>
<td>3.2</td>
</tr>
<tr>
<td>Clay</td>
<td>1.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Silt</td>
<td>2.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sand</td>
<td>8.3</td>
<td>-</td>
<td>3-4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 8.1 Methods for soil chemical analysis. Source: B. Wolfgramm

Table 8.2 Coefficients of variation (CV) of the double measured samples (LT samples) compared to lab-own and international CVs. NRCS stands for Natural Resources Conservation Service from the United States Department of Agriculture and has been taken as an international standard. An asterisk is placed behind values which are achieved with different methods than those used in the ICRAF laboratory. For fields marked with a dash ( - ) data was not available. Source: B. Wolfgramm
The resulting CVs of this procedure can be seen in table 8.2.

It is obvious that the ICRAF lab-own CVs are partly better than the calculated CVs for the Tajik soil samples. But this is no cause for concern, because the laboratory normally analyses African soils that show a lot of differences to Tajik soils and because the discrepancy between the CVs is not too large.
9. Preprocessing of spectral data

The raw spectra files were converted to columnar ascii files using the program STable.exe. A graphical analysis led to the exclusion and remeasuring of obviously erroneous spectra probably due to a detector failure. For every soil sample the two measurements were averaged and labelled in Microsoft Excel.

Several preprocessing steps were needed to prepare the spectral data for the modelling process. In a first step errors occurred during the measurement process (chapter 9.1) and wave bands prone to errors (chapter 9.2) had to be removed. The large quantity of data had to be reduced (chapter 9.3) and the data had to be transformed to extract useful information (chapter 9.4). The methods for the preprocessing steps will be specified in the following subchapters.

9.1 Correction of steps

FieldSpec Pro FR consists of three spectrometers: A VNIR- (350-1000nm), a SWIR1- (1001-1800nm) and a SWIR2-spectrometer (1801-2500nm). Each disposes of a separate fiber optic bundle with slightly different sample views. These different viewing angles can cause steps at the changeover of the three spectrometers. The nearer a sample is located to the probe the stronger the influence of the different sample views becomes. As the distance in the applied measuring setup is given by the muglight and cannot be changed, the steps have to be removed afterwards. Under the assumption of continuous curves in spectroscopy it was decided on simply removing the steps by adding the offsets at 1000nm and 1800 nm to the rest of the spectrum. Figure 9.1 shows a soil spectrum before (black colored) and after the step correction (red colored). The dashed lines indicate the spectrometer changeovers at 1000 and 1800 nm.

9.2 Omitting wave bands with low signal to noise ratio

Bands with a low signal to noise ratio (SNR) should not be used further because of their high proneness to errors. For this purpose an SNR for the used measurement setup was calculated. A soil sample was measured twenty times in series without changing its position. The differences still occurring between the measurements are the cause of the device error. The range of the repeat measurements was taken as noise and the average as signal to calculate the SNR. The resulting SNR is given in figure 9.2, smoothed with a 30 nm window.

Bands with SNR lower than 90 were removed from data set. This affected the wavelengths from 350 to 429 and 2441 to 2500 nm.
Figure 9.1  Correction of steps occurring at spectrometer changeovers at 1000 (VNIR-SWIR1) and 1800 nm (SWIR1-SWIR2) due to different sample views.

Figure 9.2  Signal to noise ratio of the applied measurement setup. Smoothed with a 30 nm window.
9.3 Data compression

The enormous quantity of data - 1465 soil samples with 2010 spectral values each - demanded data reduction. It would have been advantageous if preliminary information about the spectral features of the chemical properties had been available. In this case spectral regions of no interest could have been omitted for data reduction and the spectral resolution of 1 nm could have been retained in the regions of interest. But as specified in the chapter about molecular vibrations (chapter 3.1) spectral features are not well defined in the VIS/NIR/SWIR region for most of the chemical properties. It was therefore necessary to keep the whole spectral range. The data was resampled to a 10 nm resolution by keeping every tenth nanometer value. No averaging was used to avoid falsifications.

Figure 9.3 shows the loss of information caused by the resampling at the example of a spectral feature at 2340 nm. The peaks and minima that got lost are in the range of one thousandth of absolute reflectance and therefore not critical for the modelling process.

9.4 Scatter Correction

The differences in grain size distribution affect the spectral response. As the grain size distribution at the bottom of the Petri dish depends strongly on the manner of filling in the soil samples these differences have to be removed. This can be demonstrated by plotting the two spectra of a double measured soil sample, respectively with newly filling the samples in the Petri dishes, as is shown in figure 9.4. Although the two spectra originate from the same soil sample they show a difference of up to 6% absolute reflectance.

Among many others Shepherd et al. [57], Chang et al. [11/12] and Cozzolino et al. [14] propose the use of first derivates for scattering correction. Jarmer et al. [35] chose continuum removal. Several other methods are mentioned as multiplicative scatter correction, second derivative, mean centering, variance scaling [40].

The most promising methods, first derivative and continuum removal, were chosen for this work.

Continuum removal

The continuum was removed using the spectral library builder of ENVI. Continuum removal normalizes reflectance spectra to allow comparison of individual absorption features from a common baseline. The continuum is a convex hull fit over the top of a spectrum utilizing straight line segments that connect local spectra maxima. The first and last spectral data values are on the hull and therefore the first and last bands in the output continuum-removed data file are equal to 1.0. [18]

Figure 9.5 shows that the continuum removal is an effective way to remove the scattering effect. The two spectral measurements of the soil sample FA640203 are approximately identical. Additionally
Figure 9.3  Spectral feature at 2340 nm with 1 nm (left) and resampled 10 nm (right) resolution.

Figure 9.4  Two spectral measurements of soil sample FA640203
the continuum removal fetches important features as figure 9.6 shows. The black curve shows the correlation between the total of C and the several wavelengths. It is nearly changeless at a correlation coefficient of \( R = 0.5 \) which makes it difficult to extract useful information. By contrast the red curve showing the correlation between total C and the continuum removed wavelengths is showing well-defined features as for example the absorption feature at 2340 nm.

**First derivative**

The first derivative was performed using a Savitzky-Golay filter with a 20 nm window. Figure 9.7 shows that the first derivative is not as effective in removing the scattering effect as continuum removal. The two spectra of the soil sample FA640203 still feature differences, particularly concerning the minima and maxima. Nevertheless it was decided to work with first derivative data too because promising results were achieved therewith by Shepherd et al. [57], Chang et al. [11/12], Cozzolino et al. [14].
**Figure 9.5** Two spectral measurements of soil sample FA640203 continuum removed

**Figure 9.6** Correlation between amount of total C and the different wavelengths. Raw (black curve) and continuum removed spectral data (red curve).

**Figure 9.7** First derivative of two spectral measurements of FA640203
10. Calculation of CIE color variables

The human retina contains two groups of sensors: The rods to measure the luminance and the cones to measure color. Three types of cones form a tristimulus measuring system: a blue, a green and an orange sensor. The Comission Internationale de l’Eclairage (CIE) being responsible for color standards since 1913 defined several standard observers with exact defined color-matching functions. For each wavelength from 360 to 830 nm the sensitivity of the observer’s three cones for blue, green and orange is defined (see figure 10.1) [27]. The CIE 1964 Supplementary Standard Colorimetric Observer was utilized for this work.

The tristimulus values $X, Y, Z$ are calculated from the reflectance spectra of the soil samples using the following formulas:

\[
X = k \int_{360 \text{ nm}}^{830 \text{ nm}} P(\lambda) \bar{x}(\lambda) d\lambda \\
Y = k \int_{360 \text{ nm}}^{830 \text{ nm}} P(\lambda) \bar{y}(\lambda) d\lambda \\
Z = k \int_{360 \text{ nm}}^{830 \text{ nm}} P(\lambda) \bar{z}(\lambda) d\lambda
\]

The reflectance curves of the soil samples are first multiplied, point by point, with a curve describing the radiant power distribution of the light source to obtain absolute values ($P$). A standard spectrum for the muglight was provided by ASD and is showed in the appendix (figure A.1). The new curve is then multiplied separately with the three color-matching functions $(\bar{x}, \bar{y}, \bar{z})$ which describe the response of the human eye to light. The areas under these three curves are calculated. Finally an arbitrary factor $k$ is chosen such as the maximum of $Y$ is equal to 1.

Additionally, the tristimulus values are converted to chromaticity coordinates $x, y, z$ which are independent of the luminance:

\[
x = \frac{X}{X + Y + Z} \\
y = \frac{Y}{X + Y + Z} \\
z = \frac{Z}{X + Y + Z}
\]

[27,60]
Figure 10.1  CIE 1964 color-matching functions
11. Separation of soil types

The investigation area and with it the three test areas are not homogenous regarding soil types. On the one hand this fact was chosen on purpose to clarify if the developed models could be applied to several soil types and to investigate the variability of soils, on the other hand it was not clear from the beginning that in total as much as four types of soil occured in this area. In first modelling attempts it became clear that the data set included different populations that could not be modelled together as no good results were obtained. Therefore it was decided to separate the occurring soil types.

11.1 Occuring soil types

The soil types on which the main focus lay were the ones that formed on loess, the calcareous montane cinnamonic soil and the typical montane cinnamonic soil. The interfering soil types include high mountain meadow steppe soils, paleo soils and samples from river sediments. In figure 11.1 the different regions where these soil types approximately occur are separated.

The calcareous and montane cinnamonic soils both formed on loess. As the former developed in dry climate the lime is not washed out totally and lime-concretions occur. The latter formed in higher regions with a damper climate and the lime was washed out deeply in the soil profile. These two soil types were not separated as the concern lay in both and differences were small apart from the lime content. A typical spectrum for the loess soils is given in figure 11.2 (I). Furthermore these two soils are addressed as brown carbonate soils.

The high mountain meadow steppe soils are a kind of leptosols. They are built on grano-dioritic bedrock which occurs mainly in the changeover to the mountains but also sporadically between the loess soils where the loess cover is diminished by erosion and the bedrock shows at the surface. These soils have a very special spectrum (figure 11.2, IV) and are therefore easy to separate. The spectrum shows a negative slope between 1000 and 1400 nm and the absolute reflectance is relatively low over the whole spectrum.

The paleo soils formed during interglacial periods and have been covered with loess which conserved the soils until erosion brought them back to the surface. Although the paleo soils were built under different conditions and have different chemical and physical soil properties than the actual soils, it was very difficult to separate them. In the spectrum shown in figure 11.2 (II) no obvious difference to the loess spectrum can be observed.

The samples of river sediments were taken in the valley floor from dry riverbeds. These river sediments consist of material from the mountains which has a different composition than loess and makes it difficult to include these samples in the modelling process. The spectrum (figure 11.2, III) shows a low absolute reflectance and the slope between 1000 and 1800 nm only changes slightly compared to the other spectra.
Figure 11.1  Landscape in Faizabad test area. Separation of areas: (I) High mountain meadow steppe soils, (II) Loess area: loess soils and sporadic paleo soils and high mountain meadow steppe soils, (III) samples of river sediment on the valley floor. Photo: B. Wolfram.
Part III, Methods

11.2 Approach of separation

The separation of the four sample groups would have been difficult using only the properties of the spectral curves as was demonstrated with figure 11.2. Particularly the differences between the group of paleo soils and brown carbonate soils are evanescent. In figure 11.3 the lab photos of four soil samples each pertaining to one of the soil groups are displayed. The color differences are clearly visible, even between brown carbonate soil (first from left) and paleo soil (third from left). Therefore the calculated CIE color variables X, Y, Z, the CIE chromaticity coordinates x,y,z and the ratio x/y were used for the separation into soil groups.

For this purpose a classification tree was built in CART [10] (see chapter 12.4) using cross-validation. The samples to calibrate the tree were chosen differently. As special observations were noted in the sampling protocol examples for the leptosols on granodiorite bedrock and the samples of river sediment could easily be extracted from the data set. As the paleo soils are randomly distributed over the test areas and not simple to distinguish from the brown carbonate soils a new small data set was sampled by B. Wolfgramm in Summer 2005. Under the assistance of P. M. Sosin, an expert for paleo soils from the Soil Science Research Institute in Dushanbe, about 20 samples of paleo soil and 20 samples of brown carbonate soil were collected.

The best CART model resulted in six final nodes, two for brown carbonate samples, two for leptosol, one for paleo soil and one for samples of river sediments. Table 11.1 shows the results of the 10 fold cross validated classification tree. ‘Total cases’ describes how many samples have been classified as members of a specific soil group and ‘percent correct’ describes the percentage of correct allocated samples. Since the main aim was to determine a homogenous sample set of brown carbonate soil samples the result of 97 % correct samples is a very good result and misclassification samples among the other groups are not considered a severe restriction to the model accuracy.
Figure 11.2  Typical spectra of the four separated sample types. Brown carbonate soil (I), paleo soil (II), river sediment (III), leptosol (IV).
Part III, Methods

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Total cases</th>
<th>Percent correct</th>
<th>Leptosol N=44</th>
<th>Brown carbonate N=108</th>
<th>Paleo soil N=68</th>
<th>River sediments N=21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leptosol</td>
<td>59</td>
<td>69</td>
<td>41</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Brown Carbonate</td>
<td>103</td>
<td>97</td>
<td>0</td>
<td>100</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Paleo soil</td>
<td>63</td>
<td>92</td>
<td>1</td>
<td>4</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>River sediments</td>
<td>16</td>
<td>81</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 11.1 Result of the CART classification tree for separation of the four soil groups. Source: B.Wolfgramm

Figure 11.3 Lab photos showing color differences between four soil groups. From left to right: brown carbonate soil, leptosol, paleo soil, river sediment. Photos: B.Wolfgramm
12. Calibration methods

12.1 Selection of calibration methods

Literature research revealed that the performance of the different statistical methods can vary strongly depending on the area of application. Therefore it was determined to apply not only one but three methods. Three types of statistical methods were selected:

- A classical hyperspectral remote sensing approach: The use of continuum removed data with multiple linear regression
- An often used more complex statistical method: PLSR or PCR with 1.derivative data
- A new, rarely used method, that has achieved promising results: The regression tree software CART [10] with 1.derivative data

The decision between PLSR and PCR was taken based on a test made with a spectral data set of Kenyan soils kindly provided by C. Hett [25], as no own data was available yet at this moment. The test was carried out with the chemical soil property 'total of C' and the first derivative of spectral data, the used software was Unscrambler [56].

Figure 12.1 shows the trend of the RMSE for the validation data set for an increasing number of model components. The red curve is describing the PCR result, the black curve the PLSR result. In comparison PLSR already reaches its minimal RMSE with 7 components, PCR shows a minimal RMSE with 14 components. But the best RMSE values are at approximately 0.27 for both, and the $R^2$ values at 0.86. This observation that PLSR and PCR give comparable results corresponds to the statement of Naes et al [43]:

"Experience with NIR data has shown that PLS regression can give good prediction results with fewer components than PCR. [...]. This may in some cases lead to simpler interpretation. Using the optimal number of components in each case, however, the two methods often give comparable prediction results. Several simulation studies support this conclusion. Cases exist in practice where PLS regression is better as do cases where PCR is better, but the difference is not usually large."

So although PLS was the most frequently used calibration technique PCR seemed more appropriate for this study. On the one hand because of the comparable results that the two methods achieve, on the other hand the implementation of PLS in IDL [29] would have been more complicated and time consuming.

The Unscrambler software [56] used for the test was easy to handle but had the disadvantage that the component chosing process was hidden and could not be influenced. To avoid this disadvantage it was decided to code an own program in IDL for the use of PCR and MLR. The design of this program is specified in chapter 13. For the application ‘decision tree’ the software CART from Salford Systems was used.
12.2 Multiple linear regression

Multiple linear regression (MLR) aims to explain a dependant variable $y$ as a linear combination of $i$ independent variables $x_i$. The model can be written as:

$$y = b_0 + \sum_{i=1}^{k} b_i x_i + f$$

In order to facilitate the estimation of the regression coefficients the model has to be written in matrix form. Three vectors with $N$ elements, $y$ for the observations, $f$ for the errors and $b$ for the regression coefficients, and one $N\times K$ (where $N=$number of samples, $K=$number of predictor variables) matrix $X$ for the predictor variables are defined:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_K \end{pmatrix}, \quad f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1K} \\ 1 & x_{21} & x_{22} & \ldots & x_{2K} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{NK} \end{pmatrix}$$

The matrix form of the multiple linear regression model can now be written as:

$$y = Xb + f$$

For MLR, the equation is usually fitted to the data using the least squares estimation for $b$:

$$\hat{b} = (X'X)^{-1}X'y$$

In applications with a large number of explaining variables the variable selection process is very important as not all $x$-variables can be used for the model due to overfitting, and multicollinearity between the $x$-variables should be avoided. [16/43]
Figure 12.1  RMSE for the validation data set for different numbers of components. Red curve: PCR. Black curve: PLSR.
12.3 Principal component regression

Principal component regression is based on the same principles as the multiple linear regression. The only difference is the fact that the independent variables, in this case wavelengths, are transformed to principal components before using them as an input for the model. Principal components (PCs) are newly derived variables characterized by the following properties:

- the PCs are linear combinations of the original variables
- the PCs are uncorrelated
- the first PC captures as much as possible of the variability in all the original variables
- each additional PC accounts for as much of the remaining variability as possible

This method is highly suitable for the regression analysis of spectral data. It is due to the near-multicollinearity often found among spectral measurements. This means that some of the variables can be written approximately as linear functions of other variables. The predictor will in such cases be very unstable and lead to poor prediction performance. The near-multicollinearity can simply be shown by plotting lagged correlation of the spectral measurement of a soil sample (figure 12.2). The correlation is very high up to lag 10 (equal 100 nm). This shows that a lot of common information is present in adjacent variables which is removed by constructing the PCs, as they are totally uncorrelated.

The PCs are calculated by finding the eigenvectors of the variance matrix of the original variables. The eigenvectors are used as weight vectors for the construction of the PCs and the corresponding eigenvalues give information about how much of the original variance has been captured in each PC. [43]

This allows the explanation of the largest part of the variance of the 210 original variables using only a few PCs. A number of 20 PCs was calculated from the first derivative spectral data using the IDL function PCOMP. PCOMP calculates the PCs based on the correlation of the original variables. The explained variance of the calculated PCs for brown soil is shown in table 12.1. The first four PCs already explain 80% of the variance of the original 210 variables, all twenty PCs even 93%.

