Institut für Erd- und Umweltwissenschaften
Fernerkundung

Methodological Developments for Mapping Soil Constituents using Imaging Spectroscopy

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Geowissenschaften"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Anita Bayer

Februar 2013

Universität Potsdam

Methodological Developments for Mapping Soil Constituents using Imaging Spectroscopy

Dissertation
zur Erlangung des akademischen Grades
"doctor rerum naturalium"
(Dr. rer. nat.)
in der Wissenschaftsdisziplin "Geowissenschaften"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Potsdam

von
Anita Bayer

Februar 2013

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2013/6439/
URN urn:nbn:de:kobv:517-opus-64399
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-64399

„It all starts with the soil."

Pieter Kruger, farmer in the adjacent Baviaanskloof valley, South Africa,
who wants to migrate from goat farming to green ecology

Abstract

Recent estimates of global carbon cycle pool sizes indicate that more than 2250 Gt of carbon are permanently stored in natural land carbon reservoirs such as vegetation, soil and detritus (IPCC 2007a). Climatic variations and human activity increasingly introduce perturbations in the natural carbon cycle and cause land cover changes. The annual release of carbon by local changes of land cover is estimated as 1.6 GtC/yr. However, the uncertainty in assessments of global terrestrial carbon pool sizes and their variations is difficult because of the many processes that must be considered and their high spatial and temporal variability. It can be improved by a detailed characterization of processes determining carbon accumulation and losses on a regional scale. Especially semiarid regions are highly affected by climate change as it leads to a significant decrease in available water resources that can destabilize these regions. The resulting land cover changes cause a decrease in the major ecosystem carbon reservoirs vegetation and soil. In soils, organic carbon is one of the main constituents together with sand, clay and iron oxides, with these constituents present globally. Some of these soil constituents can be used as indicator for land cover changes.

Optical remote sensing and in particular Imaging Spectroscopy has shown the potential to assess not only general soil properties over large areas, but also soil organic carbon content. This is accomplished by taking advantage of the characteristic interactions of incident radiation and the physico-chemical properties of a soil, which allows remote sensing-based approaches to study soil properties to be transferable to large geographical regions of similar climatic conditions. The objective of this thesis is to quantify key soil parameters, including soil organic carbon, using field and Imaging Spectroscopy. Organic carbon, iron oxides and clay content are selected to be analyzed to provide indicators for ecosystem function in relation to land degradation, and additionally to facilitate a quantification of carbon inventories in semiarid soils. The semiarid Albany Thicket Biome in the Eastern Cape Province of South Africa is chosen as study site. It provides a regional example for a semiarid ecosystem that currently undergoes land changes due to unadapted management practices and furthermore has to face climate change induced land changes in the future. Decades of overgrazing by goats altered this unique ecosystem from dense shrubland with a high rate of carbon sequestration in vegetation and the peripheral soil to an open savannah-like system.

Approximately 320 km^2 of airborne HyMap hyperspectral image data was acquired over the study area, coupled with extensive field work to collect spectral and chemical ground reference data at 163 reference plots within the image area. Based on these data models are developed to quantify soil organic carbon, iron oxides and clay content. Taking account of the benefits and limitations of existing methods, the approach is based on the direct application of known diagnostic spectral features and their combination with multivariate statistical approaches. It benefits from the collinearity of several diagnostic features and a number of their properties to reduce signal disturbances by influences of other spectral features. The prediction models established with multiple linear regression techniques (MLR) provide a high correlation and are stable especially for soil organic carbon and iron oxides. Clay content models do not reach significant accuracies, neither in feature-based MLR nor in PLS, which is likely the result of its chemical determination based on particle size only. The feature-based regression approach for soil organic carbon and iron oxides provides an alternative to common statistical techniques (partial least squares regression,

PLS). However, it offers advantages in terms of providing a sound physical basis and facilitating explanatory power and transferability of the approach.

In a following step, the hyperspectral imagery is prepared for an analysis of soil constituents. The patchiness of natural vegetation, that is inherent to most semiarid landscapes, causes a large spatial heterogeneity of the study area that is also reflected in the HyMap imagery. To resolve intense spectral mixing within pixel signatures, spectral mixture analysis is performed and used to deconvolve non-homogenous pixels into their constituent components. For soil dominated pixels the subpixel information is used to remove the spectral influence of vegetation and to approximate the pure spectral signature coming from the soil. This step is an integral part when working in natural non-agricultural areas where pure bare soil pixels are rare. It is identified as the largest benefit within the multi-stage methodology, providing the basis for an successful and unbiased prediction of soil constituents from hyperspectral imagery. With the proposed approach it is possible (1) to significantly increase the spatial extent of derived information of soil constituents to areas with about 40 % vegetation coverage, so that it is close to spatially continuous in soil dominated areas, and (2) to reduce the influence of materials such as vegetation on the quantification of soil constituents to a minimum.

Subsequently, soil parameter quantities are predicted by the application of the feature-based regression model to maps of locally approximated soil signatures. Thematic maps showing the spatial distribution of soil organic carbon, iron oxides and clay content in October 2009 are produced for the study area in the Albany Thicket Biome of South Africa. An additional map sums up all uncertainties within the processing to a pixel-wise confidence assessment. The maps are evaluated for their potential to detect erosion affected areas as effects of land changes and to identify degradation hot spots in regard to support local restoration efforts. The consistency of the compiled multi-stage methodology is proven for imagery of different size and characteristics, from local subsets to the imagery of the entire study area. A regional validation carried out using available ground truth sites indicates an existing relationship, even though correlation coefficients are low. This suggests remaining factors disturbing the correlation of spectral characteristics and chemical soil constituents.

The approach is developed for semiarid areas in general and not adapted to specific conditions in the study area. All processing steps of the developed methodology are implemented in software modules, where crucial steps of the workflow are fully automated. The transferability of the methodology is shown for simulated data of the future EnMAP hyperspectral satellite. Soil parameters are successfully predicted from these data despite intense spectral mixing within the lower spatial resolution EnMAP pixels.

This study shows an innovative approach to use Imaging Spectroscopy for mapping of key soil constituents, including soil organic carbon, for large areas in a non-agricultural ecosystem and under consideration of a partially vegetation coverage. It can contribute to a better assessment of soil constituents that describe ecosystem processes relevant to detect and monitor land changes. The maps further provide an assessment of the current carbon inventory in soils, valuable for carbon balances and carbon mitigation products. The methodology provides significant benefits in particular in areas with very inhomogeneous land cover, as found in semiarid environments such as the South African study area. Future application of the approach to hyperspectral EnMAP satellite data could enable repeated monitoring and contribute to an improved characterization of soils in semiarid regions that are highly variable under climate change.

Kurzfassung

Letzten Schätzungen zufolge sind über 2250 Gt Kohlenstoff dauerhaft in terrestrischen Reservoirs wie Vegetation, Böden und biogenen Resten gebunden (IPCC 2007a). Diese terrestrischen Reservoirs unterliegen zunehmend klimatischen und anthropogenen Einflüssen, welche in den natürlichen Kohlenstoffzyklus eingreifen und Veränderungen der Landbedeckung initiieren. Der durch diese Veränderung verursachte Verlust des gebundenen Kohlenstoff wird auf 1,6 GtC/Jahr beziffert. Die Abschätzung globaler terrestrischer Kohlenstoffreservoirs und ihrer Veränderungen ist allerdings aufgrund der vielen zu berücksichtigenden Prozesse und ihrer hohen räumlichen und zeitlichen Variabilität schwierig. Sie kann vor allem durch eine genaue Beschreibung der auf regionaler Ebene wirksamen Prozesse verbessert werden. Insbesondere semiaride Naturräume sind vom Klimawandel betroffen, welcher in erster Linie eine Verknappung vorhandener Wasserreserven und eine damit einhergehende Destabilisierung der betroffenen Region bedeutet. Die daraus resultierenden Veränderungen der Landbedeckung verursachen Verringerungen der Kohlenstoffspeicher in Vegetation und Boden. In Böden ist Kohlenstoff ein Hauptbestandteil neben Sand, Ton und Eisenoxiden, wobei alle diese Bestandteile in den meisten Böden weltweit zu finden sind. Einige von ihnen können als Indikatoren für Landoberflächenveränderungen verwendet werden.

Optische Fernerkundung und im Besonderen die abbildende Spektroskopie sind etablierte Techniken, die, basierend auf der charakteristischen Wechselwirkung der solaren Einstrahlung mit den physikalisch-chemischen Eigenschaften eines Bodens, eine Abschätzung von Bodeneigenschaften wie des Kohlenstoffgehalts erlauben. Fernerkundungsbasierte Ansätze zur Analyse bestimmter Bodeneigenschaften finden deshalb eine breite Anwendung innerhalb großer geographischer Bereiche mit ähnlichen klimatischen Randbedingungen. Das Ziel dieser Arbeit ist die Quantifizierung maßgeblicher Bodeninhaltsstoffe unter Verwendung von Feld- und abbildender Spektroskopie. Dabei stehen organischer Kohlenstoff, Eisenoxide und Ton im Fokus der Betrachtung, da ihre Gehalte im Boden als Indikatoren für Landoberflächenveränderungen verwendet werden können und ihre Analyse gleichzeitig eine direkte Abschätzung des bodengebundenen Kohlenstoffreservoirs ermöglicht. Das semiaride Albany Thicket in Südafrika steht beispielhaft für einen Naturraum, der sich gegenwärtig durch nicht nachhaltige Landnutzung verändert und der voraussichtlich auch in Zukunft durch den Klimawandel bedingten Schwankungen unterliegen wird. Daher wurde es als Arbeitsgebiet ausgewählt. Jahrzehnte lange großflächige Überweidung verwandelte hier die ursprünglich dichte Buschlandschaft in ein offenes, savannenähnliches Ökosystem.

Unter Verwendung der an 163 Punkten im Feld gemessenen Reflektionsspektren und chemisch bestimmter Gehalte der obersten Bodenschicht wird ein Modell zur Bestimmung der drei ausgewählten Bodenparameter erstellt, das anschließend auf der Grundlage großflächig aufgenommener Hyperspektraldaten zur Vorhersage ihrer räumlichen Verteilung verwendet wird. Unter Berücksichtigung der Vor- und Nachteile bestehender Methoden basiert der gewählte Ansatz auf der direkten Verwendung bekannter spektraler Merkmale in Verbindung mit multivariaten Verfahren. Die Kollinearität mehrerer spektraler Merkmale und Eigenschaften hinsichtlich ihrer Form und Ausprägung wird dabei ausgenutzt, um etwaige Einflüsse weiterer spektraler Merkmale und Faktoren zu minimieren. Die mittels multipler linearer

Regressionsanalyse (MLR) erstellten Regressionsmodelle sind vor allem für organischen Kohlenstoff und Eisenoxide stabil und zeigen hohe Vorhersagegenauigkeiten. Modelle zur Vorhersage von Tongehalten erreichen keine Signifikanz, weder in der merkmalsbasierten MLR- noch in der PLS-Regressionsanalyse. Der Grund liegt vermutlich in der ausschließlich auf der Korngröße beruhenden chemischen Bestimmung des Tongehalts. Der merkmalbasierte Regressionsansatz bildet für organischen Kohlenstoff und Eisenoxide eine Alternative zu konventionellen, rein statistischen Verfahren zur Informationsextraktion (Partial Least Squares Regression, PLS). Der Ansatz ist vor allem auf Grund seiner physikalischen Grundlage sowie der Einordnung und Übertragbarkeit der Modelle den rein statistischen Verfahren überlegen.

In einem folgenden Schritt werden die hyperspektralen Bilddaten zur Analyse von Bodeneigenschaften vorbereitet. Kleinräumige Landbedeckungsänderungen, die typisch für semiaride Naturräume sind, bedingen eine nennenswerte Heterogenität innerhalb des ca. 320 km^2 umfassenden Arbeitsgebiets, welche sich in den Hyperspektraldaten widerspiegelt. Spektrale Entmischungsanalyse erlaubt auf einer Subpixel-Basis die Zerlegung nicht homogener Bildspektren in ihre spektralen Bestandteile. Dadurch kann für Pixel, die signifikante Anteile an unbedecktem Boden aufweisen, der Einfluss von Vegetationsbestandteilen auf das Spektrum herausgerechnet und die reine spektrale Signatur des Bodens in Näherung bestimmt werden. Diese Vorgehensweise stellt einen wesentlichen Bestandteil der Analyse heterogener Naturräume abseits landwirtschaftlicher Flächen dar, wo reine Bodensignaturen zumeist nur kleinräumig zu finden sind. Innerhalb der mehrstufigen Methodik trägt sie wesentlich zu einer verbesserten Beschreibung bei, da sie die Basis einer erfolgreichen und akkuraten Vorhersage von Bodeneigenschaften auf der Grundlage hyperspektraler Bilddaten bildet. Der Ansatz erlaubt zum einen eine deutliche Steigerung der Ausdehnung des Gültigkeitsbereichs, für den Bodeneigenschaften vorhergesagt werden können, sodass dieser, wo offener Boden über Vegetation dominiert, einer flächenhaften Abdeckung nahe kommt. Zum anderen wird der Einfluss von Fremdmaterialien wie Vegetation auf die Bestimmung minimiert.

Daran anknüpfend erfolgt die Vorhersage von Bodeninhaltsstoffen durch die Übertragung des merkmalbasierten Regressionsmodells auf die Bilddaten lokal extrahierter Bodensignaturen. Die räumliche Verteilung von organischem Kohlenstoff, Eisenoxiden und Tongehalten wie sie sich im Oktober 2009 im südafrikanischen Albany Thicket darstellte, wird in thematischen Karten erfasst. Eine zusätzliche Karte gibt dabei für jedes Pixel, basierend auf den während der Prozessierung wirksamen Einflussfaktoren, ein Maß für dessen Zuverlässigkeit an. Die Karten werden hinsichtlich ihres Potentials ausgewertet, Bereiche zu erkennen, die in Folge von Landbedeckungsänderungen von Erosion betroffen sind. Dies bildet eine wesentliche Grundlage zur Unterstützung lokaler Wiederaufforstungsbemühungen. Die erstellte Methodik wird mit Hilfe von Bilddaten unterschiedlicher Größe und Charakteristika auf ihre Konsistenz hin geprüft. Eine regionale Validierung erfolgt auf der Grundlage der vorhandenen Referenzmessungen. Sie zeigen einen existierenden Zusammenhang, wenngleich die Korrelationskoeffizienten niedrig ausfallen. Dies deutet auf einen bleibenden Einfluss von Faktoren hin, die sich störend auf den Zusammenhang zwischen spektralen und chemischen Eigenschaften der Böden auswirken.

Während der Entwicklung der mehrstufigen Methodik wird Wert darauf gelegt, dass die gewählten Ansätze zur Anwendung in semiariden Räumen im Allgemeinen geeignet sind und nicht auf spezifischen Gegebenheiten innerhalb des Arbeitsgebiets beruhen. Die einzelnen Prozessierungsschritte sind in Softwaremodulen umgesetzt, in denen die wichtigsten Schritte voll automatisiert sind. Eine Übertragung der Methodik erfolgt auf simulierte EnMAP-Satellitendaten. Trotz des Auftretens intensiver spektraler Mischungen innerhalb der 30 x 30 m großen EnMAP-Pixel lassen sich auch aus diesen Daten Bodeninhaltsstoffe vorhersagen.

Die vorliegende Arbeit zeigt einen innovativen Ansatz zur Verwendung abbildender Spektroskopie zur Kartierung wichtiger Bodeneigenschaften, unter anderem von organischem Kohlenstoff. Dabei erfolgt die Analyse über große Gebiete und in einem nicht landwirtschaftlich geprägten Naturraum, in dem eine teilweise Bedeckung mit Vegetation zu berücksichtigen war. Die Methodik liefert einen Beitrag zur verbesserten Abschätzung ökosystemrelevanter Bodeneigenschaften, die zur Erkennung und Überwachung von Landbedeckungsänderungen verwendet werden kann. Sie bietet vor allem Vorteile in Regionen, die eine inhomogene Landbedeckung aufweisen, wie es für semiaride Räume typisch ist. Die erstellten Kartenprodukte erlauben zudem die mengenmäßige Abschätzung vorhandener Kohlenstoffspeicher im Boden, eine Information, die für Kohlenstoffbilanzierungen und Projekte zur Kohlenstoffspeicherung von Relevanz ist. Die anvisierte Anwendung des Ansatzes auf Hyperspektraldaten des EnMAP-Satelliten würde wiederholte Kartierungen ermöglichen und damit zu einer Verbesserung der großflächigen Kartierung von Bodeninhaltsstoffen in semiariden Räumen in Zeiten des Klimawandels beitragen.

Contents

List of appendices

List of figures

List of tables

Abbreviations

AF	Absorption feature
AIC	Akaike Information Criterion
ANN	Artificial neuronal networks
ASCII	American Standard Code for Information Interchange
ASD	Analytical Spectral Devices Inc.
ATCOR	Atmospheric and Topographic Correction
AVIRIS	Airborne Visible/Infrared Imaging Spectrometer
BRDF	Bidirectional reflectance distribution function
BS	Bare soil
CASI	Compact Airborne Spectrographic Imager
CBD	Citrate-Bicarbonate-Dithionite extraction
CCBA	Climate, Community and Biodiversity Assessment
CDM	Clean development mechanism
CHRIS	Compact High Resolution Imaging Spectrometer
CIR	Colored infrared imagery
C_{org}	Soil organic carbon
CV	Cross-Validation
DEM	Digital Elevation Model
DLR	Deutsches Zentrum für Luft- und Raumfahrt e.V., German Aerospace Center
EeteS	EnMAP end-to-end Simulation Tool
EM	Endmember
EnMAP	Environmental Mapper
ENVI	Environment for Visualizing Images
ERC	Ecological Restoration Capital
ESA	European Space Agency
Fe_d	Iron oxides as determined by dithionite extraction
GCP	Ground Control Point
GFZ	German Research Centre for Geosciences, Potsdam, Germany
HF	Hull feature
HISUI	Hyperspectral Imager Suite
HyMap	Hyperspectral Mapper
HyspIRI	Hyperspectral Infrared Imager
IAS	Iterative adaptive spatial-spectral filtering
IDL	Interactive Data Language
IPCC	Intergovernmental Panel on Climate Change
IS	Imaging Spectroscopy
ISMA	Iterative spectral mixture analysis
MARS	Multivariate adaptive regression splines
µMESMA	Approach for multiple endmember spectral mixture analysis
MIR	Mid infrared spectral range (3000 to 50000 nm)

MLR	Multiple linear regression
MSC	Multiple Scatter Correction
MVA	Multivariate analysis
NASA	National Aeronautics and Space Administration
NDVI	Normalized Differenced Vegetation Index
NIR	Near infrared spectral range (approximately 700 to 1400 nm)
NPV	Non-photosynthetic active vegetation
NSMI	Normalized Soil Moisture Index
OLS	Ordinary least squares regression, see MLR
OSP	Orthogonal subspace projection
PARGE	Parametric Geocoding
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares regression
PPI	Pixel Purity Index
PRESENCE	Participatory Restoration of Ecosystem SErvices and Natural Capital, Eastern Cape
PV	Photosynthetic active vegetation
R^2	Determination coefficient
RA	Regression analysis
RGB	Red-green-blue band combination to producing true color imagery
RMSE	Root mean square error
RPD	Ratio of performance to deviation
(SA	Spectral angle)
SMA	Spectral mixture analysis
SMACC	Sequential Maximum Angle Convex Cone
SMGM	Soil Moisture Gaussian Model
SNR	Signal to noise ratio
SNV	Standard Normal Variate
SR	Soil Residuum
SRTM	Shuttle Radar Topography Mission
SSEE	Spatial-spectral endmember extraction
Stdev	Standard deviation
STRP	Subtropical Thicket Restoration Program
STEP	Subtropical Thicket Ecosystem Planning
SVM	Support Vector Machines Regression
SWIR	Short wave infrared spectral range (approximately 1400 to 3000 nm)
UNFCCC	United Nations Framework Convention on Climate Change
UTM	Universal Transverse Mercator
UV	Ultraviolet wavelength range
VCA	Vertex component analysis
VCS	Voluntary Carbon Standard
VIS	Spectral range of visible light (approximately 400 to 700 nm)

Symbols

A	Area of an absorption feature measured between normalized continuum and spectral curve
AS	Asymmetry of an absorption feature
c	Speed of light in vacuum, $2.99 \cdot 10^8$ m/s
d_{max}	Maximal depth of an absorption calculated relative to the continuum
$d_{\lambda\ lit}$	Absorption at characteristic wavelength according to respective literature
e	Error term
E	Error matrix
f_{EM}	Cover fraction of an endmember
h	Planck constant, $6.63 \cdot 10^{-34}$ J/s
j	Spectral variable
k	Boltzmann constant, $1.38 \cdot 10^{-23}$ J/K
k	Spectral variable
n	Number of samples
m	Number of variables
$r_{\rho c}$	Mean reflectance of the convex hull of a spectrum in a defined wavelength range
$S_{left/right}$	Left and right shoulders of an absorption feature
sl_ρ	mean slope of the spectral curve in a defined interval
$sl_{\rho c}$	Mean slope of the convex hull of a spectrum in a defined wavelength range
T	Temperature
T	Threshold
w	Width of an absorption feature
x	Independent variables in regression analysis, predictor variables
X	Matrix of x variables
y	Dependent variables in regression analysis, response variables
Y	Matrix of y variables

Greek symbols

α_λ	Absorbance
ε	Emissivity
λ	Wavelength
λ_{dmax}	Wavelength of maximal absorption
λ_{lit}	Characteristic wavelength of an absorption feature according to respective literature
ρ_{EM}	Endmember reflectance spectrum
ρ_λ	Reflectance
ρ_S	Pixel reflectance spectrum
ρ_{SR}	Reflectance of the soil residuum
σ	Standard deviation
σ_n	Normalized standard deviation
σ_c	Stefan-Boltzmann constant, $5.67 \cdot 10^{-8}$ W/m^{2}K^4
τ_λ	Transmission

Indices

BS	Bare soil
Cal	Calibration
j	Spectral variable
k	Spectral variable
Max	Maximal
Mean	Average
Min	Minimal
NPV	Non-photosynthetic active vegetation
PV	Photosynthetic active vegetation
SR	Soil residuum
Val	Validation

1 Introduction and objectives

1.1 From soil status to its value for the characterization of land surface processes

"Soil" is the biologically active uppermost layer of the Earth's surface. Its material is formed by the weathering of rocks conditioned by the environmental factors climate, vegetation, organisms, topography and parent material (Schachtschabel et al. 2002). Developed soils consist of a mineral structure with organic compounds arranged in distinct layers. They are the great store- and warehouse for essential nutrients, water and air, which are necessary for the development of vegetation (Greenland 2006). A soil's chemical attributes are mainly defined by its constituent components such as organic matter and mineral particles. In addition, physical soil attributes are for instance soil texture, pore volume or soil color. Together, these characteristics indicate the soil's properties, functions and finally its fertility for plant growth (Schachtschabel et al. 2002). The physical and chemical properties determine various feedback mechanisms between each other. For instance, soil organic carbon content initially plays a major role for nutrient supply and beside that also affects the physical characteristics of a soil regarding soil structure and aggregate stability and water retention (Huntington 2006). Soils are the most basic of all natural resources, and with their functions they support life in numerous ways (Lal 2006).

Land surface processes and the role of soil organic carbon

In times that are facing challenges like a rapidly growing world population, food insecurity and climate change, the importance of soils rises above its meaning as medium for plant growth, engineering foundation or industrial raw material (Lal 2006). To meet the challenges, soils now are supposed to maximize the long-term productivity, act as a buffer for environmental changes and as a sink for atmospheric carbon and other greenhouse gases (Lal 2006). However, soil health may be dramatically affected by land degradation resulting of human interventions or climate change (Greenland 2006). The process of land degradation reduces biodiversity and productivity and causes increased soil erosion, loss of fertile soils and the spread of deserts. About 70 % of all dry areas of the world are affected by land degradation (Eswaran et al. 2001). Especially in semiarid regions, large-area land cover changes coupled with degradation and soil erosion are frequent phenomena which are often the result of long periods of inappropriate management practices. The assessment and monitoring of an ecosystem's function and ongoing processes related to land degradation and erosion is a vital task. Therefore, the characterization of an ecosystem's soil condition and its spatial and temporal changes are important indicators and chemical and physical soil attributes can serve as tracers to monitor such phenomena.

Organic carbon (C_{org}) is probably the most important parameter of soils and is frequently addressed by researchers to characterize and monitor land changes (e.g. Gomez et al. 2008b, Hill and Schütt 2000, Selige et al. 2006, Stevens et al. 2008 and 2010). This results from the strong impact of soil degradation and

erosion on carbon concentrations on the one hand and from the multiple positive feedback mechanisms provided by soil organic carbon on the other hand. While the presence of soil organic carbon is even capable to mitigate droughts (Tiessen at al. 1994), reduced C_{org} contents make soils prone to accelerated erosion. Thus, the spatial variability of soil organic carbon can be an indicator for the detection of areas of a stable soil development in comparison to areas of active erosion (Birkeland 1990) as the first step in monitoring land changes.

Soils in the global carbon cycle

In terrestrial ecosystems, carbon is retained in live biomass, decomposing organic matter (detritus) and soils. It is exchanged between these systems and the atmosphere as carbon dioxide through photosynthesis and soil respiration and as organic carbon through decomposition and combustion (IPCC 2000). Globally, approximately 80 % of the carbon permanently stored in terrestrial carbon pools is contained in soils up to a depth of 1 m (IPCC 2000). This even reaches up to 95 % in semiarid environments (IPCC 2000). Therefore, soils are the major terrestrial carbon stock, and large fluctuations in their carbon inventories affect global budgets. For instance, the loss of terrestrial carbon stocks due to land use changes is estimated with 1.6 Gigatons of carbon per year (GtC/yr, figure 1-1). Though this estimation is listed as one key uncertainty in the description of global carbon pools, since it strongly depends on processes operative on local scales (IPCC 2007a).

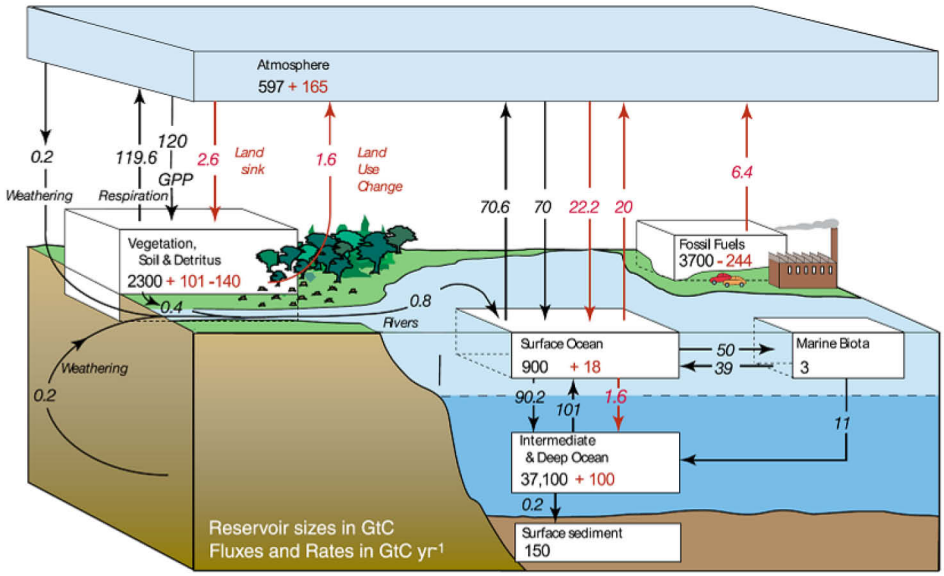


Figure 1-1: The global carbon cycle for the 1990s shows the main carbon reservoirs in GtC and annual carbon fluxes in GtC/yr (IPCC 2007a, changes in pool sizes are adapted from Sabine et al. 2004).

One of the main options for carbon mitigation identified by the IPCC is the sequestration of organic carbon in soils (Metz et al. 2007). Estimates of the global potential of C_{org} sequestration through the restoration of degraded lands, adoption of conservation tillage and other systems of sustainable management are around 0.9 ± 0.3 GtC per year (Lal 2004). Besides the effect of reducing the atmospheric greenhouse gas concentrations, the storage of carbon in soils provides vital environmental benefits for the soil itself and the ecosystem as a whole, as it is shown above. An assessment of carbon inventories in soils is required for carbon balances and carbon mitigation projects. Bellon-Maurell and McBratney (2011) describe the requirements for C_{org} assessments for carbon sequestration projects as two-fold, since it has to be both accurate and low cost.

Monitoring of land changes and carbon inventories by remote sensing

The variation of physical and chemical soil parameters in space and time is decisive when characterizing land surface processes and soil carbon inventories, but analysis costs are often a limiting factor to address soil variability especially in large-area applications. Optical remote sensing and in particular Imaging Spectroscopy (IS) has the potential to supply the spatial information needed for a regional-scale analysis of land surface processes and carbon inventories in soils. It takes advantage of the characteristic interaction of matter and radiation that result of physicochemical properties of a soil. The radiation interacts with the upper 50 µm of the soil surface (e.g. Ben-Dor et al. 1999), hence reflecting the characteristics of the uppermost topsoil. Linking this with the fact that soil organic carbon and other major soil constituents are constantly found in almost every soil, remote sensing based models for soil analysis are applicable for large ecosystems. To this effect, the IPCC (2007b) identified "improved remote sensing technologies for analysis of soil carbon sequestration potential and land-use change" as one of the key sectoral mitigation technologies for the coming decades. In this regard, Ben-Dor et al. (2008 and 2009) give an overview of state-of-the-art case studies applying Imaging Spectroscopy for soil studies. They found that soil parameters were determined in regard to soil degradation and erosion, soil type mapping, soil water content and soil contamination. They note, however, that the widespread application of IS for soil analysis is not achieved despite the proven potential of the technique so far.

1.2 Land degradation in the semiarid Albany Thicket Biome, South Africa

The semiarid Albany Thicket Biome in the Eastern Cape Province of South Africa is chosen as study site (figure 1-2). It provides an example for a semiarid ecosystem that currently undergoes land changes due to unadapted management practices and according to climate change scenarios modeled for this biome further has to face high dynamic land changes (Hoare et al. 2006). Due to the unique characteristics of the local vegetation type, the region has been identified as highly important for carbon sequestration projects, in particular for future carbon sequestration under the Clean Development Mechanism (CDM) as part of the Kyoto Protocol. These aspects highlight the relevance of the selected region for multiple processes and applications. Within the Albany Thicket Biome an area of approximately 320 km^2 is selected, which is representative on regional scale and serves as an example for larger scales.

"Thicket" is a unique vegetation type located in the semiarid valleys that cover about 17 % of the Eastern Cape Province (Lechmere-Oertel et al. 2005). It is unusual for a semiarid environment since an almost complete cover of dense vegetation is present in pristine conditions. A high component of the succulent shrub *Portulacaria afra* accounts for large proportions of carbon sequestered in biomass and peripheral soils. In the Albany Thicket Biome, land degradation processes are observed and monitored and today more than 70 % of the vegetation units are considered as "moderately" to "severely" degraded (Lloyd et al. 2002). Intensive goat farming since the early 1900s resulted in the transformation of wide parts of dense Thicket shrubland to an open savannah-like system dominated by a mixture of open soils and dry vegetation. This process is accompanied by a severe loss of biodiversity and ecosystem carbon stocks in vegetation and soils (Lechmere-Oertel et al. 2005). High contrast land cover changes are found at fence lines between pasture and game farms, showing pristine and transformed vegetation in direct neighborhood. An example is

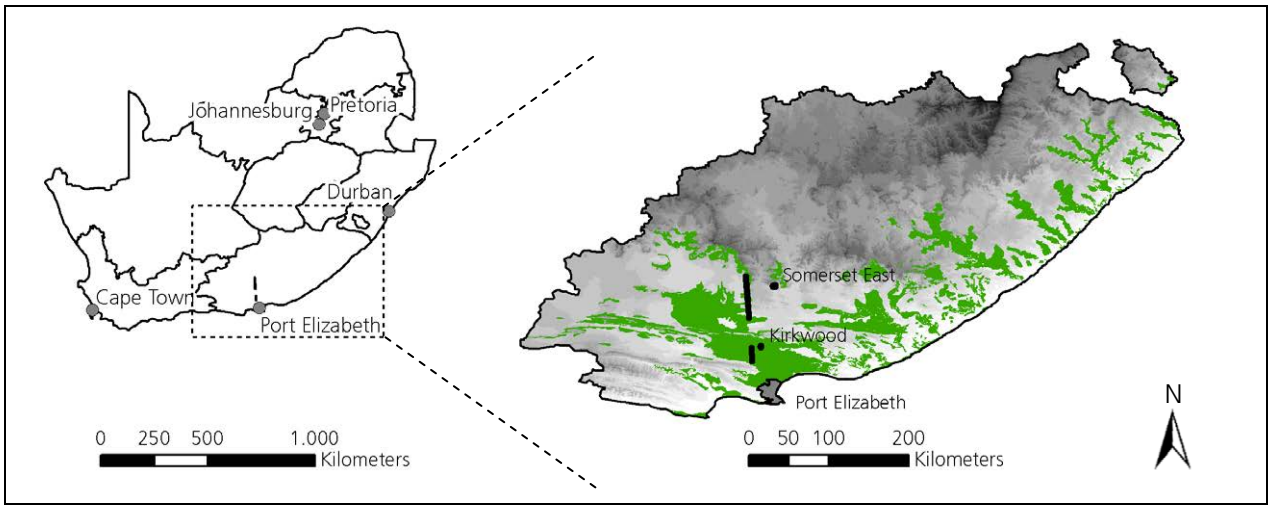


Figure 1-2: Location of the approximately 320 km² wide study area within South Africa and its Eastern Cape Province. The green areas show the distribution of Albany Thicket vegetation within the Eastern Cape (Geodata provided by the PRESENCE Network 2009).

shown in figure 1-3. The value of an intact ecosystem to provide enhanced ecosystem services was realized in the past years, and projects, such as the Subtropical Thicket Restoration Project (STRP), were initiated by governmental departments providing the technological basis for a restoration of degraded landscape. The restoration of local farm portions today is realized by trading gathered carbon credits in the free market. No mechanism is found that would provide a source of funding enabling a large-area restoration of Thicket vegetation so far.

Figure 1-3: A fence line marks the borderline between nearly pristine vegetation on a game farm and highly degraded vegetation on a stock farm in the Albany Thicket Biome, South Africa (Picture taken in October 2009).

1.3 Study objectives

In the presented context, this thesis aims at quantifying topsoil contents of key soil parameters including soil organic carbon on the basis of airborne hyperspectral remote sensing data of the Albany Thicket Biome, South Africa. Thereby indicators for ecosystem degradation and a quantification of carbon inventories in soils are provided. The mapping of soil constituents within the selected ecosystem of 320 km^2, where land cover, surface and soil parameters are subject to change, makes great demands on the followed approach. Considering the still limited availability of mid-infrared spectroscopy, the study is dedicated to the use of optical hyperspectral data covering the visible to shortwave infrared spectral range.

Selection of investigated key soil constituents

Soil organic carbon, iron oxides and clay content emerge as parameters that are both of significance in the semiarid study area and spectrally active in the selected wavelength range. The diagnostic spectral features that the three considered parameters have in the VIS to SWIR spectral domain (ca. 400 to 2500 nm) can be attributed to electronic transition and charge transfer processes of specific chemical groups (e.g. Clark et al. 1990, Hunt 1977, Stoner and Baumgardner 1981). This enables a physically-based modeling of these parameters. The role of soil organic carbon for soil stability, function and as a direct indicator for land degradation is evaluated in the previous sections. Soil iron oxides immediately reflect soil formation and furthermore play a major role in a soil's geochemical cycle, by having sorptive properties, acting as buffer in oxidation-reduction reactions and stabilizing soil structure (Qafoku and Amonette 2006). Iron oxides are directly linked to soil erosion because they are enriched whenever other soil components such as organic materials are eroded. A soil's clay mineral content affects nutrient retrieval as a result of a large and charged surface of the clay minerals, the ability of a soil to absorb water and also the binding of humic substances (Greenland 2006). Soil organic carbon is the most important among the three soil parameters considered in this study because of its outstanding role for land surface processes in particular related to degradation. Therefore, soil organic carbon is in the main focus of the analysis and treated before iron oxides and clay content.

Limitations of existing approaches

The feasibility to determine topsoil contents of the three selected parameters based on diagnostic characteristics reflected in a soil's spectrum was described in previous studies (e.g. Ben-Dor and Banin 1994, 1995a and b, Chang et al. 2001, Demattê and Garcia 1999, Viscarra Rossel et al. 2006). The Beer's Law provides the physical basis for such analysis (e.g. Clark 1999), proposing a correlation between the content of a certain soil constituent and the spectral features diagnostic for it.

The most frequently used technique to quantify soil properties from reflectance spectra is partial least squares regression (PLS, e.g. Gomez et al. 2008a and 2008b, Selige et al. 2006, Stevens et al. 2008 and 2010, Viscarra Rossel et al. 2006). It decomposes the full spectral information based on statistical factors before they are correlated with the soil attributes. However, the technique faces drawbacks in terms of significance and transferability resulting from the perfect statistical adaptation of a PLS model to its base dataset. A specific soil constituent can be modeled properly, but the correlation may be based on a second variable indirectly linked to the investigated parameter instead of physical properties or even result from random noise. The transferability of PLS models is impaired first when the models are transferred to datasets showing different sensor-specific characteristics (band positions, signal-to noise ratio, etc.) and second when

they are transferred to datasets from other locations showing different environmental conditions (surface characteristics, soil texture, geology, etc.). In consequence an established calibration model that works well for one sample population or study site might be unacceptable for another and a model re-calibration may be necessary (Chang et al. 2001, Stevens et al. 2008). This limits the application of statistically based models such as PLS derived from local observations to a regional scale as is aspired in this study.

Another approach is the direct application of spectral absorption features or band indices. Because the spectral features provide a direct physical basis, this is often seen as more robust, hence allowing for a better transferability to regions of similar environmental conditions (e.g. Mulder et al. 2011). Although broad regions of a soil spectrum measured in the VIS to SWIR range are non-specific as a result of overlapping absorptions and additional influences of non-significant factors (e.g. surface condition, sensor noise), there remain diagnostic spectral features which can be used for a quantification of soil constituents. Various characteristics of spectral absorption features have been directly used for a quantification of soil properties (e.g. Bartholomeus et al. 2008, Gomez et al. 2008a, Hill and Schütt 2000, Lagacherie et al. 2008, Richter et al. 2009). However, regression models built on one or a few spectral features are often not sufficiently robust for a practical application to a wide variety of soils (Viscarra Rossel et al. 2006).

The patchiness of natural vegetation that is inherent to most semiarid landscapes causes a large spatial heterogeneity of the study area, which is reflected in hyperspectral observations in terms of mostly non-pure spectral signatures in image pixels. This limits the usage of Imaging Spectroscopy for soil applications, because the relationships between soil constituents and reflectance are disturbed by the presence of other materials, mostly vegetation (Mulder et al. 2011). The traditional approach is to limit the image information to soil dominated pixels that comprise only a small amount of vegetation before the analysis (Ben-Dor et al. 2002). Soil constituents for the excluded areas are interpolated afterwards from the non-vegetated areas. Within the South African study area, the intense spectral mixing of surface materials mostly hamper the occurrence of pure bare soil pixels. Thus, a limitation of the analysis to areas with limited vegetation coverage (e.g. 10 %) would result in a very small spatial extent of useable pixels. On the other hand, an increase of the vegetation fraction included in the analyzed pixels significantly biases soil constituent predictions, since the vegetation introduces spectral features over the entire VIS to SWIR range. Appropriate approaches applying spectral unmixing techniques to remove spectral influence of vegetation were described in particular for agricultural areas (e.g. Bartholomeus et al. 2011), where spectral mixtures comprised of two endmembers could be well described and modeled. Within the non-agricultural South African study area, surface materials vary spatially resulting in highly complex mixtures.

Study objectives and methodical approach

Based on the identified need for further research and considering future developments in this field, five research aspects arise that are investigated in the present study.

The fundamental objective is the development of a comprehensive model using VIS to SWIR spectroscopy that allows for a robust quantification of topsoil contents of the key soil constituents organic carbon, iron oxides and clay. To achieve this, a regression model is developed based on spectral and chemical field ground reference measurements. It is based on the combined application of several diagnostic spectral features and their unique properties (e.g. absorption band depth, width, slope in a given wavelength range). This approach thereby benefits from the collinearity of diagnostic spectral features to overcome the fact that the significance of each feature might be weakened by overlapping absorptions of other properties and non-significant

factors. The developed soil prediction models with the direct application of spectral features provide a sound physical basis, they are considered as simplistic but nonetheless robust.

In a second stage, the subpixel information of hyperspectral imagery is used for an approximation of the spectral signal coming from the soil for a natural non-agricultural area with small-scale variances in land cover, such as in the Thicket Biome. A survey of the HyMap sensor from 2009 (spectral range 440 to 2500 nm, ground resolution of 3.3 m) covering 320 km^2 in the Albany Thicket Biome provides the basis for this analysis. The image pixels show intense spectral mixtures resulting from the heterogeneous land cover. Non-homogenous pixel spectra are decomposed into their pure surface materials in spectral mixture analysis. The soil signature comprised in the spectra of mixed pixels is approximated for soil dominated pixels, and the spectral influence of vegetation is removed. The application of the methodology to image data of a large and non-agricultural site with non-homogenous land coverage and with a ground resolution of several meters is described.

The first two developments are combined into a regional mapping of key soil constituents for the Albany Thicket Biome, South Africa. Therefore, the developed regression models are applied to the image data of derived soil signatures. The confidence of the predictions is assessed in a pixel-wise index. The accuracy of predicted contents is evaluated and the spatial coverage of the predicted soil constituents is compared to traditional approaches.

The contribution of spatial maps of selected key soil constituents to detect areas susceptible to land changes and to support restoration efforts in the Albany Thicket Biome, South Africa, is evaluated. It is tested if the produced maps of soil constituents can be used to differentiate areas with stable progressive soil development on the one hand and erosion affected areas on the other hand. This is related to the identification of degradation hot spots, which should be selected first for restoration. It is discussed if the derived information can provide an assessment of the current inventory and status of soil carbon stocks, valuable for carbon balances and carbon mitigation projects.

The transfer of the methodology to future satellite hyperspectral systems is tested. Future hyperspectral satellite sensor systems, for instance the EnMAP (Stuffler et al. 2007), HISUI (Ohgi et al. 2010) and HyspIRI (e.g. Hook et al. 2012) systems, provide the technical infrastructure for large-area and multi-temporal surveys. In this context the application of the developed methodology to simulated data with the specifications of the EnMAP hyperspectral satellite is tested.

With these aspects, the main focus of this study lies in the new development of a robust methodology to predict soil constituents, the extension of existing approaches towards an extraction of reliable soil signatures from IS to natural areas that show intense small-scale variances in land cover and finally to identify the benefit of soil constituent maps for application in the Albany Thicket Biome.

Structure of the study

The first section describes the background of the study in terms of the principles of spectroscopy, the state of the art in soil spectroscopy and spectral properties of soil organic carbon, iron oxides and clay and other materials relevant in this context (section 2). A special focus is set to studies applying spectral feature-based methods for soil quantification. Factors affecting the analysis of soil spectra are discussed. Thereafter, the study area is introduced with its geomorphic properties and the current development of the ecosystem

(section 3.1). Reference measurements in field and laboratory, the airborne HyMap and simulated EnMAP satellite data as datasets available for this study are described, their pre-processing is presented and their quality is assessed in sections 3.2 and 3.3. The methodical part begins with an evaluation of multivariate methods to study soil properties and the discussion of their suitability for the current application (section 4). An adequate methodology is selected that is expected to provide an accurate and robust quantification of the considered soil constituents. The implementation of its multiple stages is explained in detail in section 5, beginning from the development of a feature-based regression model, the approximation of subpixel soil pixel signatures based on the hyperspectral imagery and finally the prediction of soil organic carbon, iron oxides and clay content for the South African study area. Results of this analysis are presented for every stage in section 6, ending with the assessment of the accuracy and placement of the approach in regard to previous investigations. The influence of the size and characteristics of the imagery, the proposed workflow is applied to, is estimated. The results are evaluated in regard to the assessment of land degradation and carbon related land surface processes in the Thicket Biome, South Africa. The study is summarized in section 7, where main findings are discussed and an outlook of further suggested developments and of possible applications of the methodology are given. The following appendix includes supplementary details of data pre-processing, the technical realization of each step within the workflow and finally of the processing results including further statistics.

2 Optical remote sensing for soil quantification

Remote sensing has proven to be a time- and cost-efficient method for large-area mapping of the environment. It allows a synergistic view combining sectorial monitoring approaches. The soil layer develops as transition zone between atmosphere, biosphere, hydrosphere and solid lithosphere and provides the basis for plant growth. It is the most important layer for energy and nutrition flows and thereby of key importance for landscape analysis as is shown in section 1.1. A variation of soil physical and chemical properties in time and space results of the many factors influencing soil formation (see section 1.1). Recent developments of multi- and hyperspectral airborne and satellite based sensor systems make available a range of data that can be used for detailed continuous analysis of the earth's surface. Spectroscopy as a section of remote sensing can be defined as "the study of light as a function of wavelength that has been emitted, reflected or scattered from a solid, liquid or gas" (Clark 1999). It can be used to identify and to characterize material in its quality or quantity, or to investigate physical processes from within the atmosphere down to solid earth and its various environments.

The following section provides an introduction to physical fundamentals and radiation mechanisms. The natural sources of radiation, basic mechanisms of interaction between radiation and matter as well as the resulting spectral characteristics of natural surfaces important for remote sensing in general and spectroscopy in particular are given below. Reflectance properties of soil constituents and adjacent materials like vegetation are discussed and evaluated.

Fundamental radiation mechanisms

Solar radiation is the source for optical remote sensing and spectroscopy. The total energy the sun emits resulting of its temperature of around 5800 K can be approximated by Stefan-Boltzmann's law. It gives the total energy M released by a black body as the ideal source of radiation (emissivity ε of 1) at a temperature T. Black body radiation M_λ at wavelength λ and temperature T is given by Planck's law.

Stefan-Boltzmann law: $$M = \varepsilon \sigma_c T^4 \quad [W] \qquad (2\text{-}1)$$

Planck's law: $$M_\lambda = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1} \quad [W/m^2\mu m] \qquad (2\text{-}2)$$

where the Boltzmann constant σ_c is $5.67 \cdot 10^{-8}$ W/m^2K^4, the Planck constant h is $6.63 \cdot 10^{-34}$ J/s, the Boltzmann constant k is $1.38 \cdot 10^{-23}$ J/K, and the speed of light in vacuum c is approximated with $2.99 \cdot 10^8$ m/s. When passing the atmosphere, solar radiation is attenuated by wavelength-dependent scattering and absorption by molecules, mainly water vapor, oxygen, ozone, carbon dioxide and aerosols (see figure 2-1). Solar radiation at sea level is only a fractional amount of top of atmosphere radiation depending on its wavelength.

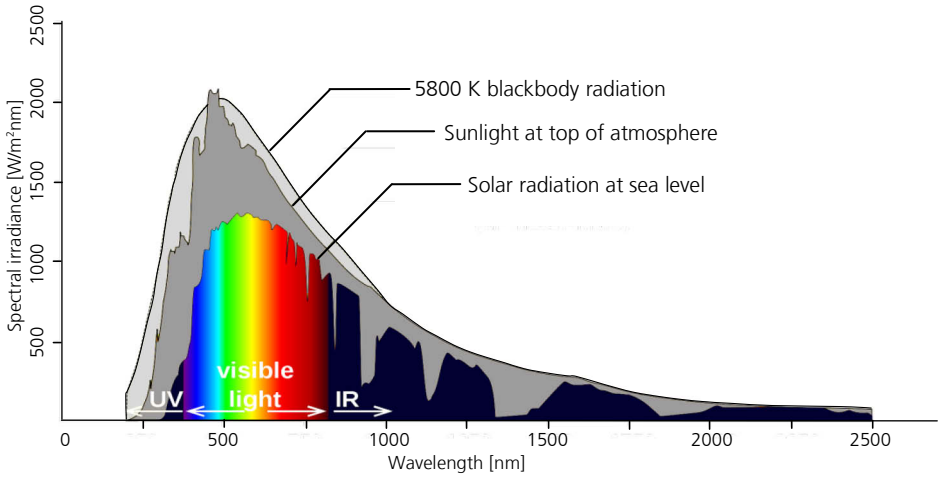


Figure 2-1: Solar irradiance at top of atmosphere and at sea level compared to blackbody radiation at 5800 K illustrating atmospheric attenuation (source: Wikipedia 2011 modified after Kronberg 1985).

In the same way, the solar radiation reflected from the earth's surface and detected by an optical sensor system is disturbed and partly extinguished by atmospheric compounds. During atmospheric correction of airborne or satellite remote sensing data, at-sensor radiance is converted into top-of-canopy reflectance and these disturbances are considered to some extent (see section 3.2.2). Nevertheless, certain wavelength ranges remain where radiometric signals are highly disturbed and with this not usable for earth observation. Atmospheric water vapor for example causes disturbances in wavelength ranges around 1400 and 1900 nm (see sections 3.2.1 and 3.2.2). The signal-to-noise ratio and thus the quality of remote sensing data is best where the earth's atmospheric transmittance and incident radiation are high, like in the visible range. The low level of incident radiation in the SWIR causes long sensor integration times and thereby a decrease in the signal-to-noise ratio of optical sensors (compare figure 2-1).

Interaction of radiation and surfaces

When sunlight hits the earth's surface, three processes describe the general interaction of electromagnetic radiation and matter. A material can take up a certain amount of energy (absorption a), change the direction of the incident radiation without a transition of energy (reflection r), or the radiation can completely or in parts pass the material (transmission t) also without changes in the energy of the radiation. The energy budget of the incident radiation (E_λ) is divided in these three quantities.

$$E_\lambda = E_{a,\lambda} + E_{r,\lambda} + E_{t,\lambda} \qquad [W/m^2] \qquad (2\text{-}3)$$

By division of each term by E_λ, dimensionless coefficients can be derived which are material properties and independent of incident radiant energy. They can be given as percent or range between 0 and 1.

$$1 = \alpha_\lambda + \rho_\lambda + \tau_\lambda \qquad [-] \qquad (2\text{-}4)$$

with absorbance α_λ, reflectance ρ_λ and transmission τ_λ. Their sum equals 1, but their magnitude is not constant and varies with wavelength according to the physical properties of the material. A spectrum displays the quantities of reflection, absorption and transmission as function of the wavelength. In this study, the term spectrum refers to a reflectance spectrum.

Photons entering an absorbing medium are absorbed according to Beer's law (e.g. Clark 1999).

$$I = I_0 e^{-c\varepsilon_\lambda d} \qquad [-] \qquad (2\text{-}5)$$

where I_0 is the intensity of incident light before passing the sample, I the transmitted light's intensity, c is the concentration of the absorbing molecule [mol/l], ε_λ is a material specific spectral extinction coefficient [l/mol cm] and d is the path length the photon travels through the medium. The absorption of radiation increases the energy budget of the absorber, resulting in an increase of its temperature. In its original meaning, Beer's Law is valid for transparent substances like fluids and gases. By approximation it is also applicable to opaque materials and the interaction of radiation with i.e. soil surfaces (Ben-Dor et al. 1999). This makes it a fundamental principle allowing the quantitative determination of bio- and geochemical parameters by the use of reflectance spectroscopy.

The way radiation is reflected from a surface depends on the surface's characteristics. Flat surfaces that act like a mirror reflect light with the same angle, at which the radiation is irradiated (specular reflection, figure 2-2a). In contrast, an ideal Lambertian surface reflects incident radiation uniformly diffuse in each direction (figure 2-2b). Most natural surfaces behave in a way in between these two extremes and reflect more specular or diffuse depending on surface roughness, the wavelength of incident radiation and incident angle (figure 2-2c and d). A homogenous surface like a smooth soil surface with small unevenness tends to diffuse reflection while the orientation of particles at a surface, e.g. when a physical crust is present, can lead to an increase of the portion of specular reflected light. The bidirectional reflectance distribution function (BRDF) describes the reflection of a surface for every viewing angle.

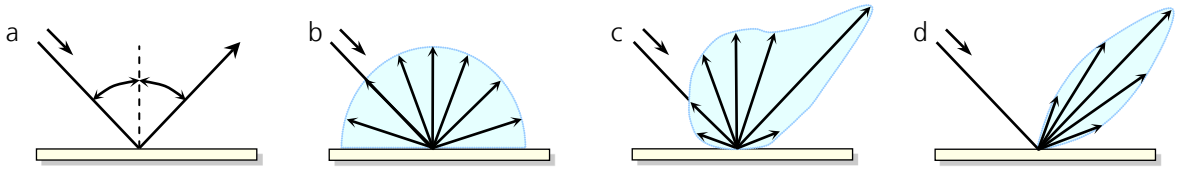


Figure 2-2: Types of surface reflection – specular (a) and diffuse reflection (b), mixed reflectance dominated by diffuse (c) and specular reflectance (d).

If light hits a non-homogenous surface that shows different surface materials, it is reflected, absorbed and transmitted with a mixture of the properties of all materials present. This mixture can only be described as linear when there is no multiple scattering between different components. To allow for a mathematical description of the light reflected from and absorbed by surfaces built up of several materials showing different physico-chemical properties, however, linear mixture is assumed also in this study (compare section 5.2.1). The combined signal is composed of the fractional area of the component times the spectrum of each component (see Clark 1999).

Optical remote sensing investigates the reflected light in the wavelength ranges of the visible to shortwave infrared light. Materials composing land surfaces are predominantly opaque and transmission is low. Thus, the effect of wavelength-dependent absorption processes allows to derive information about the chemistry of a material by measuring its reflectance spectrum. In the following, the absorption processes relevant to characterize soil inventories are described in detail.

2.1 Principles of spectroscopy – mechanisms of absorptions

The great potential of reflectance spectroscopy, including material identification and also the derivation of quantitative information, is the result of two major types of absorption processes which produce spectral features in the VIS to SWIR spectral range. They can be distinguished by their location in the spectrum and based on parameters describing their shape. The following overview is based on the works of Clark (1999), Hunt and Salisbury (1970), and Hunt (1977). These studies describe in detail the principles of interaction between radiation and matter used in spectroscopy of minerals and natural compounds.

Electronic processes

Absorption of radiation energy can cause an atom or ion electron to change from one discrete energy level to another higher one (electronic transitions), which results in the absorption of the specific quantum of energy the electron needs for the jump, and thus radiation at a specific wavelength. The electrons change back to the original or another lower energy level because the higher energy level is not stable. Thereby, a photon is emitted, usually happening at longer wavelengths in the mid-infrared range (emission or thermal spectroscopy). Absorption bands due to electronic transitions usually produce broad bands that occur mostly in the UV range but extend into the visible. There are several other electronic mechanisms that cause absorption, like conduction bands, but they are not prevailing when focusing on soil reflectance properties.

Crystal field electronic transitions occur when the stimulated electron is located in a crystal field, most common in transition elements (e.g. Ni, Cr, Fe). Absorbed energy levels depend on the valence state of the ion (e.g. Fe^{2+} and Fe^{3+}) and the crystal structure. For instance Fe^{2+} produces a crystal field absorption around 1000 nm in many minerals like olivines, pyroxenes etc.. The states of prevailing iron transition absorptions are given in detail in Hunt and Salisbury (1970).

Absorptions are referred to as charge transfer bands when the absorbed energy causes an electron to migrate between ions or between ions and ligands. Charge transfer often happens between adjacent ions of the same metal in different valence states. Since high energies are required to stimulate such transfer, the absorptions are usually hundred to thousand times more intense than those of crystal field effects and the spectral features are mostly located in the ultraviolet wavelength region. An intense charge transfer band of iron is located in the UV range, of which VIS-SWIR spectra only picture the right shoulder which causes the red color of iron oxides.

Vibrational processes

The other process leading to significant absorption of radiation in the VIS to SWIR region results from the vibration of molecules. The frequency of a molecule's vibration depends on the number and mass of the atoms present in it, their geometrical arrangement and the strength of the bonds between the atoms inside the molecule and other outside located atoms. Every single atom has several vibration modes resulting in even more modes of possible vibrations for the entire molecule. They are called fundamentals. Most materials produce fundamentals at wavelength longer than 2500 nm in the mid- and far-infrared range. When energy of two or more quanta of the fundamental mode is brought into the system, the vibration can jump to multiples of the original fundamental frequency and overtone absorptions occur at or close to

half/third/etc. of the wavelength of the specific fundamental feature. Combination tones result from the interaction of two fundamentals or overtone vibrations. Fundamental bands are usually strongest, while features get weaker with higher overtones and combinations. VIS to SWIR reflectance spectroscopy includes exclusively second and third overtones and combinations of soil minerals and materials, predominantly triggered by molecules of H_2O and OH like in clays, CO_3 like in carbonates and organics like CH, NH, etc.. Compared to electronic transition bands, vibrational processes produce very sharp bands. They are strong in the mid to short wave infrared and decline in intensity and frequency of occurrence towards the NIR to VIS range.

Analysis of absorptions from reflectance spectra

For analysis of individual absorption features, the continuum removal technique (or hull normalization) is often applied in spectroscopy. A soil reflectance spectrum measured by an appropriate sensor generally can be split up into a continuum and individual absorption features. The continuum can be seen as background or overall reflectance onto which other spectral features resulting from electronic or vibrational processes are superimposed. Continuum removal isolates spectral features from the trend of background reflection and thus allows for an analysis and intercomparison of spectral absorptions (see figure 2-3b). The normalized reflectance ($\rho_{CR\lambda}$) is calculated for each wavelength from the original reflectance (ρ_λ) divided by the reflectance of the continuum ($\rho_{C\lambda}$).

$$\rho_{CR\lambda} = \frac{\rho_\lambda}{\rho_{C\lambda}} \qquad \text{(Clark 1999)} \qquad [-] \qquad (2\text{-}6)$$

A mathematical definition of the continuum itself can be found in Clark and Roush (1984). The depth of the absorption d_{max} is then defined relative to the continuum (van der Meer and de Jong 2001):

$$d_{max} = 1 - \frac{\rho_{\lambda_{max}}}{\rho_{C\lambda_{max}}} \qquad \text{(Clark 1999)} \qquad [-] \qquad (2\text{-}7)$$

with $\rho_{\lambda max}$ being the reflectance of the wavelength of maximum absorption and $\rho_{C\lambda max}$ as the reflectance of the continuum at the same wavelength.

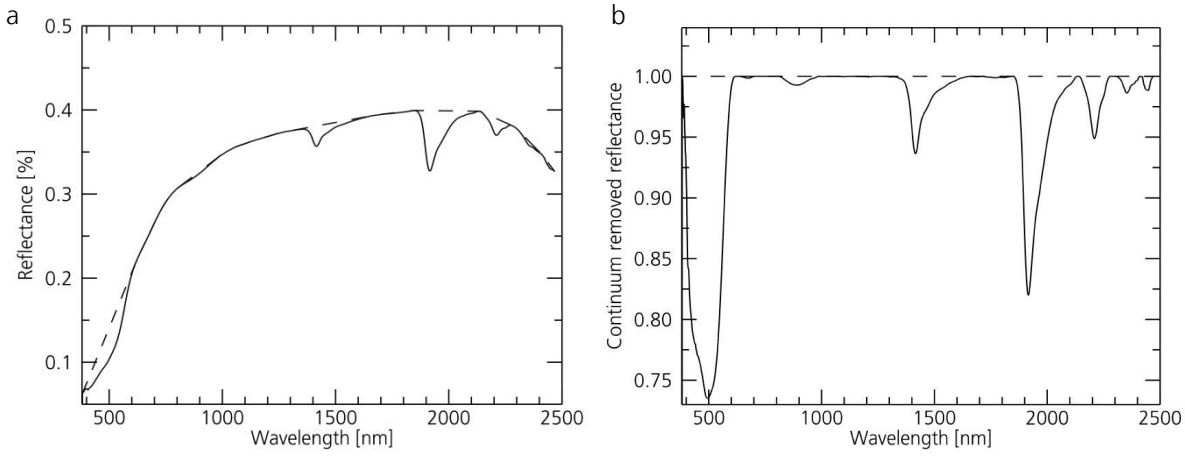


Figure 2-3: Analysis of absorption features by continuum removal. Reflectance spectrum of an exemplary soil spectrum measured in the laboratory and its derived hull as normal (a) and as continuum removed (normalized) reflectance plot (b).
~ Reflectance spectrum, - - hull spectrum.

2.2 Spectroscopy to study soil properties

Over the last decades remote sensing techniques emerged as a time- and cost-effective method for spatially continuous mapping of the earth's surface (e.g. Ben-Dor et al. 2009, Viscarra Rossel et al. 2011). To gain detailed soil information, multi-spectral remote sensing is rather limited, especially when quantitative information is needed and high spectral resolution data are required. Latest developments of hyperspectral sensor systems register hundreds of spectral bands in the visible to shortwave infrared range that compose a continuous spectrum. Spectroscopy thus brings a new dimension to remote sensing data (van der Meer and de Jong 2001).

Spectroscopy for soil applications is based on the analysis of distinct or broad spectral features occurring in the VIS to SWIR spectral domain (ca. 400 to 2500 nm) that can be attributed to electronic transition and charge transfer processes of specific chemical groups of soil compounds (see section 2.1). The potential of spectral information in this wavelength range for quantitative analysis of soil properties is well known. After first studies, e.g. Hunt and Salisbury (1970) and Hunt et al. (1971), investigated the spectral response of natural pure mineral and rock samples, spectroscopic techniques were also applied to study soils and Stoner et al. (1980) provided a (first) soil spectral atlas (Ben-Dor et al. 2008). Since then and especially with the development of new field and also airborne sensors with a better signal-to-noise ratio, many applications of soil spectroscopy have been developed. It is seen as a rapid, less-expensive, non-destructive and sometimes presumably more accurate method than conventional soil analysis by various authors (e.g. Viscarra Rossel et al. 2006). Although reflectance properties of individual soil components are well known, it is not yet possible to accurately model the reflectance of a soil (surface) as a mixture of several components, since in real world soil reflectance changes with temporally and spatially highly variable soil conditions like surface roughness, moisture content, the presence and condition of plant residues and a number of other factors (van der Meer and de Jong 2001).

Various studies compare the performance of VIS/NIR/SWIR to mid-infrared spectroscopy (MIR) for soil applications and suggest better performance of MIR spectroscopy since the signals are more intense and specific as they are the fundamental bands of the molecular vibrations (Bellon-Maurell and McBratney 2011, Reeves III et al. 2002, Viscarra Rossel et al. 2006). Nonetheless, the availability and applicability of MIR field and airborne sensors are still limited and has mainly been restricted to laboratory environments (Bellon-Maurell and McBratney 2011). Considering this, the present study is dedicated to the use of VIS/NIR/SWIR spectroscopy. Accordingly, the following review and the discussion of spectral characteristics of soil properties is limited to the wavelength range between about 400 to 2500 nm.