12.4 Regression tree

CART, a program by Salford Systems, was utilized for building the decision tree models. This software is based on the decision tree procedure developed by renowned Berkeley and Stanford Statisticians in 1984. CART is used for data analysis in a large field of applications from economics, medical diagnostics, pharmacy, sociology to engineering and environmental studies. It can be applied to tasks of data mining, classification and regression.

Shepherd et al. have used CART in studies for predicting chemical soil properties from VIS/NIR/
### Table 12.1
Explained variance for the principal components of spectral data of brown soils.

<table>
<thead>
<tr>
<th>PCs</th>
<th>Explained Variance</th>
<th>Sum of E.V.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.PC</td>
<td>40 %</td>
<td>40 %</td>
</tr>
<tr>
<td>2.PC</td>
<td>24 %</td>
<td>64 %</td>
</tr>
<tr>
<td>3.PC</td>
<td>9 %</td>
<td>73 %</td>
</tr>
<tr>
<td>4.PC</td>
<td>7 %</td>
<td>80 %</td>
</tr>
<tr>
<td>5.PC</td>
<td>3 %</td>
<td>83 %</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>20.PC</td>
<td>0.3 %</td>
<td>93 %</td>
</tr>
</tbody>
</table>

**Figure 12.2**
Autocorrelation of wavelength values for a soil sample in 10 nm resolution for lags 0 to 30. Starting at 420 nm.
SWIR spectra in Madagaskar and have gained promising results [57].

A decision tree is a flow chart which represents a classification system or predictive model. The tree is structured as a sequence of simple questions, and the answers to these questions trace a path down the tree. The end point determines the classification or prediction made by the model. CART uses binary splits to grow a decision tree that divides each parent node into exactly two child nodes by posing yes/no questions at each decision node. CART searches for questions that split nodes into relatively homogeneous child nodes, such as a group of samples with high and one with low total carbon content. As the tree evolves, the nodes get increasingly homogeneous. The mean values monitored in the terminal nodes of a tree are used as predicted values. An example for a CART single tree is given in figure 12.3.

The problem now is that a single decision tree has a finite number of end nodes and can therefore not predict continuous data accurately. The tree used to predict the total C content in figure 12.4 for example consists of 8 terminal nodes and is able to predict 8 different values only for total C resulting in large prediction errors.

In order to gain more precise results the ‘committee of experts’ can be used in CART, a function that combines separate CART trees into a single predictive engine. The fact that single decision steps can’t be followed anymore in this black box model is a big disadvantage. Table 12.2 shows how the results of different numbers of trees to combine were compared. For every chemical soil property a number of trees from 10 to 100 was tested. The CART settings were the following: Combine method ‘bagging’, linear combinations allowed, minimum node size 2/1. For validation the same hold-out validation data set (chapter 14.1) as for MLR and PCR was used.
Table 12.2 Results of validation data set for different soil organic C CART models

<table>
<thead>
<tr>
<th>N of trees</th>
<th>R²</th>
<th>RMSE</th>
<th>RMSPE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.70</td>
<td>0.412</td>
<td>53 %</td>
<td>0.349</td>
</tr>
<tr>
<td>25</td>
<td>0.72</td>
<td>0.395</td>
<td>53 %</td>
<td>0.353</td>
</tr>
<tr>
<td>50</td>
<td>0.71</td>
<td>0.404</td>
<td>56 %</td>
<td>0.360</td>
</tr>
<tr>
<td>75</td>
<td>0.73</td>
<td>0.397</td>
<td>54 %</td>
<td>0.351</td>
</tr>
<tr>
<td>100</td>
<td>0.74</td>
<td>0.390</td>
<td>54 %</td>
<td>0.351</td>
</tr>
</tbody>
</table>

Figure 12.3 Example for a CART single tree for total C (%) with four terminal nodes. A sample with values for wavelength 440 nm > 0.0075 (first derivative) and for wavelength 1330 < 0.0007 (first derivative) ends up in terminal node 3 and obtains the value 4.64 for total C.

Figure 12.4 Predicted values for total C modelled by a CART single tree with eight terminal nodes plotted against measured value.
13. Development of calibration program in IDL

For multiple linear regression and principal component regression a calibration program was built in IDL. The source code is noted in the appendix. The line numbers of program parts explained in this chapter are given in accolades, for example {4-6}.

As input data a calibration and a validation data set are used. The data separation was made with random numbers into two third calibration and one third validation data. Further information on this topic is given in chapter 14.1.

Shepherd et al. [50/57] have used transformations for the chemical variables such as Box-Cox [7] to obtain normally distributed data which improved results. This was integrated as an option in the calibration program. The program lists the result of a chi square test on normal distribution for common transformations. The most accurate one can be selected by the user {128-238}.

The variables for the calibration model are selected using a stepwise procedure (chapter 13.1). A step includes either adding a new variable to the model, removing a variable or removing outlying samples. After each step the model assumptions (chapter 13.2) are checked, the model is tested (chapter 13.3) and validated (chapter 13.4). The stepwise procedure ends if none of the remaining variables contributes significantly to the calibration model.

The program generates two outputs: a log file with exact information on every taken step and several plots. Examples will be given in the following three subchapters. These outputs can additionally be saved as txt and postscript files.

13.1 Variable selection

A user-specified number of variables, default set to 20, is preselected by maximizing the explained variance. Then these variables are added to the model following a stepwise procedure: A step includes either adding a new variable to the model, removing a variable or removing outlying samples. In order to assure that every variable contributes significantly to the model and to avoid overfitting, a two tailed t-test is applied to the regression coefficients after each step. Every regression coefficient is tested wheter it is significantly different from 0. In other words the null hypothesis $H_0: b_j = 0$ is tested against the alternative hypothesis $H_1: b_j \neq 0$. The test statistic [16]

$$ t = \frac{b_j}{s \sqrt{C_{ii}}} \quad \text{where} \quad C = (X'X)^{-1} \quad s = \sqrt{\frac{\text{SSE}}{n - p - 1}} $$

follows a t-distribution with $df = n - p - 1$ degrees of freedom with n: number of samples and p: estimated coefficients, SSE: sum of squared errors. As the level of significance alpha = 0.05 was chosen. The following example of an output for a pH-model with 3 explaining variables shows that
all coefficients are significantly different from 0. The calibration program only uses the critical value for df = 100 but gives the actual df, in this example 122. If the actual df differs too much from 100 the critical value has to be adjusted in the source code.

<table>
<thead>
<tr>
<th>variable</th>
<th>coefficient</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>42.5930</td>
<td>11.5207</td>
</tr>
<tr>
<td>1860</td>
<td>-53.5700</td>
<td>-8.16808</td>
</tr>
<tr>
<td>x</td>
<td>-20.3947</td>
<td>-4.49429</td>
</tr>
<tr>
<td>1700</td>
<td>28.3053</td>
<td>3.74914</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(df=122)</td>
<td>critical t-value: 1.987</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(for alpha=0.05, df=100)</td>
<td></td>
</tr>
</tbody>
</table>

{513-518, 759-785}, [13,16]

13.2 Verification of model assumptions

The assumptions for the least squares estimation of the regression coefficients have to be verified to assure that the least squares criterion is a bias free estimator and therefore appropriate for the data to be modelled. These assumptions include:

(1) The Errors $E_i$ are independent
(2) The expectation is $\epsilon (E_i) = 0$
(3) They all have the same variance $\text{var} (E_i) = \sigma^2$
(4) They follow a normal distribution
(5) No outliers bias the estimation

Various residual plots are used for the verification of the assumptions. Because the exact errors are unknown the residuals are used instead.

(1) Assumption of Independence

If the observations follow a certain order of time or space the residuals have to be plotted that way. In this work two orders are present:

- the order of the chemical analysis of the samples which is not available as the analysis was carried out externally.
- the order of the spectrometer measurements of the samples which can be reconstructed. However, it is not really influential in the model building process because only about a fifth of the total sample set is involved.

This assumption can thus not be checked.

[51]
(2)/(3) Assumptions expectation $\varepsilon (E_i) = 0$ and constant variance

Checking this assumptions requires to consider with which quantities the variance could be connected. Then the residuals are plotted against these quantities. The most frequent violation of this assumption is that with increasing value of the target variable the variance increases. This can be examined in the Tukey-Anscombe plot in which the residuals are plotted against the fitted values. If a trend is observed and the residuals are not continously distributed around zero, the assumptions are not achieved.

In the left plot of figure 13.1 the negative residuals increment with increasing fitted value. Furthermore the residuals are plotted against the dependant variables $x_i$.

(4) Assumption of normal distribution

Normal distribution of the residuals is not immediately neccessary for the least squares estimator, but for the applied F- and t-tests. This assumption must be checked by using a normal plot and by applying a chi square test.

For the normal plot the residuals may be normalized by subtracting the arithmetic mean and dividing by the standard deviation. Now the quantiles of these normalized residuals are plotted against the quantiles of the standard normal distribution (tabulated values). In this work, two percent steps were used. The plotted points should form a straight line if the residuals are normally distributed.

The $\chi^2$-test compares the expected frequency with the theoretical frequency. On this account the data has to be grouped. The standardized residuals were divided into four groups with the limits:

- mean – 1 standard deviation
- mean
- mean + 1 standard deviation

Null hypothesis/alternative hypothesis:

$H_0$: The sample comes from a normally distributed population

$H_1$: The sample does not come from a normally distributed population

As level of significance 5% was chosen. With four classes and two estimated parameters (mean and standard deviation) the degree of freedom resulted in 1. The critical chi square value $\chi^2_{\text{critical}} = 3.841$ was looked up.

The test value was calculated using the formula
Figure 13.1  Tukey Anscombe plots. Left: assumptions violated, variance not constant. Right: Assumptions achieved, no trend apparent.
Figure 13.2  Normal plots. Left: not normally distributed. Right: Normally distributed.
\[ \chi^2 = \sum_{i=1}^{k} \left( \frac{x_i - t_i}{t_i} \right)^2 \]

where \( x_i \): observed frequency in group i, \( t_i \): theoretical frequency in group i.

If \( \chi^2 \geq \chi^2_{critical} \) then \( H_0 \) is rejected. The sample does not come from a normally distributed population. If \( \chi^2 < \chi^2_{critical} \) there is no reason to reject \( H_0 \).

\{351-381\}, [52]

(5) Detection of influential observations

The least squares estimation is sensitive to outliers. Observations from another population, for example different soil types, or observations with large measurement errors can falsify the estimation of the regression coefficients. To detect such influential observations the cook’s statistic is utilized.

First the hat matrix \( P \) must be calculated:

\[ P = X (X^T X)^{-1} X^T \]

The diagonal entry \( p_{ii} \) of \( P \) measures the leverage of the observation i. Observations with high \( p_{ii} \) values are outliers in the space of x-variables but not stringently influential observations. If an outlier in the x-variable space lies exactly on the estimated regression line then omitting this observation would not alter the estimation. Therefore the cook’s statistic includes the residuals in order to investigate the effect on the estimation:

\[ C_i = \left( \frac{p_{ii}}{1 - p_{ii}} \right) \frac{\hat{e}_i^2}{p} \]

The first term is the ratio (variance of the ith predicted value)/(variance of the ith residual), \( \hat{e}_i \) is the ith studentized residual (residual divided by it’s standard error).

\{462-481, 787-813\}

[13/16]

13.3 F-tests

Two F-tests were applied to the model. One F-test for the overall regression equation and one for
comparing models of different sizes.

The F-test for the overall regression equation verifies if a statistically significant regression is obtained. The null hypothesis $H_0$: all $b_i = 0$ is tested against the alternative hypothesis $H_1$: not all $b_i = 0$. The test value

$$F_{overall} = \frac{SSR(p-1)}{S^2}$$

where $SSR = b'X'Y - \frac{(\sum Y)^2}{n}$

is treated as an $F(p-1, n-p)$ variate with $p$: number of estimated coefficients and $n$: number of samples. As level of significance $\alpha = 0.01$ was used. [16],[490-500]

The F-test for comparing models tests if the actual model explains significantly more of the variance of the dependent variable than the previous model with one explaining variable less. Overfitting of the model can thereby be avoided. The test value follows an $F$-distribution with $(p+1-k)$ and $(n-p-1)$ degrees of freedom

$$F_{comparing} = \frac{[SSE(RM) - SSE(CM)](n-p-1)}{SSE(CM)(p+1-k)}$$

where $SSE$: Sum of squared errors, $RM$: reduced model, $CM$: complete model, $p$: number of explaining variables in complete model, $k$: $p$ - number of explaining variables in reduced model.

[504-511][13]

### 13.4 Model verification

The model verification is a highly important step in the modelling process. It has to be verified if the formula used in the program are implemented correctly. A number of tests were carried out for this purpose:

- The criteria for the variable selection were checked manually to ascertain that the right variables are selected by the program.
- The same data set was applied for calibration and validation to verify if the accuracy measures result in the same values.
- The correctness of the accuracy measures was further checked by importing the modelled data in Microsoft Excel and recalculating them.
14. Model validation

14.1 Selection of calibration/validation samples

The data set of the chemically analysed soil samples was split up into two third calibration and one third validation samples using random numbers. The distribution for each chemical property was placed under a comparison between the calibration and validation data set using histograms to assure that the differences were not too big. An example is given in figure 14.1, the rest of the histograms are pictured in the appendix (figures A.3 to A.6). The scale of the y-axis is not equal for the calibration and the validation set such as to ease comparison.

The histograms for nitrogen show that the random selection split the sample set into two equally distributed parts, as the form of the histograms is nearly the same. The only difference is that the validation data set has a few samples with higher values than the calibration data set, which should not provoke any problems in the modelling process. The histograms for phosphorus present a totally different image. The distribution of the chemical property is highly special with many samples with low phosphorus content and very few samples with extremely high phosphorus content.

14.2 Accuracy measures

After a calibration equation is computed, it is essential to determine its ability to predict unknown y-values.

On the one hand the within model accuracy is calculated and on the other hand the ability to predict chemical properties is tested for the validation data set. For this purpose four measures and a plot are used. The shown examples come from a pH model with 4 explaining variables. The text output from the IDL calibration program is given below:

```
Model Calibration

multiple R square: 0.603008
RMSEc: 0.125854
RMSPEc: 2 %
MAPEc: 1 %

Model Validation

multiple R square: 0.601797
RMSEP: 0.156521
RMSPEp: 2 %
MAPEp: 2 %
```

**Multiple R²**

Multiple R² is the ratio of variance of the dependant variable which is explained by the model.
Figure 14.1  
Histograms for the calibration (left side) and validation set (right side) for nitrogen (top) and phosphorus (bottom).
Part III, Methods

$$R^2 = 1 - \frac{\sum (y_{measured} - y_{predicted})^2}{\sum (y_{measured} - mean(y_{measured}))}$$

A high $R^2$ is not immediately related to a good prediction ability as $R^2$ only shows the strength of correlation. However, the correlation between measured $y$ and predicted $y$ should follow a straight 1:1 line to make the calibration model predict accurately. Aside from this, $R^2$ can be strongly influenced by outliers. In order to investigate these disadvantages the measured values are plotted against the predicted values.

**Plot measured $y$ versus predicted $y$**

A good calibration will lead to observations falling close to a 45° straight line as shown in figure 14.2. Such a plot has the advantage that it may also be used to identify regions with different levels of prediction accuracy.

**RMSE**

RMSE is the Root Mean Square Error, a standard measure for errors.

$$RMSE = \sqrt{\frac{1}{N} \sum (Y_{predicted} - Y_{measured})^2}$$

Without knowing the distribution and location of the target variable the RMSE is not that meaningful. For example, a large difference exists if the RMSE is 1.5 for a target variable with the average of 10 or 1000. For that reason two measures for relative errors were implemented additionally. The squaring makes this measure susceptible to high errors.

**RMSPE**

The Root Mean Square Percentage Error is the relative equivalent to the RMSE.

$$RMSPE = 100 \sqrt{\frac{1}{N} \sum \left( \frac{Y_{measured} - Y_{predicted}}{Y_{measured}} \right)^2}$$

The squaring makes this measure susceptible to high errors as it was the case for the RMSE.
Figure 14.2  Measured $y$ versus predicted $y$ plot. Left: for calibration data set. Right: for validation data set.
MAPE

The Mean Absolute Percentage Error:

\[
MAPE = 100 \frac{1}{N} \sum \left| \frac{Y_{\text{measured}} - Y_{\text{predicted}}}{Y_{\text{measured}}} \right|
\]

14.3 Validation based on repeat measurements

During the spectral measurements of the soil samples two clusters were measured twice to test the repeatability of the three methods MLR, PCR and CART. These samples allow the usage of both spectral measurements for the prediction with calibration models. If the two predictions of the repeat measurements lie close together, it can be assumed that the calibration model acts stable, if not, the model has to be rejected. As measure the root mean squared distance (RMSD) was used:

\[
RMSD = \sqrt{\frac{1}{N} \sum (predicted_A - predicted_B)^2}
\]

where A,B: spectral repeat measurements of a soil sample

15. Prediction of new samples

The developed models can not be applied to any soil samples without precaution since the models were set up on a small variation of specific soils. Limits were set to distinguish samples that show properties beyond these limits and for which the developed models are therefore not applicable. As limits the minimal and maximal values for the regressors were taken of the calibration data set. Table 15.1 shows the example of the total C model, the remaining tables are given in the appendix (tables A.4 to A.8). Before a model is applied to new soil samples it is necessary to check if the spectral values of the soil samples are within the limits. This will be demonstrated in chapter 21 on the basis of the remaining samples of the test areas Faizabad and Yavan - which were not used for modelling - and of the third test area, Varzob.
<table>
<thead>
<tr>
<th>Regressor</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.105</td>
<td>0.404</td>
</tr>
<tr>
<td>2180</td>
<td>0.956</td>
<td>0.973</td>
</tr>
<tr>
<td>1490</td>
<td>0.964</td>
<td>0.990</td>
</tr>
<tr>
<td>2320</td>
<td>0.966</td>
<td>0.987</td>
</tr>
<tr>
<td>1860</td>
<td>0.965</td>
<td>0.987</td>
</tr>
<tr>
<td>1300</td>
<td>0.997</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Table 15.1 Limits for the application of total C model
Part IV

Results & Discussion
16. Results of chemical analysis

16.1 Distributions of chemical soil properties

The distributions of the chemical soil properties for the whole analysed sample set were examined using histograms and basic statistical measures as mean, minimum, maximum and various quantiles. The histograms for all chemical soil properties and the table with their statistical measures can be looked up in the appendix (figures A.7 to A.10, table A.1). In order to simplify discussion the histograms were divided into four characteristical groups for which one example each is given in figure 16.1:

- symmetrical distribution:
  The values are distributed approximately symmetrical around the mean. The histogram of total C is given as example. Furthermore, pH, silt and extractable calcium belong to this group.
- right skewed distribution:
  Example organic C. The values show a strong left skewness. The number of samples increases steeply from low values and slowly decreases towards higher values. In addition total N, CaCO$_3$, sand, exchangeable potassium and magnesium belong to this group.
- distribution with extreme outlier:
  Example exchangeable magnesium. Most of the samples show low values, a few high valued outliers appear. Extractable phosphorus and cation exchange capacity also belong to this group.
- bimodal distribution:
  Clay follows a distribution with two peaks, one at 5 to 10 % and one at 30 to 35 %.

As modelling with regression techniques does not require a specific distribution of the dependant variable, no transformations are necessary. Shepherd & Walsh [50] and Vagen, Shepherd & Walsh [57] suggest box-cox transformations [7] to obtain approximately normally distributed dependant variables as they improved results. In this study untransformed chemical properties were used as transformations did not improve modelling accuracy.

The histograms are important for evaluating the prediction ability as the least squares principle was used to estimate the calibration models. It is only logical that the calibration models are fitted best to the regions with the highest number of samples. For example this means that the prediction ability decreases with increasing values for right skewed distributions.
Figure 16.1  Histograms for chemical soil properties of the whole analysed sample set. From top to bottom: total C as example for a symmetrical distribution, organic C as example for a right skewed distribution, exchangeable magnesium as example for a distribution with extreme outlier, clay as example for a bimodal distribution.
16.2 Correlations between chemical soil properties

Table 16.1 shows the correlations between the chemical soil properties. Between most of the chemical soil properties only low correlation can be observed. But there are some correlations that are strikingly high:

- Correlation between total N and SOC: $R^2 = 0.76$
  Total N is for the largest part bounded in organic matter, therefore the concentrations of total N and soil organic C are showing this high correlation [21/59].

- 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 (SOC) 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$_3$.

- Correlations between the three fractions clay, silt, sand: $R^2 > 0.86$
  These correlations can be easily explained through the fact that the three fractions complement one another to 100 percent.

- Correlation between CEC and Ca: $R^2 = 0.98$
  This high correlation originates in the calculation of the CEC (cation exchange capacity) where extractable Ca, Mg, and K were summed. As Ca is accountable for the highest part of CEC, the correlation is very low between CEC and respectively Mg and K.
Table 16.1  Correlations ($R^2$) between chemical soil properties

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total C</td>
<td>0.11</td>
<td>0.02</td>
<td>0.07</td>
<td>0.53</td>
<td>0</td>
<td>-0.02</td>
<td>0</td>
<td>-0.01</td>
<td>0</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0</td>
</tr>
<tr>
<td>SOC</td>
<td>0.76</td>
<td>-0.18</td>
<td>0.04</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0.29</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total N</td>
<td>-0.27</td>
<td>-0.12</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.25</td>
<td>-0.02</td>
<td>0.01</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>pH</td>
<td>0.28</td>
<td>0</td>
<td>-0.06</td>
<td>0</td>
<td>-0.08</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>CaCO$_3$</td>
<td>0</td>
<td>-0.03</td>
<td>0</td>
<td>-0.07</td>
<td>-0.01</td>
<td>0</td>
<td>-0.01</td>
<td>-0.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Extr. P</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Exch. Ca</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.98</td>
</tr>
<tr>
<td>Exch. Mg</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Exch. K</td>
<td>0</td>
<td>-0.01</td>
<td>0</td>
<td>-0.02</td>
<td>0</td>
<td>0.92</td>
<td>0.92</td>
<td>0</td>
<td>0</td>
<td>0.86</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clay</td>
<td>0.92</td>
<td>0.92</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Silt</td>
<td>0.86</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sand</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
17. Results of soil color quantification

In table 17.1, statistical measures are given for the CIE color coordinate X for each test area. The statistical measures for the remaining color variables can be looked up in the appendix (table A.3). The table shows that the measures are comparable for the three test areas and that no large differences are visible. This indicates that the three test areas are relatively homogeneous regarding soil color.

Figure 17.2 shows the soil samples with maximal and minimal value for the CIE color coordinate Z which corresponds to blue color.

17.1 Correlation between CIE color variables and chemical soil properties

In table 17.2 the correlation between the CIE color variables and the chemical soil properties are given. CaCO$_3$ particularly shows a high correlation ($R^2 = 0.55$) with CIE X. The more CaCO$_3$ a soil contains the brighter it appears. This result corresponds with the observation of Jarmer et al. [31] with $R^2 = 0.90$ for CIE x but is considerably lower. A relatively high correlation ($R^2 = 0.40$) is also noticeable between total C and CIE X, a fact which can be explained by the high correlation ($R^2 = 0.53$) between total C and CaCO$_3$ (chapter 16). Apart from this no significant correlations could be found.
<table>
<thead>
<tr>
<th>Soil property</th>
<th>CIE X</th>
<th>CIE Y</th>
<th>CIE Z</th>
<th>CIE x</th>
<th>CIE y</th>
<th>CIE z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total C</td>
<td>0.38</td>
<td>0.40</td>
<td>0.40</td>
<td>-0.15</td>
<td>0.16</td>
<td>0.11</td>
</tr>
<tr>
<td>SOC</td>
<td>-0.07</td>
<td>-0.05</td>
<td>-0.03</td>
<td>-0.05</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>Total N</td>
<td>-0.14</td>
<td>-0.12</td>
<td>-0.08</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>pH</td>
<td>0.22</td>
<td>0.21</td>
<td>0.21</td>
<td>-0.02</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>CaCO_3</td>
<td>0.55</td>
<td>0.53</td>
<td>0.49</td>
<td>-0.07</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Extr. P</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0</td>
<td>-0.03</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Exch. Ca</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.02</td>
<td>0.02</td>
<td>-0.01</td>
<td>-0.03</td>
</tr>
<tr>
<td>Exch. Mg</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.01</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>Exch. K</td>
<td>0.04</td>
<td>-0.04</td>
<td>-0.02</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Clay</td>
<td>0</td>
<td>-0.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Silt</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sand</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.03</td>
<td>-0.01</td>
</tr>
<tr>
<td>CEC</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.02</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

**Table 17.1**  Statistical measures for the CIE color coordinate X for the three test areas Faizabad, Yavan and Varzob. SD: Standard deviation.

**Figure 17.2**  Photos of soil samples with minimal (left) and maximal (right) CIE Z value (blue). Photos: B. Wolfgramm

**Table 17.2**  Correlations ($R^2$) between CIE color variables and chemical soil properties.
18. Result of soil type separation

In table 18.1, the number of samples falling into the different soil classes are given per test area. Faizabad test area shows the highest percentage of paleo soils, a fact that could be confirmed by P. M. Sosin, a paleo soil expert from the SSRI in Dushanbe (oral communication). The samples indicate that the Yavan area has a high number of samples of river sediments. This is due to the location of the test area: A good part lies in the valley floor where river sediments prevail. The Varzob test area is the most homogenous as it has the smallest percentage of non-brown-carbonate soils. The number of samples attributed to the class of brown carbonate soils is 1116 out of 1405. The models established represent only these samples.

18.1 Influence on the distribution of the chemical soil properties

The distributions of most chemical properties were not affected by excluding the three soil classes leptosol, paleo soil and samples of river sediments. The range and approximate distribution function were maintained. Only calcium and therewith the cation exchange capacity and sand fraction showed differences to the original distributions. The histograms of the whole data set (left side) and the adjusted data set (right side) are given in figure 18.1. The few samples with values for calcium higher than 50 were classified as leptosols and therefore removed. This results in a more or less symmetrical histogram. As the calcium values were used to calculate the cation exchange capacity, the same effect can be observed there. Samples with sand fraction higher than 70 % were classified mainly as river sediments, a fact that can be simply explained by the higher proportion of sand in river sediments. The histograms of the remaining chemical soil properties and a table with the statistical measures for the reduced data set of the brown carbonate soils are given in the appendix (figures A.7 to A.10, table A.2).

18.2 Impact on model accuracy

In order to show the impact of the soil type separation on the model accuracy validation measures for the validation data set for MLR models are compared in table 18.2. Models for total C, total N, pH and CaCO$_3$ were compared for the complete data set and for the reduced data set of the brown carbonate soils. Soil type separation definitely improves the model accuracy. The RMSE for the pH model for example even halves from 0.304 to 0.157.
<table>
<thead>
<tr>
<th>Soil class</th>
<th>brown carbonate</th>
<th>leptosol</th>
<th>paleo soil</th>
<th>river sediment</th>
<th>area total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faizabad</td>
<td>457</td>
<td>52</td>
<td>99</td>
<td>16</td>
<td>624</td>
</tr>
<tr>
<td>Yavan</td>
<td>283</td>
<td>27</td>
<td>8</td>
<td>51</td>
<td>369</td>
</tr>
<tr>
<td>Varzob</td>
<td>376</td>
<td>10</td>
<td>23</td>
<td>3</td>
<td>412</td>
</tr>
<tr>
<td>Soil class total</td>
<td>1116</td>
<td>89</td>
<td>130</td>
<td>70</td>
<td>1405</td>
</tr>
</tbody>
</table>

Table 18.1 Result of the CART classification tree applied to whole data set. Number of samples given falling into different classes. Source: B Wolfgramm

Figure 18.1 Histograms of calcium, CEC and sand for the whole data set (left side) and for the reduced data set of the brown carbonate soils (right side).
19. Comparison of applied statistical methods

19.1 Comparison of model accuracy

In table 19.1, the most suitable models for the three methods MLR, PCR and CART are compared regarding the validation data set. The fact that from the original analysed 13 chemical soil properties only 6 appear in this table is striking. For the other chemical properties no adequate models could be built. This is specified in chapter 20.6. The following conclusions can be drawn from table 19.1:

- **organic C, CaCO$_3$**
  
  For PCR no adequate models could be developed. In both cases the MLR model shows explicit advantages over the CART model. All accuracy measures are better for the MLR model. Especially the relative error measures RMSPE and MAPE show the large difference between the two models.

- **total N**
  
  For total N the three models show comparable results, the MLR model slightly exceeding the others.

- **total C**
  
  The PCR model performs best for total C followed by the MLR model.

- **pH**
  
  CART and MLR show comparable results for pH. PCR is a little less accurate.

Recapitulating, MLR thus shows considerable advantages over PCR and CART regarding the model accuracy.

19.2 Comparison of model complexity

If two methods give comparable results regarding the accuracy, the less complex one will be chosen, as it can be retraced and explained better and is less susceptible to errors.

MLR is the least complex of the three methods: The regressors are continuum removed wavelengths that can even be compared to literature. Table 19.1 shows that the MLR models reach their maximum number of components already at 6. PCR is more complex because the individual regressors are a product of all continuum removed wavelengths and all samples. Therefore comparisons with literature are difficult and the influence of the individual wavelengths cannot be simply retraced anymore. CART is the most complex model. As several trees (25-100) were combined to gain the results in table 19.1, no information on the individual separation criterias in the tree nodes are accessible.
Table 19.1 Comparison of MLR, PCR and CART models. N: abbreviation for number of regressors (MLR,PCR) or number of trees (CART).

<table>
<thead>
<tr>
<th>Soil property</th>
<th>Method</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>RMSPE</th>
<th>MAPE</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>total C</td>
<td>MLR</td>
<td>0.67</td>
<td>0.609</td>
<td>0.486</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR</td>
<td>0.73</td>
<td>0.039</td>
<td>0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td></td>
<td></td>
<td>18 %</td>
<td>16 %</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18 %</td>
<td>13 %</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11 %</td>
<td>10 %</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13 %</td>
<td>18 %</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.71</td>
<td>4.906</td>
<td>0.72</td>
<td>4.630</td>
<td></td>
</tr>
<tr>
<td>organic C</td>
<td>MLR</td>
<td>0.81</td>
<td>0.330</td>
<td>37 %</td>
<td>24 %</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>PCR</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td>0.74</td>
<td>0.390</td>
<td>54 %</td>
<td>35 %</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>37 %</td>
<td>24 %</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18 %</td>
<td>37 %</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18 %</td>
<td>54 %</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18 %</td>
<td>54 %</td>
<td>100</td>
</tr>
<tr>
<td>total N</td>
<td>MLR</td>
<td>0.73</td>
<td>0.039</td>
<td>0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR</td>
<td>0.78</td>
<td>0.033</td>
<td>0.035</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td>0.80</td>
<td>0.035</td>
<td>0.035</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13 %</td>
<td>13 %</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22 %</td>
<td>22 %</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22 %</td>
<td>22 %</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22 %</td>
<td>22 %</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.71</td>
<td>4.906</td>
<td>0.72</td>
<td>4.630</td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>MLR</td>
<td>0.61</td>
<td>0.157</td>
<td>2 %</td>
<td>2 %</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>PCR</td>
<td>0.49</td>
<td>0.166</td>
<td>2 %</td>
<td>1 %</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td>0.55</td>
<td>0.156</td>
<td>2 %</td>
<td>2 %</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 %</td>
<td>2 %</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 %</td>
<td>2 %</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 %</td>
<td>2 %</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.61</td>
<td>4.630</td>
<td>0.71</td>
<td>4.630</td>
<td></td>
</tr>
<tr>
<td>CaCO$_3$</td>
<td>MLR</td>
<td>0.72</td>
<td>4.630</td>
<td>30 %</td>
<td>21 %</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>PCR</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CART</td>
<td>0.60</td>
<td>5.488</td>
<td>71 %</td>
<td>34 %</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 19.2 Impact of soil type separation on model accuracy. Comparison of $R^2$ and RMSE of the validation data of MLR models for the complete dataset and the reduced data set of the brown carbonate soils at the examples of total C, total N, pH and CaCO$_3$. 

Table 18.2 Impact of soil type separation on model accuracy. Comparison of $R^2$ and RMSE of the validation data of MLR models for the complete dataset and the reduced data set of the brown carbonate soils at the examples of total C, total N, pH and CaCO$_3$. 

Table 18.1 Comparison of MLR, PCR and CART models. N: abbreviation for number of regressors (MLR,PCR) or number of trees (CART).
19.3 Comparison of repeatability

As already mentioned in chapter 7 the two plots FA64 and YA24, in total 29 samples after soil separation, were measured twice with the spectrometer to determine the methods’ repeatability. The second measurements of the two plots were handled as a totally new data set, the whole data preparation process was performed separately from the other data. The best models for the three methods MLR, PCR and CART were now applied to this repeat data set and the results were compared to the modelled results of the first spectral measurements of these soil samples. The repeatability of the methods was determined using the root mean squared difference RMSD between the two modelling results of the soil samples. The smaller the RMSD, the nearer the two predictions of the soil samples lie and the better the repeatability of the method is.

Table 19.2 shows the RMSD values for the different methods and soil properties. It is striking that the highest RMSD is achieved through PCR respectively. MLR and CART show comparable RMSD values except for the SOC model where CART is a lot better and the CaCO₃ model where MLR is better. The high RMSD values for PCR are connected to the calculation of the principal components. Every sample has an influence on the calculation of the principal components of the other samples. This means that if a part of the data set is excluded for the calculation, the samples achieve different values for the principal components than if the whole data set is used. This problem does not occur using continuum removed data as the continuum removal is applied to each sample separately. PCR is thus not suitable for this application in which additional samples should be predicted in future.

Figure 19.1 underlines this observation: The two predictions A and B for total C for the repeat measurements are plotted against each other and should approximate the 1:1 line if the repeatability is warranted. This is the case for the MLR and CART models, but not for PCR. The predictions for the PCR model follow a steeper line and some of the repeat samples are even predicted negative.

19.4 Selection of most suitable statistical method

The evaluation of the statistical methods MLR, PCR and CART in the previous three chapters resulted in the following conclusions:

- PCR is not suitable for this application, as new data can not be predicted by the developed models.
- CART achieves good results, sometimes even better than MLR, but the disadvantage of the model complexity overvails.
- MLR provides the most constant results. Furthermore, the model is the least complex and can be retraced best.
<table>
<thead>
<tr>
<th>Soil property</th>
<th>MLR</th>
<th>PCR</th>
<th>CART</th>
</tr>
</thead>
<tbody>
<tr>
<td>total C</td>
<td>0.28</td>
<td>2.72</td>
<td>0.24</td>
</tr>
<tr>
<td>SOC</td>
<td>0.26</td>
<td>-</td>
<td>0.15</td>
</tr>
<tr>
<td>total N</td>
<td>0.01</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>pH</td>
<td>0.05</td>
<td>0.57</td>
<td>0.06</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>0.70</td>
<td>5.29</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figure 19.1: The two predictions of total C for the spectral repeat measurements plotted against each other. From top to bottom: PCR model, MLR model, CART model

Table 19.2: RMSD for MLR, PCR and CART models resulting from repeatability determination

The two predictions of total C for the spectral repeat measurements plotted against each other. From top to bottom: PCR model, MLR model, CART model
Part IV, Results

These reasons led to the decision not to choose the best method for every single soil property separately but to use MLR for all soil properties. This allows to compare the different models among each other, too.

In the following chapter the resulting MLR models are described and discussed more precisely.

20. Resulting calibration models

An overview of the accuracy of the resulting MLR calibration models is shown in tables 20.1 and 20.2. Total C, SOC and total N could be modelled with high accuracy whereas the pH and CaCO3 models are less accurate.