Investigations in laboratory and field environments

Many studies showed that a wide range of soil constituents can be derived under laboratory and field conditions using different modeling approaches. Reflectance spectroscopy in particular has been used to estimate soil constituents like soil organic matter, clay, iron oxides and carbonates (e.g. Bartholomeus et al. 2008, Ben-Dor and Banin 1994, Ben-Dor et al. 1997, Brown et al. 2006, Chang and Laird 2002, Dalal and Henry 1986, Demattê and Garcia 1999, Genu and Demattê 2006, Gomez et al. 2008a and b, Lagacherie et al. 2008, Selige et al. 2006). Further, elementary contents like N, K, Ca, Al or P (e.g. Chang et al. 2001, Selige et al. 2006, Udelhoven et al. 2003, Viscarra Rossel et al. 2006), Cd and Zn in contaminated soil material (Kooistra et al. 2001), soil properties like CEC, pH and texture (e.g. Ben-Dor and Banin 1995b,

Brown et al. 2006, Sudduth and Hummel 1993, Viscarra Rossel et al. 2006), soil moisture and salinity (e.g. Ben-Dor et al. 2002, Dehaan and Taylor 2002, Haubrock et al. 2008, Metternicht and Zinck 2003) and sulphate minerals (e.g. Cloutis et al. 2006) were spectrally determined. Individual studies propose a global soil characterization using VIS/NIR spectral libraries (e.g. Brown et al. 2006, Viscarra Rossel 2008b).

From point to image

Imaging Spectroscopy (IS) adds a spatial domain to point spectroscopy. Therewith, it is able to provide up-to-date and spatially continuous data of large areas which can be used for soil analyses. The transfer of methods established on point spectral data to imagery bears great potential but also introduces problems, such as (i) dealing with data of a lower spatial and spectral resolution and signal-to-noise level due to a short integration time over the target area, (ii) atmospheric absorptions interfering with the spectral measure, (iii) large datasets, (iv) BRDF effects and (v) spectral mixture within pixels (Ben-Dor et al. 2009, Mulder et al. 2011). Additionally when studying soil surfaces, the spatial variation of surface parameters like surface roughness, soil texture, moisture, vegetation coverage, degree of crusting, etc. induces a further spectral variability that may influence the study of a soil property (Ben-Dor et al. 2008). The interaction of incident radiation with a soil surface is limited to the very first 50 µm as there is very little penetration of incident radiation through the opaque soil medium (e.g. Ben-Dor et al. 1999). Thus, IS observations comprise only the topsoil layer and they do not provide a complete solution for pedologists who wish to know the soil profile variations (Ben-Dor et al. 2009). Nevertheless, the uppermost soil layer surveyed by IS may contain valuable information of soil chemical and physical parameters, of which many can be directly or indirectly related to the active land surface processes (refer to section 1.1).

Recent developments using IS for soil applications

The last decades were dedicated to the increasing use of Imaging Spectroscopy using airborne and minor also satellite sensors to study soil properties. Today a number of imaging spectrometers exist providing the necessary specifications for high quality surveys, such as the AVIRIS, CASI, HyMap or HySpex sensors. Hyperspectral satellites are available with Hyperion on the EO-1 platform (NASA, since 2000) and CHRIS on board of Proba-1 (ESA, since 2001). Future satellite systems are developed with the EnMAP (Stuffler et al. 2007), HISUI (Ohgi et al. 2010) and HyspIRI systems (e.g. Hook et al. 2012). Van der Meer et al. (2012) give an overview of sensors currently used for soil and mineral mapping and planned missions.

Ben-Dor et al. (2008 and 2009) provide summaries of essential applications and key studies applying airborne and satellite sensors to study soil properties. IS was for instance used to map the spatial distribution of soil organic carbon on different agricultural areas in central Europe by Bartholomeus et al. (2011, AHS-160), Stevens et al. (2008 and 2010, AHS-160), Selige et al. (2006, HyMap) and in semiarid regions in south-eastern Spain by Schwanghart und Jarmer (2011, HyMap), in Israel by Ben-Dor et al. (2002, DAIS-7915) and from tropical soils in Brazil by Galvao et al. (1998, AVIRIS). The enrichment of iron oxides was derived in Israelian sand dunes by Ben-Dor et al. (2006, CASI) and in southern Spain by Richter (2010a, HyMap). Clay and carbonate content were studied in French vineyards by Gomez et al. (2008a, HyMap) and Lagacherie et al. (2008, HyMap). The distribution of different clay minerals in Colorado, US, was mapped by Chabrillat et al. (2002, AVIRIS and HyMap). Soil moisture and salinity were mapped by Ben-Dor et al. (2002, DAIS-7915), Dehaan and Taylor (2002, HyMap) and Haubrock (2009, HyMap). IS for soil applications often was applied to small areas and mostly homogenous land cover types, i.e. agricultural areas or analysis was limited to soil dominated areas. A number of studies report a decrease in the accuracy of soil parameter estimations when moving from laboratory (L) to field (F) and finally airborne (A) sensors (e.g. Lagacherie et

al. 2008 (L, F, A), Kooistra et al. 2003 (L, F), Stevens et al. 2008 (L, F, A), Gomez et al. 2008a (L, A), Gomez et al. 2008b (F, Satellite). Best results in soil and mineral mapping have been obtained from areas where this material occurs purely at the surface and is not masked or influenced by other materials such as vegetation (e.g. Chabrillat et al. 2002). The widespread use of Imaging Spectroscopy to study soil properties is constrained by the given sensor characteristic factors and the interference of the signal with surface parameters (see above). Accordingly, the use of IS for soil applications was constituted by Ben-Dor et al. (2009) as "not yet … fully implemented".

Methods for soil spectra analysis

Different methods have been used for the analysis of soil spectra and their correlation to chemical constituents. In general, they can be divided in first classical spectroscopic methods that directly apply spectral features for soil quantification, and second multivariate techniques. Approaches belonging to the first group apply a single or a linear combination of spectral features or band indices. In general, the concentration of a given constituent is correlated with the depth of an absorption band (van der Meer et al. 2012), thereby these approaches provide a physical basis. First studies that specifically use absorption band parameters include the relative absorption-band-depth approach of Crowley et al. (1989), the spectral feature fitting of Clark et al. (1990) and the Tricorder/Tetracorder detection systems (Crowley and Swayze 1995, Clark et al. 2003) all working on continuum removed reflectance spectra (see section 2.1). Recent studies applying spectral feature or band indices for soil quantification from IS are given in section 2.3, since they are in the focus of this study. The question why spectral features can directly be applied although affected by the overlapping of multiple features or influences of non-significant factors is discussed in section 2.4. Furthermore multivariate techniques were applied to extract the soil information "hidden" in the complex mixtures of soil constituents accounting for a soil's spectral signal. The multivariate methods that were used in spectroscopy, their benefits and limitations are in detail discussed in section 4.1. Here the focus is set to studies applying these techniques for deriving soil related information.

Ben-Dor et al. (2002) might have been the first to apply multivariate calibration techniques to map soil parameters (C_{org}, moisture and salinity) from airborne DAIS-7915 data using a multiple regression of several wavelengths for each soil characteristic. A review that focuses on the investigation of soil properties with various multivariate statistical approaches and in different spectral domains (VIS, NIR and MIR) is given by Viscarra Rossel et al. (2006). They noticed that amongst all techniques so far applied to quantify soil properties partial least squares regression (PLS, 12 of 20 collected studies) and principal component regression (PCR) are the most frequently used. Both techniques are applied to construct predictive models when there are many predictor variables (i.e. spectral bands) that are highly correlated. The spectral information is decomposed into spectral vectors that are subsequently correlated to soil attributes. In this process, PLS accesses soil chemical information, while PCR does not (see section 4 for details). PLS was used e.g. by Bartholomeus et al. (2011), Gomez et al. (2008a and b), Schwanghart and Jarmer (2011), Selige et al. (2006), Stevens et al. (2008, 2010) and Viscarra Rossel et al. (2006) and PCR among others by e.g. Chang et al. (2001) and Islam et al. (2003). Subordinated, techniques such as a multiple linear regression of specific bands (MLR, e.g. Ben-Dor and Banin 1995a and b, Ben-Dor et al. 2002, Genu and Demattê 2006, Selige et al. 2006) and complex methods like artificial neuronal networks (ANN, e.g. Viscarra Rossel and Behrens 2010), support vector machines (SVM, e.g. Stevens et al. 2010) or multivariate adaptive regression splines (MARS, e.g. Shepherd and Walsh 2002) have been implemented (Viscarra Rossel et al. 2006). Although multivariate techniques such as PLS are valuable tools, they must face some important drawbacks in terms of significance and transferability that were addressed in section 1.3. For example, Stevens et al. (2010) concluded that PLS calibrations are very sensitive to changing environmental conditions and thus not

well suited when dealing with large datasets that include a certain variability of factors affecting soil reflectance (such as soil types, surface conditions etc.). They compared the ability of PLS and SVM regressions for SOC predictions for a Luxembourgian study site (≈ 420 km², ≈ 300 reference samples). While PLS provided only moderate calibration/validation accuracies (R^2_{Cal}: 0.57 and R^2_{Val}: 0.59), SVM regression performed better (R^2_{Cal} 0.97, R^2_{Val} 0.74) considering the entire datasets. However, a trend of the SVM model to over-fitting was suggested. Both methods provided good results (R^2_{Cal} 0.91-0.99) when working on a local scale defined by principal soil types or agro-geological regions.

2.3 Reflectance properties of soil constituents

Generally, soil spectral behavior is a function of soil composition and soil condition. The spectral signatures shown in a reflectance spectrum as a function of wavelength, result from the electronic transitions of atoms and vibrational modes of atoms (see section 2.1). The major chemical components of soil minerals - oxygen, silicon, and aluminum – do not have diagnostic absorption features in the VIS to SWIR range but can be well identified in the thermal wavelength region. Though vital soil components that occur as accessories in sometimes low concentrations are spectrally active in the visible, near and shortwave infrared region. Figure 2-4 shows laboratory spectra of soils dominated by high contents of organic carbon, iron oxides and clay minerals.

The following overview is dedicated to spectral properties of the soil constituents that are spectrally active in the VIS to SWIR range. A special focus is set on the three parameters chosen for this study. They are reviewed in order of their importance identified in section 1.3 and not according to their chemical concentrations in soils (mostly clay > iron > C_{org}). In addition, materials and parameters affecting soil reflectance, e.g. vegetation, are further discussed to provide the basis for the development of the methodology in this study. Examples of recent studies that directly use various spectral characteristics for the delineation of soil properties, especially ones using IS are given. The chemical characterization of components and their function for soils and the ecosystem are taken from Rice (2006, Soil organic carbon), Qafoku and Amonette (2006, iron oxides) and Schachtschabel et al. (2002). The description of their reflectance characteristics is mainly based on the works of Clark et al. (1990), Clark (1999), and van der Meer and de Jong (2001).

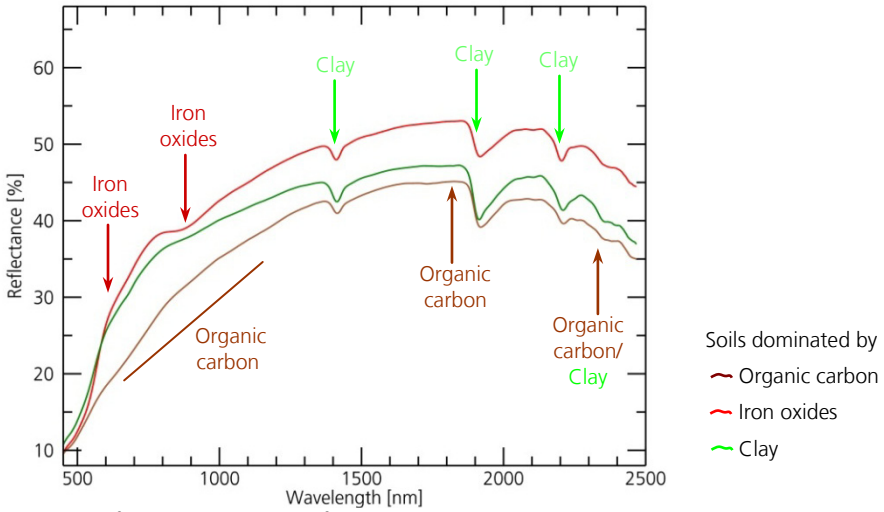


Figure 2-4: Laboratory reflectance spectra of soils dominated by organic carbon, iron oxides and clay (own measurements).

2.3.1 Soil organic carbon

Soil organic carbon includes any dead biomass added from vegetation as litter decays to the soil, and all its conversion products. The rate of organic matter accumulation varies strongly depending on climate, organisms, parent plant material, composition stage, water content and time. Afterwards, chemical, biological and physical processes decompose the material into its compounds which releases nitrogen, sulphur and phosphorus, all of them essential nutrients needed by soil organisms. As result of the input and subsequent decomposition of organic material, the chemistry of organic carbon components is very complex, ranging from simple sugars, amino acids and proteins to cellulose, hemicellulose, fat, starch, wax, lignin and tannin. Soil organic carbon is the primary supplier of nutrients for plants and microorganism and determines important soil characteristics such as the cation exchange capacity, soil acidity and others. Besides the obvious impact of soil organic carbon on the chemical properties of a soil, organic substances help to establish a stable aggregate structure together with mineral components and due to the poriferous structure of organic compounds provide a great capacity for water retention. Its value as indicator for assessing ecosystem indicators is addressed in section 1.1.

Organic carbon is a perfect property to be assessed by IS, because it is mainly concentrated in the topsoil horizon that is exposed to the surface (Ben-Dor et al. 2009). Because of the complex chemistry of organic materials, the corresponding spectral features are rather numerous and particularly small. Major and pure organic components like cellulose, starch and lignin (see Elvidge 1990) are related to the functional groups O-H, C-H and N-H. Their fundamental features occur in the mid- to thermal infrared range, but their overtone and combination features spread nearly over the entire NIR to SWIR region (Ben-Dor et al. 1997). Some stronger absorptions occur around 1730 nm (first overtone of C-H stretch) and 2330 nm (combination band of O-H stretch, see figure 2-4), while smaller features are located near 1150, 1670, 1765, 2070, 2110, 2140, 2170, 2190, 2280, 2310 and 2390 nm (see also Ben-Dor et al. 1997). Also typical water bands around 1400 and 1900 nm can be caused by water in organic components. Because the various absorption features of organic components are sometimes close to each other they may overlap across the NIR and SWIR spectral region.

Besides these distinct absorption features, an increase in soil organic carbon content affects the shape of the spectral curve. It leads to a decrease in the reflectance over the entire spectrum and especially in the visible and near-infrared region due to the darkness of humic acid (see figure 2-4) and causes the spectrum

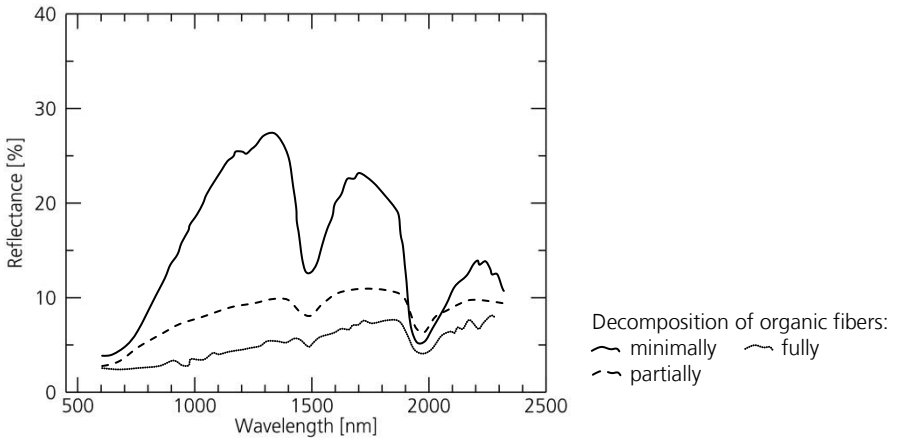


Figure 2-5: Averaged reflectance spectra of organic soils of different composition levels (according to Stoner and Baumgardner 1981, redrawn).

to become more concave in this spectral domain (e.g. Bartholomeus et al. 2008, Baumgardner et al. 1985, Hill and Schütt 2000). Nevertheless, the impact of this effect depends on organic carbon content and the decomposition level of the plant residues (figure 2-5, Stoner and Baumgardner 1981). Baumgardner et al. (1985) found that this spectral effect of organic carbon is only clearly observable besides the features of other soil constituents when the organic carbon content exceeds 2 % (20 g/kg).

As a result of the importance of soil organic carbon for soil stability and fertility and in the course of the global initiatives to quantify ecosystem carbon stocks and their changes (refer to section 1.1), there have been very intense efforts of using spectroscopy for these purposes resulting in various studies dealing with soil organic carbon quantification under laboratory (e.g. Bartholomeus et al. 2008, Ben-Dor et al. 1997, Chang and Laird 2002, Dalal and Henry 1986, Hummel et al. 2001, Viscarra Rossel and Behrens 2010, Vohland et al. 2011) and field conditions (e.g. Hill and Schütt 2000, Sudduth and Hummel 1993) and also from hyperspectral imagery (e.g. Ben-Dor et al. 2002, Stevens et al. 2008 and 2010, Bartholomeus et al. 2011, Gomez et al. 2008b, Galvao et al. 1998, Selige et al. 2006). For instance Ben-Dor et al. (1997) studied the spectral behavior of high organic matter compost materials during their decomposition process and categorized the observed changes of the spectral shape in the VIS/NIR and also in the SWIR into groups representing typical organic components such as lignin, cellulose, etc.. Elvidge (1990) described the spectral response of chemical components of dry plant materials, which is also relevant when considering soil organic components since their biochemical composition and thus spectral signature depends on the source material and its decomposition stage. The quantification of soil organic carbon using IS in semiarid and arid landscapes is complicated by the generally low concentrations found (< 15 % in mineral soils) and further influenced by signal mixtures caused by vegetation or litter covering the soil surface (Ben-Dor et al. 2009).

A number of studies apply PLS techniques for soil organic carbon predictions. Among these, Stevens et al. (2008) investigated C_{org} concentrations in cultivated areas in Belgium with reliable results on laboratory and field scale. In a subsequent study of 2010, they predicted organic carbon concentration for a dataset of around 420 km^2 of agricultural fields in Luxembourg including different soil types and concentrations using PLSR and also SVM regression and found local calibrations more accurate and useful than global ones. Bartholomeus et al. (2011) advanced the efforts on the quantification of C_{org} in the Belgian Lorraine region to fields that are partially vegetated. The study is referred to more in detail in section 5.2.2. Schwanghart and Jarmer (2011) assessed the relationship between C_{org} patterns derived from HyMap data and topography for a semiarid region in Southern Spain. Gomez et al. (2008b) seem to have been the first to predicted soil organic carbon from Hyperion satellite data using PLSR techniques, even with a medium prediction accuracy.

The estimation of soil organic carbon apart from multivariate techniques was done using individual VIS/NIR/SWIR bands (e.g. Ben-Dor et al. 2002, Selige et al. 2006), but mostly based on the overall decrease in reflectance in the VIS/NIR region (see Bartholomeus et al. 2008, Kooistra et al. 2003). Bartholomeus et al. (2008) tested the performance of several spectral indices to retrieve soil carbon and found good correlations in the VIS range. The shape of peak-normalized reflectance spectra between 350 and 1400 nm was used by Hill and Schütt (2000) to predict organic carbon concentrations in arid south-eastern Spain from field spectra and multispectral imagery by fitting a third order polynom and a subsequent regression applying the fit's coefficients. They finally used the spatial soil carbon distribution to map areas of soil loss and accumulation areas.

2.3.2 Iron oxides

Pedogenic iron oxides are formed during weathering of iron bearing primary rocks and minerals and occur as strongly pigmented clay sized particles. If oxygen is present during the forming process, trivalent iron oxides are formed, while bivalent oxides develop in anaerobic conditions, thereby goethite (FeOOH), respectively hematite (Fe_2O_3), being the most common iron oxides. Goethite is the prevalent pedogene iron oxide because of its high stability and many ways it can be formed so that it occurs in soils of all climatic regions. Hematite forms under relatively high temperatures of tropic and subtropic soils and leads to a typical reddish soil color.

The importance of iron oxides for soils results in particular from their positive charge. It provides sorptive properties and buffers the oxidation-reduction activity of soils, which plays a major role in the geochemical cycle of many elements having agronomic or environmental significance. Iron oxides further are able to build trace elements in their lattice when they are formed and to attach heavy metals to their surface, thereby limiting the metal's mobility in soils. The attachment of positively charged iron ions to the permanent negative charge of clay mineral layers promotes the formation and stabilization of soil aggregates. Soil iron oxide content often ranges between 0.2 and 20 % iron oxides, depending on soil texture, rock parental material and the status of pedogenesis. Thus, iron oxides content is a direct indicator for soil formation and an important parameter for soil classification (Ben-Dor et al. 2008).

Hematite and goethite are the only soil constituents that have absorption features in the visible range, thus highly affecting soil color (see figure 2-4). Yellowish-brown colored soils occur to have an increased goethite content while reddish soils are high in hematite. The spectral features result of electronic transition processes of the bi- and trivalent iron ions, where hematite generally is correlated to absorption of the ferric iron ion (Fe^{3+}) and goethite of the ferrous iron ion (Fe^{2+}). Main absorption features occur near 650 to 700, 860 to 900, and 1000 to 1100 nm. Some weaker absorption features are located at 400 and 550 nm (see also e.g. Baumgardner et al. 1985, Hunt et al. 1971, Morris et al. 1985). Small shifts in wavelength position of these features are reported due to the substitution of aluminum for iron. The strong decrease of soil reflectance towards the blue wavelength region is caused by an intense charge transfer absorption centered in the blue to ultraviolet region of which only the right shoulder extends into the VIS range (see also Hunt et al. 1971, Morris et al. 1985).

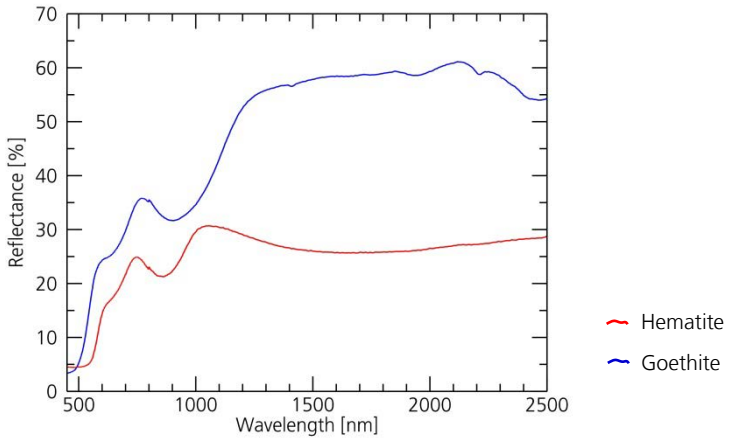


Figure 2-6: Reflectance spectra of the major iron oxides (according to Grove et al., 1992, particle size < 45 µm).

Iron oxides were studied in the laboratory (e.g. Ben-Dor and Banin 1994, Chang et al. 2001, Richter et al. 2009, Torrent et al. 1983), field (e.g. Palacios-Orueta and Ustin 1998) and using IS (Ben-Dor et al. 2006, Galvao et al. 1998). Richter et al. (2009) quantified free iron oxides content based on the band depth of the 900 nm absorption feature and evaluated the spectral influence of variable soil textures on prediction accuracy in a laboratory set-up. Best prediction accuracies were found for sand-dominated samples and a slightly lower for soils dominated by silt and clay. The enrichment of iron oxides in Israelian sand dunes (rubification) was mapped by Ben-Dor et al. (2006) from imagery of an airborne sensor. They applied the Redness-Index ($693nm^2/(477nm \cdot 556nm^3)$) to predict mostly very low concentrations (< 0.5 % Fe_d) with reasonable accuracy.

2.3.3 Clay minerals

Clay minerals refers to a group of layered alumosilicates that predominate the clay-sized fraction of soils (< 2 µm, Barton and Karathanasis 2006). They form as an alteration product of silicates, particularly feldspar, and thus are found in most soils worldwide. The chemical structure of clay group minerals consists of two or three platy layers of silicate tetrahedra and aluminum octahedra layers. The frequently occurring mineral kaolinite is formed as first product in the alteration process as a simple two-sheet mineral type and is predominant in tropical soils. It has a low capacity for cation absorption and exchange and thus soil fertility due to its structure. Montmorillonite (smectite group) and illite are the most common members among the large variety of three-sheet minerals. Cation exchange capacity is highest for montmorillonite as it happens to expand when the mineral hydrates. In soils clay minerals contribute to nutrient retrieval as they retain and exchange nutrient cations at their large and mostly negatively charged surface (Greenland 2006). They also play a prominent role in the water budget of a soil due to their ability to absorb water. Furthermore, clay minerals are able to build very stable compounds with humic substances and thus prevent them from rapid decomposition.

For the spectral identification of the large variety of clay minerals, the SWIR part of the spectrum is the most useful (see figure 2-4 and figure 2-7). An Al-OH overtone and O-H combination band around 2200 nm is most diagnostic for clays in general and thus is visible in nearly every soil spectrum. Kaolinite clays shows fine structured absorptions with a well-known doublet feature at around 2160 and 2208 nm, while montmorillonite and illite show a single band absorption around 2200 nm (see figure 2-7, see also e.g. Grove et al. 1992, Hunt 1982, Viscarra Rossel and Behrens 2010). This feature shows small shifts in the wavelength of maximum absorption depending on the type of clay mineral, its exact chemical composition and its decomposition level. Some existing studies link the shift of the 2200 nm feature for instance of montmorillonites to longer wavelength to an increasing Ca or decreasing Na content. Only illite shows additional absorption bands near 2340 and a weaker one near 2450 nm, that is absent in the spectra of montmorillonite and kaolinite (figure 2-7, see also e.g. Chabrillat et al. 2002, Viscarra Rossel and Behrens 2010). Small features in the VIS/NIR range can be caused by Fe^{2+} ions, which substitute Al in the lattice of several clays. Kaolinite minerals also show a strong hydroxyl (O-H) band at 1400 nm and a weak 1900 m feature due to the lack of bound water. Despite that montmorillonite and extenuated also illite clay minerals typically comprise bound water which results in strong water bands at both 1400 and 1900 nm. Even though the last-mentioned features are very strong and would support the spectral identification of clays, they are not useable in a field environment because they are masked by atmospheric water.

A soil's dominant grain size additionally affects the overall reflectance of soil samples and as such the trend of the spectral continuum (refer to section 2.1). In an aggregation of smaller grains there are more surface reflections compared to grain internal paths where light is absorbed according to Beers Law (Clark 1999). Thus, a decreasing particle size and a higher clay-sized fraction lead to a brighter soil surface which is measured by spectroradiometers as an increase in soil reflectance in the VIS and also NIR region (see also e.g. Ben-Dor et al. 2003). For instance Hunt and Salisbury (1970) reported changes in reflectance of about 5 % due to changes in particle size. In addition, they noted that these changes occur without altering the position of diagnostic spectral features.

Studies, where Imaging Spectroscopy was used to quantify soil clay content were done in an laboratory set-up e.g. by Ben-Dor and Banin (1995b), Kooistra et al. (2001), Viscarra Rossel and Behrens (2010) and Viscarra Rossel et al. (2006), in a field environment e.g. by Waiser et al. (2007) and using airborne scanner e.g. by Chabrillat et al. (2002), Gomez et al. (2008a), Lagacherie et al. (2008) and Selige et al. (2006). Chabrillat et al. (2002) applied the varying shape of the 2200 nm feature of smectite, illite and kaolinite clay minerals to detect and map their spatial distribution from airborne AVIRIS and HyMap data of Colorado, USA. Despite significant vegetation coverage they were able to identify regions of exposed pure smectitic clays that are prone to swelling and cause engineering problems in the rapidly growing urban area. Gomez et al. (2008a) and also Lagacherie et al. (2008) successfully applied the continuum removed band depth of the 2200 nm feature to quantify clay content. However, Gomez et al. found a slightly reduced performance compared to PLSR.

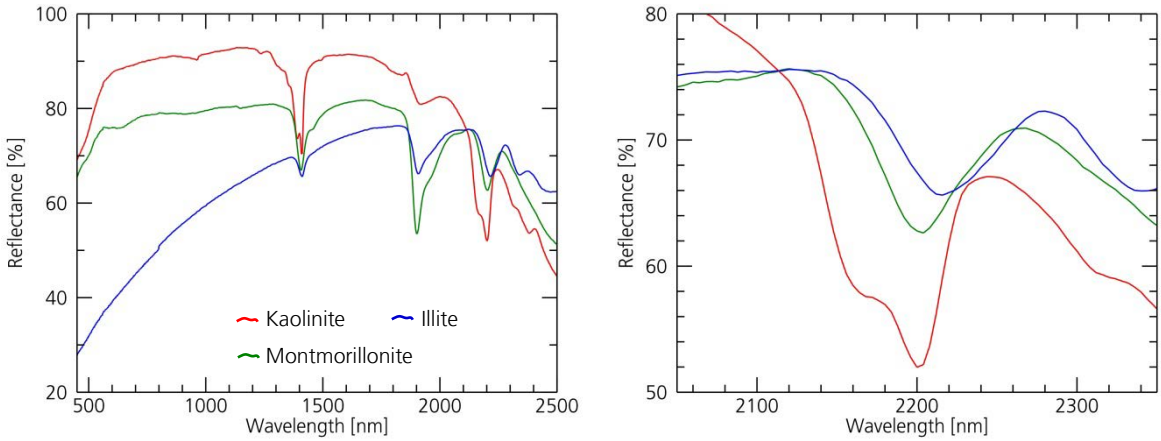


Figure 2-7: Reflectance spectra of typical clay minerals (according to Grove et al., 1992, particle size < 45 µm).

2.3.4 Other soil constituents and relevant materials

In addition to the three considered soil constituents, a number of materials comprising soils or occurring accessorily in them are spectrally active in the VIS to SWIR region. From these soil water and carbonates are considered in detail. Also vegetation is included in the survey as it is relevant for signal separation during spectral unmixing.

Soil moisture

Soil water or moisture can be present as absorbed water bound to molecules or particle surfaces or as capillary condensed water in small pores. It is the source for water requiring surface processes like plant transpiration and the refill of ground water reservoirs. Soil water spectrally can be well identified by two strong absorption features near 1400 nm (overtone band of hydroxyl group) and near 1900 nm (combination band of the water molecule) and some weaker features in the NIR range at around 960 and 1100 nm (see also e.g. Baumgardner et al. 1985, Hunt and Salisbury 1970). The presence of both of the strong features indicates free water in the soil or water bound in mineral, while the appearance of the 1400 nm without the 1900 nm indicates O-H groups bound to another material, e.g. clays (Baumgardner et al. 1985, Hunt and Salisbury 1970). As mentioned beforehand, the two features are masked by disturbances introduced by atmospheric water vapor.

Soil water content has an additional effect on the overall reflectance causing a broad absorption in the VIS which the human eye visualizes as a darker soil color. The absorption extends into NIR and SWIR ranges, thus resulting in a decrease of the overall albedo (Baumgardner et al. 1985, Lobell and Asner 2002). In order to detect soil moisture in spectral ranges unaffected by atmospheric attenuation and overlaps with other features, several authors proposed a better suitability of the SWIR range for an assessment of soil moisture (e.g. Haubrock et al. 2008, Lobell and Asner 2002, Whithing et al. 2004). This leaded to the implementation of two often used soil moisture indices - SMGM (Soil Moisture Gaussian Model) and NSMI (Normalized Soil Moisture Index). With SMGM Whiting et al. (2004) offered a robust technique that applies the fundamental water absorption at 2800 nm which is beyond the limit of usually applied instruments but can be applied to fit an inverted Gaussian function centered at 2800 nm to the continuum of the logarithmic soil spectra. The NSMI applies the bands around the water combination band at 1900 nm and was used by Haubrock et al. (2008) for mapping of soil moisture in an area of varying environmental factors. However, a complete correction of the effect of soil moisture on spectral features of other soil properties is not yet possible (Ben-Dor et al. 2009).