In the following subchapters the single models will be discussed. Reasons for the failure with distinct chemical soil properties will be given in chapter 20.6. The model assumption plots and test results for the models are given in the appendix (figures A.11 to A.15). A normal distribution of the residuals was not totally obtained in most cases. Nevertheless, the F- and t-tests were used, but rather as diagnostics than as hard tests.
<table>
<thead>
<tr>
<th>Soil property</th>
<th>Calibration data set</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R²</td>
<td>RMSE</td>
<td>RMSPE</td>
</tr>
<tr>
<td>Total C</td>
<td>0.85</td>
<td>0.348</td>
<td>13 %</td>
</tr>
<tr>
<td>SOC</td>
<td>0.74</td>
<td>0.354</td>
<td>72 %</td>
</tr>
<tr>
<td>Total N</td>
<td>0.72</td>
<td>0.030</td>
<td>30 %</td>
</tr>
<tr>
<td>pH</td>
<td>0.61</td>
<td>0.126</td>
<td>2 %</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>0.62</td>
<td>5.870</td>
<td>163 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Soil property</th>
<th>Validation data set</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R²</td>
<td>RMSE</td>
<td>RMSPE</td>
</tr>
<tr>
<td>Total C</td>
<td>0.76</td>
<td>0.436</td>
<td>16 %</td>
</tr>
<tr>
<td>SOC</td>
<td>0.81</td>
<td>0.330</td>
<td>37 %</td>
</tr>
<tr>
<td>Total N</td>
<td>0.83</td>
<td>0.030</td>
<td>47 %</td>
</tr>
<tr>
<td>pH</td>
<td>0.61</td>
<td>0.157</td>
<td>2 %</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>0.72</td>
<td>4.630</td>
<td>30 %</td>
</tr>
</tbody>
</table>

Table 20.1: Accuracy measures for calibration data set of the MLR models
Table 20.2: Accuracy measures for validation data set of the MLR models
20.1 Calibration model for total C

The final calibration equation for total C consists of six regressors: a color variable and five continuum removed wavelengths. The regressors and its coefficients are given in table 20.3.

The appearance of the CIE color variable X could be expected as total C is strongly correlated ($R^2 = 0.73$) with the amount of CaCO$_3$. The relative amount of CaCO$_3$ influences soil brightness substantially and therewith the color variables.

The regressor 2320 nm is a consequence of the correlation between total C and CaCO$_3$, too. Jarmer et al. [31] describe the spectral range 2300 - 2350 nm as a strong diagnostic absorption band of carbonates. This absorption band will be specified in chapter 20.5.

According to Jarmer et al. [31] and Hunt & Salisbury [28], the regressors 1860 and 2180 nm are weaker absorption bands of carbonates. They also appear in the calibration equations of SOC, total N and pH.

For the two regressors 1490 and 1300 nm no confirmation could be achieved through literature.

Figure 20.1 shows two charts for the measured versus the predicted amounts of total C, for the calibration data on the left, for the validation data on the right. Their points approximate the 1:1 line well and the differences between the measured and the predicted values seem to be constant over the whole range.
Table 20.3  Regressors and coefficients for the MLR model for total C

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-416.183</td>
</tr>
<tr>
<td>X</td>
<td>8.621</td>
</tr>
<tr>
<td>2180</td>
<td>185.116</td>
</tr>
<tr>
<td>1490</td>
<td>-109.733</td>
</tr>
<tr>
<td>2320</td>
<td>-131.953</td>
</tr>
<tr>
<td>1860</td>
<td>94.079</td>
</tr>
<tr>
<td>1300</td>
<td>383.166</td>
</tr>
</tbody>
</table>

Figure 20.1  Measured versus predicted charts for total C. Left: calibration data set. Right: validation data set.
20.2 Calibration model for soil organic C

The calibration model for soil organic carbon consists of the three wavelengths 1870, 1530 and 2180 nm. The regression coefficients are given in table 20.4.

As mentioned in chapter 4, there is a relation between organic C and the soil brightness. It is therefore striking that no color variable appears in the model. This can be explained by the fact that there is only little organic C in the researched Tajik soils and the influence on soil color is thus small.

The two regressors 1870 and 2180 nm are weak absorption bands of carbonates [4/28/31]. The third regressor wavelength, 1530 nm, is an absorption of N-H stretch first overtone [1], and thus originates in the correlation between total C and total N.

Figure 20.2 shows two charts for the measured versus the predicted amount of SOC, for the calibration data on the left, for the validation data on the right. Their points approximate the 1:1 line quite well. It seems as if the model predicts rather too low values for samples with SOC higher than 2.0 %.
### Table 20.4
Regressors and coefficients for the MLR model for SOC

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-142.662</td>
</tr>
<tr>
<td>1870</td>
<td>168.152</td>
</tr>
<tr>
<td>1530</td>
<td>-105.07</td>
</tr>
<tr>
<td>2180</td>
<td>87.955</td>
</tr>
</tbody>
</table>

### Figure 20.2
20.3 Calibration model for total N

The calibration model for total N is composed of three regressors: the wavelengths 1870, 1530 and 900 nm. The regression coefficients are given in table 20.5. The fact that the first two regressor wavelengths 1870 and 1530 nm are the same that appear in the calibration equation of SOC attracts attention. The 1870 nm absorption is a weak band of carbonates. It seems to contain more information on total N than the absorption at wavelength 1530 nm due to N-H stretch first overtone [1] as it was chosen as the first regressor. This is caused by pseudo correlation. For the third regressor, 900 nm, no confirmation could be achieved through literature.

Figure 20.3 shows two charts for the measured versus the predicted amount of total N: for the calibration data on the left, for the validation data on the right. Their points approximate the 1:1 line well. In the validation data set one outlier is visible which is predicted too high. For all samples with measured total N higher than 1.6 % the model predicts too low values. This originates in the distribution of the measured total N amounts: The majority of the chemically analysed samples have total N amounts between 0.5 and 1.5 % and the model is fitted best to this region.
Table 20.5 Regressors and coefficients for the MLR model for total N

<table>
<thead>
<tr>
<th>Regressor</th>
<th>Regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-6.294</td>
</tr>
<tr>
<td>1870</td>
<td>12.700</td>
</tr>
<tr>
<td>1530</td>
<td>-8.832</td>
</tr>
<tr>
<td>900</td>
<td>2.909</td>
</tr>
</tbody>
</table>

Figure 20.3 Measured versus predicted charts for total N. Left: calibration data set. Right: validation data set.
20.4 Calibration model for pH

The final calibration equation for pH consists of four regressors: 1860 nm, CIE chromaticity coordinate \(x'\), 1700 nm and 720 nm. The coefficients are given in table 20.6.

The regressor 1860 nm is a weak absorption band of carbonates [4/28/31]. This could be connected to the buffering function of CaCO\(_3\) [21] but no information was found on the exact relation. No confirmation on the remaining three regressors could be achieved as no other modelling attempts for pH could be found in literature. The reason is probably that pH is determinable quite simply and cheaply with chemical methods.

Figure 20.4 shows two charts for the measured versus the predicted pH: for the calibration data on the left, for the validation data on the right. Their points don’t approximate the 1:1 line accurately, a fact which can be noted observing the R\(^2\) values at 0.61.
Table 20.6  Regressors and coefficients for the MLR model for pH

Figure 20.4  Measured versus predicted charts for pH. Left: calibration data set. Right: validation data set.
20.5 Calibration model for CaCO$_3$

The calibration model for CaCO$_3$ consists of three regressors: The CIE color coordinate X and the two wavelengths 2340 and 2380 nm. The coefficients are given in table 20.7. Due to its bright color, CaCO$_3$ is positively correlated with soil brightness which explains the appearance of the color coordinate. The two wavelength regressors lie very close to each other. The ASD introduction to NIR describes the two features as following [1]:

- 2335 nm: C-H stretch/C-H deformation
- 2380 nm: C-H stretch/C-C stretch combination

Furthermore, the spectral region at 2300 - 2350 nm is known as a strong diagnostic absorption band of carbonates [4/28/31]. Jarmer et al. [31/35] found a high correlation between the absorption depth at 2350 nm and the amount of inorganic C. Figure 20.5 (left) shows the absorption of CaCO$_3$ at 2340 nm at the example of four spectra of soils with differing amount of CaCO$_3$. The absorption gets deeper with increasing amount of CaCO$_3$. In figure 20.5 (right) the amount of CaCO$_3$ for all chemically analysed soil samples is plotted against their continuum removed wavelength 2340 nm. The four soil samples of figure 20.5 (left) are marked with red color. A high correlation is visible and confirms the observations above.

Figure 20.6 shows two charts for the measured versus the predicted CaCO$_3$; for the calibration data on the left, for the validation data on the right. They points don’t approximate the 1:1 line well which results in the low R$^2$ values with 0.62, respectively 0.72.
<table>
<thead>
<tr>
<th>Regressor</th>
<th>Regression coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-795.649</td>
</tr>
<tr>
<td>X</td>
<td>112.194</td>
</tr>
<tr>
<td>2340</td>
<td>-544.065</td>
</tr>
<tr>
<td>2380</td>
<td>1323.09</td>
</tr>
</tbody>
</table>

Table 20.7 Regressors and coefficients for the MLR model for CaCO₃

Figure 20.5 Absorption feature of CaCO₃ at 2340 nm. Left showed at the example of four soil samples with 1.5 (I), 11 (II), 28.8 (III) and 35.9 % (IV) of CaCO₃. Right amount of CaCO₃ plotted against the continuum removed wavelength 2340 nm for all chemically analysed samples. The red points correspond with the soil samples in the left figure.

Figure 20.6 Measured versus predicted charts for CaCO₃. Left: calibration data set. Right: validation data set.
20.6 Modelling problems with remaining soil properties

As already mentioned, no calibration models could be developed for the chemical soil properties 'extractable phosphorus', 'exchangeable calcium', 'magnesium' and 'potassium', the fractions 'clay', 'silt' and 'sand', and the 'cation exchange capacity'. Poor calibrations for extractable soil properties have also been reported by Udelhoven et al. [54]. They were able to model the total amounts of Ca, Mg, K, F, Mn, but not the amounts of K, P, Mg available for plants. The findings of Shepherd and Walsh [50] are contradictory: they modelled exchangeable Ca and Mg with good accuracy. Successful modelling of the fractions silt, clay and sand [11/14] and of the cation exchange capacity [8/11/57] was obtained in other studies.

These different conclusions suggest that modelling a specific soil properties depends on the following points:

- Enough information on the soil property must be available in the VIS/NIR/SWIR spectrum: Figure 20.8 shows the correlation coefficient between the continuum removed wavelengths and the percentage of clay. The correlation is never higher than absolute $R = 0.30$. This is a very low correlation and it is not astonishing that modelling is not possible with this marginal information.

- The adequate statistical method must be chosen to extract this information: The examples given above show that specific soil properties could be modelled with one method and with the other not, as Shepherd and Walsh [50] and Udelhoven et al. [54] used different statistical methods. Figure 20.7 shows the plots for the measured versus the predicted percentage of clay, for the calibration data on the left and for the validation data on the right. The points do not approximate the 1:1 line but are strongly scattered. It seems possible that non-linear regression methods could lead to improvements with this chemical property as the points follow a non-linear relationship. However, this issue was not further evaluated in this study.

- The exact composition of soils: This could also be a reason why specific soil properties could be modelled in some studies but not in others. The composition of soils depends strongly on the geographic conditions. This could make modelling of specific soil properties possible due to pseudo-correlation in some type of soils and in others not.
Figure 20.7  Measured versus predicted charts for percentage of clay. Left: calibration data set. Right: validation data set.

Figure 20.8  Correlation coefficient R between the continuum removed wavelengths and the percentage of clay.
21. Applying models to new data

In table 21.1, the number of samples are given that remain after applying the model limits (chapter 15) to the two test areas Faizabad and Yavan. The number of samples that had to be excluded is negligibly low. This demonstrates that the variability of the soil samples within the Faizabad and Yavan area has been perfectly covered when selecting representative samples for chemical analysis through principal components.

The result of the prediction for the remaining samples from Faizabad and Yavan area is given in figure 21.1. It shows the histogram for chemically analysed SOC on the left and the histogram of the predicted SOC on the right. The characteristics of the two histograms are approximately the same, the only difference is that one negative prediction has occurred. This seems to be an outlier which cannot be predicted properly by the model.

The ability of the model to predict soil samples from different areas is tested by predicting the chemical properties of soil samples from the Varzob area which have not been involved in the model building process. Table 21.2 shows the number of samples that remain after applying the model limits. The percentages of samples within model limits strongly decrease compared to the Faizabad/Yavan areas from nearly 100 % to 90 % for total N or even 60 % for total C. Further soil samples would have to be chemically analysed and included in the model building process to make predictions of the excluded samples possible.

Figure 21.2 shows the improvement achieved with applied model limits. The chemically analysed samples for the Faizabad and Yavan area are given in the first histogram as comparison. The second histogram shows the prediction for all brown carbonate samples from the Varzob area and the third the predictions for the brown carbonate Varzob samples within the model limits only. The result has improved a lot: the extreme negative predictions are diminished after applying the model limits. Nevertheless, negative predictions still do occur, a problem that cannot be overcome without analysing some of these samples and including them in the model building process. This step was not taken in this particular research. The histograms for the predictions of the remaining chemical soil properties are given in the appendix (figures A.16 to A.20).
Table 21.1  Samples within model limits for the Yavan and Faizabad test areas, given in absolute numbers and percentage.

Figure 21.1  Histograms for SOC for Faizabad and Yavan area. Left: chemically analysed samples. Right: predictions of samples within model limits.

<table>
<thead>
<tr>
<th>Model</th>
<th>Faizabad samples within model limits</th>
<th>Yavan samples within model limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>absolute number</td>
<td>percentage</td>
</tr>
<tr>
<td>Total C</td>
<td>445</td>
<td>98 %</td>
</tr>
<tr>
<td>SOC</td>
<td>452</td>
<td>&gt; 99 %</td>
</tr>
<tr>
<td>Total N</td>
<td>454</td>
<td>&gt; 99 %</td>
</tr>
<tr>
<td>pH</td>
<td>451</td>
<td>99 %</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>455</td>
<td>&gt; 99 %</td>
</tr>
</tbody>
</table>
Part IV, Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Varzob samples within model limits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>absolute number</td>
</tr>
<tr>
<td>Total C</td>
<td>223</td>
</tr>
<tr>
<td>SOC</td>
<td>317</td>
</tr>
<tr>
<td>Total N</td>
<td>343</td>
</tr>
<tr>
<td>pH</td>
<td>328</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>335</td>
</tr>
</tbody>
</table>

Table 21.2  Samples within model limits for the Varzob test area, given in absolute numbers and percentage.

Figure 21.2  Histograms for SOC for Varzob area. Top: chemically analysed samples from the Faizabad and Yavan area for comparison. Middle: predictions of all Varzob samples. Bottom: predictions of Varzob samples within model limits.
22. Area-wide modelling of soil properties using satellite imagery

The original idea was to attempt an area-wide modelling of the chemical soil properties after modelling the single soil samples. Several studies currently deal with mapping soil condition in landscapes using satellite imagery. Vagen, Shepherd and Walsh [57] developed a soil fertility index, integrating ten commonly used agronomic indicators of soil fertility and successfully calibrated it to local conditions, allowing the spatial representation of soil fertility based on remote sensing satellite imagery. Udelhoven, Jarmer and Hill [55] modelled inorganic C using simulated hyperspectral satellite data of AVIRIS, HyMap and DAIS 7915. Hill and Schütt [26] successfully mapped soil organic matter in a LANDSAT TM image using spectral unmixing.

The following satellite images were available for this study:

- Landsat 7 ETM+ image from 22.08.2000
- Three ASTER images covering the three test areas from 22.08.2000

The satellite images are from the summer, the dryest season in Tajikistan, which makes influence of vegetation on the soil signal as low as possible. The collected soil samples and their modelled chemical soil properties should have been used as ground truth and the bands of the satellite images as independent variables for the modelling process. Additionally, the CIE color coordinates should have been calculated from the satellite imagery using similar methods as described by Mathieu, Pouget et al. [39].

However, the attempt was not followed up in this study as the chances of success were low, in fact due to the following reasons:

- The inaccurate spectral resolution of the satellite images: The spectrometer data for modelling of the soil samples had a spectral resolution of 10 nm and a total number of bands of over 200 (see chapter 9.3). The satellite images have broader band widths of 40 to 270 nm (Landsat 7 ETM+) and 40 to 100 nm (ASTER) and an explicitly lower total number of bands of 6 (Landsat 7 ETM+) and 9 (ASTER) in the VIS-NIR-SWIR region. The use of hyperspectral data would be more adequate for this application but was not accessible due to high costs.

- The inaccurate geometric resolution of satellite images (ASTER 15 to 30 m; Landsat 30 m) compared to the high areal variability of soils.

- Accuracy of georeferencing of the satellite images: Five GPS measured control points were used to determine the accuracy of georeferencing. An RMSE of 53 m in x- and 12 m in y-direction resulted which makes the allocation of ground truth to image data difficult.

- Influence of atmosphere: the intention was to remove atmospheric influence using the atmospheric correction program ATCOR3 within the software Geomatica [20]. Nevertheless, uncertainties remain as the atmospheric conditions at the time of image recording have to be estimated and the digital elevation model (a combination of SRTM and digitalized contours of Russian topographic maps) is affected by inaccuracies.

- Differences between measurements in field (soil crust) and the specially prepared soil samples (drying, grinding, sieving) in laboratory (see chapter 6.2).
23. Comparing model accuracies with other studies

A comparison with other studies is difficult as in most papers insufficient information is given on the used data and the achieved results. Furthermore a pure comparison of results is only of low significance: different statistical methods have been used, the investigation areas strongly differ in location, dimension and soil conditions and different numbers of samples have been used for modelling. Despite all these differences, it is interesting to see if the models give comparable results last but not least due to the attempts of building models for global use [8].

The models for SOC and for total N were chosen for comparison as they are the most frequently modelled chemical soil properties. Nevertheless, only two studies were found with sufficient information for a reasonable comparison.

Table 23.1 shows the comparison of the SOC model to a study of Shepherd and Walsh [50] and one of Chang and Laird [12]. Shepherd and Walsh have been using an extensive sample set from eastern and southern Africa of around 1000 soil samples for modelling. The range of SOC in the African soils and in the Tajik soils is quite similar, a fact which makes the results better comparable. The RMSEP and the $R^2$ are nearly the same for the African ($R^2 = 0.80$, RMSEP = 3.1) and the Tajik model ($R^2 = 0.81$, RMSEP = 3.3). Certainly, it isn’t possible to conclude that the two modelling methods ‘multivariate adaptive regression splines’ [38] and ‘multiple linear regression’ have exactly the same ability to predict SOC, as African and Tajik soils are composed differently. It still shows that comparable results can be achieved using different methods. The comparison with Chang and Laird is more difficult because a different validation method was used (cross-validation), the samples have been artificially mixed in the laboratory and the range of SOC is far larger than in the other two works. The RMSEP approximately doubles compared to the other studies, a fact which is due to the higher range of SOC in these samples.

Table 23.2 shows the comparison of the total N model to a study of Hett [25] and the study of Chang and Laird [12] already used above. The three models show comparable results although different methods were used. The RMSEP varies between 0.30 and 0.44, and is again connected to the range of chemical property in the modelled soils: the best result is achieved for the lowest range of total N.
### Table 23.1
Comparing SOC model (see chapter 19.2) with SOC models developed by Shepherd and Walsh [50] and Chang and Laird [12]. RMSEP: Root mean squared error of prediction. N calibration/validation: Number of samples used for calibration/validation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.80</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>RMSEP [g kg$^{-1}$]</td>
<td>3.1</td>
<td>6.20</td>
<td>3.3</td>
</tr>
<tr>
<td>Minimum [g kg$^{-1}$]</td>
<td>2.3</td>
<td>15.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Maximum [g kg$^{-1}$]</td>
<td>55.8</td>
<td>144.9</td>
<td>46.0</td>
</tr>
<tr>
<td>N calibration</td>
<td>~ 600</td>
<td>76</td>
<td>128</td>
</tr>
<tr>
<td>N validation</td>
<td>~ 300</td>
<td>32</td>
<td>55</td>
</tr>
<tr>
<td>Statistical method</td>
<td>Multivariate adaptive regression splines (MARS)</td>
<td>Partial least squares regression</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>Validation method</td>
<td>1/3 hold out random samples for validation</td>
<td>Cross-validation</td>
<td>1/3 hold out random samples for validation</td>
</tr>
<tr>
<td>Origin of soil samples</td>
<td>East/South Africa diverse soils</td>
<td>Sample mixing in laboratory</td>
<td>Two test areas of 10 x 10 km in western Tajikistan</td>
</tr>
</tbody>
</table>

### Table 23.2
Comparing total N model (see chapter 19.3) with total N models developed by Hett [25] and Chang and Laird [12]. RMSEP: Root mean squared error of prediction. N calibration/validation: Number of samples used for calibration/validation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.84</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>RMSEP [g kg$^{-1}$]</td>
<td>0.44</td>
<td>0.36</td>
<td>0.30</td>
</tr>
<tr>
<td>Minimum [g kg$^{-1}$]</td>
<td>0.4</td>
<td>1.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Maximum [g kg$^{-1}$]</td>
<td>7.1</td>
<td>5.5</td>
<td>3.8</td>
</tr>
<tr>
<td>N calibration</td>
<td>~ 400</td>
<td>76</td>
<td>128</td>
</tr>
<tr>
<td>N validation</td>
<td>~ 300</td>
<td>32</td>
<td>55</td>
</tr>
<tr>
<td>Statistical method</td>
<td>regression tree (CART)</td>
<td>Partial least squares regression</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>Validation method</td>
<td>Cross-validation</td>
<td>Cross-validation</td>
<td>1/3 hold out random samples for validation</td>
</tr>
<tr>
<td>Origin of soil samples</td>
<td>Test area of 10 x 10 km in Kenya</td>
<td>Sample mixing in laboratory</td>
<td>Two test areas of 10 x 10 km in western Tajikistan</td>
</tr>
</tbody>
</table>
24. Conclusions

24.1 Research question and specific objectives

In this chapter, the research question and the specific objectives of this study will be answered and discussed, and results will be summarized. Possible improvements in methodology will be noted to show constraints that could be avoided in a future application.

The main research question was the following:

What chemical soil properties can be predicted from spectral information for the loess areas in western Tajikistan?

The research lead to the following results:

- Very good results were 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 SOC ($R^2 = 0.81$, RMSEP = 3.30 g kg$^{-1}$)

- Good results were achieved 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, cation exchange capacity and the fractions silt, clay and sand.

The specific objectives of this study were the following:

Building a soil spectral library of a sample set including around 1500 sub- and topsoil samples from three similar test areas.

The soil spectral library was successfully built and is available on the digital appendix as Excel file. Two critical notes have to be made which could eventually be improved in a future application: The soil samples could be grinded to a smaller size than 2 mm for spectral analysis in order to reduce the scattering differences. Furthermore, it has to be a main concern whether the data reduction from a 1 nm to a 10 nm resolution is appropriate. A higher resolution prolongs the computing time but could lead to slight improvements regarding model accuracy.

Selection of around 250 representative samples for chemical analysis and for calibration and validation of the models.

The variability of the soil samples within the Faizabad and Yavan area was perfectly covered by selecting representative samples for a chemical analysis through principal components, be-
cause the application of the model limits on the remaining samples of the two test areas led to
the exclusion of less than 1 percent of the samples.
The used random selection of calibration and validation samples does not result in independant
data sets. As the spatial structure of calibration and validation samples is important, randomly
selected independant hold-out sites should be used for validation instead [9].

**Testing the ability of different statistical methods for building models to predict chemical soil
properties from spectral information.**

The three different statistical methods 'principal component regression' (PCR), 'regression
tree' (CART [10]) and 'multiple linear regression' (MLR) were used and compared to each
other with the following results:

- The results of the complex methods PCR and CART were less accurate than the results of the
  simpler and better traceable method MLR.

- PCR cannot be used for an application in which new samples from outside of the basic sample
  set have to be modelled.

It can therefore be recommended to use the classical method MLR with continuum removed
spectral data. Especially the possibility to compare absorption features directly to other studies
makes this method very useful. But as the strong points of the different statistical methods can
vary depending on the application it is not advisable to rely on one method only without testing
it against others.

Two critical remarks should be made regarding the methodology:
- The use of cross-validation in the calibration process gives more stable results and should
  therefore be used for future studies.
- Further statistical methods should be included in the testing process: Non-linear regression
  methods and multivariate adaptive regression splines (MARS) as they could make modelling
  of further chemical soil properties possible.

**Evaluating the integration of soil color in the model building process.**

Soil color was quantified using the CIE color system. Its integration resulted in an explicit im-
provement of model accuracy with chemical soil properties that are connected with soil color
and can therefore be strongly recommended for future applications. The use of other auxiliary
predictors such as soil properties that can be determined quickly and inexpensively - for ex-
ample pH - could also improve calibration results [9]. This was not further evaluated in this
particular study.
Applying the models to new data and a new area.

The calibration models were fitted to a selection of samples from two of the three test areas. Limits for the model components were set based on this selection. A new sample should lie within these limits in order to be predicted properly by the calibration models. The prediction of the remaining samples from these two test areas gave plausible results. The application on the third test area which was not involved in the model building process revealed the following difficulties:

- more samples were off limits.
- Some of the within limit samples could nevertheless not be plausibly predicted and obtained negative values. This factor could simply be avoided by involving some of these samples in the model building process.

The fact that the predictions of new samples could only be checked on plausibility is disadvantageous. The real ability of the models to predict chemical properties for new soil samples could only have been verified if some of these samples had been chemically analysed.

Comparing the results to other studies.

A comparison with other studies is difficult as it can not be distinguished if differences in model accuracy come from the different statistical methods or if they are caused by the different soil compositions. It is nevertheless interesting to see if comparable results are achieved in spite of these differences. A further difficulty resulted from the insufficient information on the used data set and the achieved accuracies given in other studies which made comparisons impossible. Due to these limitations only the SOC and the total N models could be compared to two other studies (see chapter 23). The achieved results are very similar despite of the differences in methodology.

Furthermore, it was analyzed whether problems with modelling certain soil properties also appeared in other studies. Problems with extractable and exchangeable chemical properties can also be observed in some other studies but there are also examples where modelling was possible with good accuracy. The same applies for the fractions clay, silt and sand. The reasons for these contradictory results could be the following:

- The used statistical method: A different method can make information on a soil property accessible.
- The soil composition: On the one hand the spectral features of different soil properties could overlap and make modelling difficult, on the other hand modelling can be made possible through pseudo correlation depending on the composition of soils.
The exact reason could be evaluated if the statistical methods that lead to a successful modeling of these soil properties in other studies would be applied to the Tajik data set.

24.2 Outlook

Chances for area-wide modelling by means of satellite images were low in this study and this approach was therefore not followed up. The main reasons were the inaccurate georeferencing and the upscaling problem regarding the spectral and spatial resolution of the LANDSAT and ASTER images. If these problems could be overcome, for example by using hyperspectral data and more accurate digital elevation models, the prospects would be promising. Other studies have shown that it is possible to retrieve distinct soil properties as soil organic matter [26] and inorganic C [55] from satellite images. This would make monitoring possible of soil conditions with low time and effort.

The scale of application is a very important point in modelling soil properties. If models can be developed for local scale only, the effort is very high as for each new area a new model has to be calibrated. A study of Brown et al. [8] attempts to develop models for a global scale of application with promising results. This is surprising, considering the fact that in Tajikistan the different soil groups in the three small test areas of 10 x 10 km already weakened the accuracy of the predictions considerably.

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 potential of the many different statistical methods used in this and other studies should be evaluated further. In my opinion, the classical method of MLR with continuum removed data has the advantage that it simplifies comparisons between different studies and uses physically defined absorption bands whereas some more complex methods - for example multivariate adaptive regression splines [38] - seem to have the ability to model more soil properties. For modelling of key properties in soil fertility as organic and inorganic C the MLR method seems far more appropriate. For other soil properties such as extractable and exchangeable soil parameters more complex methods must be used.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASD</td>
<td>Analytical Spectral Devices [1]</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and regression tree [10]</td>
</tr>
<tr>
<td>CDE</td>
<td>Centre for Development and Environment, University of Berne</td>
</tr>
<tr>
<td>CEC</td>
<td>Cation exchange capacity</td>
</tr>
<tr>
<td>CIE</td>
<td>Comission Internationale de l’Eclairage</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of variation</td>
</tr>
<tr>
<td>ENVI</td>
<td>Environment for visualizing images</td>
</tr>
<tr>
<td>FAO</td>
<td>Food and Agriculture Organization, United Nations</td>
</tr>
<tr>
<td>ICRAF</td>
<td>International center for research in agroforestry</td>
</tr>
<tr>
<td>IDL</td>
<td>Interactive data language [29]</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean absolute percentage error</td>
</tr>
<tr>
<td>MARS</td>
<td>Multivariate adaptive regression splines [38]</td>
</tr>
<tr>
<td>MIR</td>
<td>Mid-infrared electromagnetic radiation (3 - 15 _m)</td>
</tr>
<tr>
<td>MLR</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>NCCR</td>
<td>National Centres of Competence in Research, Switzerland</td>
</tr>
<tr>
<td>NIR</td>
<td>Near-infrared electromagnetic radiation (0.7 - 1.4_m)</td>
</tr>
<tr>
<td>PC</td>
<td>Principal component</td>
</tr>
<tr>
<td>PCR</td>
<td>Principal component regression</td>
</tr>
<tr>
<td>PLSR</td>
<td>Partial least square regression</td>
</tr>
</tbody>
</table>
RMSE  Root mean squared error

RMSPE  Root mean squared percentage error

RSL  Remote Sensing Laboratories, Department of Geography, University of Zurich

SNR  Signal to noise ratio

SOC  Soil organic carbon

SWIR  Shortwave-infrared electromagnetic radiation (1.4 - 3_m)

VIS  Visible electromagnetic radiation (0.35 - 0.7_m)
List of figures

Figure 2.1  Map of Tajikistan and neighbouring countries. The red square is indicating the position of the clipped LANDSAT image in figure 2.2. 17

Figure 2.2  Clipping of LANDSAT 7 image from 22.08.2000 showing the three test areas Yavan (I), Faizabad (II), Varzob (III). 17

Figure 2.3  Monthly precipitation in the two test areas Faizabad (black) and Yavan (grey) averaged over the years 1988 to 2002. Source: B.Wolfgramm 19

Figure 2.4  Erosion in the loess deposits of western Tajikistan. From erosion rills (above) to deep gullies (below). 20

Figure 4.1  Munsell Color System, simplified. Circle: The 10 main hues for maximal chroma and value 5. Five chroma steps for hue 5YR(Yellow-Red). 29

Figure 6.1  Sampling design for Faizabad test area. Source: B.Wolfgramm 35

Figure 6.2  Soil sample preparation. Top to bottom: sampling pit, drying of samples. Photo: B.Wolfgramm 37

Figure 6.3  Soil sample preparation. Top to bottom: grinding, sieving. Photo: B.Wolfgramm 38

Figure 6.4  Soil sample preparation. Top to bottom: storage of the prepared samples in paperbags, storage of the remaining soil material in plastic bags. 39

Figure 7.1  Measurement setup for spectral analysis. Left side: muglight with soil sample in Petri dish. Right side: spectrometer with interface. 41

Figure 7.2  Lab photo of soil sample FA550302 with paperbag. Photo: B.Wolfgramm 43

Figure 8.1  Soil organic matter (%) for the 10 laboratory test samples. Two subsamples measured by each laboratory. Bottom: Subsamples A and B for each soil laboratory plotted against each other. SSRI (square), GIUB (circle), GIPROZEM (triangle). Source: B.Wolfgramm 45

Figure 9.1  Correction of steps occuring at spectrometer changeovers at 1000 (VNIR-SWIR1) and 1800 nm (SWIR1-SWIR2) due to different sample views. 51

Figure 9.2  Signal to noise ratio of the applied measurement setup. Smoothed with a 30 nm window. 51

Figure 9.3  Spectral feature at 2340 nm with 1 nm (left) and resampled 10 nm (right) resolution. 53
Figure 9.4 Two spectral measurements of soil sample FA640203

Figure 9.5 Two spectral measurements of soil sample FA640203 continuum removed

Figure 9.6 Correlation between amount of total C and the different wavelengths. Raw (black curve) and continuum removed spectral data (red curve).

Figure 9.7 First derivative of two spectral measurements of FA640203

Figure 10.1 CIE 1964 color-matching functions

Figure 11.1 Landscape in Faizabad test area. Separation of areas: (I) High mountain meadow steppe soils, (II) Loess area: loess soils and sporadical paleo soils and high mountain meadow steppe soils, (III) samples of river sediment on the valley floor. Photo: B. Wolfgramm

Figure 11.2 Typical spectra of the four separated sample types. Brown carbonate soil (I), paleo soil (II), river sediment (III), leptosol (IV).

Figure 11.3 Lab photos showing color differences between four soil groups. From left to right: brown carbonate soil, leptosol, paleo soil, river sediment. Photos: B. Wolfgramm

Figure 12.1 RMSE for the validation data set for different numbers of components. Red curve: PCR. Black curve: PLSR.

Figure 12.2 Autocorrelation of wavelength values for a soil sample in 10 nm resolution for lags 0 to 30. Starting at 420 nm.

Figure 12.3 Example for a CART single tree for total C (%) with four terminal nodes. A sample with values for wavelength 440 nm > 0.0075 (first derivative) and for wavelength 1330 < 0.0007 (first derivative) ends up in terminal node 3 and obtaines the value 4.64 for total C.

Figure 12.4 Predicted values for total C modelled by a CART single tree with eight terminal nodes plotted against measured value.

Figure 13.1 Tukey Anscombe plots. Left: assumptions violated, variance not constant. Right: Assumptions achieved, no trend apparent.

Figure 13.2 Normal plots. Left: not normally distributed. Right: Normally distributed.

Figure 14.1 Histograms for the calibration (left side) and validation set (right side) for nitrogen (top) and phosphorus (bottom).
Figure 14.2  Measured y versus predicted y plot. Left: for calibration data set. Right: for validation data set.

Figure 16.1  Histograms for chemical soil properties of the whole analysed sample set. From top to bottom: total C as example for a symmetrical distribution, organic C as example for a right skewed distribution, exchangeable magnesium as example for a distribution with extreme outlier, clay as example for a bimodal distribution.

Figure 17.2  Photos of soil samples with minimal (left) and maximal (right) CIE Z value (blue). Photos: B.Wolfgangm

Figure 18.1  Histograms of calcium, CEC and sand for the whole data set (left side) and for the reduced data set of the brown carbonate soils (right side).

Figure 19.1  The two predictions of total C for the spectral repeat measurements plotted against each other. From top to bottom: PCR model, MLR model, CART model

Figure 20.1  Measured versus predicted charts for total C. Left: calibration data set. Right: validation data set.

Figure 20.2  Measured versus predicted charts for SOC. Left: calibration data set. Right: validation data set.

Figure 20.3  Measured versus predicted charts for total N. Left: calibration data set. Right: validation data set.

Figure 20.4  Measured versus predicted charts for pH. Left: calibration data set. Right: validation data set.

Figure 20.5  Absorption feature of CaCO$_3$ at 2340 nm. Left showed at the example of four soil samples with 1.5 (I), 11 (II), 28.8 (III) and 35.9 % (IV) of CaCO$_3$. Right amount of CaCO$_3$ plotted against the continuum removed wavelength 2340 nm for all chemically analysed samples. The red points correspond with the soil samples in the left figure.

Figure 20.6  Measured versus predicted charts for CaCO$_3$. Left: calibration data set. Right: validation data set.

Figure 20.7  Measured versus predicted charts for percentage of clay. Left: calibration data set. Right: validation data set.

Figure 20.8  Correlation coefficient R between the continuum removed wavelengths and the percentage of clay.