Carbonate Minerals and other accessory soil minerals

Calcite ($CaCO_3$) and dolomite ($CaMg(CO_3)_2$) are the major carbonate minerals in rocks and soils. They are either inherited from the parent rock material or are formed by precipitation during soil formation (Ming 2006). Carbonates influence soil pH, ad- and desorption of charged compounds and soil cementation (Ming

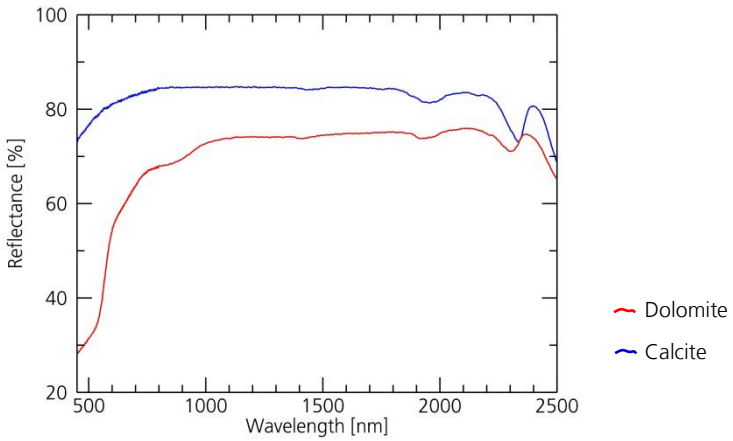


Figure 2-8: Reflectance spectra of the major carbonate minerals (according to Grove et al., 1992, particle size < 45 µm).

2006). The two carbonate minerals cause a strong absorption peak around 2345 nm due to CO_3 overtone vibrations and some weaker bands near 1860, 1990 and 2140 nm (figure 2-8, see also e.g. Hunt 1977). Exact band positions vary once more with composition (Hunt et al. 1971). Recent studies by Gomez et al. (2008a) and Lagacherie et al. (2008) successfully apply the continuum removed band depth of the 2345 nm feature to quantify soil carbonate content.

Other accessory soil minerals like phosphates, halogenides, etc. also have diagnostic features in the VIS to SWIR range but their significance for soil function is limited unless they occur in locally prominent deposits, which is not relevant for an investigation of the working area.

Vegetation

Depending on climate and a number of local factors, soils are to some extent overgrown by vegetation. Vegetation is generally found as green and wet and thus photosynthetically active on the one side and dry non-photosynthetic vegetation on the other side, and of course all forms in between the two endmembers. As plants are built of some basic materials, there are similar spectral signatures that allow the identification of vegetation in general. Photosynthetic active vegetation is characterized by a low reflection in the visible due to chlorophyll absorption and a high reflectance in the NIR region (red edge), which is an indicator for the plant vitality. Pigments such as chlorophyll a and b absorb strongest in the blue and red wavelength regions (around 450 and 650 nm), leading to higher reflectance in the green range. The NIR to SWIR region is dominated by vibrational absorptions of liquid water. Further spectral features vary in wavelength position and intensity according to pigmentation, physiological structure and water content. Variances in them can be used to discriminate different vegetation types or even species. The spectrum of dry non-photosynthetic vegetation is rather flat in the VIS to NIR range but shows numerous overtone and combination absorptions in the SWIR. Major features of O-H, C-H, and N-H assigned to cellulose, holocellulose, lignin and starch are around 1690, 1720, 1780, 2050 to 2140, 2180, 2270, 2300, 2330 and 2380 nm (see also e.g. Elvidge 1990). These features are also inherent to green vegetation but mostly masked by the strong water bands. Elvidge (1990) gives an extensive summary of vegetation compound's absorptions and plant spectra of different phenology stages. Besides the reflection of the present plant species' individual components, the reflection properties of a vegetation canopy depend on the spatial distribution of vegetated and non-vegetated spots and the material in between.

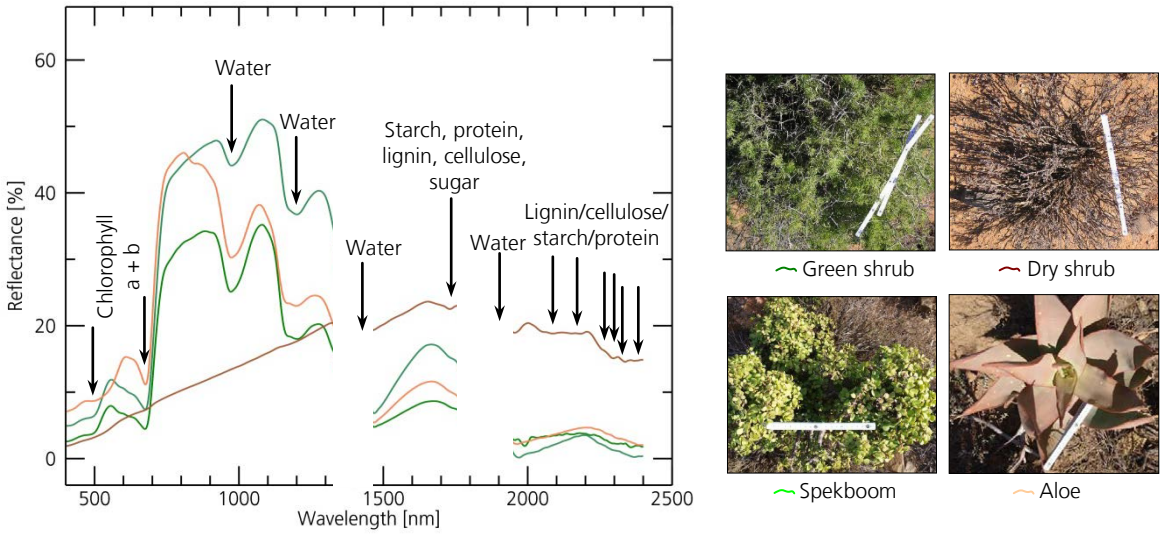


Figure 2-9: Field reflectance spectra of typical plants growing in the Albany Thicket Biome, South Africa (examples from field measurements in July 2009). Major water bands are masked.

2.3.5 Soil physical factors affecting reflectance

Soil reflectance is triggered by its predominant mineral and organic components, but highly influenced by physical factors like soil particle size, surface conditions, etc. These properties mainly influence the soil's albedo and only secondarily introduce absorption features.

Soil particle size

The distribution of soil particle size is a crucial soil property that affects e.g. soil structure, nutrient and water holding capacity and a soils susceptibility to erosion (Geering and Bing So 2006). It is still difficult to determine spectral effects of soil texture changes. Generally, as the grain size becomes larger, the path of the light through the grain becomes larger and more light is absorbed according to Beer's law, whereby reflectance decreases and absorption features become more pronounced (see also e.g. Baumgardner et al. 1985). Hunt et al. (1971) separated the effect of particle size on soil albedo of two groups of minerals. For mostly bright minerals such as silicates and carbonates it was noted that decreasing particle size results in an increase in overall reflectance. Minerals like oxides and hydroxides that often have a low reflectance show a reversed behavior – decreased particle size leads to a decrease in reflectance. The difference was explained by the balance between specular and volume components, which composes the total reflectivity for mat surfaces. Nonetheless, as soils comprise of both of these groups but predominantly bright silicate minerals, decreasing particle size in soils leads to an increased reflectance (see figure 2-10, see also e.g. Stoner and Baumgardner 1981). Sand particles in soils are usually built from very mature sandstones or quartz rich parent material and thus mostly comprise of quartz as it is very resistant to the weathering process and accumulates as alteration residuum. Pure quartz does not have any spectral features in the VIS to SWIR region. It can only be identified due to significant lattice impurities or fluid inclusions that produce characteristic features in this spectral region. Besides of quartz, clay sized particles are predominantly made up of clay minerals which spectral features in the NIR and SWIR range were presented in section 2.3.3.

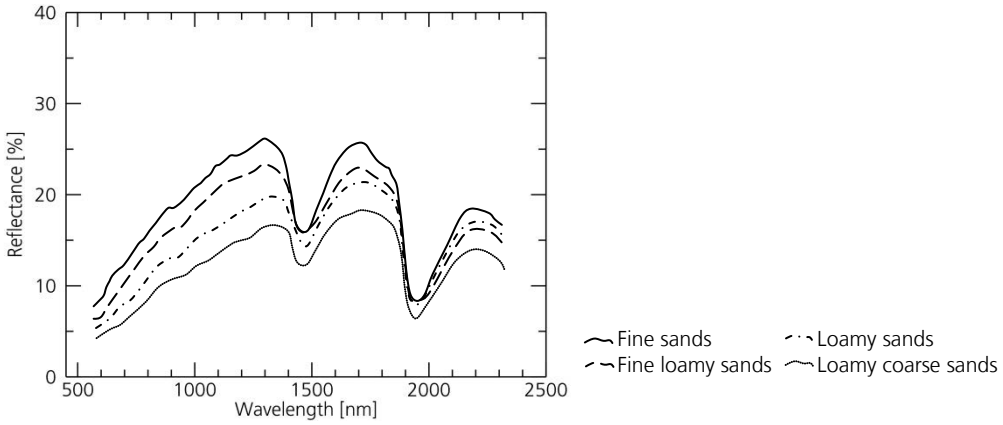


Figure 2-10: Averaged reflectance spectra of soils differentiated by dominant particle size (according to Baumgardner et al. 1985, redrawn).

Aggregates

The spectral influence of the soil particle size can be masked by the ability of a soil to build aggregates. Most of the light falling into the interspaces of large and irregularly shaped aggregates is completely extinguished, so that aggregates contribute to a lower reflection (Baumgardner et al. 1985). Clay-rich soils

aggregate more and thus appear as rougher and darker surfaces. Aggregates form by root and microorganism activity whereas the presence of organic compounds, iron oxides and clay minerals that provide large and mostly charged surfaces facilitates bindings between components to larger aggregates.

Soil salinity

Soils get affected by salt where there is an influence of saline concentrations in soils, saline groundwater or where land is irrigated what causes increased soil degradation and erosion (Dehaan and Taylor 2002). Since the major salt mineral halite is featureless in the VIS/NIR/SWIR region, in previous studies salt affected areas have been mapped according to corresponding features influencing soil reflectance such as e.g. surface roughness or the occurrence of sulphate minerals or salt-tolerant plants (Dehaan and Taylor 2002, Metternicht and Zinck 2003). Soil salinity is not considered as an influencing factor since there is no influence of salt observable in the study area.

Physical soil crusts

A soil physical crust is a thin layer of low permeability at the soil surface that forms as result of raindrop impacts on the uncovered soil surface (McIntyre 1958). This phenomenon is common in arid and semiarid areas where a protecting vegetation cover is reduced. Water infiltration is significantly reduced upon crusted soils and the surface water runoff increased. The formation of soil crusts in general and their appearance throughout the study area are described in detail in section 3.1.2. Spectrally, crusted soils appear brighter as the portion of fine particles at the surface is increased and also the portion of specular reflected light increases due to the smoothed surface and clay absorption features become more accentuated (de Jong et al. 2011). This correlates with a decrease of the significance of absorptions features. The spectral behavior of semiarid soils during stepwise crust formation using a rainfall simulator with varying levels of rain energy was investigated by several authors (Ben-Dor et al. 2003, Eshel et al. 2004, Goldshleger et al. 2002). They found that crusts can be spectrally discriminated from the non-crusted samples by stronger clay features and a higher baseline reflection. Ben-Dor et al. 2003 attributed the detected spectral changes to changes in particle size distribution and mineralogical composition of the surface material in the process of crust formation. Even if the differences in airborne spectra are reported to be small by de Jong et al. (2011), they were able to map the distribution of crusted and uncrusted soils for a semiarid area in southern France.

2.4 Consequences for the retrieval of soil information by using spectroscopy

The number of chemical and physical soil properties and in addition external factors that have an effect on soil reflectance determine important influences and sometimes also limitations for the analysis of specific soil constituents. This section sums them up and evaluates how the important information contained in soil spectra can be used for the quantitative determination of specific soil constituents.

Factors influencing the significance of soil spectra

Complicating factors for the use of soil spectroscopy result (1) from ambiguities within the spectral signals itself, (2) changes in the relationship between absorption features and chemical contents, (3) are introduced

by the acquisition of the spectra, especially in a field environment or from airborne or satellite platforms, and (4) may result from the determination of soil chemical reference contents.

The previous section presented an extract of the most important spectral characteristics found in soil spectra. Thereby, it became apparent that most spectral features are not unique within a soil spectrum, because distinct features may overlap and also unspecific broad features occur (e.g. absorption in the VIS range due to soil moisture, organic carbon and particle size). This leads to a reduction in prediction accuracy that one has to be aware of. Resulting ambiguities between overlapping features can be minimized if some factors logically can be excluded based on local conditions (e.g. absence of carbonates). Mathematical models, such as partial least squares regression analysis, can help to decompose the spectral information and discriminate between diagnostic and non-significant characteristics so that only the diagnostic information may be correlated with soil properties (Viscarra Rossel and Behrens 2010). The correlation between the chemical content and the size of corresponding diagnostic absorption features is presumed to follow a linear trend. This in approximation applies for low material concentrations. Nonetheless, especially for higher contents, an intensity saturation of absorption features was described in previous studies (e.g. Torrent et al. 1983 for iron oxides absorptions).

Reflectance measurements from an airborne or satellite platform introduce further factors that need to be addressed by an adequate handling before any soil related analysis. They can be grouped to (1) sensor dependent factors, (2) atmospheric disturbances and (3) surface cover interferences. Sensor related inaccuracies may result for instance from a low signal-to-noise level, optical shifts or geometrical rectification (Ben-Dor et al. 2008). They are not further discussed in this context, because they are required to be corrected within pre-processing of the raw data. Atmospheric gas and aerosol attenuation is corrected by appropriate radiative transfer models (e.g. Richter 2010b). It is critical as it may leave features in the image spectra that can mistakenly be held for diagnostic soil spectral features or at least influence spectral feature properties (Ben-Dor et al. 2008). A highly accurate atmospheric correction and pre-processing of the imagery is needed to discriminate between these effects and the sometimes weak diagnostic spectral features in soils. A second component inherent to remote sensing and resulting from data acquisition are surface cover interferences. Depending on the spatial resolution of the sensor system, spectral mixture of variant intensity occurs within the present surface materials which is represented in the image pixels' spectra. According to FAO (1994) about 56 % of the land areas are covered by miscellaneous types of vegetation. Where vegetation is dense it masks the features of the underling soil and prevents quantitative soil analysis. In contrast, where vegetation coverage is partial, to some extent the chemical and physical properties of the underlying soil can be resolved (Murphy and Wadge 1994). In the spectral mixture, green vegetation mostly affects the spectral region from 700 to 1300 nm resulting from the steep increase in reflectance, whereas dry vegetation in particular introduces spectral features in the SWIR region (Ben-Dor et al. 1999). Then, a careful separation of soil and vegetation signals, notably of mixing pixels of hyperspectral imagery, is needed before a quantitative analysis of soil.

The correlation of spectral and chemical information provides the basis of soil spectroscopy. This includes the fact that only the thin surface layer of about 50 µm is viewed by spectrometers but more soil material contributes to the chemical determination of the constituent's contents. The influence of this discrepancy is neglectable for soil constituents that are evenly distributed within the topsoil matrix. This in approximation applies for soil organic carbon and iron oxides, because their variances in the topsoil layer are small in relation to the size of the thickness of the layer the soil sample is taken from, which in this study is about 1 cm. However, this introduces problems when a small-scale layering is present at the surface, which applies when soil crusts are present. Then this results in a weakening of the correlation between the spectral

signature and the determined chemical content and may be problematic in particular for the determination of soil clay content.

Successful delineation of selected soil constituents from hyperspectral imagery

Despite these limiting factors, soil parameters can be quantified from spectra, which was shown in numerous previous studies (see references given in section 2.2). This could be achieved by the combination of appropriate pre-processing techniques, such as radiometric and atmospheric correction, providing accurately calibrated imagery as the basis and the application of sophisticated analysis methods, such as feature extraction and multivariate regression techniques (see section 2.2).

In the same way, in this study adequate techniques are applied to separate useful from non-significant information so that absorption features can directly be used for analysis which provides a physical base to developed prediction models. Some of the described feature ambiguities could be resolved by the conditions present in the study area. In addition, the proposed method makes use of the collinearity between several spectral features diagnostic for one soil constituent to reduce the influence of overlapping spectral features or by non-significant factors. Influencing data related factors such as atmospheric disturbances are addressed in data pre-processing. The spectral influence of non-soil materials such as vegetation on analyzed image spectra is removed in a separate processing step. It is shown that following this approach, the information contained in soil spectra can be used to develop quantitative prediction models for selected soil constituents that are regionally applicable.

3 Study area and available data

The study area is located near the city of Port Elizabeth in the Eastern Cape Province of South Africa (33.0°S/25.3°E). The approximately North to South trending transect is 80 x 4 km and was selected within the Albany Thicket Biome (see below) so that it includes the highest variance of Thicket vegetation classes (STEP vegetation classification after Vlok et al. 2003). The transect comprises two sections as it is crossed by the ranges of the Kleinwinterhoek Mountains that are part of Addo Elephant National Park. The northern section (57.2 x 4.1 km) lies between the two small towns Somerset East and Pearston to the South. Terrain elevation changes from 930 (North) to 300 m a.s.l (South). Small mountain ranges result in an undulating landscape crosscut by small rivers that only bear water after heavy rains. Thicket vegetation usually covers the gently dipping slopes of these mountain ranges, leaving the hill tops and flat plains only sparsely vegetated. The southern section (22.5 x 4.1 km) with a terrain elevation between 130 (North) and 400 m a.s.l. (South) comprises the flat plains of the Sundays River Valley and in the very South the foothills of the Grootwinterhoek Mountains. Vegetation coverage increases to the South where the undulating area is less accessible.

3.1 General description and landscape evolution

Climate

The region is semiarid with annual rainfalls between 200 and 400 mm and daily mean temperatures ranging from 13 to 25°C (ARC-ISCW 2011, Hoare et al. 2006, Lechmere-Oertel et al. 2005). The average rainfalls are evenly distributed over the year since it is located in an area of changeover between winter rains in the southwest and summer rain dominated areas in the northeast, with yearly variances corresponding to the dominant influence. Droughts of several months do occur frequently. A climate diagram showing the local climate of the study area was calculated as average of three stations situated around the transect area including mean daily temperatures and rainfall data of the years 2008 to 2010 (figure 3-1a, data provided by Agricultural Research Council, Institute for Soil, Climate and Water, Pretoria, South Africa). It clearly shows a dominant impact of the summer rain affected areas in those years, with 2009 being a year of a drought during the winter months. During a re-visit of the area in May 2011 we found the ecosystem had a significant change after weeks of heavy rains (figure 3-1b).

Landuse

Commercial goat pastoralism in the Eastern Cape started before the 20th century (Hoare et al. 2006), although it has intensified since the 1930's (Mills and Cowling 2006, Mills and Fey 2004a). These days, stock farming predominantly with angora goats is still the main source of income. Although, many farms have converted part or all of their properties into game farms that are exclusively used for hunting tourism.

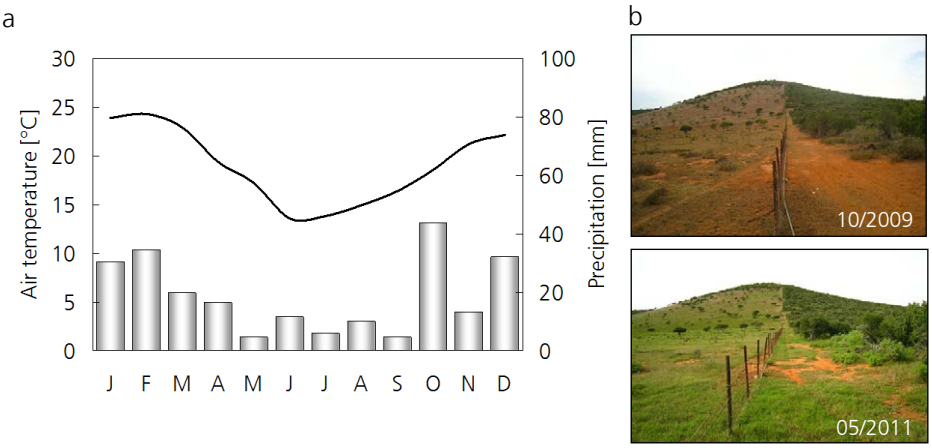


Figure 3-1: Climatic conditions of the study area in the Albany Thicket Biome, South Africa. a: Mean temperature (solid line) and average precipitation (blocks) generated as average of daily data of three weather stations close to the study area of the years 2008 to 2010 (data provided by ARC-ISCW 2011). b: Changes in vegetation status due to differences in rainfall shown at a fence line contrast in the southern transect. The evergreen dense Thicket vegetation hardly changes.

Geology and geomorphology

The following review is limited to the geologic units comprising the surficial geology of the study area and thus the units that are relevant for soil generation. Details on their formation and development can be found in Johnson et al. (2006). Figure 3-2 gives a zoom in on the geologic map containing the two transects (CGS 2010).

The soil parent material of the northern study area is built of sediments of the Karroo Supergroup (Ecca and Beaufort Groups, Late Carboniferous to Middle Jurassic). The predominant part of the northern section consists of very similar, cyclically alternating grey mudstones and fine to medium-grained lithofeldspathic sandstones (Adelaide subgroup, as part of the Beaufort Group, Johnson et al. 2006). The thickness of the layers varies from some decimeters to tens of meters (figure 3-3), with thicker, sometimes lenticular sandstone units building the hill tops. The very south of the northern section is built by sediments of the Ecca Group (Waterford Formation), which are alternating mudstones and very fine-grained, lithofeldspathic sandstones (Johnson et al. 2006).

The geology underlying the southern section is more complex with sediments of the Late Mesozoic Uitenhage Group in the north and the Palaeozoic Cape Supergroup in the south (CGS 2010). The sediments of the Uitenhage Group (Kirkwood Formation, Early Cretaceous) consist of coarse- to medium-grained lithic sandstones interbedded with thick siltstones and mudstones (Shone 2006). Conglomeratic horizons can occur. Dark siltstones and fine-grained sandstones of the Cape Supergroup (Bookeveld group, Early Ordovician to Early Carboniferous, ~500 to 330 Ma) cover the southern half of the southern transect

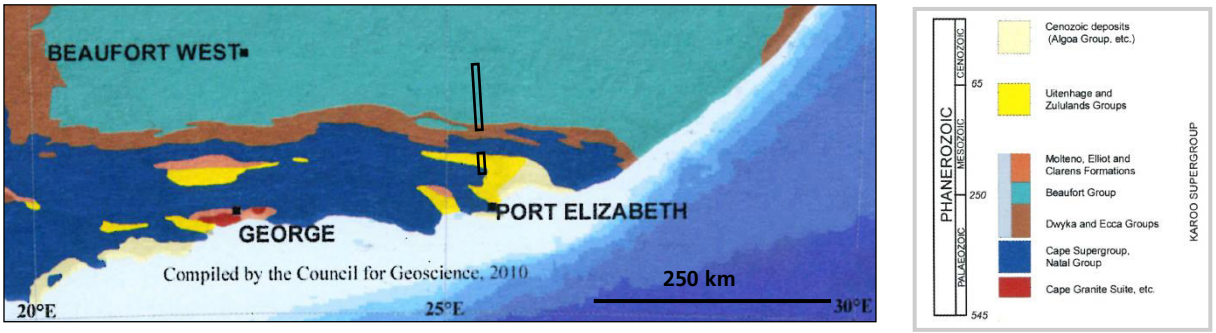


Figure 3-2: Subset of the simplified Geology of South Africa around the study area (CGS 2010, 1 : 1 000 000).

(Thamm and Johnson 2006). The sandstones are more resistant to weathering processes, and thus they produce positive topographic features which are here the softly undulating, vegetated hills that rise to the Grootwinterhoek Mountains in the south.

The sediments of the Karroo Supergroup strata (entire northern section) and the older Cape Supergroup (southern part of southern section) were deformed in the Cape orogeny (Late Permian to Early Triassic). This process tilted the strata and introduced faults striking approximately east-west and small local folds and thrusts (figure 3-3, Newton et al. 2006). Low grade metamorphism occurred during the compression as the rocks are located distal to the central collision zone far in the SW. This has resulted in shales and quartzites (Newton et al. 2006).

In the northern transect dolerite intrusions occur in the sedimentary strata. They formed as dykes which are connected to large flood basalt intrusions and associated with the break up of the supercontinent Gondwana (Early Jurassic, Duncan and Marsh 2006). An interconnected network of vertical dykes, horizontal sills, and sheets developed. These rocks today typically form the resistant hill caps comprising softer underlying sedimentary strata (GCS 2010, figure 3-4a and b). In the study area, the dolerites appear over a length of about 11 kilometers as separate patches covering mainly the hill tops or ridges.

Figure 3-3: Alternating mudstones and sandstones of the Karroo Supergroup (Beaufort Group) comprise the parent rock material of the predominant part of the northern transect area. Within the Cape Orogeny the strata were tilted and transformed to meta-sedimentary rocks

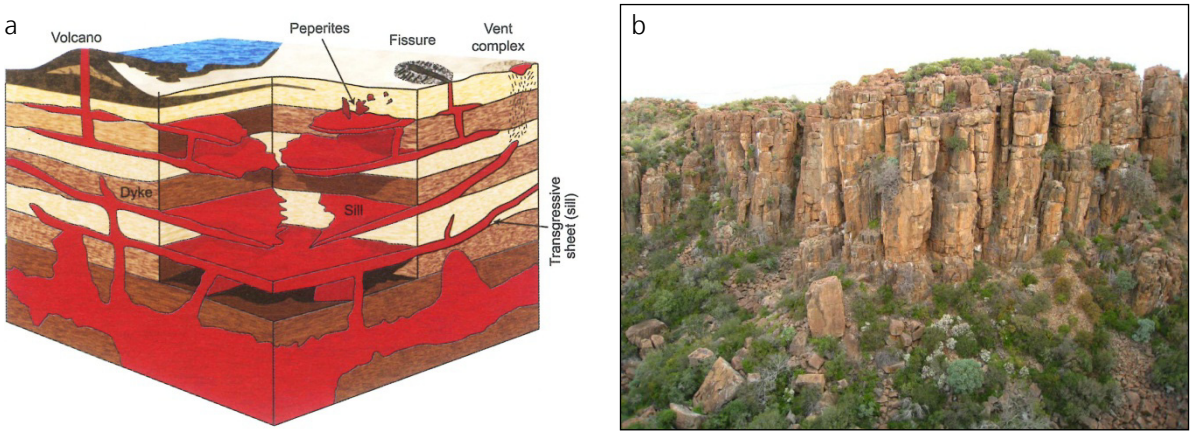


Figure 3-4: Early Jurassic magmatic intrusions as dykes and sills (a) divide the sedimental deposits into compartments (CGS 2010). Exposed piled dolerite columns in the Camdeboo National Park near Graff-Reinet (b).

3.1.1 Recent development of the ecosystem

The description of vegetation types and units follows the South African national classification according to Mucina and Rutherford (2006). Thus, the term Albany Thicket is used for the biome ecological unit (Hoare et al. 2006), the study area is located in. This unit was previously called „Valley Bushveld" (Acocks 1975) received biome status in 1996, and was called the Subtropical Thicket Biome (Low and Rebelo 1996). Vlok et al. (2003) classified 112 distinct types of Thickets within the Subtropical Biome, but his classification was not adapted in the national classification und the name was changed to Albany Thicket Biome. To reduce confusion this thesis will refer to Thicket vegetation in general and to the Albany Thicket Biome as biome ecological unit.

Pristine and transformed vegetation in the Albany Thicket Biome

In pristine conditions, the vegetation of the Albany Thicket Biome consists of a dense canopy (up to 70 %) of largely evergreen semi-succulent and thorny shrubs and low trees (0.5-3 m), making it a very unusual vegetation type for a semiarid environment (Vlok et al. 2003, Mills and Fey 2004a). The Thicket Biome shows unique structural and functional characteristics and its own highly biodiverse flora with about 20 % of the species endemic to it (Vlok and Euston-Brown 2002). The endemic leaf-succulent shrub *Portulacaria afra*, more commonly referred to as spekboom (Africaans for "fat tree" or "elephant food"), is an keystone species unique to Thicket vegetation (figure 3-5). *P. afra* can be locally dominant, with the vegetation type called Spekboom Thicket or Spekboomveld. Spekboom shrubs can switch between C_3 photosynthesis when water is sufficiently available and C_4/CAM photosynthesis during dry periods (Guralnick et al. 1984). This allows a highly efficient use of water and thus, a relatively high productivity and biomass for areas with limited rainfalls and periodic droughts (Guralnick et al. 1984). With this *P. afra* accounts for the fixation of large amounts of carbon in its biomass and the peripheral soils (Lechmere-Oertel et al. 2005).

The original extension and density of Thicket vegetation is primarily conditioned by local microclimate (frost

Figure 3-5: Shrubs of the locally dominant species *P. afra* are usually dense with thick branches and leafs (a, c). It produces small pink flowers usually in spring (b, photo taken in October). Re-growing spekboom after it has been browsed starts from the remaining stem and thick branches (d). Field photos were taken in 2009 (a-c) and 2011 (d).

free), rainfall, slope and aspect and prevailing fire-regime (Cowling and Vlok 2005, Vlok et al. 2003). Only where rainfall is low enough spekboom Thicket is able to displace other woody Thicket vegetation. In the second instance, associated abiotic factors (lithology, soil structure and a high nutrient availability) and biotic interactions (termites, small and large indigenuous mammal activity) determine the spatial distribution of Thicket clumps and solid Thicket within the Albany Thicket Biome (Lloyd et al. 2002). Spekboom Thicket is largely restricted to the semiarid valleys of the Eastern Cape Province, South Africa, that occur in the northern section of the study area (Lloyd et al. 2002). The pristine form comprises *P. afra*, *Rhigozum obovatum*, *Ptaeroxylon obliquum* and *Cussonia spicata* in large clumps that built a continuous mosaic with karroid shrubs (*Becium burchellianum*, *Walafrida geniculata*, *Pentzia incana*) and grasses (*Cymbopogon plurinodis*, *Aristida congesta*, *Eustachys mutica*) (Hoare et al. 2006). Locally, *Euphorbia caerulescens* (see figure 3-6e) builds a mosaic with Thicket shrubs and solitary trees (very often in the south of the northern transect). In the Sundays River Valley of the southern study section Thicket vegetation differs from Spekboom Thicket in having a sparse tree component (Arid Thicket, Lloyd et al. 2002). Woody shrubs (*Gymnosporia polyacantha, Rhigozum obovatum*) and a well established grass layer dominate the landscape, with only a sporadic occurrence of *P. afra* (figure 3-6f). The plains of the Sundays River are sparsely vegetated, while the foothills are covered by tall, dense Thicket. After periods of rain, changes in the vegetation occur most notably in the grass and herbs layer and woody shrubs (see figure 3-1b), while spekboom plants as evergreen species hardly show any alteration. For a detailed description of species and taxa of the Albany Thicket Biome please see Hoare et al. (2006) and Vlok et al. (2003).