Figure 21.2  Histograms for SOC for Varzob area. Top: chemically analysed samples from the Faizabad and Yavan area for comparison. Middle: predictions of all Varzob samples. Bottom: predictions of Varzob samples within model limits.
Figure A.1  VIS-NIR-SWIR spectrum of muglight 136
Figure A.2  Transmission of Duran glass in the VIS-NIR-SWIR spectral region 136
Figure A.3  Histograms of the soil properties total C, SOC, total N and pH. Calibration data set on the left, validation data set on the right. 137
Figure A.4  Histograms of the soil properties CaCO$_3$, extractable P, exchangeable Ca and Mg. Calibration data set on the left, validation data set on the right. 138
Figure A.5  Histograms of the soil properties exchangeable K, clay, silt and sand. Calibration data set on the left, validation data set on the right. 139
Figure A.6  Histogram of the soil property cation exchange capacity. Calibration data set on the left, validation data set on the right. 140
Figure A.7  Histograms of the soil properties total C, SOC, total N and pH. Before (left side) and after (right side) the soil type separation. 141
Figure A.8  Histograms of the soil properties CaCO$_3$, extractable P, exchangeable Ca and Mg. Before (left side) and after (right side) the soil type separation. 142
Figure A.9  Histograms of the soil properties exchangeable K, clay, silt and sand. Before (left side) and after (right side) the soil type separation. 143
Figure A.10 Histogram of the soil property cation exchange capacity. Before (left side) and after (right side) the soil type separation. 144
Figure A.11 Model assumption plots for the total C model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance. 145
Figure A.12 Model assumption plots for the SOC model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance. 146
Figure A.13 Model assumption plots for the total N model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance. 147
Figure A.14 Model assumption plots for the pH model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance. 148
Figure A.15 Model assumption plots for the CaCO$_3$ model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance. 149
Figure A.16 Histograms for the soil property total C. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits. 150
Figure A.17  Histograms for the soil property SOC. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.

Figure A.18  Histograms for the soil property total N. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.

Figure A.19  Histograms for the soil property pH. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.

Figure A.20  Histograms for the soil property CaCO₃. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.
List of tables

Table 3.1  Extract of approximate wavelengths of some common functional groups between 2300 and 2500 nm [44].  29
Table 8.1  Methods for soil chemical analysis. Source: B.Wolfgramm  47
Table 8.2  Coefficients of variation (CV) of the double measured samples (LT samples) compared to lab-own and international CVs.  47
Table 11.1  Result of the CART classification tree for separation of the four soil groups. Source: B.Wolfgramm  62
Table 12.1  Explained variance for the principal components of spectral data of brown soils.  67
Table 12.2  Results of validation data set for different soil organic C CART models  69
Table 15.1  Limits for the application of total C model  81
Table 16.1  Correlations ($R^2$) between chemical soil properties  87
Table 17.1  Statistical measures for the CIE color coordinate X for the three test areas Faizabad, Yavan and Varzob. SD: Standard deviation.  89
Table 17.2  Correlations ($R^2$) between CIE color variables and chemical soil properties  89
Table 18.1  Result of the CART classification tree applied to whole data set. Number of samples given falling into different classes. Source: B.Wolfgramm  91
Table 18.2  Impact of soil type separation on model accuracy. Comparison of $R^2$ and RMSE of the validation data of MLR models for the complete dataset and the reduced data set of the brown carbonate soils at the examples of total C, total N, pH and CaCO$_3$.  93
Table 19.1  Comparison of MLR, PCR and CART models. N: abbreviation for number of regressors (MLR,PCR) or number of trees (CART).  93
Table 19.2  RMSD for MLR, PCR and CART models resulting from repeatability determination  95
Table 20.1  Accuracy measures for calibration data set of the MLR models  97
Table 20.2  Accuracy measures for validation data set of the MLR models  97
Table 20.3  Regressors and coefficients for the MLR model for total C  99
Table 20.4  Regressors and coefficients for the MLR model for SOC  101
Table 20.5  Regressors and coefficients for the MLR model for total N  103
Table 20.6  Regressors and coefficients for the MLR model for pH  105
Table 20.7  Regressors and coefficients for the MLR model for CaCO₃  107
Table 21.1  Samples within model limits for the Yavan and Faizabad test areas, given in absolute numbers and percentage.  111
Figure 21.1  Histograms for SOC for Faizabad and Yavan area. Left: chemically analysed samples. Right: predictions of samples within model limits.  111
Table 21.2  Samples within model limits for the Varzob test area, given in absolute numbers and percentage.  112
Table 23.1  Comparing SOC model (see chapter 19.2) with SOC models developed by Shepherd and Walsh [50] and Chang and Laird [12]. RMSEP: Root mean squared error of prediction. N calibration/validation: Number of samples used for calibration/validation.  117
Table 23.2  Comparing total N model (see chapter 19.3) with total N models developed by Hett [25] and Chang and Laird [12]. RMSEP: Root mean squared error of prediction. N calibration/validation: Number of samples used for calibration/validation.  117
Table A.1  Statistical measures for the soil properties of the chemically analysed soil samples before the soil type separation.  155
Table A.2  Statistical measures for the soil properties of the chemically analysed soil samples after the soil type separation, i.e. of the brown carbonate soil samples only.  155
Table A.3  Statistical measures for the CIE color variables subdivided into the three test areas Faizabad, Yavan and Varzob.  156
Table A.4  Limits for application of models for total C, SOC, total N, pH and CaCO₃  157
References


[34] JARMER, T., UDELHOVEN, T., 2001: Der Einsatz der Reflexionsspektrometrie zur Quantifizierung von organischem und anorganischem Kohlenstoff als Betrag zum „Precision Farming“. Geoid 1/2001, pp. 32-34.


Appendix

Additional figures

**Figure A.1** VIS-NIR-SWIR spectrum of muglight

**Figure A.2** Transmission of Duran glass in the VIS-NIR-SWIR spectral region
Figure A.3  Histograms of the soil properties total C, SOC, total N and pH. Calibration data set on the left, validation data set on the right.
Figure A.4  Histograms of the soil properties CaCO$_3$, extractable P, exchangeable Ca and Mg. Calibration data set on the left, validation data set on the right.
Figure A.5  Histograms of the soil properties exchangeable K, clay, silt and sand. Calibration data set on the left, validation data set on the right.
Figure A.6   Histogram of the soil property cation exchange capacity. Calibration data set on the left, validation data set on the right.
Figure A.7  Histograms of the soil properties total C, SOC, total N and pH. Before (left side) and after (right side) the soil type separation.
Figure A.8  Histograms of the soil properties CaCO$_3$, extractable P, exchangeable Ca and Mg. Before (left side) and after (right side) the soil type separation.
Figure A.9  Histograms of the soil properties exchangeable K, clay, silt and sand. Before (left side) and after (right side) the soil type separation.
Figure A.10  Histogram of the soil property cation exchange capacity. Before (left side) and after (right side) the soil type separation.
Figure A.11   Model assumption plots for the total C model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance.
Figure A.12  Model assumption plots for the SOC model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance.
Figure A.13  Model assumption plots for the total N model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance.
Figure A.14  Model assumption plots for the pH model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance.
Figure A.15 Model assumption plots for the CaCO$_3$ model. From top to bottom: Normal plot, Tukey-Anscombe plot, cook distance.
Figure A.16  Histograms for the soil property total C. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.
Figure A.17  Histograms for the soil property SOC. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.
Figure A.18  Histograms for the soil property total N. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.
Figure A.19  Histograms for the soil property pH. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.
Figure A.20  Histograms for the soil property CaCO$_3$. From top to bottom: Chemically analysed soil samples after soil type separation, predictions for the Yavan/Faizabad samples within model limits, predictions for the Varzob samples within model limits.
### Additional tables

#### Table A.1
Statistical measures for the soil properties of the chemically analysed soil samples before the soil type separation.

<table>
<thead>
<tr>
<th>Soil property</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
<th>0.025p</th>
<th>0.25p</th>
<th>0.5p</th>
<th>0.75p</th>
<th>0.975p</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total C</td>
<td>2.94</td>
<td>0.25</td>
<td>5.77</td>
<td>1.08</td>
<td>0.92</td>
<td>2.15</td>
<td>3.06</td>
<td>3.69</td>
<td>4.84</td>
<td>%</td>
</tr>
<tr>
<td>SOC</td>
<td>1.09</td>
<td>0.02</td>
<td>4.60</td>
<td>0.71</td>
<td>0.12</td>
<td>0.63</td>
<td>0.98</td>
<td>1.36</td>
<td>3.05</td>
<td>%</td>
</tr>
<tr>
<td>Total N</td>
<td>0.13</td>
<td>0.02</td>
<td>0.38</td>
<td>0.06</td>
<td>0.03</td>
<td>0.08</td>
<td>0.12</td>
<td>0.16</td>
<td>0.29</td>
<td>%</td>
</tr>
<tr>
<td>pH</td>
<td>8.05</td>
<td>7.01</td>
<td>8.66</td>
<td>0.25</td>
<td>7.27</td>
<td>7.98</td>
<td>8.09</td>
<td>8.21</td>
<td>8.40</td>
<td>log H</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>19.41</td>
<td>0.37</td>
<td>41.40</td>
<td>9.37</td>
<td>1.63</td>
<td>11.58</td>
<td>20.75</td>
<td>27.52</td>
<td>32.00</td>
<td>%</td>
</tr>
<tr>
<td>Extr. P</td>
<td>5.75</td>
<td>0.13</td>
<td>212.20</td>
<td>16.25</td>
<td>0.40</td>
<td>1.40</td>
<td>2.35</td>
<td>4.89</td>
<td>28.98</td>
<td>mg/kg</td>
</tr>
<tr>
<td>Exch. Ca</td>
<td>12.56</td>
<td>0.70</td>
<td>67.80</td>
<td>6.25</td>
<td>4.20</td>
<td>10.44</td>
<td>12.10</td>
<td>14.00</td>
<td>17.82</td>
<td>me/100g</td>
</tr>
<tr>
<td>Exch. Mg</td>
<td>0.88</td>
<td>0.12</td>
<td>7.10</td>
<td>0.68</td>
<td>0.30</td>
<td>0.50</td>
<td>0.70</td>
<td>1.00</td>
<td>2.26</td>
<td>me/100g</td>
</tr>
<tr>
<td>Exch. K</td>
<td>0.35</td>
<td>0.08</td>
<td>1.09</td>
<td>0.18</td>
<td>0.10</td>
<td>0.21</td>
<td>0.30</td>
<td>0.46</td>
<td>0.75</td>
<td>me/100g</td>
</tr>
<tr>
<td>Clay</td>
<td>28</td>
<td>4</td>
<td>50</td>
<td>10</td>
<td>6</td>
<td>26</td>
<td>31</td>
<td>34</td>
<td>41</td>
<td>%</td>
</tr>
<tr>
<td>Silt</td>
<td>50</td>
<td>10</td>
<td>87</td>
<td>13</td>
<td>20</td>
<td>44</td>
<td>51</td>
<td>56</td>
<td>82</td>
<td>%</td>
</tr>
<tr>
<td>Sand</td>
<td>22</td>
<td>6</td>
<td>82</td>
<td>14</td>
<td>8</td>
<td>14</td>
<td>18</td>
<td>24</td>
<td>66</td>
<td>%</td>
</tr>
<tr>
<td>CEC</td>
<td>13.79</td>
<td>1.85</td>
<td>68.21</td>
<td>6.23</td>
<td>5.02</td>
<td>11.58</td>
<td>15.13</td>
<td>15.69</td>
<td>18.84</td>
<td>me/100g</td>
</tr>
</tbody>
</table>

#### Table A.2
Statistical measures for the soil properties of the chemically analysed soil samples after the soil type separation, i.e. of the brown carbonate soil samples only.

<table>
<thead>
<tr>
<th>Soil property</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
<th>0.025p</th>
<th>0.25p</th>
<th>0.5p</th>
<th>0.75p</th>
<th>0.975p</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total C</td>
<td>2.94</td>
<td>0.25</td>
<td>5.77</td>
<td>1.08</td>
<td>0.92</td>
<td>2.15</td>
<td>3.06</td>
<td>3.69</td>
<td>4.84</td>
<td>%</td>
</tr>
<tr>
<td>SOC</td>
<td>1.09</td>
<td>0.02</td>
<td>4.60</td>
<td>0.71</td>
<td>0.12</td>
<td>0.63</td>
<td>0.98</td>
<td>1.36</td>
<td>3.05</td>
<td>%</td>
</tr>
<tr>
<td>Total N</td>
<td>0.13</td>
<td>0.02</td>
<td>0.38</td>
<td>0.06</td>
<td>0.03</td>
<td>0.08</td>
<td>0.12</td>
<td>0.16</td>
<td>0.29</td>
<td>%</td>
</tr>
<tr>
<td>pH</td>
<td>8.05</td>
<td>7.01</td>
<td>8.66</td>
<td>0.25</td>
<td>7.27</td>
<td>7.98</td>
<td>8.09</td>
<td>8.21</td>
<td>8.40</td>
<td>log H</td>
</tr>
<tr>
<td>CaCO₃</td>
<td>20.91</td>
<td>0.40</td>
<td>41.40</td>
<td>9.37</td>
<td>1.66</td>
<td>14.40</td>
<td>23.65</td>
<td>28.10</td>
<td>32.47</td>
<td>%</td>
</tr>
<tr>
<td>Extr. P</td>
<td>6.34</td>
<td>0.30</td>
<td>212.20</td>
<td>18.75</td>
<td>0.63</td>
<td>1.50</td>
<td>2.70</td>
<td>5.18</td>
<td>26.97</td>
<td>mg/kg</td>
</tr>
<tr>
<td>Exch. Ca</td>
<td>12.23</td>
<td>0.70</td>
<td>19.80</td>
<td>2.78</td>
<td>6.96</td>
<td>10.70</td>
<td>12.10</td>
<td>13.90</td>
<td>16.89</td>
<td>me/100g</td>
</tr>
<tr>
<td>Exch. Mg</td>
<td>0.86</td>
<td>0.20</td>
<td>5.50</td>
<td>0.56</td>
<td>0.35</td>
<td>0.50</td>
<td>0.70</td>
<td>1.00</td>
<td>2.08</td>
<td>me/100g</td>
</tr>
<tr>
<td>Exch. K</td>
<td>0.36</td>
<td>0.09</td>
<td>1.09</td>
<td>0.19</td>
<td>0.11</td>
<td>0.22</td>
<td>0.33</td>
<td>0.48</td>
<td>0.75</td>
<td>me/100g</td>
</tr>
<tr>
<td>Clay</td>
<td>29</td>
<td>4</td>
<td>43</td>
<td>9</td>
<td>6</td>
<td>28</td>
<td>31</td>
<td>34</td>
<td>41</td>
<td>%</td>
</tr>
<tr>
<td>Silt</td>
<td>54</td>
<td>16</td>
<td>87</td>
<td>11</td>
<td>38</td>
<td>50</td>
<td>53</td>
<td>57</td>
<td>83</td>
<td>%</td>
</tr>
<tr>
<td>Sand</td>
<td>17</td>
<td>6</td>
<td>68</td>
<td>7</td>
<td>8</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>32</td>
<td>%</td>
</tr>
<tr>
<td>CEC</td>
<td>13.45</td>
<td>1.85</td>
<td>22.13</td>
<td>2.75</td>
<td>8.29</td>
<td>11.84</td>
<td>13.23</td>
<td>15.20</td>
<td>17.88</td>
<td>me/100g</td>
</tr>
<tr>
<td>Varzob</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.101</td>
<td>0.097</td>
<td>0.015</td>
<td>0.505</td>
<td>0.409</td>
<td>0.060</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>0.359</td>
<td>0.285</td>
<td>0.057</td>
<td>0.528</td>
<td>0.415</td>
<td>0.082</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.211</td>
<td>0.189</td>
<td>0.030</td>
<td>0.514</td>
<td>0.412</td>
<td>0.074</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.054</td>
<td>0.042</td>
<td>0.008</td>
<td>0.003</td>
<td>0.001</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Yavan</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.127</td>
<td>0.097</td>
<td>0.014</td>
<td>0.499</td>
<td>0.396</td>
<td>0.053</td>
</tr>
<tr>
<td>Max</td>
<td>0.351</td>
<td>0.285</td>
<td>0.053</td>
<td>0.551</td>
<td>0.415</td>
<td>0.087</td>
</tr>
<tr>
<td>Mean</td>
<td>0.236</td>
<td>0.189</td>
<td>0.034</td>
<td>0.515</td>
<td>0.411</td>
<td>0.074</td>
</tr>
<tr>
<td>SD</td>
<td>0.051</td>
<td>0.042</td>
<td>0.008</td>
<td>0.006</td>
<td>0.003</td>
<td>0.004</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Faizabad</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.105</td>
<td>0.084</td>
<td>0.016</td>
<td>0.504</td>
<td>0.396</td>
<td>0.056</td>
</tr>
<tr>
<td>Max</td>
<td>0.404</td>
<td>0.328</td>
<td>0.064</td>
<td>0.542</td>
<td>0.417</td>
<td>0.080</td>
</tr>
<tr>
<td>Mean</td>
<td>0.240</td>
<td>0.191</td>
<td>0.033</td>
<td>0.518</td>
<td>0.411</td>
<td>0.071</td>
</tr>
<tr>
<td>SD</td>
<td>0.048</td>
<td>0.040</td>
<td>0.008</td>
<td>0.007</td>
<td>0.003</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Table A.3  Statistical measures for the CIE color variables subdivided into the three test areas Faizabad, Yavan and Varzob.
<table>
<thead>
<tr>
<th>Limits total C model</th>
<th>Regressor</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.105</td>
<td>0.404</td>
<td></td>
</tr>
<tr>
<td>2180</td>
<td>0.956</td>
<td>0.973</td>
<td></td>
</tr>
<tr>
<td>1490</td>
<td>0.964</td>
<td>0.990</td>
<td></td>
</tr>
<tr>
<td>2320</td>
<td>0.966</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>1860</td>
<td>0.965</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>1300</td>
<td>0.997</td>
<td>0.999</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Limits SOC model</th>
<th>Regressor</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1870</td>
<td>0.953</td>
<td>0.980</td>
<td></td>
</tr>
<tr>
<td>1530</td>
<td>0.977</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td>2180</td>
<td>0.965</td>
<td>0.973</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Limits Total N model</th>
<th>Regressor</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1870</td>
<td>0.953</td>
<td>0.980</td>
<td></td>
</tr>
<tr>
<td>1530</td>
<td>0.977</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td>900</td>
<td>0.977</td>
<td>0.999</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Limits pH model</th>
<th>Regressor</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1860</td>
<td>0.965</td>
<td>0.987</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>0.503</td>
<td>0.520</td>
<td></td>
</tr>
<tr>
<td>1700</td>
<td>0.984</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>720</td>
<td>0.938</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Limits CaCO₃ model</th>
<th>Regressor</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.105</td>
<td>0.404</td>
<td></td>
</tr>
<tr>
<td>2340</td>
<td>0.956</td>
<td>0.984</td>
<td></td>
</tr>
<tr>
<td>2380</td>
<td>0.992</td>
<td>0.999</td>
<td></td>
</tr>
</tbody>
</table>