Contemporary vegetation patterns are the effect of human activity and domestic herbivory (Hoare et al. 2006). Spekboom plants are consumed by both wild and domestic herbivores but are preferred by goats and highly vulnerable to disturbances. Intensive goat farming resulted in the loss of *P. afra* and the transformation of original dense Thicket to an open Savannah-like system, which is dominated by a layer of grasses, dwarf karroid shrubs and remnant trees (Mills et al. 2005b, Lechmere-Oertel et al. 2005, figure 3-6b). The removal of the original vegetation leads to the spread of certain endemic species, such as *Aloe ferox* and *striata* (although the first mentioned is an indicator species for Spekboom Thicket according to Vlok and Euston-Brown 2002, see figure 3-6g, h) and also allows for the invasion of alien plants (e.g. *Opuntia ficus-indica* in figure 3-6i and *Agave americana* see figure 3-6i, j). This process of ecosystem change is accompanied by a severe loss of biodiversity and ecosystem carbon stocks (Lechmere-Oertel et al. 2005, Mills and Fey 2003). Figure 3-6c to f show examples for the present vegetation patterns in both sections of the study area with locally varying levels of degradation. The removal of vegetation first causes a decreased enrichment of the soil surface with nutrients coming directly or indirectly from the vegetation and has wide impacts on the soil and its function. The exposure of the soil surface makes it vulnerable to rapid weathering, increases erosion risk and the tendency of soils to crust (see below). This sealing clay layer, formed by rain drop impacts, directly reduces soil infiltrability and water retention capacity, increases water runoff, and has further effects on the nutrient balance of soils. The net loss of carbon as a result of degradation was estimated to be approximately 4.0 kg/m^2 in soils to a depth of 50 cm and 4.5 kg/m^2 in biomass (Mills and Cowling 2006, Mills et al. 2005a).

Restoration efforts and cooperation

In this region more than 70 % of the vegetation units show a moderate to severe level of degradation (Lloyd et al. 2002). Since restoration of transformed Albany Thicket does not occur spontaneously after the herbivores are removed (Mills and Cowling 2010, Vlok et al. 2003), it can only be achieved by extensive restoration programs. A clear practical understanding of key inventories required to reverse Thicket

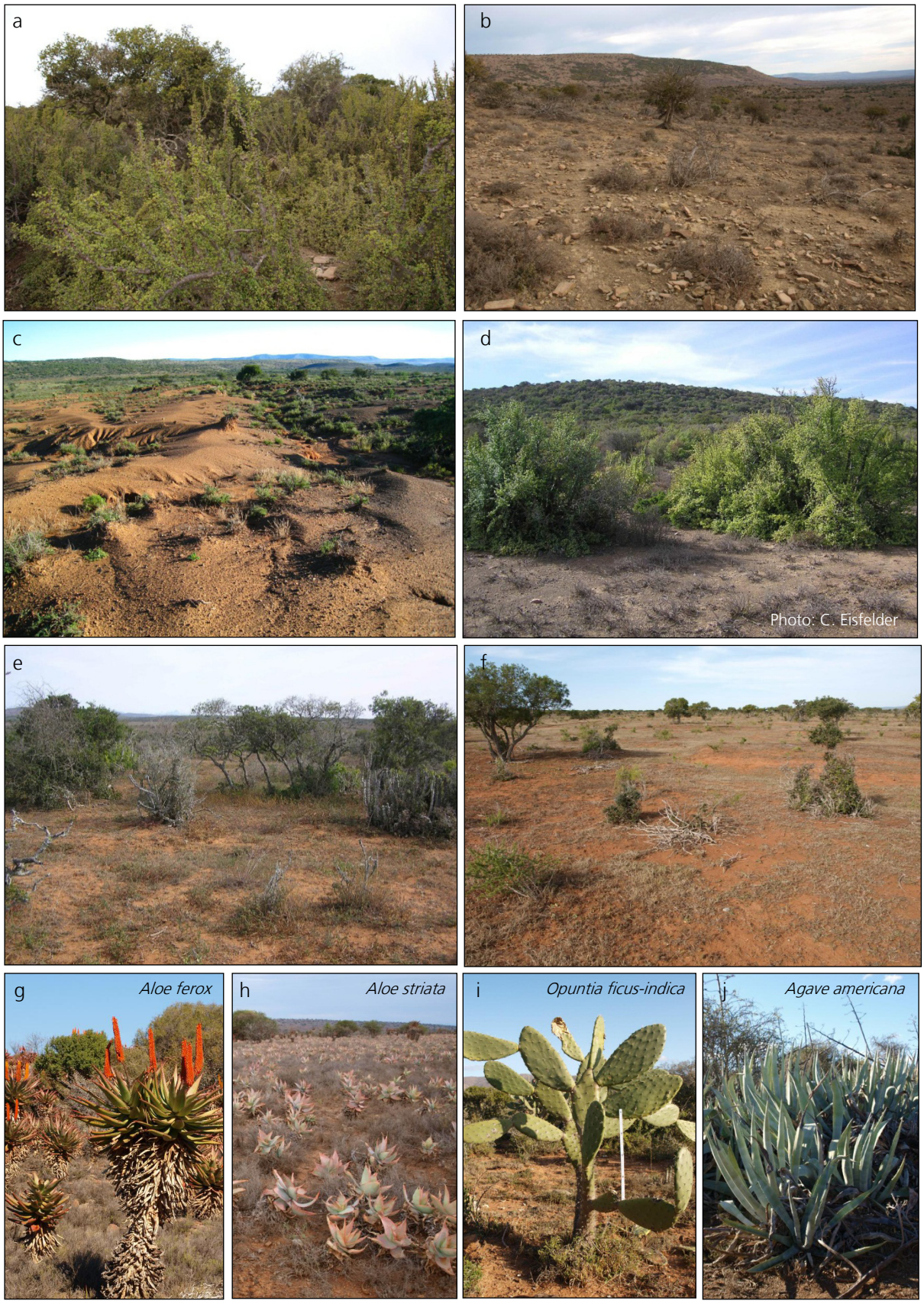


Figure 3-6: Field photos of different types and stages of Albany Thicket vegetation: Pristine (a) and transformed (b) Spekboom Thicket. Vegetation patterns of the northern (c, d) and southern study area section (e, f). Endemic plants and invasive alien species spread by the removal of vegetation (g-j). All photos were taken in July and October 2009, only c is from May 2011).

degradation is still in discussion but most approaches suggest *P. afra* as the critical component (Powell et al. 2006a and b), as it can grow from cuttings with a high rate of around 30 cm/yr and has an unusual ability to fix large amounts of carbon in semiarid environments. At present, this seems the only cost-effective method to enable the restoration of wide areas of degraded Thicket (Powell et al. 2006a), although this would not return the full complement of species to Thicket ecosystems.

Mills and Cowling (2006) investigated potential rates of carbon sequestration by planting of *P. afra* cuttings in a field experiment that took place only a few kilometers from the southern transect area. They found a total sequestration of 11 kg C/m^2 over 27 years, with an average rate of carbon sequestration of 0.42 ± 0.08 kg C/m^2yr (4200 kg C/ha yr). This rate is between 14 and 42 as high as gross rates generally given for dry and warm regions (100-300 kg C/ha yr according to Lal 2002). With the recreation of biodiversity, soils get stabilized, water retention capacity improves and landscape aesthetics valuable for tourism can be returned back to a non-degraded ecosystem as further benefits of ecosystem restoration (e.g. Marais et al. 2009).

As a product of several precursor projects, in 2004 the Subtropical Thicket Rehabilitation Program (STRP) was initiated by the South African Department of Water Affairs and Forestry to draft a conservation strategy for Subtropical Thickets. The program focused in particular on the potential role of carbon sequestration as a funding source for Albany Thicket rehabilitation (Mills et al. 2007, Marais et al. 2009, Powell 2009). The Clean Development Mechanism (CDM) implemented within the 1997 Kyoto Protocol provides the basis for a trade of carbon sequestrated in appropriate projects in developing countries for a fee (carbon credits). However, the requirements for projects to be accredited for the global carbon market are stringent, making the realization of such projects difficult. There are plans to acquire the certification under the Climate, Community and Biodiversity Assessment (CCBA) Standard. A number of local institutions acting under several integrated groups including university scientists, governmental and private institutions are working on this (e.g. Rhodes Restoration Research Group, ERC, PRESENCE, etc.).

The restoration of degraded Albany Thicket landscapes is currently is accomplished for carbon credits traded exclusively in the free market as implemented as a conclusive voluntary carbon offsets project under the Voluntary Carbon Standard (VCS). A pilot project is currently in action for one farm portion larger than 2500 hectares and located in direct neighborhood to the northern transect. It is densely planted with spekboom cuttings and will be kept isolated from further interventions (manmade and also by wild and domestic herbivores) for 30 years (see www.ercap.co.za).

The PRESENCE network (Participatory Restoration of Ecosystem Services and Natural Capital, Eastern Cape) is a valuable partner for this study. It is set up as a learning network that acts as a multi-stakeholder platform to provide transdisciplinary social-ecological research and supports the implementation of best management practices aiming in the restoration of "living landscapes" in South Africa (for more information please see www.livinglandscapes.co.za). PRESENCE activities are focused on the Albany Thicket Biome with a pilot region in the Baviaanskloof Mega-Reserve, approximately 150 km southwest of the study area, and an adjacent learning village. The PRESENCE network supported the field data acquisition for this study, provided geo-data information and provided a platform for discussions about the improved use of remote sensing techniques in Thicket research and possible applications of the outcomes of this study.

3.1.2 Predominant soils

The underlying sandstones, mudstones and less extensive also siltstones have resulted in the development of loamy and sandy soils (Cambisols, Luvisols and Regosols, see Kompter 2012, Mills and Cowling 2006, Mills et al. 2005b and others). In the southern section also calcareous soils occur. Soil development is limited due to the low annual rainfall amount. In addition it is assumed to be further reduced by accelerated sheet erosion in the last decades. The present soils are shallow usually with a thickness less than 50 cm. A soil survey in 2011, where in total 32 profiles were taken, revealed deeper average profile depths in the southern part (average depth of 52 cm in southern transect and 32 cm in the northern transect as average of 21, respectively 11 soil profiles, see Kompter 2012). This indicates a higher stage or a more intense soil development in the southern transect compared to the northern transect. Rock outcrops of sandstone and mudstones are prevalent in all topographic units (figure 3-7a and b). Soil colors are variant, though yellowish to brown soils dominate the northern and reddish soils the southern part of the study area (own research and Kompter 2012).

The 2011 soil survey further located thick soil horizons with depths over 1 m only in river plains and also in alluvial fans at the foothills. Their deeper structure is locally revealed at erosional scars attached to fluvial channels in the flat plains. The deep soils are sub-divided by cemented horizons (figure 3-7c), which are usually carbonate free. Only in the southern study area cemented calcaric horizons can locally be very thick and weather positively (Johnson et al. 2006, figure 3-7d).

Soil crusts

Under dense Thicket canopy a thick litter layer usually covers the dark, carbon-rich soils (Mills et al. 2005b). The exposure of soils when vegetation is diminished reduces the litter layer by reduced input and further enables the erosion of the organic topsoil layer and the formation of soil physical crusts. Soil crusts are formed by rain drop impacts on the uncovered soil surface that cause (1) a disintegration of aggregates at the surface and (2) a dispersion of the clay minerals, which are washed with infiltrating water into and

Figure 3-7: Unique features and structures of the soils of the study area (photo b of July 2009 and a, c and d of May/June 2011).
a: The outcrop of sandstone in a small runoff channel shows the extremely thin soil layer developed on a hill top (northern section), b: Erosional scar cut in layered mudstones within the northern section, c: Erosion structures attached to fluvial channels show cemented horizons (northern section), d: Cemented calcaric horizon in the sediments of the Kirkwood Formation (southern section).

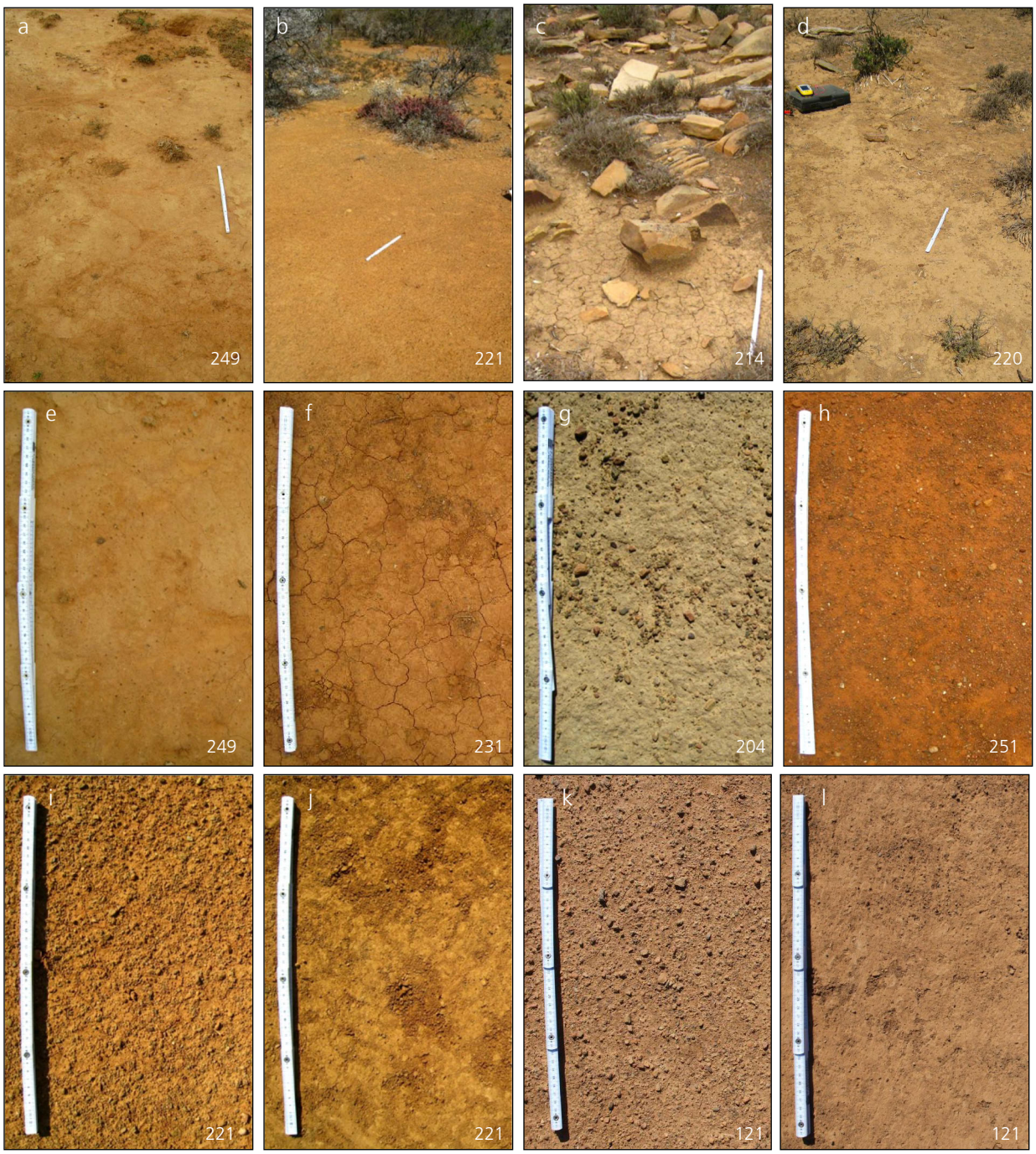


Figure 3-8: Soil surface conditions in different surroundings (a-d) and in detail (e-l, photos of July/October 2009, length of measuring stick of 40 cm, sample site numbers are given).
a, e: Nearly uncovered, strongly crusted surface. b, i: Small stones overlying a bare soil with crusted surface layer. c: Open spot of soil with drying cracks in a surrounding dominated by sandstone rocks (10-40 cm) and surface litter. d: Soil influenced by minor litter and minor residues of loose grass. e: Bare soil surface with few baked-in stones. f: Bare soil with drying cracks. g: Small stones (<0.8 cm) made of sandstones and mudstone overlying a grayish soil surface. h: Deep red soil with stones of several millimeters size. i-l: Two examples of soil surfaces where overlying stones were removed to reveal the uncovered soil surface. Small holes remain.

deposited into pores (McIntyre 1958). Finally, a thin surface layer of low permeability develops, sealing the soil surface by an extremely reduced infiltration rate (McIntyre 1958). Thus, low-intensity rainfalls result in strong surface water runoff and soil erosion. Crusting is also linked to soil organic carbon content as the organic compounds are necessary for aggregate stability. A decrease in soil carbon content directly causes

an increased risk of crusting (Mills and Fey 2003). Silt and fine sand are commonly cited as the main size fractions that promote crusting (Bradford and Huang 1992), making the study area soils highly vulnerable to crusting when exposed. Mills and Fey (2004b) investigated the relationship of vegetation cover and the tendency of different South African soils to crust, with one of the sites in the Albany Thicket Biome (soils at eight fence line contrasts were sampled). They found that infiltration in Savannah soils, in which are similar to degraded Thicket areas, was about half and crusting tendency was correspondingly higher than in undisturbed Thicket sites.

Biological soil crusts are absent in the northern transect section, but do occur only frequently in the most northern part of the southern section in areas that are periodically irrigated for stock farming (properties of the Correctional Service, Kirkwood).

Appearance of soil surfaces

The surfaces of the present soils are heavily influenced by vegetation cover and underlying geology. Constrained by this, they occur as bare surfaces -crusted or uncrusted-, with a litter or stone overlay (figure 3-8). Strongly crusted soil surfaces occur where vegetation is rare and soils are mostly uncovered, often at alluvial fans and erosion ridges in fluvial plains existing throughout the study area (figure 3-6c). As relicts of the crust formation process and depending on the prominent soil size fractions, small stones (mm to cm) usually coated with soil material can cover soil surfaces (figure 3-8i, k). Close to vegetation litter and organic residues are prevalent, but due to the semiarid climate the amount is low and leaves are small (figure 3-8c, d). Stones overly soil surfaces mostly where the soils are only thinly developed and the influence of the underlying geologic sequences is high. On massive bedded sandstones often rock fragments remain that are several tens of centimeter (figure 3-8c), while in finer layered sandstone units (or when the surface is strongly weathered) smaller sandstone fragments, millimeter to centimeter in size, can cover the soil surface in various proportions. Soils developed on mudstones are mostly covered with fine stones (1 to usually 5 mm), as weathering product of the mudstones. The removal of an existing stone coverage for spectral measurements in some cases resulted in small wholes remaining in the surface (see section 3.2.1, figure 3-8i-l).

3.2 Data acquisition and processing

The following section deals with the datasets available and the pre-processing workflow applied in this study. Further details of pre-processing of laboratory and field spectra and HyMap imagery are included in appendix B. A comparison of the characteristics of the available spectral datasets is given in table 3-1.

Table 3-1: Specifications and comparison of spectral datasets.

	ASD spectroradiometer measurements Portable device for point measurements, Analytical Spectral Devices Inc.	**HyMap airborne image data** Airborne based scanner system (whiskbroom), HyVista Corporation, Australia	**EnMAP satellite image data** German satellite mission, launch planned for 2016 simulated data generated by EeteS (Guanter et al., 2009)
Ground sampling distance	covered footprint of approx. 50 cm	Resampled pixel size of 3.3 x 3.3 m	Resampled pixel size of 30 x 30 m
Spectral coverage[1]	400 – 2500 nm	427 – 2486 nm	420 – 2450 nm
Spectral bands[1]	2150	128	228
Full width half maximum	1.4 – 2 nm	15 – 20 nm	8.1 (±1)– 12.1 (±1.5) nm
Spectral sampling interval	1 nm	13-17 nm	6 – 10 nm
Field of view (FOV)	25° (bare fiber)	Total FOV 61.3°	FOV: ±1.06°
Instantaneous field of view (IFOV)		Across track IFOV: 2.5 mrad Along track IFOV: 2.0 mrad	IFOV: 7.63"
Signal-to-Noise ratio (SNR)		VIS/NIR: 1200:1 NIR (1400-1900): 900:1 SWIR (1900-2400 nm): 600:1	VIS/NIR: 500:1 SWIR: 150:1
Further information	www.asdi.com	www.hyvista.com	www.enmap.org
Applied pre-processing steps	• Correction of SpectralonTM reflectance and detector jumps • Averaging of singular measurements • Filtering • Field dataset: removal of water bands	• Orthorectification • Atmospheric correction • Band selection • Band interpolation and spectral polishing • Leveling of flight lines • Iterative adaptive spatial filtering • Masking of artificial objects	• Spatial resize • Band selection • Band interpolation and spectral polishing

[1]Before pre-processing

3.2.1 Field and laboratory reference data

During two sampling campaigns a total of 163 natural, non-agricultural sites were sampled for ground truth (96 in June 2009 and 67 in September/October 2009). Owing to a lack of detailed soil or geologic information of the area of interest prior to the sampling campaigns, preparatory analysis to support the selection of appropriate sampling sites was not possible. Thus, the sampled areas are randomly distributed over the two study areas taking into account all topographic units and geological parent materials visible on-site. In the southern area there are 44 sites in the Kirkwood vicinity (approx. 92 km^2) and 119 in the northern part near Somerset East (approx. 235 km^2). Figure 3-9 shows the location of the sampled sites within the two transect parts. Most of the samples are taken in areas where bare soil is the dominant land cover fraction. However, some isolated sparsely vegetated spots in vegetation dominated regions are also included. Using this sampling scheme, all regions in the study area that are reachable by foot or by car in an adequate time are represented. Hence, it is assumed that the sites cover the complete variance of soils present.

In the selected areas, the surveyed sites are chosen to represent the typical conditions of the surrounding area. Location specific information (GPS coordinates, terrain information, land use, etc.), field spectra of the exposed soils, soil samples, photographs and additional information characterizing location and surface conditions are collected for the sample site with its center spot and its surrounding in two raster sizes. A raster size of 1 x 1 m is applied for a small-scale description of the center spot where spectral measurements and soil samples are taken, including a detailed description of the surface conditions (especially soil crusts

and stone coverage). To take into consideration the direct surrounding of the center spot, the fractional land cover is assessed in an area of 10 x 10 m around (see below). This is done also in regard to mixing phenomena caused by HyMap's pixel size. Table A-1 in the appendix summarizes the information collected at each sample site.

Field spectral measurements

Spectroradiometric measurements are taken using a portable FieldSpec Pro 2 spectrometer (Analytical Spectral Devices, Inc.). The instrument measures radiance in the reflective solar domain from 400 to 2500 nm in sampling intervals between 1.4 and 2 nm, interpolated to 1 nm bandwidth. Measurements are taken with the bare fiber, resulting in a field of view of 25° and a maximum in collected light. They are directly converted to reflectance using a Spectralon™ white panel as a reference. In the field, five measurements are taken for each site with an internal sampling rate of 25 single measurements. Measurement height is kept constant at 1.15 m, resulting in a footprint of approximately 50 cm diameter. For sites where the undisturbed bare soil fraction at the center spot is very small, an alternative height of 0.5 m (22 cm diameter footprint) is applied. If small stones (mm to cm range) covered the surface, measurements are carried out first on the unchanged surface and then with the stones carefully removed. Of the 163 sites, 125 include field spectral measurements, where the remaining sites do not owing to poor weather conditions.

Soil sampling and collection of land cover information

Soil samples of the representative site include the top layer to a maximum depth of 0.5 to 1.5 cm, because the very first soil layer contains the material triggering the reflectance signal recorded by field or imaging spectrometers. If physical crusts are present at the surface, the sampling is limited to the crust. Two equal soil samples are taken from an area of at most around 1 m^2 at the center spot of each site. Each one sample is used for soil chemical analysis and the other for laboratory spectral measurements.

The fractional land cover is assessed in a 10 by 10 m area around the surveyed center spot. The fraction of bare soil, covering stones, litter and vegetation according to life form (grass, herbs, shrubs and trees) and photosynthetic stage (green, mixed, dry) is collected, denoted in percent. The grading scheme for vegetation types is adapted from BIOTA biodiversity monitoring programs operating in Africa (see www.biota-africa.org). Only spekboom shrubs are discriminated at species level, as the possible identification of these plants in hyperspectral imagery is an aspect of further research (see Table A-1).

Soil profile information

In 2011 from 32 sites within the study area (21 located in the northern and 11 in the southern transect), soil profile measurements are collected. This is done as basis for a spatial identification of soil types and finally to evaluate the homogeneity of soils within the study area. In addition to the site specific information, soil and surface description and land cover assessment as given in Table A-1, soil samples are collected in different depth intervals with a soil auger up to the maximal depth that can be reached (surface, 0.5/1.5-5 cm, 5-15 cm, 15-25 cm, 25-45 cm and 45-65 cm). Samples from each depth are mixed and homogenized before subsamples for chemical analysis are taken.

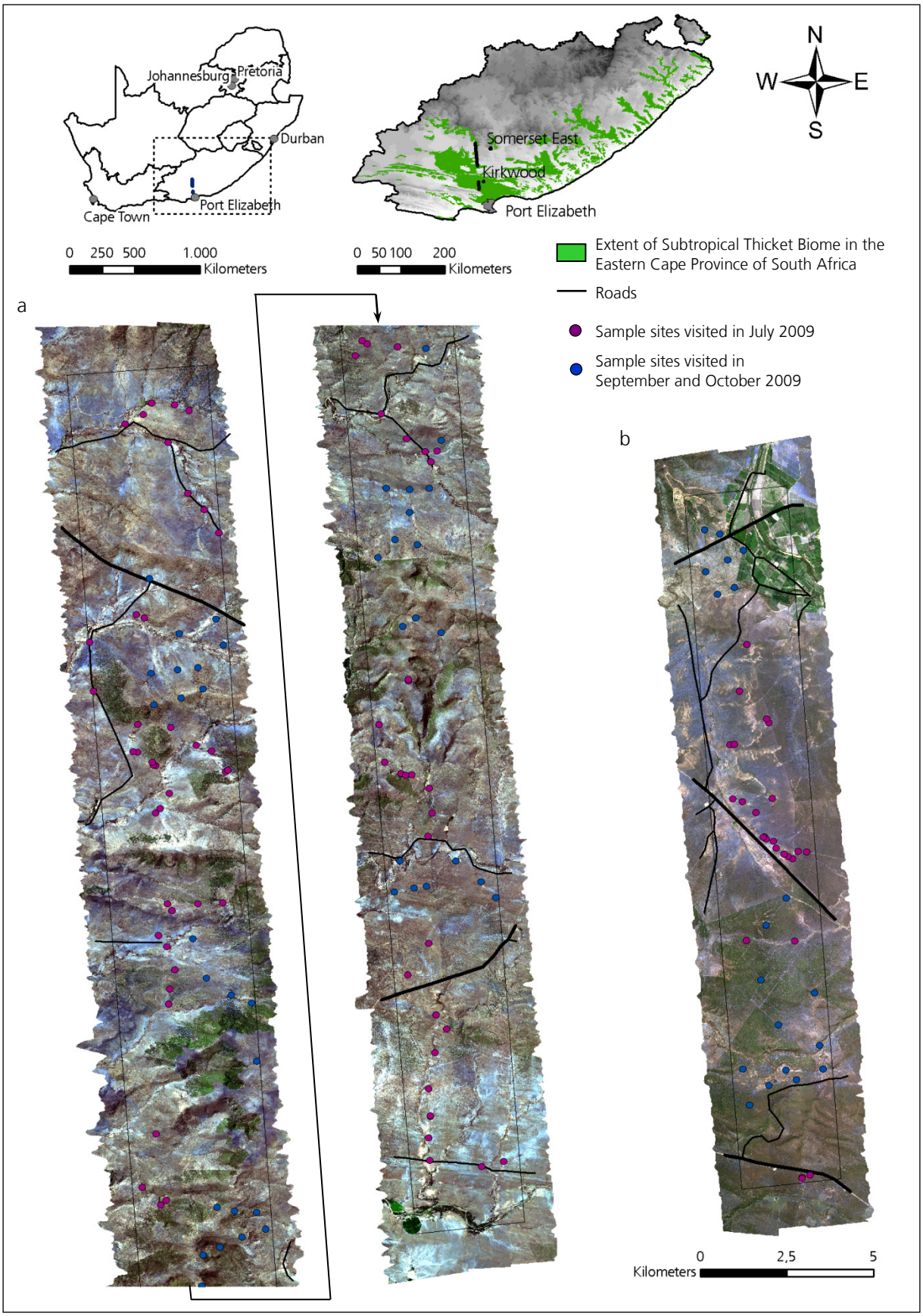


Figure 3-9: Location of ground reference plots in the northern (a) and southern part (b) of the study area located in the Eastern Cape Province, South Africa, that are collected in the two field campaigns. The HyMap imagery acquired in 2009 is used as background.

Soil quantitative chemical analysis

Air-dried soil samples are grinded and sieved by 2 mm, prior to be analyzed for organic carbon, iron oxides and clay using methods consistent with other studies. The Walkley-Black method is used to determine soil organic carbon content (Walkley and Black 1934). A citrate-bicarbonate-dithionite extraction method (CBD) described by Mehra and Jackson (1958) is applied for the determination of free iron oxides. This selective extraction technique determines the concentration of the most common pedogenic iron oxides goethite and hematite based on their solubility under specific conditions (Qafoku and Amonette 2006). This content is further referred to the content of free iron oxides. Particle size is determined in five fractions by pipette analysis (Gee and Bauder 1986). Chemical analysis is provided by the Central Analytical Facility as division of the University of Stellenbosch, South Africa. The accuracy of replicability is given with 0.5 % for each method.

Laboratory spectral measurements

Soil samples are spectrally measured in the laboratory under controlled conditions to generate a reflectance dataset that allows studying the spectral characteristics of soil properties without influences by atmospheric conditions, illumination, surface conditions and radiation reflected by neighboring material as are present in an outdoor environment. Measurements are made in the DLR spectroscopic laboratory, which is internally covered with matt black pond foil. This foil was tested to have a continuously dark reflection and no significant absorption features (Figure B-2a in the appendix).