Table A.4  
Limits for application of models for total C, SOC, total N, pH and CaCO₃.
**IDP codes**

**CIE color parameter calculation**

```idl
; RSL Programs and Utilities
; ---------------------------------------------------------------------
; NAME:  cie_xyz.pro
; PURPOSE: This procedure calculates the CIE color coordinates X,Y,Z and the chromaticity coordinates x,y,z.
; INPUTS: 
; - ascii file containing the spectrum of the illuminating light source between 360 and 830 nm in 1 nm steps
; - ascii file containing the CIE color matching functions in 1 nm steps
; - ascii file containing the spectra of the samples in 1 nm steps between 360 and 830 nm
; OUTPUTS: ascii-file containing the calculated color parameters for each sample in a column: X,Y,Z,x,y,z
; Modification History: July 27, 2006 by Bruno Seiler, RSL, University Zurich

PRO  cie_xyz

; path = /data/apex_paf/bseiler/src/idl, get_path = workpath

cmf = pickfile(Title = 'Select ascii-file containing color matching functions')
cmf = rapparr(cmf,/nolabel)

spectra = pickfile(Title = 'Select ascii-file containing spectra')
spectra=rapparr(spectra,title)

lightrad = pickfile(Title = 'Select ascii-file containing muglight radiance')
lightrad = rapparr(lightrad,/nolabel)

resultarr=fltarr(n_elements(spectra[*,0]),6)

ymax=0

FOR i=0,469 DO ymax=ymax+lightrad[0,i]*cmf[0,i]

FOR i=0,(n_elements(spectra[*,0])-1) DO BEGIN
  x=0
  y=0
  z=0
  FOR j=0,469 DO BEGIN
    result=spectra[i,j]*lightrad[0,j]*cmf[0,j]
    x=x+result
  ENDFOR
  resultarr[i,0]=x/ymax
  FOR k=0,469 DO BEGIN
    result=spectra[i,k]*lightrad[0,k]*cmf[1,k]
    y=y+result
  ENDFOR
  resultarr[i,1]=y/ymax
  FOR l=0,469 DO BEGIN
    result=spectra[i,l]*lightrad[0,l]*cmf[2,l]
    z=z+result
  ENDFOR
  resultarr[i,2]=z/ymax
```

```
; adding chromaticity coordinates x, y, z to resultarr
resultarr[3] = x / (x + y + z)
resultarr[4] = y / (x + y + z)
resultarr[5] = z / (x + y + z)
ENDFOR

labels = title[0:n_elements(spectra[*,*])-1]

wapparr, resultarr, 'XYZ.txt', [labels]
END

PC calculation

; RSL Programs and Utilities
;-----------------------------------------------------------------
; NAME: pcc.pro
; PURPOSE: This procedure calculates the principal components for a data set.
; INPUTS: ascii file containing the data: samples in rows, variables in columns
; OUTPUTS: -pccoefficients.txt: contains the coefficients describing the influence of the old variables on each PC
; -pc.txt: contains the newly calculated principal components
; -percentvariance.txt: contains the rate of the explained variance for each PC
; MODIFICATION HISTORY: July 27, 2006 by Bruno Seiler, RSL, University Zurich
;----------------------------------------------------------------------------

PRO pcc

pcdata = pickfile(Title='Select whole dataset for calculation of principal components')

pcdata = rapparr(pcdata)

; set number of principal components
xvalue, ['Specify the number of pcs to be calculated:'], pcnum, continue, /window, /number, default=20

; calculating the principal components
;m=n_elements(pcdata[*,0])
m=n_elements(pcdata[0,*])
means = fltarr(m)
means = TOTAL(pcdata, 2,/double)/n
FOR i=0, (m-1) DO pcdata[i,*] = pcdata[i,*] - means(i)
pccalibr = PCOMP(pcdata, coefficients=coefficients, nvariables=pcnum, variances=variances)

wapparr, coefficients, 'pccoefficients.txt'
wapparr, pccalibr, 'pc.txt'
wapparr, variances, 'percentvariance.txt'
END
Calibration program

1 ; RSL Programs and Utilities
2 ; -------------------------------------------------------------
3 ;
4 ; NAME: calibration.pro
5 ;
6 ; PURPOSE: Calculates calibration equations to predict chemical properties based on spectral data.
7 ;
8 ; INPUTS:
9 ; two ascii-files containing chemical data and spectral data in columns:
10 ; - two third for model calibration
11 ; - one third for model validation
12 ;
13 ; KEYWORD PARAMETERS:
14 ;   group: group of dialog messages or child widgets (heritable)
15 ;
16 ; OUTPUTS:
17 ;  -logcal.txt: logfile containing the shell-text-output of the program
18 ;  -if option postscript is selected: postscript output of all plots
19 ;
20 ; MODIFICATION HISTORY:
21 ;   July 27, 2006 by Bruno Seiler, RSL, University Zurich
22 ;
23 ;----------------------------------------------------------------------------
24 ;calibration program using multiple linear regression
25 
26 PRO calibration
27
28 !except=0
29 ;*****************************************
30 ;setting inputfiles and variables
31 ;*****************************************
32
33 calibrpath = pickfile(Title = 'Select ascii-file for Calibration')
34 calibr = rapparr(calibrpath, title)
35
36 validpath = pickfile(Title = 'Select ascii-file for Validation')
37 valid = rapparr(validpath)
38
39 ;set column number of the dependant variable
40 xvalue, ['Enter the column number of the dependant variable:'], nrdep, continue, /window, /number, default=1
41 nrdep=nrdep-1
42
43 ;set column number of the first independant variable
44 xvalue, ['Enter the column number of the first independant variable:'], nrindep, continue, /window, /number, default=1
45 nrindep=nrindep-1
46
47 ;set number of independant variables for multiple linear regression
48 xvalue, ['Enter the number of wavelengths for multiple linear regression:'], ncorr, continue, /window, /number, default=5
49
50 ;set number of model variables in input ascii-file
51 var=n_elements(calibr[*,0])-nrdep
52
53 ;set number of samples in input calibration ascii-file
54 csamples=n_elements(calibr[0,*])
55
56 ;set number of samples in input validation ascii-file
57 vsamples=n_elements(valid[0,*])
58
59 ;set output: screen or screen and postscriptfile
60 xvalue, ['Do you want output on screen only or additionally as postscript("PS")?'], o, continue, /window, default='screen'
61
62 IF o EQ 'screen' THEN output=0
63 IF o EQ 'PS' THEN output=1
64
65 ;logfile
66 openw, log, 'log_mlrfs.txt', /get_lun
67
68
; descriptive statistics of calibration and validation data set
;---------------------------------------------------------------

vstatmeas=fltarr(3)
cstatmeas=fltarr(3)

vstatmeas[0]=mean(valid[nrdep,*])
vstatmeas[1]=max(valid[nrdep,*])-min(valid[nrdep,*])
vstatmeas[2]=stdev(valid[nrdep,*])
cstatmeas[0]=mean(calibr[nrdep,*])
cstatmeas[1]=max(calibr[nrdep,*])-min(calibr[nrdep,*])
cstatmeas[2]=stdev(calibr[nrdep,*])

FOR n=0,1 DO BEGIN
  IF n EQ 0 THEN lun=-1
  IF n EQ 1 THEN lun=log
  printf, lun,'**
  printf, lun,'* Descriptive statistics for calibration and validation data set *
  printf, lun,'**
  printf, lun,'   calibration data set   --------------------
  printf, lun,'   mean ',cstatmeas(0)
  printf, lun,'   range',cstatmeas(1)
  printf, lun,'   stdev',cstatmeas(2)
  printf, lun,'   n   ',string(csamples)
  printf, lun,'**

  printf, lun,'   validation data set   -------------------
  printf, lun,'   mean ',vstatmeas(0)
  printf, lun,'   range',vstatmeas(1)
  printf, lun,'   stdev',vstatmeas(2)
  printf, lun,'   n   ',string(vsamples)
  printf, lun,'**
ENDFOR

;---------------------------------------------------------------

; y-variable transformation
;---------------------------------------------------------------

y=calibr[nrdep,*]
transy=fltarr(12,csamples)
transvy=fltarr(12,vsamples)
vy=valid[nrdep,*]

FOR i=0, (csamples-1) DO BEGIN
  transy[4,i]=(y(i)^2-1)/2
  transy[5,i]=(y(i)^0.5-1)/0.5
  transy[6,i]=(y(i)^0.3-1)/0.3
  transy[7,i]=(y(i)^0.1-1)/0.1
  transy[8,i]=(y(i)^(-0.1)-1)/(-0.1)
  transy[9,i]=(y(i)^(-0.3)-1)/(-0.3)
  transy[10,i]=(y(i)^(-0.5)-1)/(-0.5)
  transy[11,i]=(y(i)^(-2)-1)/(-2)
ENDFOR

transvy[0,*]=vy
transy[1,*]=alog(y)
transy[2,*]=sqrt(y)
transy[3,*]=vy

FOR i=0, (vsamples-1) DO BEGIN
  transvy[4,i]=(vy(i)^2-1)/2
  transvy[5,i]=(vy(i)^0.5-1)/0.5
  transvy[6,i]=(vy(i)^0.3-1)/0.3
  transvy[7,i]=(vy(i)^0.1-1)/0.1
  transvy[8,i]=(vy(i)^(-0.1)-1)/(-0.1)
  transvy[9,i]=(vy(i)^(-0.3)-1)/(-0.3)
  transvy[10,i]=(vy(i)^(-0.5)-1)/(-0.5)
  transvy[11,i]=(vy(i)^(-2)-1)/(-2)
ENDFOR

;---------------------------------------------------------------
transystand=standardize(transy)

; expected frequency
p1=0.1578
p2=0.3422
p3=0.3413
p4=0.1587

tfreq=fltarr(1,4)
tfreq[0,0]=p1*csamples
tfreq[0,1]=p2*csamples
tfreq[0,2]=p3*csamples
tfreq[0,3]=p4*csamples

; counting frequency of four groups with borders +1Stdev, Mean, -1 Stdev, in this case : 1, 0, -1
chi2=fltarr(12)

FOR k=0,11 DO BEGIN
  freq=fltarr(4)
  FOR i=0, (csamples-1) DO BEGIN
    IF transystand[k,i] LT -1 THEN freq[0]=freq[0]+1
    IF (transystand[k,i] LT 0 AND transystand[k,i] GT -1) THEN freq[1]=freq[1]+1
  ENDFOR
  FOR l=0,3 DO chi2[k]=chi2[k]+(freq[l]-tfreq[0,l])^2/tfreq[0,l]
ENDFOR

transname=[\'NONE\',\'LN\',\'\'SQR\'\',\'\'SQRT\'\',\'\'SQUARE\'\',\'\'BOXCOX(2)\'\',\'\'BOXCOX(0.5)\'\',\'\'BOXCOX(0.3)\'\',\'\'BOXCOX(0.1)\'\',\'\'BOXCOX(-0.1)\'\',\'\'BOXCOX(-0.3)\'\',\'\'BOXCOX(-0.5)\'\',\'\'BOXCOX(-2)\'\']

ENDFOR

FOR n=0,1 DO BEGIN
  IF n EQ 0 THEN lun=-1
  IF n EQ 1 THEN lun=log

  printf, lun, \'*
  printf, lun, \'*****************************************
  printf, lun, \'* Choose transformation for y-variable *
  printf, lun, \'*****************************************
  printf, lun, \'chi square test values for different transformations of y variable (test on normal distribution)
  printf, lun, \'transformation    test value
  printf, lun, \'---    -----------
  FOR i=0,11 DO printf, lun, \'+string(transname(i))+string(chi2(i))
  printf, lun, \'---    -----------

  printf, lun, \'critical value for alpha=0.05 and df=1: 3.8411
ENDFOR

; choosing transformation
xvalue, [\'Choose transformation:\'], transf, continue, window, window, default=1

FOR n=0,1 DO BEGIN
  IF n EQ 0 THEN lun=-1
  IF n EQ 1 THEN lun=log

  printf, lun, \'
  printf, lun, \'choosing transformation
  printf, lun, \'trans=\'+transf+'-1
  printf, lun, \'y=\'+transy+'\'
  printf, lun, \'specified\'

  FOR n=0,1 DO BEGIN
    IF n EQ 0 THEN lun=-1
    IF n EQ 1 THEN lun=log

    printf, lun, \'
    printf, lun, \'chosen transformation for y-variable: \'+transname(transf)
  ENDFOR
  printf, lun, \'

  ;*****************************************
  ; choosing wavelengths for forward stepwise multiple linear regression
  ;*****************************************

  cmultiplecorr= fltarr(ncorr)
  mnamechosen=strarr(ncorr)
  mvarchosen=intarr(ncorr)
  chosen=fltarr(var)

  printf, \'
  printf, \'b=\'+fltarr(var)
  printf, \'
  printf, \'c=\'+fltarr(var)
first=1

chemical=transpose(y)

FOR j=0, (ncorr-1) DO BEGIN
  b(*)=0
  c(*)=0
  FOR j=0,(var-1) DO BEGIN
    IF chosen(j) EQ 0 THEN BEGIN
      IF first EQ 0 THEN s=cspec, calibr[j+nrindep, *]
      IF first EQ 1 THEN s=calibr[j+nrindep, *]
      b(j)=m_correlate(s, chemical)
    ENDIF
  ENDFOR
  FOR f=0, (var-1) DO IF NOT finite(b(f)) THEN b(f)=0
  c=reverse(sort(abs(b)))
  d=c(0)
  IF first EQ 1 THEN cspec=calibr[d+nrindep, *]
  IF first EQ 0 THEN cspec=[cspec, calibr[d+nrindep, *]]
  cmultiplecorr(i)=d
  mnamechosen(i)=title(d+nrindep)
  chosen(d)=1
  mvarchosen(i)=d+nrindep
  first=0
ENDFOR

---

predict=fltarr(1, vsamples)
validpearson= fltarr(ncorr)
rmsep=fltarr(ncorr)
meanpred=fltarr(ncorr)
rangepred=fltarr(ncorr)
stdevpred=fltarr(ncorr)
regrcoeff=fltarr(ncorr, ncorr+1)
sse=fltarr(ncorr)
see=fltarr(ncorr)
mapec=fltarr(ncorr)
mapep=fltarr(ncorr)
rmspec=fltarr(ncorr)
rmspep=fltarr(ncorr)
residuals=fltarr(50)
cook=fltarr(ncorr, csamples)
f=f=fltarr(ncorr)
fc=f=fltarr(ncorr, 3)
t=t=fltarr(ncorr, ncorr+1)
times=0
pcorr=fltarr(ncorr, ncorr)

first=1

FOR j=0, (ncorr-1) DO BEGIN
  predict[*]=0
  IF first EQ 0 THEN BEGIN
    FOR j=1, i DO BEGIN
      x=mvarchosen(j)
      cspec=[cspec, calibr[x, *]]
      vspec=[vspec, valid[x, *]]
    ENDFOR
  ENDIF
  IF first EQ 1 THEN BEGIN
    x=mvarchosen(0)
    cspec=calibr(x, *]
    vspec=valid(x, *]
  first=0
  ENDIF
ENDFOR

result=regress(cspec, transpose(y), sigma=omega, const=const, yfit=yfit, ftest=ftest, mcorrelation=mochoice)

writing regression coefficients in array
regrocoeff[0]=const
FOR j=0, DO regrocoeff[j+1]=result[j]

;calculating model diagnostics
;***********************
336 ; calculating sum of squared errors for f-test
337 sse(i)=total((serror-yfit)^2)
338 ; calculating the standard error of estimate (\textit{RMSE})
339 see(i)=sqrt((mean(serror^2)))
340 ; calculating MAPE
341 mapec(i)=100*mean(abs(y-yfit)/y)
342 ; calculating RMSPE
343 rmspec(i)=100*sqrt(mean(((y-yfit)^2)/(y^2)))
344 ; testing the residuals on normal distribution with chi-squared test
345 p1=0.1578
346 p2=0.3422
347 p3=0.3413
348 p4=0.1587
349 tfreq=fltarr(4)
350 tfreq(0)=p1*n_elements(y[0,:])
351 tfreq(1)=p2*n_elements(y[0,:])
352 tfreq(2)=p3*n_elements(y[0,:])
353 tfreq(3)=p4*n_elements(y[0,:])
354 ; normalization of residuals
355 sserrorstand=standardize(serror)
356 ; calculating test value
357 chisq=0
358 FOR l=0, (n_elements(y[0,:])-1) DO BEGIN
359 IF sserrorstand(l) LT -1 THEN tfreq(0)=tfreq(0)+1
360 IF (sserrorstand(l) LT 0 AND sserrorstand(l) GT -1) THEN tfreq(1)=tfreq(1)+1
361 IF (sserrorstand(l) LT 1 AND sserrorstand(l) GT 0) THEN tfreq(2)=tfreq(2)+1
362 IF sserrorstand(l) GT 1 THEN tfreq(3)=tfreq(3)+1
363 ENDFOR
364 FOR l=0,3 DO chisq=chisq+(tfreq(l)-tfreq(l))^2/tfreq(l)
365 ; critical value for alpha=0.05 and degree of freedom=1: 3.841
366 FOR n=0,output DO BEGIN