The soil samples used for laboratory spectral analysis are pre-treated the same way as for chemical analysis. After drying the soils at room temperature, the soil material is broken down into its natural particles to dissolve aggregates resulting from soil crusts. Sieving by 2 mm removed small stones and larger plant residues that have not yet decomposed to soil organic matter. The soil samples are measured in the laboratory using the same ASD FieldSpec Pro 2 spectroradiometer used in the field. The measurement set-

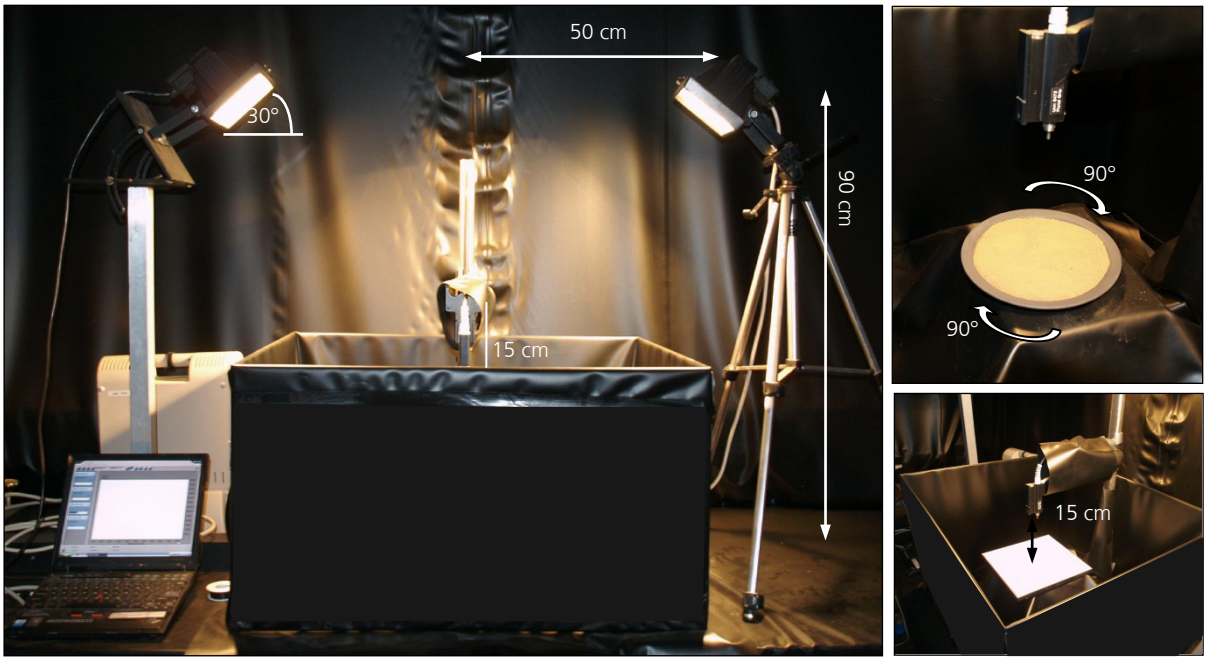


Figure 3-10: Set-up for laboratory spectral measurements (Photos: E. Kompter).

up applied in the laboratory is shown in figure 3-10. Two Quartz halogen lamps with a power of 300 W each illuminate the sample at a zenith angle of 30°. The measurements are taken using a bare fiber (field of view of 25°) with a distance of 15 cm between fiber and sample resulting in a measured spot of about 14 cm in diameter. A dish covered with a mat black 3M varnish is used as tray to minimize reflection of the surrounding on the measurement (see Figure B-2b for a reflection spectrum of the tray). After bringing up the soil powder on the dish, the surface is carefully smoothed. For this, the powder is covered with a paper and slightly compressed with a flat plate. To reduce illumination effects of the soil surface, after five measurements with an internal spectral resampling rate of 80, the sample is turned by 90°. Thus, in total 1600 single measurements (4 x 5 x 80) are resampled to collect an average laboratory spectrum for each of the 163 samples.

Processing of field and laboratory spectra

Field and laboratory spectrometer datasets are corrected for the specific reflectance of the applied Spectralon™ reference panel (reflectance spectrum see Figure B-1 in the appendix) and spectral jumps between the three ASD sensor detectors, which is done additively taking the middle detector (SWIR-1) as

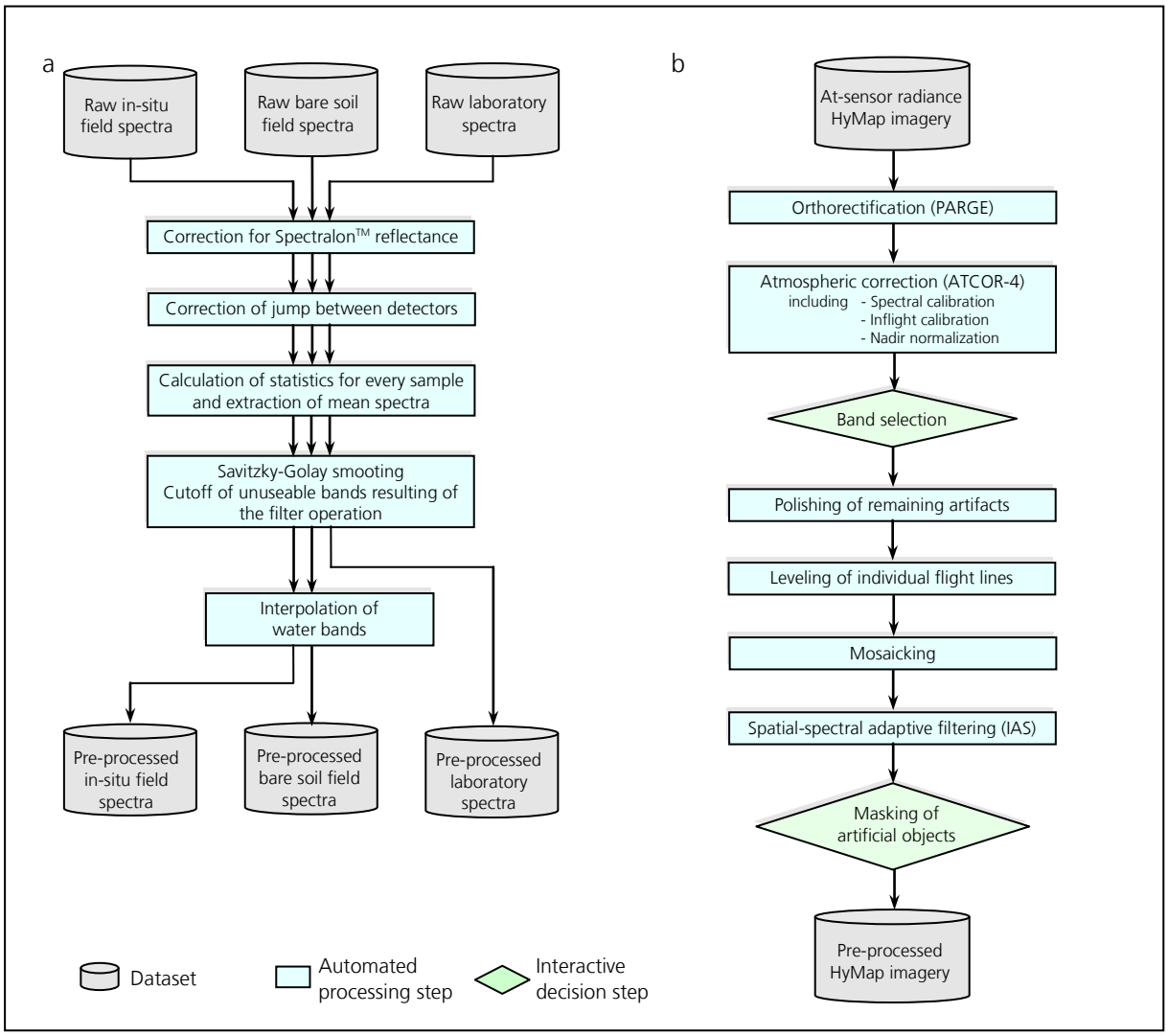


Figure 3-11: Overview of processing steps applied for the pre-processing of the three soil spectral libraries measured in the field and laboratory (a) and for the airborne HyMap hyperspectral imagery (b).

a reference. For all samples mean spectra and statistical deviations are calculated. Savitzky-Golay filtering (Savitzky and Golay 1964) is applied to the averaged spectra with a filter size of 32 and a polynom degree of 4 to remove small instrument inaccuracies and remaining noise. For the field dataset, ranges of atmospheric water bands are interpolated from 1345 to 1445 nm and 1800 to 1960 nm. Signals beyond 2450 nm are excluded due to increasing noise. Figure 3-11a gives an overview of the pre-processing steps for spectra measured in field and laboratory.

Finally, there exist four point spectral libraries of field and laboratory measurements: (1) the mean spectra of the 125 field sites of the in-situ unchanged soil surface eventually including small overlying stones (< 1 cm) referred to as "in-situ field spectra". (2) Mean field spectra of the same 125 sites that include predominantly bare soil measurements ("bare soil field spectra"). For 53 plots small stones overlaid the bare soil surface and can be removed for a second measurement. All other plots already show an uncovered bare soil surface (43 plots) or removal of stones was not possible because they were deeply mixed with the soil (29 plots). (3) "Other field spectra" like vegetation samples including spekboom, stones, etc. and (4) "laboratory spectra" of all 163 sampled sites.

The range of all spectral datasets used in this study is fixed from 450 to 2450 nm. This range is covered by common spectrometers, it allows the application of a Savitzky-Golay filter of a large kernel size to field and laboratory spectrometer measurements and also covers HyMap's spectral range after the removal of poorly performing bands (resulting in a range of 457 to 2456 nm). The final spectral libraries of point spectral data used for the calibration of prediction models are resampled to the chosen HyMap resolution of 116 bands (see below) to enable a direct comparison of ASD measurements and HyMap imaging spectrometer data.

3.2.2 Airborne hyperspectral imagery

High quality hyperspectral image data of the HyMap sensor were acquired over the study area on 14th October 2009. The HyMap sensors operated by HyVista Corporation, Australia, scan the 440 to 2500 nm spectral region with 126 bands, resulting in average spectral resolutions between 13 and 17 nm depending on the wavelength range. Specified signal to noise ratio (SNR) is indicated to be higher than 600:1. For detailed specifications of the HyMap sensors it is referred to Cocks et al. (1998). The overflight was completed during cloudless weather conditions over a 1.5 hour period around solar noon. Flight altitudes are chosen about 2050 m above ground for the northern and about 1700 m for the southern part. This results in a ground resolution of 3.3 m for the entire dataset. Northern and southern part of the study area are recorded separately in three flight lines of 57.2 km and 22.5 km, respectively. Total coverage of the two areas is around 320 km^2. Table B-5 in the appendix lists the flight parameters of all flight lines. Appendix B.2 gives details of the processing of the HyMap imagery.

Processing of HyMap data

When aiming for quantitative analysis of hyperspectral image data, the correction of atmospheric attenuation is critical as it may leave features that can mistakenly be held for soil spectral features or influence the determination of feature properties (Ben-Dor et al. 2008). Thus, the at-sensor radiance data are atmospherically corrected and pre-processed according to DLR's recommended workflow. Pre-processing steps include an accurate geocoding of the image data, which is due to the operation of the sensor from a non-stabilized platform. Orthorectification is performed using the PARGE procedure for geo-

correction of airborne data (Schläpfer and Richter 2002). Besides the internal GPS data during data acquisition, it uses ground control points selected based on Google Earth and a medium resolution digital elevation model available from the SRTM mission (spatial resolution of 30 m). The HyMap imagery is projected onto an UTM coordinate system (zone 35 South, datum WGS-84). Typical accuracy of this procedure is usually better than ± 3 pixels (approximately 10 m). Atmospheric effects such as scattering and absorption as well as terrain illumination effects are corrected with the ATCOR-4 procedure (Richter 2010b). This includes inflight calibration which is based on field reference measurements to enhance the accuracy of spectral signals. A large homogenous bare soil area is used for correlation of ground and airborne measurements due to missing invariant artificial targets in both parts of the study area. Nadir normalization, as implemented in the ATCOR workflow, is applied to account of BRDF effects and reduce cross-track illumination effects. Of HyMap's 126 bands, 11 are classified as reduced quality and removed. These include the first two bands at 427 and 442 nm, the band at 889 nm as overlap between the first two detectors, six bands influenced by water absorption (1956 to 2050 nm) and two bands at the far end of the short wave infrared range (2471 and 2486 nm). The bands close to the detection borders are removed due to the low level of incoming radiation and thus lower SNR. All subsequent work is applied using the remaining 115 bands. Small remaining spectral artifacts are locally smoothed using a 3 x 3 filter kernel (spectral polishing, see Table B-8 in the appendix). They occur most notably in flight line 2 but also extenuated in all other flight lines.

The three individual flight lines of each section are leveled to each other with an approach described by Rogge et al. (2012), which corrects for brightness and radiometric differences (e.g. resulting of atmospheric correction) across the flight lines. This resulted in a seamless mosaic of the three HyMap flight lines that is used as basis for the generation of quantitative maps. Appendix B.2 gives details of the leveling procedure.

As a last pre-processing step, the HyMap imagery is spatially filtered with an iterative adaptive spatial filtering approach (IAS, Rogge and Rivard 2010). The IAS minimizes random noise and local intra-class spectral variability while retaining edges. This supports the automatic extraction of image endmembers (see section 5.2.1). The filter iteratively smoothes pixel spectra according to their spectral similarity to the adjacent pixels. The filter is designed to adapt the filter kernel and the weights given for the filter operation according to the spectral similarity of the pixels in the filter kernel. Appendix B.2 gives technical details of the filter operation. Figure 3-12 illustrates the effect of the IAS filter operation on a portion of the farm Gonakraal showing different land cover conditions (see section 6.2 for overview of location and characteristics of this subset). Figure 3-12a gives an example of a narrow river bed as a highly variant area. Here, pixel spectra are not smoothed because the contrast, i.e. spectral variance, between image pixels is high. Despite that, the effect of the smoothing operation is highly visible in uniform areas (figure 3-12b), where spectral variance between pixels is low and it can be assumed that differences between pixel spectra are mostly a result of random noise. Due to the adaptive filter kernel, singular pixels showing a high spectral contrast to their surroundings, such as isolated trees, are retained by the filter operation (figure 3-12c). A further influence or a manipulation of finally delineated soil information resulting from this filter operation can be neglected as several tests indicate.

Clearly visible artificial objects and rural settlements like farm buildings, barns, water ponds, etc. are masked before further subsequent analysis as they can influence endmember selection or unmixing statistics. Figure 3-11b provides a summary of the pre-processing steps of hyperspectral imagery.

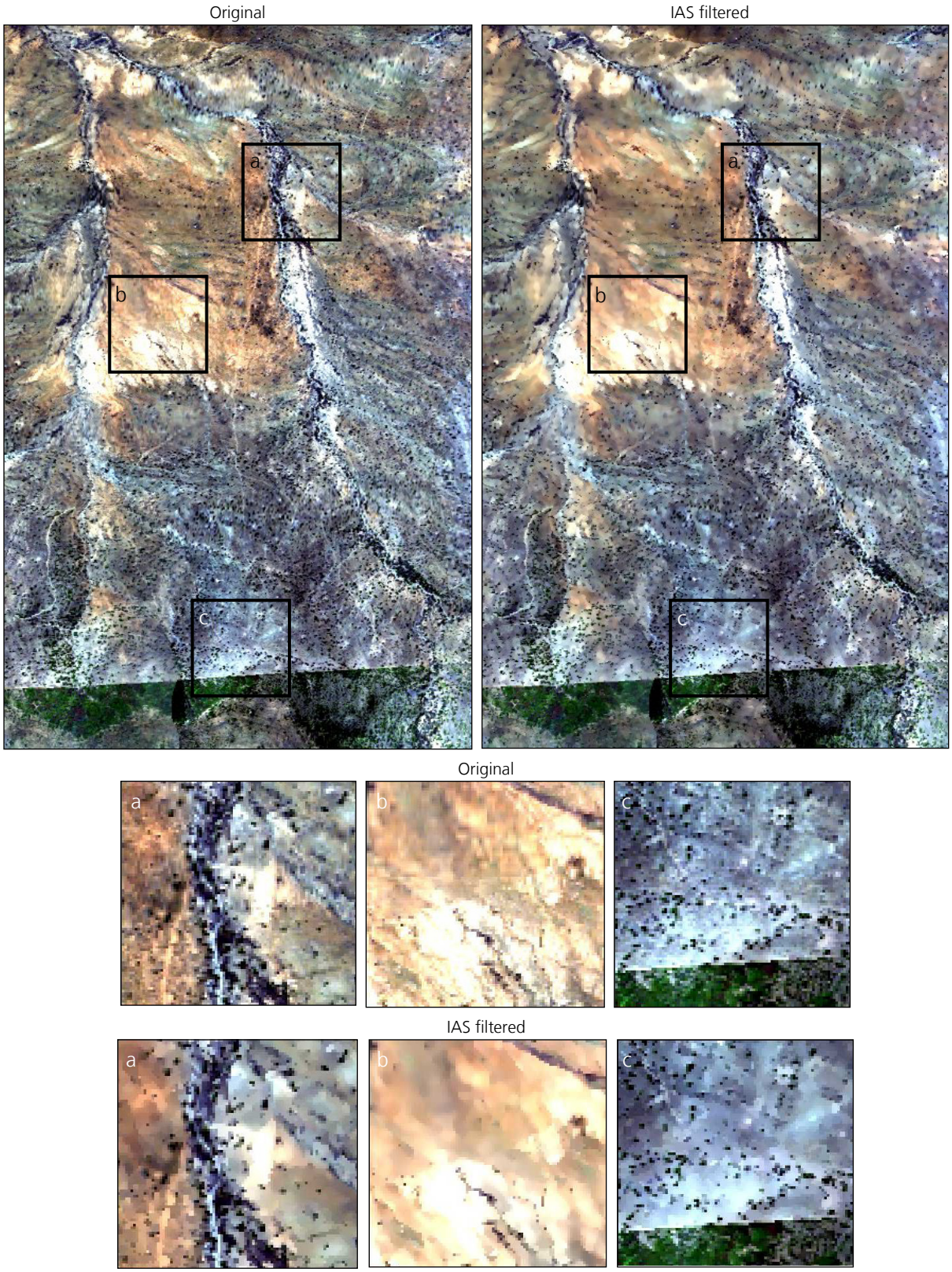


Figure 3-12: Effect of iterative adaptive spatial filter (IAS) applied to the HyMap hyperspectral imagery of the South African study area (subset of the farm Gonakraal, see section 6.2).

3.2.3 Simulated EnMAP datasets

The launch of the EnMAP satellite as German hyperspectral satellite mission is scheduled for 2016. As it will provide a new generation of spatially extensive high quality hyperspectral imagery on a timely and frequent basis, it is of further interest in this study to develop the approach for a possible application to the EnMAP data. Based on the HyMap imagery, simulated data with the specifications of the future EnMAP satellite are generated with the Image Simulation Tool for the EnMAP mission (EeteS) developed by the German Research Centre for Geosciences (Guanter et al. 2009). It was developed for an optimization of the instrument specifications by quantifying the effects of different instrument and environmental parameters on scene characteristics. The scene simulator generates EnMAP-like data in an automatic way under a set of user-driven instrumental and scene parameters, among others based on airborne hyperspectral imagery. A detailed description of the simulation tool and the processing steps is given in Guanter et al. (2009). This tool is used to simulate realistic EnMAP data for two subsets of the northern part of the study area covering areas of 1.8 x 4.4 km (subset farm "Gonakraal") and 5.5 x 6.6 km (subset farm "Wintersfontein"), respectively.

The simulated datasets are delivered as atmospherically and geometrically corrected level 2 data. Pixels at the direct border areas are excluded as they show a mixing signal influenced by the background signal which results of the simulation process. Overlapping bands are reduced and spectral bands around the large atmospheric water disturbances are interpolated. Remaining artifacts due to atmospheric water vapor are interpolated in three NIR regions as they may influence the spectral determination of iron oxides. The entire image spectra are spectrally smoothed with a 3 x 3 band sized filter. Additionally, the first six bands are excluded as this spectral range is not covered by the HyMap sensor. This pre-processing steps results in simulated EnMAP data with a spectral coverage of 455 to 2446 nm with 198 spectral bands that are used for further analysis. Details on the individual pre-processing steps are given in appendix B.3.

3.3 Qualitative description of input datasets

This section gives an outline of the quality assessment of the available datasets. Statistics are given for soil chemical analysis, the three soil spectral libraries, the airborne HyMap imagery and simulated EnMAP data. The qualification of the data for analysis aiming at the spectral feature-based delineation of soil constituents is assessed. This is relevant for the evaluation of the final results.

Analysis of spectra measured in field and laboratory

The different measurement conditions are reflected in the statistics of the spectral libraries (figure 3-13). For example, the mean reflectance of the spectra measured in the laboratory is higher than of the two field datasets owing to laboratory sample preparation, e.g. homogenizing and sieving, that minimizes shadow effects caused by soil aggregates or extraneous material (figure 3-13a). For the in-situ field measurements, average soil reflectance is reduced by small stones overlying the soil surface. Bare soil field spectra in particular present large extremes in maximum and minimum spectra (figure 3-13c). Where physical soil crusts are present, the smooth soil surface in combination with a reduced size of the particles at the soil surface increases soil reflectance. Thus, the reflectance range of individual samples measured in a field environment exceed the reflectance range of the corresponding laboratory spectra. The controlled

laboratory conditions result in low standard deviations, whereas the spectral variance of the field measurements is stretched by influences such as surface condition and illumination effects (figure 3-13b, c). A higher standard deviation of this dataset compared to corresponding bare soil field measurements shows that the presence of stones also introduced a further variability to the measurements.

The field and laboratory spectra show typical spectral features inherent to most soils, such as features around 500 (iron oxides), 900 (iron oxides), 2200 nm (clay) and 2350 nm (C_{org} and clay). Additional features are present but less pronounced or they do not appear in every spectrum.

Chemical reference data

Figure 3-14a provides the statistics of the soil chemical analysis. The analysis shows the presence of sandy and loamy soils with generally low organic carbon contents. Overall the histograms show that the distributions are asymmetric and skewed to lower concentrations.(figure 3-14a). The C_{org} contents are relatively low with an average of 1.21 % C_{org}. Only two samples exceed 3.5 % and show high contents of 5.7 and 5.9 %, with the first sample taken in an area affected by droppings of grazing animals, whereas the second sample site had a partial grass covering the bare soil surface that may have caused an increased

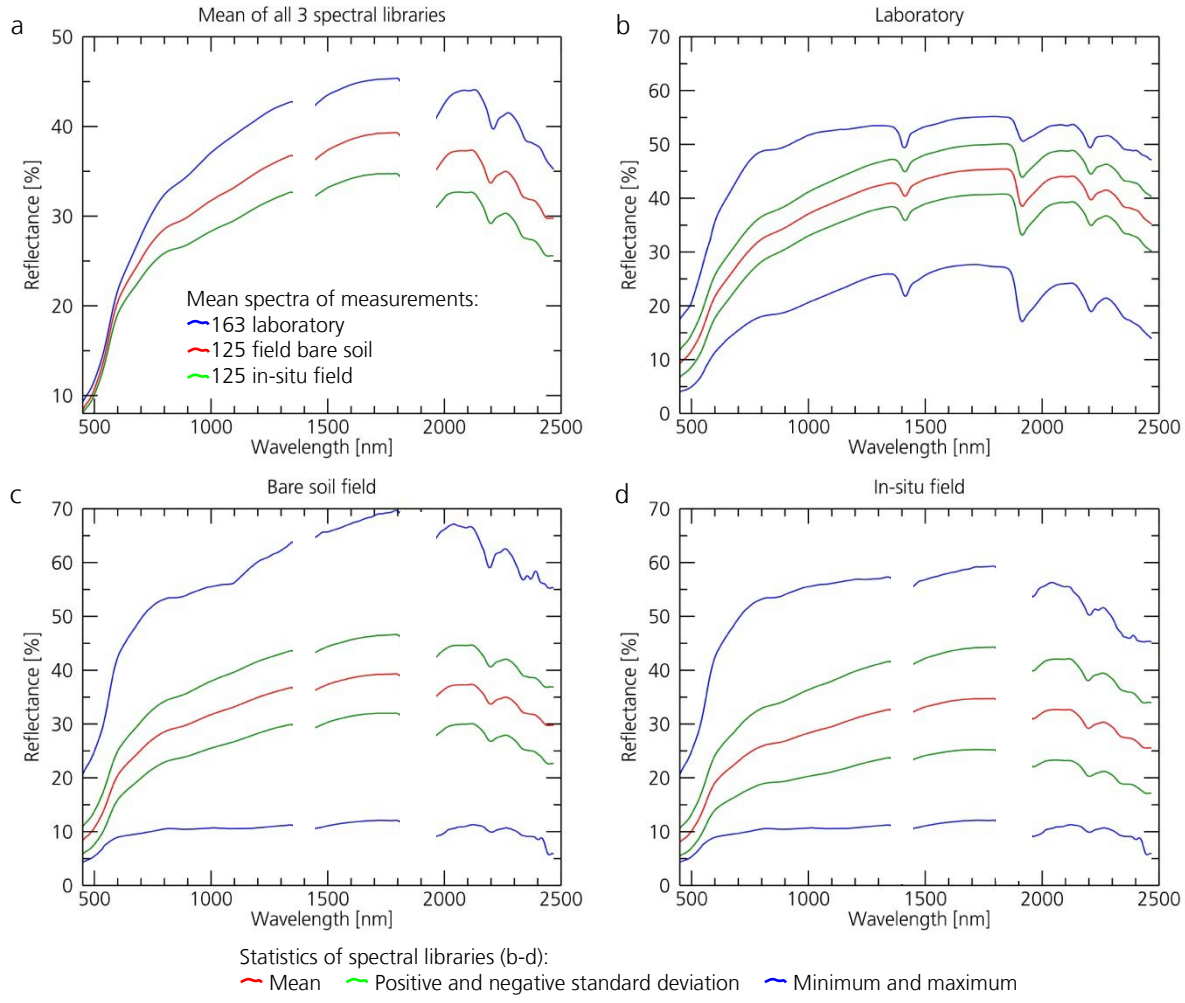


Figure 3-13: Overview of soil spectral datasets with mean spectra of all three spectral libraries (a), and the statistics of laboratory (b), field bare soil (c) and in-situ field measurements (d).

carbon input to the soil. The measured concentrations are below average contents for grasslands (about 7 % C_{org}) and also for agricultural soils (about 2 % C_{org}) which were given by Kuntze et al. (1988). Other studies carried out in the Albany Thicket Biome report soil organic carbon concentrations about 3.7 % in open degraded areas and about 5.6 % C_{org} under a plant canopy of *Portulacaria afra* (Mills and Fey 2004a). Iron oxides, which primarily are the result of weathering processes on the underlying sand- and mudstones, range up to 10.6 %. No studies are known where soil iron oxides concentrations were measured in Thicket soils. Schachtschabel et al. (2002) report iron oxides contents in most soils being in the range between 0.2 and 20 % Fe_d. Correlations between the analyzed soil parameters are low (figure 3-14b), though a weak correlation (R^2 of 0.20) exists between clay and iron content. This is often reported since pedogenic iron oxides mainly occur as clay sized particles (Qafoku and Amonette 2006). The detailed analysis results of each sample are listed in Table A-2

The texture triangle (figure 3-15) shows the presence of sandy soils with different amounts of loam (FAO 2006). Within this range, field surveys show that the distribution of clay and sand in overlying soils are directly related to the underlying parental material. Clay and sand concentrations increase where mudstones, respectively sandstones, occur close to the surface. However, the differences are not distinct enough to define two texture classes.

All samples were used for further processing, even though the chemical contents of some samples differ from the distribution of the rest of the population (see figure 3-14a). The population is tested for outlier samples in cross-validation as part of the model development.

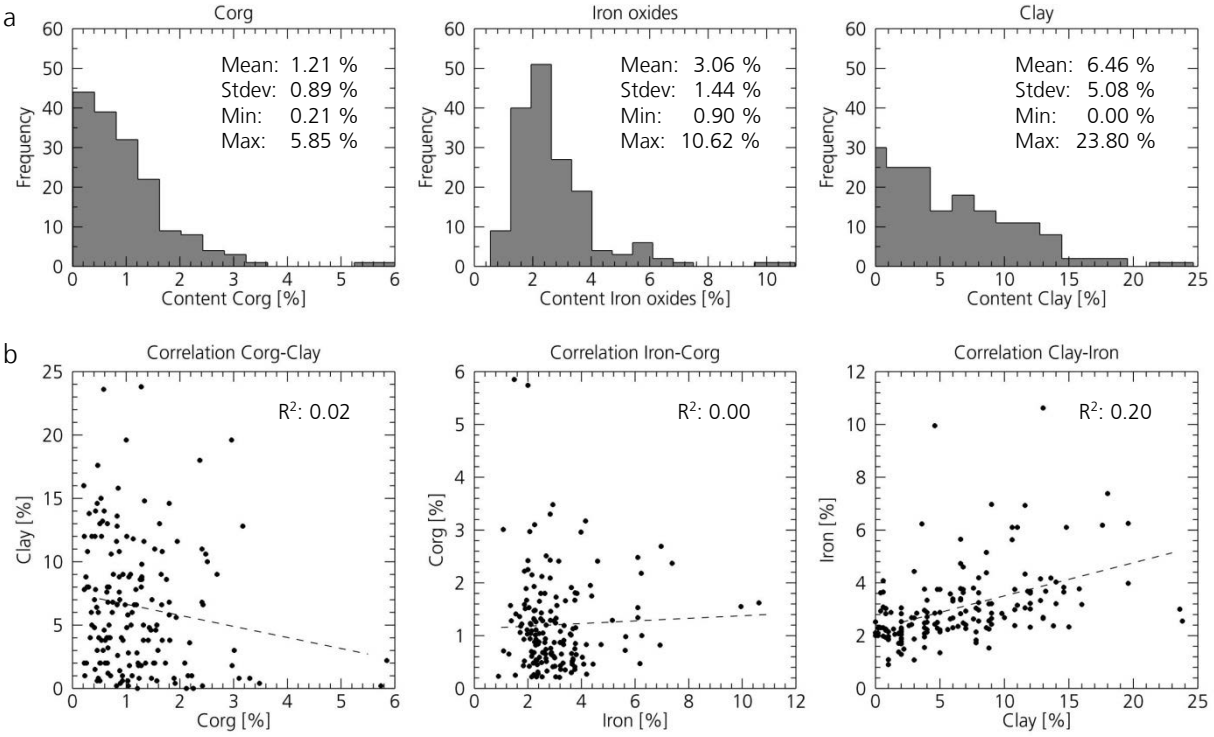


Figure 3-14: Histogram distribution with statistical parameters for 163 topsoil samples collected in the Albany Thicket Biome, South Africa, analyzed for soil organic carbon, iron oxides and clay content (a). Correlation plots for the three soil constituents (b).

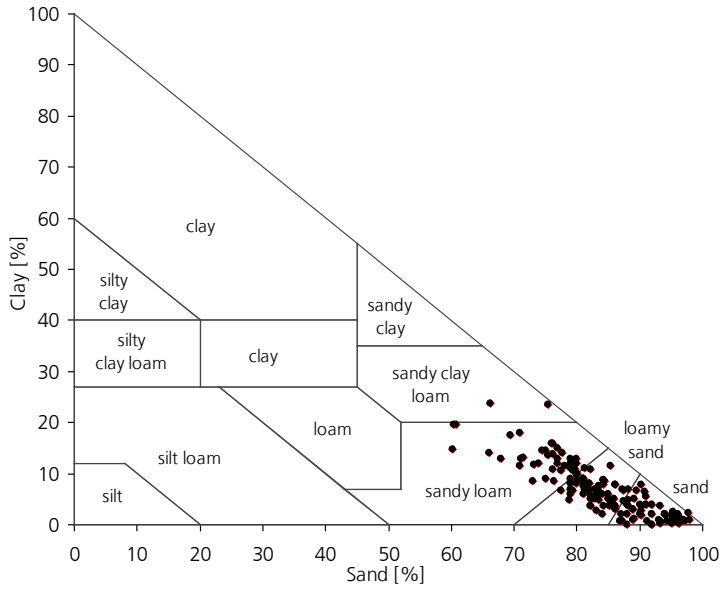
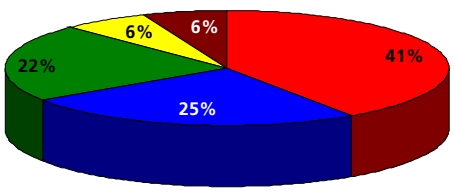


Figure 3-15: Soil texture triangle shows the presence of sandy soils with different amounts of loam (classification according to FAO 2006).

Soil type classifications

Detailed results of the soil survey conducted in 2011 including soil type classification according to the international WRB (IUSS Working Group WRB, 2007) and a South African classification are given in Kompter 2012. Some main results relevant in the context of this study are summarized in appendix A.3. In the following only the results of the soil classification are presented.

In general, soil profiles are shallower in the northern than in the southern transect (average depth of 32 in the northern and 52 cm in the southern transect, see Table A-4). Deeper soils are found on lower slope segments and on river floodplains, while profiles are shallow on hill tops and erosive upper slopes. Leptosol is the predominant soil type particularly in the northern transect (41 % of profiles, see figure 3-16). These are shallow soils with a depth below 25 cm and a strong influence of the parental material (gravely structure, high amount of rock fragments, see IUSS Working Group WRB 2007 for details). Luvisols are characterized by clay accumulation in the subsoil layer. Soils with significant alteration of the parental material and beginning formation of subsurface horizons are classified as Cambisols. Subsidiary also Regosols showing an unstructured soil profile and Calcisols (only in southern transect) indicating the presence of carbonates occur.



Soil type	Soil types identified from 32 profiles within the study area		
	North	South	Total
Leptosol	12	1	13
Luvisol	5	3	8
Cambisol	3	4	7
Calcisol	-	2	2
Regosol	1	1	2

Figure 3-16: Soil types classified based on 32 profiles taken in 2011 in the study area.

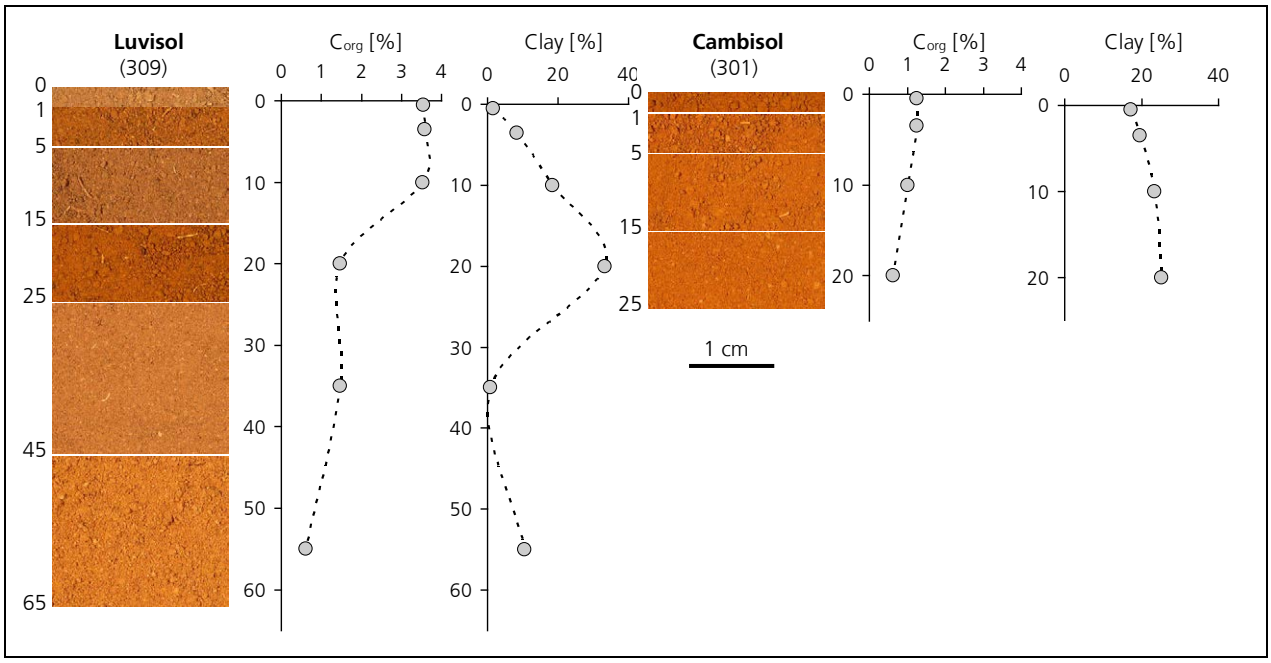


Figure 3-17: Soil profiles of a Luvisol and a Cambisol generated from dry bulk samples taken in defined depth intervals. Corresponding chemical reference for soil organic carbon and clay content is given.

Figure 3-17 shows an example of the profile analysis of one Luvisol with the distinct clay accumulation horizon around 20 cm and one Cambisol with a well developed soil structure, though shallow depth. The increase of soil clay content in the Luvisol below 35 cm is an indicator for mudstones as parental material. Both profiles show the reduction of C_{org} with increasing depth. The soil survey shows the presence of soils in different development stages, parental material and sediment regimes. Nevertheless information about soil types does not exist spatially continuous and thus can not be included within the further workflow.

Airborne HyMap imagery

The quality of the airborne HyMap imagery is discussed based on its orthorectification and radiometric calibration. Accuracy of orthorectification is evaluated relative to maps provided by Google Earth reference as the only digitally available reference. Errors in position are below 2 to 3 pixels which is around 6.5 to 10 m in x and y direction, respectively.

The ATCOR calibrated radiometry usually is evaluated by the comparison of field reference and image spectra of homogenous artificial targets. The study area comprises only natural areas which are highly variable from pixel to pixel, with no artificial areas which show homogenous spectral behavior over an area of at least 3 x 3 pixels are found in the imagery. The comparison of field and image spectra for this reason can only be done based on the soil spectra taken in the field. However, most of the corresponding image spectra are influenced by surrounding vegetation, shadows or other materials (e.g. stones) which are mostly darker, which results in most image spectra having lower albedo compared to the field spectra (figure 3-18). Differences between field and corresponding image spectra are mostly within the natural variance of the image spectra or close to it. Larger differences occur only for individual points (e.g. last plot in figure 3-18). Bands close to atmospheric absorption bands are similar in image and field spectra.

Spikes due to decalibrated bands could be removed in the pre-processing. The image spectral signatures which are smooth throughout the imagery, cover reasonable ranges, show characteristic spectral features

and, within the given constraints, match well to ground reference targets, indicate a good radiometric quality of the HyMap imagery. However, continuative tests could not be run due to the lack of invariant targets that can be used for proper comparison to ground reference data. For further details on inflight validation of airborne data see e.g. Strobl et al. (1997).

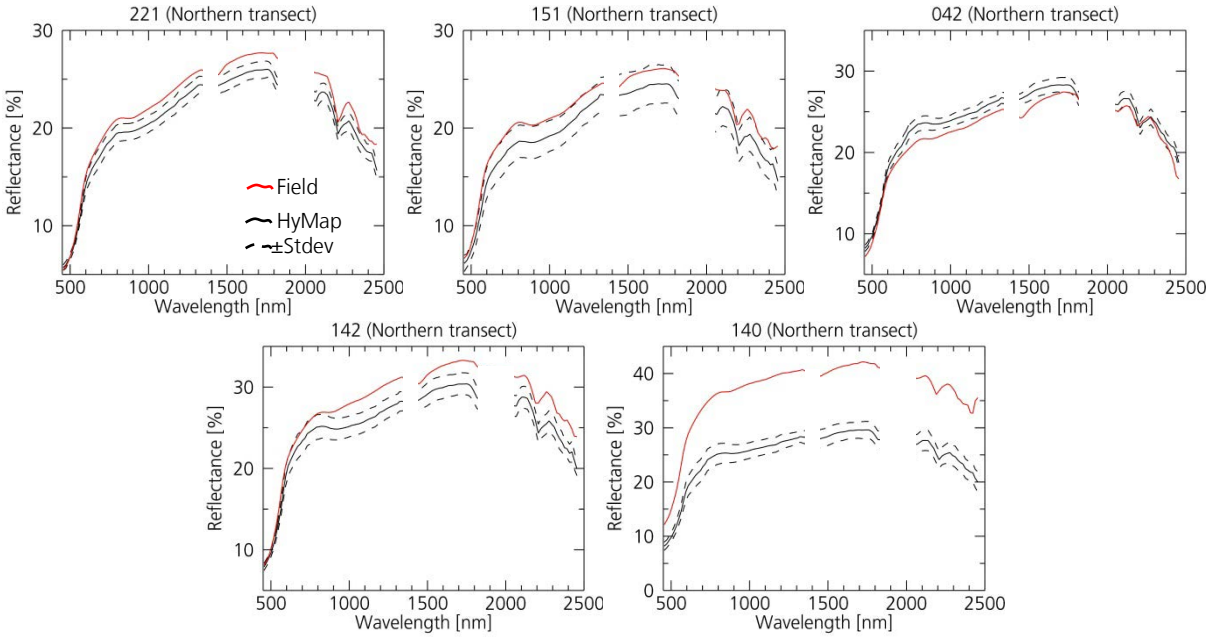


Figure 3-18: Comparison of field measurements and airborne HyMap spectra for three exemplary ground reference sites. For HyMap spectra the natural variance is assessed by the standard deviation of the spectra in a 3 x 3 pixel area around the spot where the field spectrum is taken.

Simulated EnMAP satellite data

The EnMAP satellite scene simulation tool has been applied since 2009 for the generation of a number of EnMAP-like images under varying environments, atmospheric and instrumental configurations based on different hyperspectral and multispectral data (see Guanter et al. 2009). A discussion of the performance of the simulation tool can be found in Guanter et al. (2009).

Similar to the HyMap imagery, the accuracy assessment of the simulated EnMAP hyperspectral imagery is problematic, owing to the lack of large homogenous ground reference targets. Thus, simulated EnMAP spectra are compared to airborne HyMap spectra which are averaged for the areas covered by the EnMAP pixels. However, this includes positioning inaccuracies, since the area of one EnMAP pixel (30 x 30 m) may be slightly shifted form the corresponding 9 x 9 HyMap pixels (area of 29.7 x 29.7 m). Figure 3-19 gives examples of EnMAP and corresponding HyMap spectra for three example areas of relatively homogenous land coverage.

The comparison of simulated EnMAP to HyMap spectra provided good agreement. Thus, it is assumed that with the application of the well established EnMAP simulation tool, realistic EnMAP-like data could be generated based on the airborne HyMap imagery.

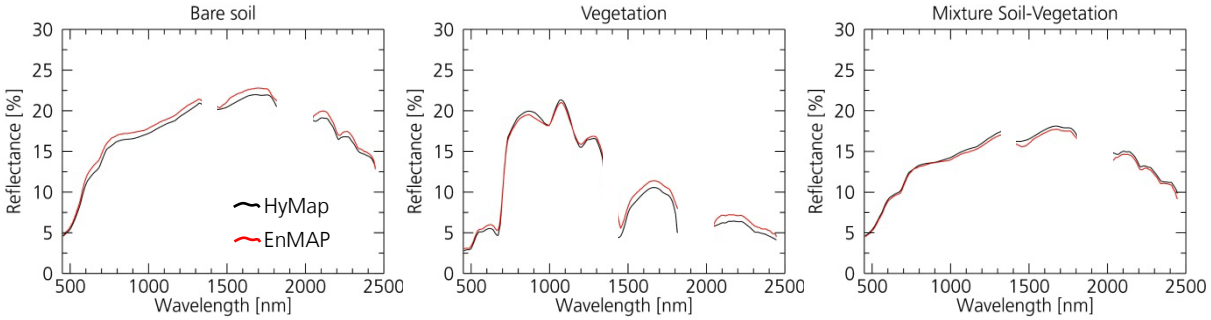


Figure 3-19: Comparison of simulated EnMAP and airborne HyMap spectra for three exemplary sites from subset Gonakraal. HyMap spectra comprised in the distinct EnMAP pixels were averaged.

4 Multivariate modeling to study soil constituents

This section provides an introduction to multivariate regression analysis, the background and the selection of an adequate multivariate approach to model soil constituents based on spectra. The characteristics and benefits of regression methods used in spectroscopy are briefly discussed. Finally, the approach, which is expected to be the most qualified to meet the objective of this study, is selected and a methodology is developed. An alternative statistical approach often used in previous studies for quantification of soil constituents using spectroscopy is described and tested for comparison.

Multivariate analysis (MVA) in general is applied for instance to reduce large data matrices, separate useful information from noise or to classify information. Multivariate regression analysis as one application of multivariate techniques are applied to find a mathematical model that describes the behavior of two groups of variables. Usually, one group of variables is called dependent or response variables (y), which are the command variables to be modeled, whereas the independent of predictor variables (x) are used to model y. Often, the x variables are parameters which are easy to measure compared to the command variables y. For modeling, a number of observations (or samples) of x and corresponding y variables are needed for regression analysis.

The need of multivariate analysis for soil spectroscopy

In section 2.3 it is shown that a soil's spectrum is the result of the entirety of spectral characteristics produced by its major and accessorily chemical constituents and all its physical properties. The complexity of soil materials and their spectral properties overlapping in a soil's VIS-SWIR spectrum make wide parts of it non-specific. In addition soil reflectance is influenced by changing physical soil properties (e.g. differences in soil texture, moisture, surface condition with degree of crusting, roughness and vegetation coverage) and sensor-specific variances such as random noise. These factors are referred to in the following with "non-significant" influences on reflectance spectra. This induces changes in soil reflectance that are not directly related to the soil constituents under consideration or to soil compositional properties at all. These two factors often permit the usage of spectral analysis techniques and spectral matching algorithms for a quantitative investigation of a soil's constituents (Ben-Dor et al. 2008 and 2009). Multivariate methods are used to mathematically extract the "hidden" information so that it may be correlated with measured soil properties (Viscarra Rossel and Behrens 2010). The basic principle of multivariate analysis (Martens and Naes 1989) allows to intrinsically model a defined constituent without knowing from first hand how it influences the spectra. As such the method handles non-selectivity in spectra, and one constituent can be modeled based on a signal that is overlaid by another constituent or interferent.

In spectroscopy multivariate analysis can be used to describe and model the relationship which exists between the spectrum of a material and its chemical and/or physical properties (compare section 2.3). The combination of spectroscopy with multivariate techniques provides a cost-effective tool and can be used for spatially continuous mapping of soil properties over large areas using airborne spectrometer data (see section 2.2). In these cases y is usually one variable such as the chemical content of a soil constituent.

Accordingly, x is given by a number of spectral characteristics, which are determined from the spectra, either in a predefined way (e.g. absorptions at given wavelength) or by statistical means within the regression analysis. In most of the cases, the application of MVA in spectroscopy presumes a linear relationship between the reflectance spectrum and component concentrations (Martens and Naes 1989).

Three steps of model development

Multivariate regression analysis is conducted in three steps (based on Kessler 2007).

1. In an initial calibration step, a model is set up based on a sample population (training set) and the model parameters (regression coefficients) are determined. The size and distribution of the training dataset has to be representative for both the variation of the soil constituent being considered and also the variation in the spectra (Martens and Naes 1989). Additionally, the number of samples (samples represent spectra in our application) in the training dataset must be larger than the number of variables used in the regression equation, which results in a over-determined equation system that can be solved in regression analysis according to the partial-least squares principle. To calibrate the regression model one of the methods for regression analysis presented below can be used.

2. In the next step, the validation, the model performance for future predictions of unknown data is assessed. This may be conducted using an independent set of samples (test set) which samples were not part of the training set (external validation) or by using the same samples used for model calibration (internal validation). In case of an external validation, the selection and characteristics of the samples of the test set influences the validation result. For an internal validation, a method called cross-validation is often applied. The initial dataset (n samples) is partitioned in complementary subsets (e.g. n-10 samples for training, and 10 samples for test set), a model is developed based on the training subset and its performance is evaluated using the test subset. This is done as long as each sample is once predicted as part of a test set. Results of the cross-validation (R^2_{CV}, $RMSE_{CV}$, RPD_{CV}) are calculated by comparing the predicted with the observed y values.

3. In a third step of regression analysis, the calibrated and validated regression model is used to predict unknown y variables from another set of x variables, for example the hyperspectral survey data.

The calibration and validation accuracy of a regression model is assessed using common accuracy indicators, such as the coefficient of determination (R^2) as correlation between measured and modeled y variables, the ratio of performance to deviation (RPD) and the root mean squared error (RMSE). The accuracy indicators are defined in appendix C.

4.1 Methods for multivariate regression analysis

Among the many methods applied for regression analysis the most often used are multiple linear regression (MLR), principle component regression (PCR) and partial least squares (PLS) regression. The result of all these methods is a regression equation which gives the relationship between the response variable y and the

predictor variables x. Some recent studies using the presented regression techniques to study soil properties are given in section 2.2.

Multiple linear regression (MLR)

MLR, also named ordinary least squares (OLS), uses a linear relationship to describe the behavior of one response variable (y) and a given number of independent predictor variables (x). The MLR model for y can be described as follows:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m + e = \sum_{j=1}^{m} b_jx_j + e \quad \text{(Kessler 2007)} \quad [-] \quad (4\text{-}1)$$

with b_0 as constant and $b_{0\ldots m}$ as the regression coefficients of the m predictor variables x. The residual or error term e is the deviation of the modeled response variables from the measured ones. This term cannot be modeled since it may, for instance, result from measurement uncertainties, e.g. noise. In regression analysis, the coefficients b are determined based on multiple observations of x and corresponding y variables. The number of observations of x and y variables (number of samples n) must be higher than the number (m) of predictor variables (x, e.g. spectral characteristics). Only then the analysis results in an over-determined linear equation system, which is solved using a least squares approach so that the sum of the squares of the residuals is minimal. MLR is a simple regression method and used to establish a linear model which describes the relationship between one response variable and multiple predictor variables that preferably are uncorrelated. The approach is limited by the number and selection of the predictor variables x. Because of n > m (no. samples > no. predictor variables), the application of an entire reflectance spectrum with for instance a thousand bands is not possible, because this would require at least 1 more than a thousand observations/samples.

Principle component regression (PCR)

PCR is a combination of principal component analysis (PCA) and MLR. From the initial data matrix (X as matrix of x variables), the principal component factors, or eigenvectors (P), are derived as latent variables, which are a linear combination of the initial variables but weighted with scores (T) according to their variance (see figure 4-1). X can be written as

$$X = TP^T + F \quad \text{(Kessler 2007)} \quad [-] \quad (4\text{-}2)$$

with P^T as the transposed matrix of the principal components and the matrix F of the error term.

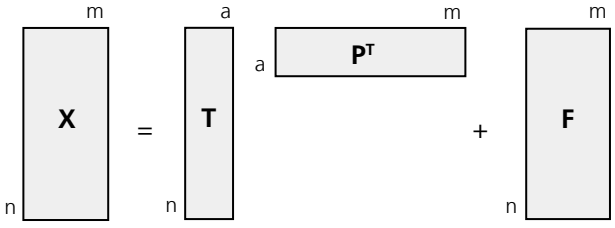


Figure 4-1: Graphical illustration of the relationship of matrices in the decomposition of the x variables in principle component regression analysis shown in eq. 4-2 (after Kessler 2007).
Abbreviations: n: number of samples, m: number of spectral bands, a: number of principle component factors used to model the matrix of x variables, see text for further abbreviations.

The scores of the highest principal component factors (T), accounting for most of the variance of the original data are subsequently correlated to the response variables (Y) in a MLR. The basic MLR equation (see eq. 4-1) in PCR can be rewritten as:

$$Y = qT + F \qquad \text{(Kessler 2007)} \quad [-] \qquad (4\text{-}3)$$

with the regression coefficients q and the matrix F of the error term. Since PCR applies derived PC factors, it can handle a multicollinearity of the input data and thus full spectra. A drawback of this method is that during decomposition of the initial data into PC factors potential relationships between x and y variables may get lost since the actual response variable y is not used during the decomposition and the information related to y may be included not until higher PC's. Thus, the threshold used to determine the number of PCs to be included in the analysis may be of high importance.

Partial least squares regression (PLS)

PLS is a very variable tool developed by Wold et al. (1983) and currently the most often used regression method and a standard tool for chemometrics (see also Martens and Naes 1989). In contrast to PCR, the decomposition and regression is performed in a single step. This means that the response variables y, i.e. soil chemical information, are used for a decomposition of the predictor variables x and the determination of the PLS factors. A PLS calibration can be broken down into separate iterative steps that apply calibration and prediction using MLR (Haaland and Thomas 1988). Figure 4-2 depicts the decomposition of the matrices of x and y variables into their PLS factors (score matrices T and U) and loadings (matrices P^T and Q^T). In this process, the dimensionality of the data is minimized while the covariance between x and y variables is maximized. The PLS factors are successively selected as orthogonal factors that maximize the covariance between the remaining variance in the predictor and response variables. The best number of PLS factors to use for a model is determined in cross-validation (see above) where based on sample subsets parallel models are developed and compared.

PLS analysis is often preferred compared to MLR or PCR because PLS models explain more of the variance of the x variables with fewer factors, the factors are mostly easier to interpret and the PLS algorithm is computationally fast. It is often applied to construct prediction models when there are many predictor variables, since it adequately handles multicollinearity and as such can handle full spectra (Martens and Naes 1989, Wold et al. 2001). PLS models are seen as robust in terms of data noise and missing values (Viscarra Rossel et al. 2006).

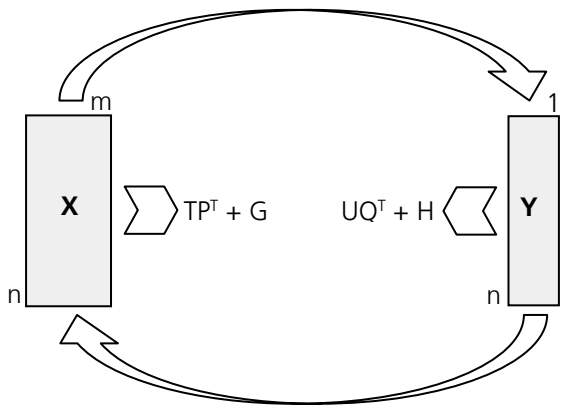


Figure 4-2: Graphical illustration of the PLS analysis and contributing matrices (adapted from Kessler 2007). Abbreviations: n: number of samples, m: number of spectral bands, G, H: error terms, see text for further abbreviations.

Other multivariate regression methods

Among other methods used to obtain multivariate calibrations in spectroscopy are data mining techniques such as multivariate adaptive regression splines (MARS), support vector machines (SVM) regression, and artificial neuronal networks (ANN). These techniques are mentioned in section 2.2. Some can model non-linear relationships but involve enhanced processing efforts.

Discussion of the performance of different multivariate regression methods

The performance of different multivariate regression methods depends on the characteristics of the selected algorithm and its suitability to model the relationship between response and predictor variables for the given sample population. It is further affected by the existence of non-significant signals in the data usually being a result of noise, differences in particle size, etc. (Vohland et al. 2011, see above). For instance, PLS models with a high number of PLS components is prone to model noise (Kessler 2007). Viscarra Rossel et al. (2006) provide a review of studies where soil attributes were predicted using different multivariate methods.

4.2 Selection of an adequate method to model soil constituents

The PLS algorithm is well-established among the various multivariate regression analysis techniques and in the past has often been used to predict various soil properties (e.g. Gomez et al. 2008a and b, Stevens et al. 2008, Udelhoven et al. 2003). However, the technique faces drawbacks in terms of significance and transferability. The high degree of adaptation induced by the selection of the latent variables and of the final model based on error minimization results in a perfect adaptation of PLS models to the characteristics of the local observations the model is built upon (see e.g. Stevens et al. 2008 and 2010). Especially models with a high number of PLS factors tend to over-fit the model data (Viscarra Rossel 2007). This results in a limited transferability, first when models are transferred to datasets showing different sensor-specific characteristics such as band positions, signal-to noise ratio, etc. and second when they are transferred to datasets eventually from other locations showing different environmental conditions such as surface characteristics, soil texture, geology, etc. In these cases in general a re-calibration of the PLS regression model is necessary. Another issue of PLS is the significance of established models. Since the factors used for model development are selected by exclusively statistical reasons, a specific variable can be modeled properly but the correlation may instead of physical properties be based on a second variable indirectly linked to the investigated parameter or even result from random noise.

These factors cause an established calibration model that works well for one sample population or study site might being unacceptable for another (Chang et al. 2001) and a model re-calibration may be necessary. This just the same limits the transferability of PLS models derived from local observations to a regional scale as it is aspired in this study. Similar limitations apply for black-box models such as ANN.

Feature-based regression approaches

To allow for a mapping of soil constituents for ecosystems of a regional extent, where land cover, surface and soil parameters are subject to change, more robust algorithms are needed. For these applications, approaches directly applying spectral absorption features or band indices are considered to be more suitable

since they are physically based (e.g. Mulder et al. 2011). They are considered more robust and allow for a better transferability to regions of similar environmental conditions.

Although broad regions of a soil spectrum measured in the VIS to SWIR range are non-specific as result of overlapping absorptions and additional influences of non-significant factors (see above for definition), there remain diagnostic spectral features which can be used for a quantification of soil constituents. This includes for instance, broad iron electronic transition bands in the VIS/NIR and some fundamental and overtone absorptions due to clay and organic materials in the SWIR. Various characteristics of spectral absorption features have been directly used for a quantification of soil properties (e.g. Bartholomeus et al. 2008, Gomez et al. 2008a, Hill and Schütt 2000, Lagacherie et al. 2008, Richter et al. 2009). Some selected studies are briefly discussed in section 2.3.

Various authors also tested the performance of different multivariate regression methods for the prediction of soil parameters. For instance Viscarra Rossel and Behrens (2010) applied PLS, SVM, MARS, ANN, MLR (used with quadratic instead of linear polynomials) and others to model soil organic carbon, clay content and soil pH. They found that simple MLR analysis coupled with an effective feature selection can produce comparable predictions to those of complex algorithms such as SVM and ANN. The only constraint of MLR which must be considered is that it can only model linear relationships, though individual known non-linear relationships can be transformed to a linear behavior using adequate pre-processing techniques (e.g. transformation to log, Viscarra Rossel 2007).

However, regression models built on one or only a few spectral absorption features or characteristics are not sufficiently robust for a practical application for a wide variety of soils (Viscarra Rossel et al. 2006). This is due to the mentioned disturbance of the spectral signal caused by overlapping absorptions and non-significant influences of soil physical properties or sensor-specific factors (see above). In some studies soil constituents' predictions based on individual spectral indices were described as less robust than PLS models (e.g. Bartholomeus et al. 2008).

Selection of an adequate feature-based regression approach

Coming from the general suitability of spectral feature based approaches, but considering the given signal disturbances, the combined application of several spectral absorption features and the characteristic shape of the spectral curve in a defined range, each of them diagnostic, is expected to be useful. Absorption features can be described by properties such as the absorption band depth or width and a characteristic shape of the reflectance curve for instance by the slope in a given wavelength range. According to Beer's law (see section 2.1), each of these properties correlate to the chemical content of a soil constituent in a linear relationship. Thus, they should be highly collinear when neglecting the spectral influences by overlapping absorptions and non-significant factors (see above). Nevertheless, the consideration of the spectral variance introduced by these factors results in a weakening of the multicollinearity of spectral features characteristic for the soil property under consideration so that they can be used for regression analysis.

In this study a physically based model is developed which directly applies properties of diagnostic spectral features for regression analysis. For each of the three investigated soil constituents, a set of diagnostic spectral features and various properties of these features (such as feature depth, width etc.) are applied to establish a regression model using multiple linear regression techniques. The numerous spectral properties are expected to provide robust indicators to allow for a quantification of the selected soil constituents based on spectra measured in the laboratory, field or from an airborne hyperspectral sensor. The predictability of

the described feature-based regression approach is compared to conventional PLS techniques. This is evaluated in detail based on laboratory and field spectra in Bayer et al. (2012).

With the given physical basis of the approach, the proposed methodology is suggested to be more robust in terms of its transferability compared to exclusively statistical methods such as PLS. Since it further applies multiple spectral features and their properties, it suggests the preference of this approach compared to ones that are based on a single diagnostic spectral feature or band ratio. Additionally, the established models are considered simplistic and require minimal computation due to a limited number of variables in the regression relationship.

4.3 Application of an alternative statistical regression approach

A first analysis for testing the correlation of soil spectral signatures and measured soil contents is conducted using PLS regression techniques as well established method for this purposes, as is discussed in sections 2.2 and 4.1. Therefore, the three spectral libraries (in-situ field, bare-soil field and laboratory) in both the original ASD resolution and the resolution matching the HyMap sensor are applied. The analysis is done to assess the predictability of this method for the available spectral libraries and for comparison to the developed feature-based regression approach. Detailed results of the comparison of the two approaches is given in Bayer et al. (2012). In the following, the results of the PLS analysis are presented and briefly discussed. The data and analysis given in the publication differ from the ones used in this study in (1) having 116 instead of 115 spectral bands (band at 889 nm additionally included), and (2) the reference data of 163/125 samples are separated into training and test set. However, the results of PLS analysis are presented here in consistency with Bayer et al. (2012).

PLS regression analysis is performed using the ParLes software for multivariate modeling and prediction (Viscarra Rossel 2008a). 75 % of the data (94 samples for field datasets, 123 samples for laboratory dataset) are selected by random stratified sampling based on the chemical reference of each soil constituent for training and used for model calibration, while the remaining 25 % (31, respectively 40 samples) are used as test set to validate the models.

The spectral datasets are pre-processed with one or a combination of two of in total 11 manipulation methods (transformation of reflectance (R) to log(1/R), 5 light scatter and baseline corrections such as multiplicative scatter correction (MSC), standard normal variate transform (SNV), wavelet detrending etc., calculation of 1st and 2nd derivative, mean centering and variance scale of the data and a combination of the latter two). The performance of each pre-processing setting is evaluated in leave-one-out cross-validation (CV, see section 4). For each setting, the optimal number of latent variables to be used for modeling is examined by the variation of the root mean square error (RMSE) and the Akaike Information Criterion (AIC) as a function of the number of latent variables (for a definition of AIC see appendix C). The optimal number of PLS factors is determined at local minima of RMSE and AIC within a steady trend of these two factors. The one pre-processing setting in combination with its corresponding optimal number of PLS factors performing best in cross-validation is selected and subsequently used for model calibration based on the observations of the training set. Calibrated models are further applied to the test set for validation purposes. If several pre-processing settings provided similar CV accuracies, then model calibration and validation is performed for each setting. The PLS model with highest accuracies in both calibration and validation, as well

as a minimal difference in between them is selected. This way the most significant and robust PLS prediction models are retrieved and over-fitting of the models is prevented.

Results of PLS analysis

For each soil constituent, a model is set up using the three spectral datasets and two spectral resolutions, resulting in 12 models per constituent. All samples, though split into training and test sets, are applied for modeling. Model performance is assessed for each method and dataset and compared based on the models' correlation coefficient (R^2) for predicted versus measured compositions, root mean square error (RMSE) and ratio of performance to deviation (RPD). This is done for both calibration and validation using the corresponding datasets (see Table 4-1). Goodness of predictions is evaluated following the qualitative classifications of Chang et al. (2001). These use R^2 greater than 0.80 and RPD values greater than 2.0 as indicators for excellent prediction models, R^2 between 0.50 and 0.80 and RPD values between 2.0 and 1.4 were considered to be medium quality and useful for quantitative predictions in most applications, and lastly, R^2 lower than 0.50 and RPD lower than 1.4 considered to be not useable.

Table 4-1 gives the detailed calibration and validation accuracies for prediction models developed using PLS techniques. The application of the different pre-treatment methods allowed the derivation of consistent PLS prediction models, that apply between 4 and 9 PLS factors for regression analysis. In over 40 % of the models mean centering of the data revealed to be the best pre-treatment technique. Figure 4-3 gives examples of scatter plots of measured versus calculated contents for PLS models developed based on bare soil field spectra in HyMap's spectral resolution. These would be the models to be used for a prediction of soil constituents from the HyMap imagery. Regression coefficients of established PLS models provide information about the importance of certain wavelength for the regression model. Although they are highly variable depending on the settings of each model, some of their information can be related to physical soil properties. Two examples are shown in figure 4-4.

Table 4-1: Calibration and validation accuracies of models developed with partial least squares regression (see Bayer et al. 2012). The models matching the ones that are applied for a large-area prediction of soil constituents based on HyMap imagery are highlighted (see section 6.1).