367 IF n EQ 0 THEN BEGIN
368 set_plot, 'X'
369 window, 0, xsize=750, ysize=600
370 !p.multi=[0,2,2,0,0]
371 ENDIF
372 IF n EQ 1 THEN BEGIN
373 IF times EQ 0 THEN filename=ʻplots1.psʻ
374 IF times EQ 1 THEN filename=ʻplots2.psʻ
375 IF times EQ 2 THEN filename=ʻplots3.psʻ
376 IF times EQ 3 THEN filename=ʻplots4.psʻ
377 IF times EQ 4 THEN filename=ʻplots5.psʻ
378 IF times EQ 5 THEN filename=ʻplots6.psʻ
379 IF times EQ 6 THEN filename=ʻplots7.psʻ
380 IF times EQ 7 THEN filename=ʻplots8.psʻ
381 IF times EQ 8 THEN filename=ʻplots9.psʻ
382 IF times EQ 9 THEN filename=ʻplots10.psʻ
383 IF times EQ 10 THEN filename=ʻplots11.psʻ
384 IF times EQ 11 THEN filename=ʻplots12.psʻ
385 IF times EQ 12 THEN filename=ʻplots13.psʻ
386 IF times EQ 13 THEN filename=ʻplots14.psʻ
387 IF times EQ 14 THEN filename=ʻplots15.psʻ
388 IF times EQ 15 THEN filename=ʻplots16.psʻ
389 IF times EQ 16 THEN filename=ʻplots17.psʻ
390 IF times EQ 17 THEN filename=ʻplots18.psʻ
391 IF times EQ 18 THEN filename=ʻplots19.psʻ
392 IF times EQ 19 THEN filename=ʻplots20.psʻ
393 IF times EQ 0 THEN filename=plots1.ps
394 IF times EQ 1 THEN filename=plots2.ps
395 IF times EQ 2 THEN filename=plots3.ps
396 IF times EQ 3 THEN filename=plots4.ps
397 IF times EQ 4 THEN filename=plots5.ps
398 IF times EQ 5 THEN filename=plots6.ps
399 IF times EQ 6 THEN filename=plots7.ps
400 IF times EQ 7 THEN filename=plots8.ps
401 IF times EQ 8 THEN filename=plots9.ps
402 IF times EQ 9 THEN filename=plots10.ps
403 IF times EQ 10 THEN filename=plots11.ps
404 IF times EQ 11 THEN filename=plots12.ps
405 IF times EQ 12 THEN filename=plots13.ps
406 IF times EQ 13 THEN filename=plots14.ps
407 IF times EQ 14 THEN filename=plots15.ps
408 IF times EQ 15 THEN filename=plots16.ps
409 IF times EQ 16 THEN filename=plots17.ps
410 IF times EQ 17 THEN filename=plots18.ps
411 IF times EQ 18 THEN filename=plots19.ps
412 IF times EQ 19 THEN filename=plots20.ps
413 IF times EQ 0 THEN filename=plots1.ps
414 IF times EQ 1 THEN filename=plots2.ps
415 IF times EQ 2 THEN filename=plots3.ps
416 IF times EQ 3 THEN filename=plots4.ps
417 IF times EQ 4 THEN filename=plots5.ps
418 IF times EQ 5 THEN filename=plots6.ps
419 IF times EQ 6 THEN filename=plots7.ps
420 IF times EQ 7 THEN filename=plots8.ps
421 IF times EQ 8 THEN filename=plots9.ps
422 IF times EQ 9 THEN filename=plots10.ps
423 IF times EQ 10 THEN filename=plots11.ps
424 IF times EQ 11 THEN filename=plots12.ps
425 IF times EQ 12 THEN filename=plots13.ps
426 IF times EQ 13 THEN filename=plots14.ps
427 IF times EQ 14 THEN filename=plots15.ps
428 IF times EQ 15 THEN filename=plots16.ps
429 IF times EQ 16 THEN filename=plots17.ps
430 IF times EQ 17 THEN filename=plots18.ps
431 IF times EQ 18 THEN filename=plots19.ps
432 IF times EQ 19 THEN filename=plots20.ps
433 ; Tukey-Anscombe plot
434 variables=string(i+1)
435 plot, yfit,sserrorstand, psym=5, symsize=0.7, xtitle=ʻfitʼ, ytitle=ʻresidualsʼ, $title=ʻTukey-Anscombe plot for ‘+title(nrdep)+‘ modelʼ, subttitle=variables, predictor variable(s)ʼ
a=[0,0]
b=[-100,100]

;plotting residuals against x-variables
plot, cspec[0,*], sserrorstand, psym=5, symsize=0.7, xtitle=\$xnamechosen(i)\$, ytitle=\$residuals\$
residuals, title=\'Residuals vs.  x-variable\', subtitle=variables\$+  predictor variable(s)\$
a=[0,0]
b=[-100,100]

;plotting residuals against x-variables
plot, b, a, thick=0

;normal plot with 2% quantiles
normal=[-2.05,-1.75,-1.55,-1.40,-1.28,-1.16,-1.08,-0.98,-0.84,-0.76,-0.70,\$-0.64,-0.58,-0.52,-0.46,-0.40,-0.34,-0.30,-0.24,-0.2,-0.14,-0.1,\$-0.04,0,0.06,0.12,0.16,0.22,0.26,0.32,0.36,0.42,0.48,0.54,0.6,0.66,0.72,\$0.78,0.86,0.92,1,1.1,1.18,1.3,1.42,1.6,1.8,2.1]
b=csamples/50
quant=48

a=sort(sserrorstand)
k=0
FOR l=0,quant DO BEGIN
  k=k+b
  c=a(k)
  residuals(l)=sserrorstand(c)
ENDFOR

plot, normal, residuals, /isotropic, psym=5, symsize=0.7, xtitle=\'theoretical quantiles\', ytitle=\'quantiles of residuals\', title=\'normal plot: residuals of \$+title(nrdep)\$, subtitle=variables + \$predictor variable(s)\$
xrange=[-2.5,2.5], yrange=[-2.5,2.5]
a=[-3,3]

;calculating and plotting cook distance

a=fltarr(1,n_elements(y[0,*]))
a*[*,*]=1
studres=fltarr(n_elements(y[0,*]))

x=[a,cspec]
p=x##invert(transpose(x)##x) ;hat matrix
p=diag_matrix(p) ;leverage vector

;calculating studentized residuals
s2=sse(i)/(n_elements(y[0,*])-i-1)

FOR l=0,(n_elements(y[0,*])-1) DO BEGIN
  studres(l)=sserror(l)/(sqrt(s2*(1-p(l))))
  cook[i,l]=(studres(l)^2)*p(l)/(1-p(l))/(i+1)
ENDFOR

plot, cook[i,*], psym=5, xtitle='samples\', ytitle='cook distance\', title='cook distance of \$+title(nrdep)\$ samples\', subtitle=variables + \$predictor variable(s)\$

IF n EQ 0 THEN BEGIN

;calculating test statistic for f-test (whole model)
b=[const,transpose(result)]
SS=b##transpose(x)##y-total(y)/n_elements(y[0,*])

test value, df of numerator=(i+1), df of denominator=(csamples-i-1)
f(i)=(SS/(i+1))/(s2)

;critical f-values for alpha=0.01, df of denominator=120, df of numerator=1,...,20
cf=[6.85,4.79,4.14,3.48,3.22,2.96,2.81,2.62,2.57,2.47,2.34,\$2.34,2.19,2.19,2.19,2.03,2.03,2.03,2.03,2.03]

IF i GT 0 THEN BEGIN

;calculating test statistic for f-test (comparing models)
FOR j=1, i DO BEGIN
  fcm[j,0]=(sse[i-1]-sse[i])/(i)/(sse[i]/(n_elements(y[0,*])-(i+1)))
ENDFOR

IF j GT 0 THEN BEGIN
$fcm[i,1] = i$
$fcm[i,2] = n_elements(y[0,*]) - (i+1)$

ENDIF

; t-test for regression coefficients, df=csamples-i-2
; H0: $b_i \equiv 0$
; H1: $b_i \neq 0$

$C_x = invert(transpose(x)##x)$
FOR $j=0,(i+1)$ DO $t[i,j] = b[j]/(sqrt(s2*C_x[j,j]))$

; calculating partial correlation coefficients for variables in model
IF i EQ 0 THEN $pcorr[i,i] = mcorrelation$
IF i NE 0 THEN BEGIN
FOR $j=0,i$ DO BEGIN
partial = transpose(cspec[j,*])
index = replicate(1L,i+1)
index(j) = 0L
keepcolumn = where(index EQ 1)
remove = cspec[keepcolumn, *]
$pcorr[i,j] = p_correlate(partial, transpose(y), remove)$
ENDFOR
ENDIF

; predicting values for validation data set
FOR $j=0,i$ DO predict = predict + regrcoeff[i,j+1]*vspec[j,*]
predict = predict + regrcoeff[i,0]

ENDIF

IF n EQ 0 THEN BEGIN
set_plot, 'X'
window, 1, xsize=750, ysize=300
!p.multi = [0,2,1,0,0]

ENDIF

IF transf EQ 0 THEN BEGIN
; plotting untransformed data
plot, y, yfit, psym=5, symsize=0.7, xtitle='measured', ytitle='predicted', title='Calibration: measured vs. predicted', subtitle=variables+ ' predictor variable(s)'
oplot, [0,1000], [0,1000], thick=0
plot, vy, predict, psym=5, symsize=0.7, xtitle='measured', ytitle='predicted', title='Validation: measured vs. predicted', subtitle=variables+ ' predictor variable(s)'
oplot, [0,1000], [0,1000], thick=0
ENDIF

IF transf NE 0 THEN BEGIN
IF transf EQ 1 THEN BEGIN
rey = exp(y)
reyfit = exp(yfit)
revy = exp(vy)
repredict = exp(predict)
ENDIF

IF transf EQ 2 THEN BEGIN
rey = y^2
reyfit = yfit^2
revy = vy^2
repredict = predict^2
ENDIF

IF transf EQ 3 THEN BEGIN
rey = sqrt(y)
reyfit = sqrt(yfit)
revy = sqrt(vy)
repredict = sqrt(predict)
ENDIF

IF transf EQ 4 THEN BEGIN
rey = sqrt(2*y+1)
reyfit = sqrt(2*yfit+1)
revy = sqrt(2*vy+1)
repredict = sqrt(2*predict+1)
ENDIF

IF transf EQ 5 THEN BEGIN
rey = (0.5*y+1)^2
reyfit = (0.5*yfit+1)^2
revy = (0.5*vy+1)^2
repredict = (0.5*predict+1)^2

ENDIF
IF transf EQ 6 THEN BEGIN
  rey=(0.3*y+1)^(float(10)/float(3))
  reyfit=(0.3*yfit+1)^(float(10)/float(3))
  revy=(0.3*vy+1)^(float(10)/float(3))
  repredict=(0.3*predict+1)^(float(10)/float(3))
ENDIF

IF transf EQ 7 THEN BEGIN
  rey=(0.1*y+1)^10
  reyfit=(0.1*yfit+1)^10
  revy=(0.1*vy+1)^10
  repredict=(0.1*predict+1)^10
ENDIF

IF transf EQ 8 THEN BEGIN
  rey=(-0.1*y+1)^(-10)
  reyfit=(-0.1*yfit+1)^(-10)
  revy=(-0.1*vy+1)^(-10)
  repredict=(-0.1*predict+1)^(-10)
ENDIF

IF transf EQ 9 THEN BEGIN
  rey=(-0.3*y+1)^(float(-10)/float(3))
  reyfit=(-0.3*yfit+1)^(float(-10)/float(3))
  revy=(-0.3*vy+1)^(float(-10)/float(3))
  repredict=(-0.3*predict+1)^(float(-10)/float(3))
ENDIF

IF transf EQ 10 THEN BEGIN
  rey=(-0.5*y+1)^(-2)
  reyfit=(-0.5*yfit+1)^(-2)
  revy=(-0.5*vy+1)^(-2)
  repredict=(-0.5*predict+1)^(-2)
ENDIF

IF transf EQ 11 THEN BEGIN
  rey=(-2*y+1)^(-0.5)
  reyfit=(-2*yfit+1)^(-0.5)
  revy=(-2*vy+1)^(-0.5)
  repredict=(-2*predict+1)^(-0.5)
ENDIF

;plotting measured against predicted
plot, rey, reyfit, psym=5, symsize=0.7, xtitle=ʻmeasuredʻ, ytitle=ʻpredictedʻ, title=ʻCalibration: measured vs. predictedʻ, subtitle=variables+ predictor variable(s)

plot, [0,1000], [0,1000], thick=0
plot, revy, repredict, psym=5, symsize=0.7, xtitle=ʻmeasuredʻ, ytitle=ʻpredictedʻ, title=ʻValidation: measured vs. predictedʻ, subtitle=variables+ predictor variable(s)

ENDIF

IF transf EQ 0 THEN BEGIN
  validpearson(i)=correlate(predict, vy)
  rmse=0
  FOR j=0, (vsamples-1) DO rmse=rmse+((predict[0,j]-vy(j))^2)
  rmsep(i)=sqrt(rmse/vsamples)
  mapep(i)=100*mean(abs(predict-vy)/vy)
  rmspep(i)=100*sqrt(mean(((predict-vy)^2)/(vy^2))

ENDIF

IF transf NE 0 THEN BEGIN
  validpearson(i)=correlate(repredict, revy)
  rmse=0
  FOR j=0, (vsamples-1) DO rmse=rmse+((repredict[0,j]-revy(j))^2)
  rmsep(i)=sqrt(rmse/vsamples)
  mapep(i)=100*mean(abs(repredict-revy))/revy)
  rmspep(i)=100*sqrt(mean(((repredict-revy)^2)/(revy^2))

ENDIF
ENDIF

IF output EQ 1 THEN device, /close_file

times=times+1

; printing results
FOR n=0,1 DO BEGIN

  IF n EQ 0 THEN lun=-1
  IF n EQ 1 THEN lun=log

  printf, lun, " **********************************'
  printf, lun, '* Model of '+title(nrdep)+'                *'
  printf, lun, '* number of x-variables: '+variables+' *'
  printf, lun, ' **********************************'
  printf, lun, '
  printf, lun, '  Model Calibration'
  printf, lun, '
  printf, lun, '    multiple R square: ',string(mcorrelation^2)
  printf, lun, '    RMSEc:            ',string(see(i))
  printf, lun, '    RMSPEc:       ',string(round(rmspec(i))),' %'
  printf, lun, '    MAPEc:        ',string(round(mapec(i))),' %'
  printf, lun, '
  printf, lun, '  Model Validation'
  printf, lun, '
  printf, lun, '    multiple R square: ',string(validpearson(i)^2)
  printf, lun, '    RMSEp:            ',string(rmsep(i))
  printf, lun, '    RMSPEp:       ',string(round(rmspep(i))),' %'
  printf, lun, '    MAPEp:        ',string(round(mapep(i))),' %'
  printf, lun, '
  printf, lun, '    Variables in equation'
  printf, lun, '    ---------------------'
  printf, lun, '       variable      coefficient   t-value    partial correlation'
  printf, lun, '       ---------     -----------   --------   -------------------'
  printf, lun, '       intercept',string(const),string(t[i,0])
  FOR j=0,i DO printf, lun, '       ',mnamechosen(j),'     ',string(result(j)),string(t[i,(j+1)])
  ',string(pcorr(i,j))
  printf, lun, '       ---------     -----------   --------   -------------------'
  printf, lun, '               (df=',fix(csamples-i-2-1),')'
  printf, lun, '       critical t-value:           1.987'
  printf, lun, '
  printf, lun, '    f-test: overall regression equation'
  printf, lun, '    -------'
  printf, lun, '      f-value:                  ',f(i)
  printf, lun, '       (df=',fix(i+1),',',fix(csamples-i-2),')'
  printf, lun, '      critical value:           ',cf(i)
  printf, lun, '       (alpha=0.01, df=120)'
  printf, lun, '
  IF i GT 0 THEN BEGIN
    printf, lun, '    f-test: comparing with smaller model'
    printf, lun, '    -------'
    printf, lun, '      f-value:                  ',fcm[i,0]
    printf, lun, '       (df=',fix(fcm[i,1]),',',fix(fcm[i,2]),')'
    printf, lun, '      critical value:           ',cf(i-1)
    printf, lun, '       (alpha=0.01, df=120)'
  ENDIF
  printf, lun, '
  printf, lun, '    chi square test: testing residuals on normal distribution'
  printf, lun, '    ----------------
  printf, lun, '     chi square value: ',chisq
  printf, lun, '     critical value:       3.841'
  printf, lun, '      (alpha=0.05, df=1)'
  printf, lun, '
  printf, lun, '
ENDFOR

; excluding independent variables with low t-values
xcontrol, ['Do you want to continue?'], flagname, /window, yesno

IF flagname EQ 0 THEN BEGIN
  free_lun, log
  stop
ENDIF
IF flagname EQ 1 THEN BEGIN

keepcolumn=where(t[i,1:ncorr] EQ 0 or t[i,1:ncorr] GT 1.987 or t[i,1:ncorr] LT -1.987)

IF n_elements(keepcolumn) LT ncorr THEN BEGIN

xcontrol, [' Do you want to remove predictor variables with t-values lower than critical t-value? ']

IF flagname EQ 1 THEN BEGIN

mvarchosen=mvarchosen[keepcolumn]
mnamechosen=mnamechosen[keepcolumn]
i=i-fix((ncorr-n_elements(keepcolumn))x1)
ncorr=n_elements(keepcolumn)
col=mvarchosen(0)
cspec=calibr[col,*]
vspec=valid[col,*]
ENDIF ELSE BEGIN

col=mvarchosen(0)
cspec=calibr[col,*]
vspec=valid[col,*]
ENDIF ELSE BEGIN

; excluding samples with high cook distance

xvalue, [' How many samples (with high cook distance) do you want to remove? '], remsamp, continue,/window,/number, default=0

IF remsamp NE 0 THEN BEGIN

sortcook=reverse(sort(cook[,*]))
delrow=fltarr(remsamp)
FOR k=0, (remsamp-1) DO delrow(k)=sortcook(k)
index=replicate(1L, csamples)
index[delrow]=0L
keeprow=where(index EQ 1)
y=y[*,keeprow]
calibr=calibr[*,keeprow]
cspec=calibr[col,*]
vspec=valid[col,*]
cspec=cspec[*,keeprow]
IF i EQ 0 THEN first=1
i=i-1
FOR n=0,1 DO BEGIN

IF n EQ 0 THEN lun=-1
IF n EQ 1 THEN lun=log
printf, lun, ' recalculate without samples ',fix(delrow+1)
ENDFOR

ENDIF

ENDFOR

END

END

END