		Spectral dataset	Pre-treatment	No. of factors	Calibration 94[1]/123[2] samples			Validation 31[1]/40[2] samples		
					R^2_{Cal}	$RMSE_{Cal}$	RPD_{Cal}	R^2_{Val}	$RMSE_{Val}$	RPD_{Val}
ASD spectral resolution	C_{org}	1 - In-situ field	1st derivative	6	0.79	0.45	2.16	0.53	0.52	1.30
		2 - Bare soil field	MSC + 1st derivative	6	0.82	0.41	2.38	0.65	0.42	1.61
		3 - Laboratory	Mean center + 1st deriv.	6	0.82	0.40	2.36	0.69	0.45	1.53
	Iron oxides	1 - In-situ field	Mean center	7	0.66	0.76	1.74	0.39	0.84	1.14
		2 - Bare soil field	Mean center	8	0.69	0.73	1.80	0.43	0.80	1.20
		3 - Laboratory	Log(1/R) + mean center	9	0.82	0.64	2.39	0.43	0.98	1.14
	Clay	1 - In-situ field	Mean center	7	0.23	4.34	1.15	0.02	4.63	0.97
		2 - Bare soil field	Log(1/R)	5	0.11	4.67	1.06	0.06	4.42	1.02
		3 - Laboratory	Mean center	8	0.33	4.31	1.22	0.08	4.43	1.02
HyMap spectral resolution	C_{org}	1 - In-situ field	Log(1/R) + Mean center	5	0.77	0.46	2.11	0.59	0.47	1.43
		2 - Bare soil field	Log(1/R) + 1st deriv.	5	0.82	0.41	2.38	0.62	0.45	1.51
		3 - Laboratory	Log (1/R)	9	0.79	0.43	2.20	0.79	0.32	2.11
	Iron oxides	1 - In-situ field	Mean center	7	0.67	0.75	1.75	0.43	0.76	1.26
		2 - Bare soil field	Mean center	7	0.66	0.76	1.73	0.45	0.77	1.25
		3 - Laboratory	Log(1/R) + Mean center	9	0.81	0.67	2.28	0.49	0.88	1.27
	Clay	1 - In-situ field	Mean center	4	0.14	4.58	1.09	0.01	5.06	0.89
		2 - Bare soil field	Mean center	5	0.15	4.56	1.09	0.03	4.47	1.01
		3 - Laboratory	1st derivative	8	0.37	4.16	1.27	0.10	4.43	1.02

[1]Number of samples in training and test sets for field datasets, [2]Number of samples in training and test sets for laboratory datasets.

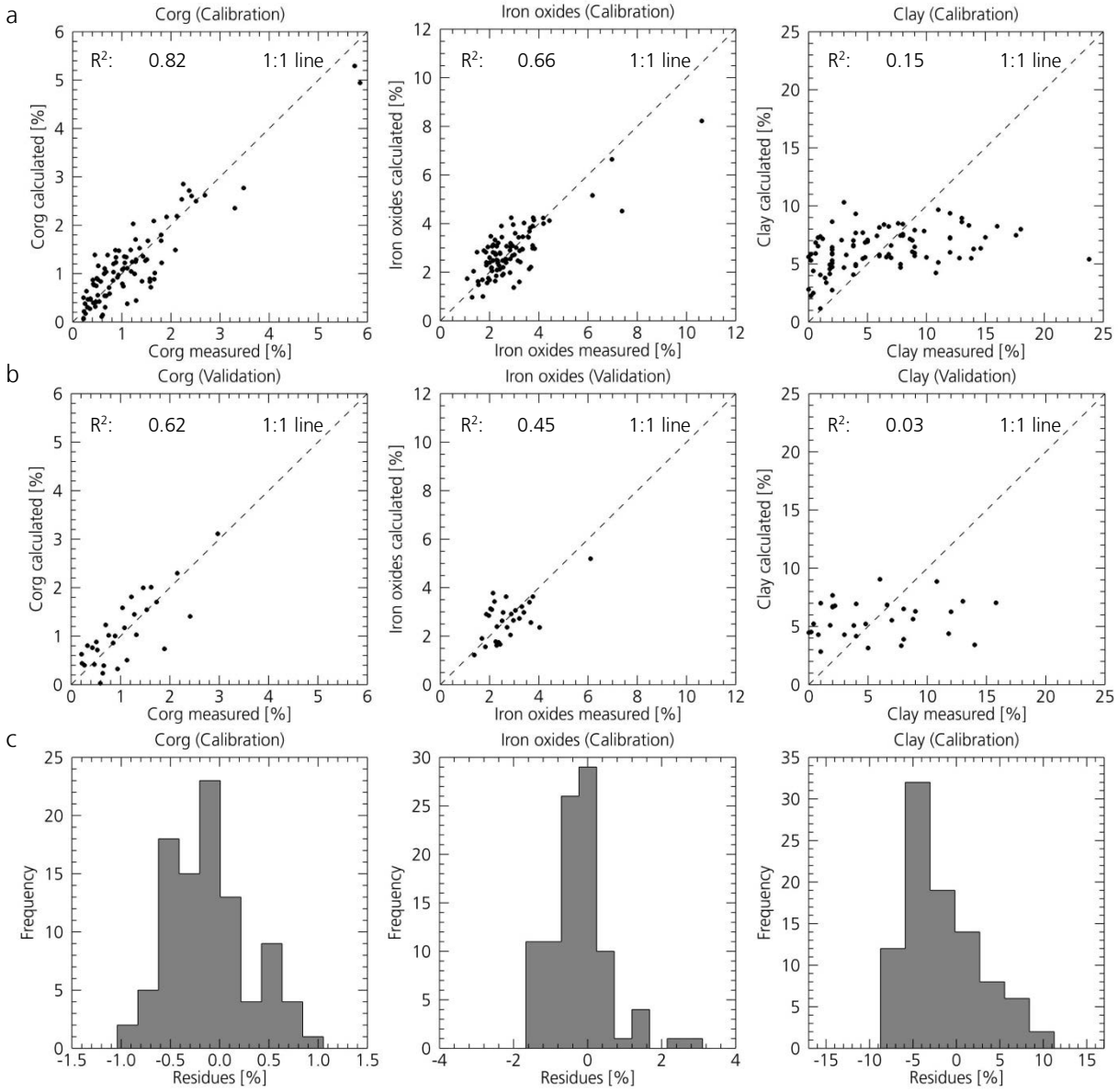


Figure 4-3: Scatter plots of measured versus calculated contents for PLS models developed in calibration (a) and validation (b) using partial least squares regression and histogram distribution of modeling residues of calibration. The models shown here are based on bare soil field spectra in HyMap's spectral resolution.

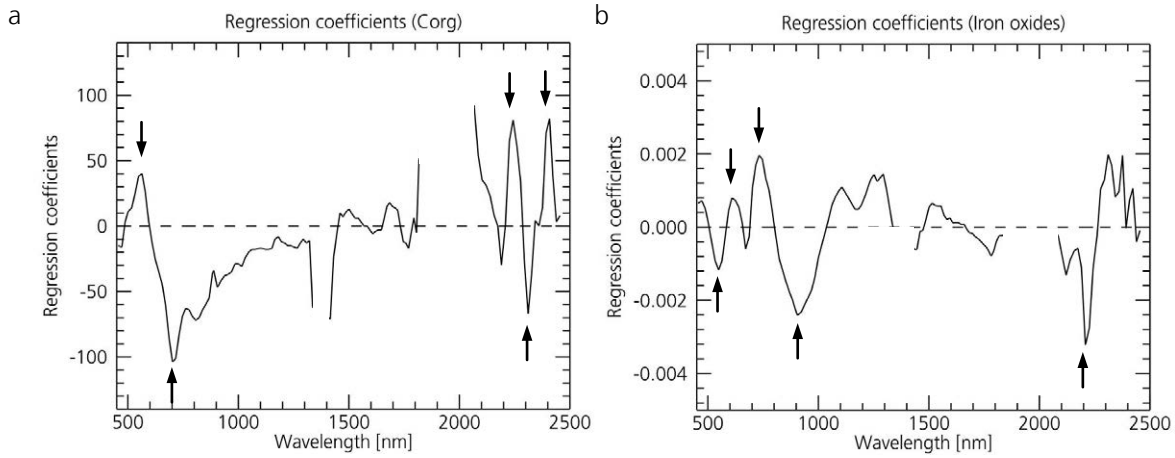


Figure 4-4: Exemplary regression coefficients of PLS models for the prediction of soil organic carbon (a) and iron oxides (b) built upon bare soil field spectra in HyMap's spectral resolution. Shape and range of spectral regression coefficients are related to the selected data pre-treatment method (A: Log(1/R) + first derivative, B: Mean center).

Soil organic carbon prediction models for the three spectral datasets are very consistent for a number of pre-treatment techniques, showing only slight differences. Calibration accuracies of the finally selected PLS models for every dataset in terms of R^2_{Cal} range between 0.77 and 0.82 and RPD_{Cal} range between 2.11 and 2.38 for both applied spectral resolutions. Both with low $RMSE_{Cal}$ slightly above 0.4 % C_{org}. Test set validation results of organic carbon models are also of good quality (R^2_{Val} between 0.53 and 0.79, $RMSE_{Val}$ between 0.32 and 0.52 % C_{org}, RPD_{Val} between 1.30 and 2.11), with highest accuracies for laboratory spectra and considerably lower for in-situ field spectra. Nonetheless, organic carbon prediction models established by partial least squares techniques indicate good performance in both calibration and validation and do not indicate model over-fitting, which can be identified by a large difference between calibration and validation performance. The regression coefficients of most soil organic carbon PLS models indicate a high importance of wavelengths in the VIS range, which may be related to soil color (see figure 4-4a for an example). Distinct features additionally occurring in the SWIR can be linked to absorptions of functional groups of soil organic carbon (see table 5-1).

For the calibration of iron oxides prediction models, consequently 7 or more PLS factors are needed to set up significant models. The PLS models show correlations (R^2_{Cal}) between 0.66 and 0.82 for all six combinations of spectral datasets and spectral resolutions. Resulting $RMSE_{Cal}$ errors between 0.64 and 0.76 % Fe_d and RPD_{Cal} between 1.73 and 2.39 indicate slightly lower prediction ability than for organic carbon. Test set validation of the established iron oxides prediction models show reduced accuracies, which is most obvious in a decrease in R^2 and RPD, while RMSE for the most models only slightly increases from calibration to validation ($RMSE_{Val}$ between 0.76 and 0.98 % Fe_d). No iron oxides model reaches an RPD_{Val} of 1.4 indicating a medium predictive model (Chang et al. 2001). PLS regression coefficients of iron oxides models (example in figure 4-4 b) show an increased importance of wavelengths in the VIS range, which can be related to electronic transition bands of the iron ions primarily the one around 900 nm (compare table 5-2). In the SWIR, the 2208 nm clay absorption shows up, representing an indirect correlation with features related to clay. In fact, a correlation of iron oxides and clay content can be identified based on chemical reference (R^2 of 0.20, see figure 3-14b).

The calibration accuracy of clay predictive models with R^2_{Cal} below 0.4 and RPD_{Cal} below 1.3 is below the accuracies of medium prediction models. RMSE are constantly above 4.0 % Clay. Among these models, the best correlations are reached for the laboratory datasets. Corresponding validation results are equally poor. Regression coefficients are highly variant. These results show that for estimations of the clay content in soils, no significant prediction model providing sufficient accuracy can be established when using partial least squares regression techniques and based on the ground reference data of the South African study site.

Conclusions of PLS analysis

These results show that partial least squares regression techniques are able to establish good prediction models for organic carbon and iron oxides but cannot model the relationship between soil clay content and the spectral signature. This is reflected in scatter plots of measured versus predicted contents and the histograms of the modeling residues (figure 4-3). Calibration and validation scatter plots are reasonable for soil organic carbon and iron oxides but for clay models indicate a systematic overestimation and underestimation of low and high clay contents, respectively. Modeling residues are close to a normal distribution, except for clay content.

5 Methodical developments for a semi-automatic quantification of key soil constituents

Based on the advantages and drawbacks of approaches that were previously used to quantify soil constituents from hyperspectral data as was discussed in section 4.2, a new and physical approach for the quantification of soil organic carbon, iron oxides and clay is developed in this study. The benefit of this approach is the application of multiple spectral features for each of the three soil constituents. The combined spectral information is used for subsequent multiple linear regression analysis which establishes a relationship between spectral and chemical reference information. The method is developed for subsequent regional (approximately 320 km^2) prediction of soil constituents from hyperspectral imagery of the South African study site.

While the previous section gives the background of the selected approach, this section presents the development of an adequate methodology to predict soil constituents using a multi-stage workflow which is summarized in figure 5-1. Automatic and interactive steps are indicated. In a first stage a regression relationship between spectral properties and chemical reference data is established (section 5.1). The second stage comprises the preparation of the HyMap hyperspectral imagery for soil related analysis (section 5.2) before in the last step soil constituents are predicted regionally and the confidence of the predictions is assessed (section 5.3).

5.1 Regression modeling using MLR and assigned spectral feature variables

In section 2.3 is shown that the presence of many soil constituents affect soil reflectance in a diagnostic way. The fundamental basis for model development is that this also applies for soil organic carbon, iron oxides and clay and that the relationship between their chemical content and their spectral features can be modeled by a linear relationship (see section 2.1). Spectral features and characteristics diagnostic for soil organic carbon, iron oxides and clay minerals are selected according to their presence in the sample spectra and respective literature. In parameterization their properties are described with "spectral feature variables" that subsequently are correlated to measured chemical contents in linear regression analysis. Validation of the regression model is performed internally using cross-validation.

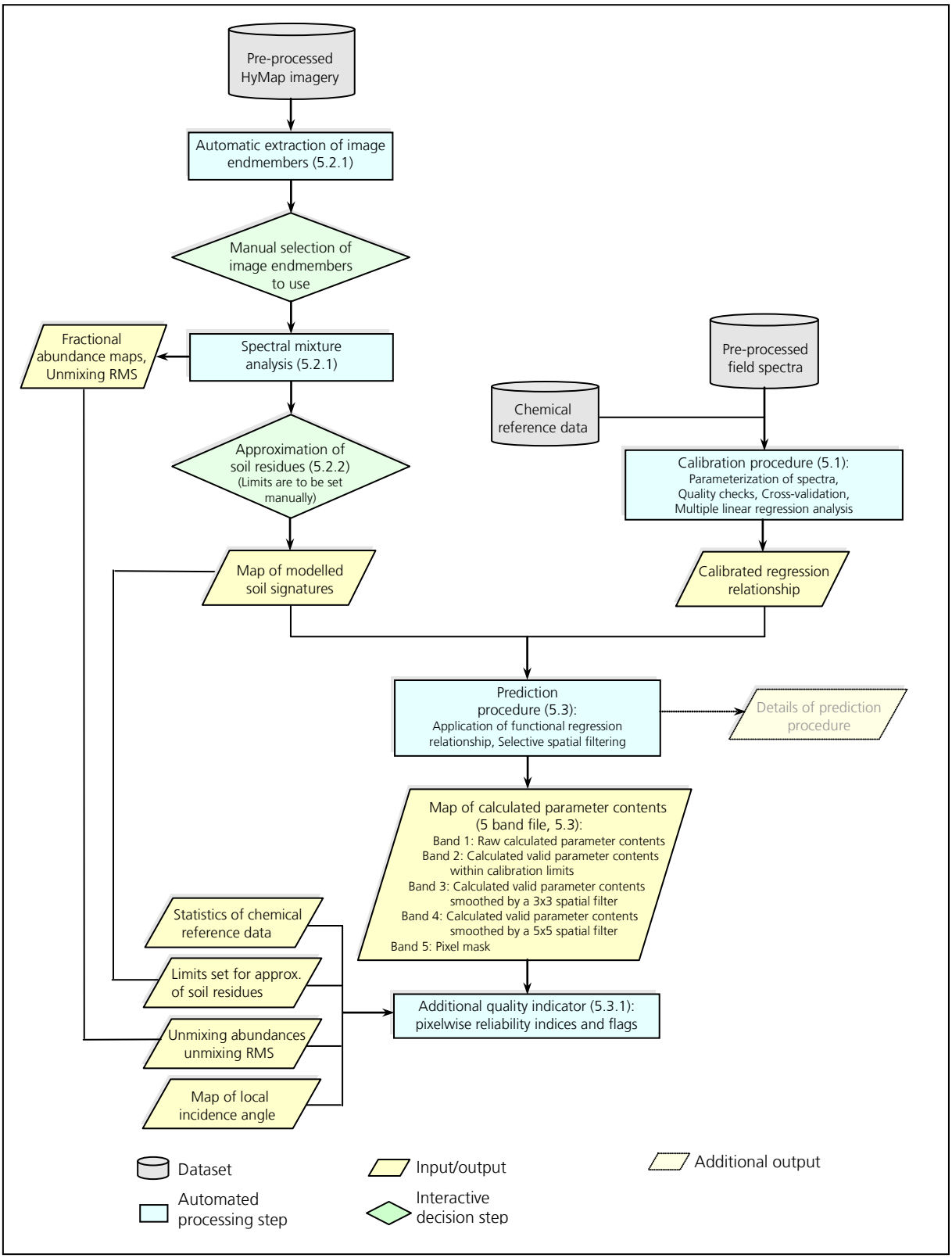


Figure 5-1: Processing workflow within the multi-stage methodology from initial spectral and chemical datasets to the final map of predicted soil parameters, including the data layers produced in between. The sections where the individual steps are explained in detail are given in parenthesis.

5.1.1 Selection of diagnostic spectral features

The selection of diagnostic spectral characteristics is based on their presence in the sample spectra and their documentation in previous studies, most notably in Ben-Dor and Banin (1994), Clark (1999), Grove et al. (1992), van der Meer and de Jong (2001) or Viscarra Rossel and Behrens (2010). From the spectral characteristics described as significant for soil organic carbon, iron oxides and clay content that are summarized in section 2.3, only the most important and strong features are selected to increase the robustness of the algorithm. Spectral features that make an impact on only few spectral bands and there only produce a minor reduction of the reflectance, are excluded from the analysis because they are assumed to be too small for the given signal-to-noise ratio and achievable accuracy of the hyperspectral image data. Features unique to one parameter are preferred but not all spectral characteristics can fulfill this requirement. This is discussed in the last paragraph in this section. Special care is taken not to include spectral characteristics in wavelength ranges where typical features of vegetation occur. This minimizes the influence of vegetation signatures potentially remaining in the pixel spectra after processing (see section 5.2).

Table 5-1 to table 5-3 list the spectral features that are selected to be used in this study for each soil constituent. Where known, the detailed cause of the absorption is specified. References to previous studies where the three soil components are measured in enriched soils or as non-synthetic mineral powders are given, to be regarded as summary of respective literature. The selected spectral characteristics are divided into three feature types: absorption features (AF), features of the spectral curve itself (CF) and of its continuum as convex hull of each spectrum (HF). With these three types it is possible to cover all specific spectral characteristics described in previous studies. The feature types are defined and explained in detail in the following parameterization section. The wavelength positions of the selected spectral features are given with a reference to studies describing it. The collected spectral features document a small spread in the wavelength positions of the maximal absorption. This is a result from the specific material and sensor used and measurement conditions applied in the different studies (see also 2.3.5). Additionally, it was reported that the exact location of an absorption may also shift from an expected location since real molecules do not behave totally harmonically (Bishop et al. 1994).

Soil organic carbon

Section 2.3.1 presented spectral characteristics previous studies found to be characteristic for molecules and combinations of organic matter. This comprises many features, but due to usually low concentrations of the absorbing material in soils, many of the features are often only weakly established. Among them, the wavelength ranges around 1730 and 2330 nm are assumed to be the most significant and because of this they are selected in this study for a delineation of organic carbon. In both wavelength regions, several weaker or stronger absorptions unique to organic matter overlap and in total lead to a noticeable absorption. They result of second and third overtones and combinations of functional groups associated with mainly C-H bands but also COO, CH_2, NH and CH_3. These groups are parts of organic compounds such as cellulose, lignin, starch, pectin, glucan, protein, wax and humic acid (Ben-Dor et al. 1997, Clark 1999, Viscarra Rossel and Behrens 2010). The intervals where continuum removal is performed which specify the ranges of spectral features to include, are between 1600 and 1815 nm and 2240 and 2410 nm, respectively. Table 5-1 lists a summary of soil organic matter related absorptions described in respective

literature included in the selected wavelength ranges. Due to the many nearly equally small absorptions overlapping in the region around 1730 nm, it is omitted to define a center wavelength giving the common position of the feature according to literature and used for the parameterization of the feature. For the second feature, center wavelength is defined at 2330 nm as this wavelength was given as the dominant position in most sources.

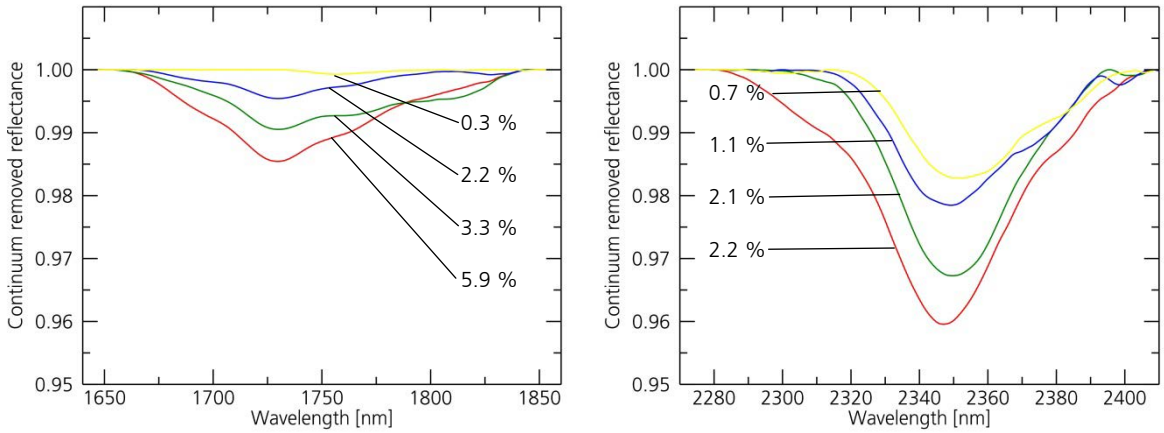


Figure 5-2: Detailed view of the soil organic carbon absorptions at 1730 nm (a) and 2330 nm (b) in continuum removed laboratory spectra. Corresponding contents of soil organic carbon are given.

Table 5-1: The spectral characteristics that are used for the delineation of soil organic carbon are given with details used for parameterization. Their wavelength position and assignment as they were described in previous studies are given.

Feature and type	Wavelength	Assignment	Reference
• **1730 nm (AF):** CR: 1600–1815 nm λ_{max}: not defined	1650 nm	2υ of aromatic CH stretch	Viscarra Rossel and Behrens (2010)
	1669 nm	2υ of aromatic CH stretch	Ben-Dor et al. (1997)
	~1700 nm	2υ of CH stretch	Clark (1999)
	1706/1754 nm	2υ of CH, alkyl doublet	Viscarra Rossel and Behrens (2010)
	1706 nm	4υ of aliphatic CH stretch	Viscarra Rossel and Behrens (2010)
	1726 nm	2υ of aliphatic CH stretch	Ben-Dor et al. (1997)
	1761/1769 nm	2υ of aliphatic CH stretch	Ben-Dor et al. (1997)
	1730 -1852 nm	4υ of methyl CH	Viscarra Rossel and Behrens (2010)
• **2330 nm (AF):** CR: 2240–2410 nm λ_{max}: 2330 nm	2275 nm	3υ of aliphatic CH stretch	Viscarra Rossel and Behrens (2010)
	2279 nm	3υ of CH_2, CH_3	Ben-Dor et al. (1997)
	~2300 nm	υ+υ of CH stretch	Clark (1999)
	2307-2469 nm	3υ of methyl CH stretch	Viscarra Rossel and Behrens (2010)
	2309 nm	3υ of aliphatic CH stretch	Ben-Dor et al. (1997)
	2331 nm	3υ of CH_2, COO	Ben-Dor et al. (1997)
	2337/2386 nm	3υ of COO, CH_3	Ben-Dor et al. (1997)
	2347 nm	3υ of aliphatic CH stretch	Ben-Dor et al. (1997)
	2381 nm	3υ of CO stretch of carbohydrates	Viscarra Rossel and Behrens (2010)
• **450–740 nm (HF):**	Decrease in reflectance in the visible range		e.g. Bartholomeus et al. (2008), Baumgardner et al. (1985), Ben-Dor et al. (1997), Hill and Schütt (2000)
• **1460–1750 nm (HF):**	Decrease in reflectance in the near to shortwave infrared range		e.g. Bartholomeus et al. (2008), Baumgardner et al. (1985), Demattê and Garcia (1999), Hill and Schütt (2000), van der Meer and de Jong (2001)

Abbreviations: AF: absorption features, HF: hull features, λ_{max}: wavelength of maximal absorption predominantly found in literature, CR: range where continuum removal is performed, υ: overtone bands, υ+υ: combination bands of fundamentals and overtones.

To take into account the general decrease in reflection of organic carbon rich soils, for instance in the visible due to the broad absorption of humic acid, also features describing the characteristic shape of the spectral curve are included. There exist several studies that employed different wavelength regions in order to describe this property (e.g. Bartholomeus et al. 2008, Ben-Dor et al. 1997). In this study, two ranges are applied to take into account the overall shape of the reflectance spectrum - one range in the visible between 450 and 740 nm and a second range in the near to shortwave infrared range between 1460 and 1750 nm is selected as HF. To reduce the influence of eventually remaining vegetation signals in the image data, the VIS range is extended from 700 nm, as was used by Bartholomeus et al. (2008), to 740 nm

Tests on the inclusion of spectral features for the delineation of soil organic carbon show that the absorptions around 2100 nm which can be related to organic soil materials and were used in several previous studies (e.g. Bartholomeus et al. 2008) are too weak to be included as diagnostic spectral feature. The spectral variables of such a feature are highly variable in regression analysis, have low regression coefficients and sometimes even show up as non-significant in appropriate quality checks (see 5.1.3). Thus, the 2100 nm feature is not included in the consideration of soil organic carbon since it would only provide a statistical adaptation within regression analysis.

This results in two AF and two HF that are used for the delineation of soil organic carbon. Some of these features can also be caused by other soil properties (compare section 2.3). Ambiguous or overlapping spectral characteristics are presented and discussed at the end of this section. Figure 5-2 depicts the selected absorptions around 1730 and 2330 nm in laboratory spectra of exemplary samples.

Iron oxides

The most common iron oxides hematite (Fe_2O_3) and goethite (FeOOH, precisely a hydroxide) are found as clay-sized particles of free iron oxides in soils. For the determination of iron oxides in soils strong absorptions occurring in the visible range around 550 nm, 700 nm and near 900 nm are used which result from electronic transition bands of the bi- and trivalent iron ions (Clark 1999, Hunt and Salisbury 1970). These include all the known features in the VIS to SWIR spectral range characteristic of hematite and goethite. A decrease in reflectance towards the ultraviolet wavelength region results from a broad Fe absorption band (see section 2.3) and is included as two features of the spectral curve. The change in reflectance in the VIS is described by the slope of the spectral curve between 550 and 590 nm (CF feature). Additionally, the shape

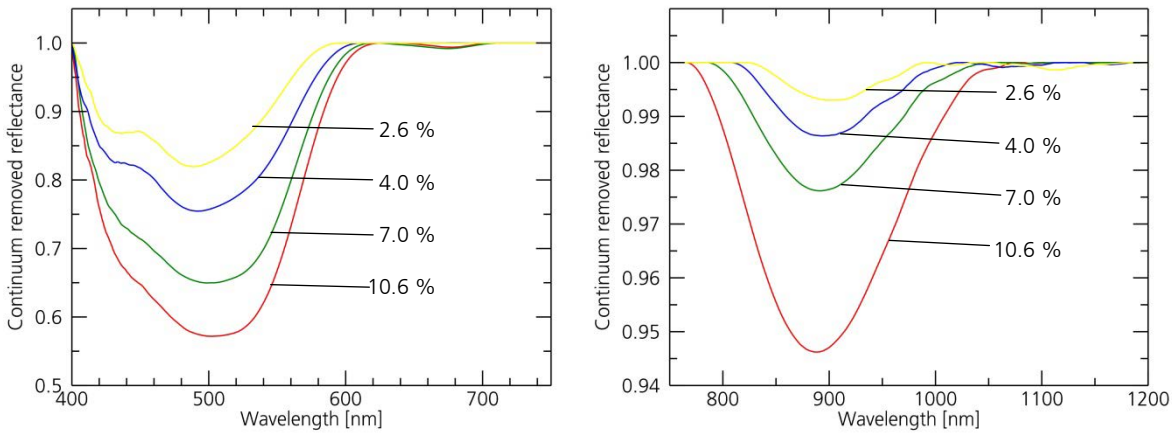


Figure 5-3: Detailed view of the considered iron oxides absorptions at 550 nm (a) and 900 nm (b) in continuum removed laboratory spectra. Corresponding contents of iron oxides (Fe_d) are given.

of the continuum over the entire VIS range is included as HF feature. This is to pick up the shape of the spectral curve without the influence of the 550 nm absorption. The spectral features selected for the determination of iron oxides thus include three AF, one CF and one HF feature. As iron is the only soil constituent triggering distinct spectral absorptions in the visible and beginning near-infrared range, the features are unique and there is no ambiguity. Only the shape of the reflectance curve in the VIS range is also included for other soil constituents (both organic carbon and clay). Figure 5-3 illustrates iron absorptions in continuum removed laboratory spectra.

Table 5-2: The spectral characteristics that are used for the delineation of iron oxides are given with details used for parameterization. Their wavelength position and assignment as they were described in previous studies are given.

Feature and type	Wavelength	Assignment	Reference
• **550 nm (AF):** CR: 450–680 nm λ_{max}: 550 nm	480 nm	ET band of goethite	Viscarra Rossel and Behrens (2010)
	~485-520 nm	ET band of Fe^{2+}	Ben-Dor et al. (1999)
	~490 nm	ET band of Fe^{3+}	Hunt and Salisbury (1970)
	503 nm	Goethite	Grove et al. (1992)
	~510 nm	ET band of Fe^{2+}	Hunt and Salisbury (1970), Hunt et al. (1971)
	529 nm	ET band of hematite	Viscarra Rossel and Behrens (2010)
	535 nm	Hematite	Grove et al. (1992)
	~550 nm	ET band of Fe^{2+}	Hunt and Salisbury (1970), Hunt et al. (1971), van der Meer and de Jong (2001)
	550 nm	ET band of hematite	Morris et al. (1985)
• **700 nm (AF):** CR: 580–800 nm λ_{max}: 700 nm	650 nm	ET band of hematite and goethite	Morris et al. (1985), Viscarra Rossel and Behrens (2010)
	665 nm	ET band of goethite	Grove et al. (1992)
	~665-700 nm	ET band of Fe^{3+}	Ben-Dor et al. (1999)
	700 nm	4υ of O–H	Viscarra Rossel and Behrens (2010)
	~700 nm	ET band of Fe^{3+}	Baumgardner et al. (1985), Hunt and Salisbury (1970), Hunt et al. (1971), van der Meer and de Jong (2001)
• **900 nm (AF):** CR: 750–1300 nm λ_{max}: 870 nm	~850 nm	Hematite	Hunt (1982)
	850-870 nm	ET band of hematite	Morris et al. (1985)
	~850-860 nm	ET band of Fe^{3+}	Ben-Dor et al. (1999)
	860 nm	ET band of Fe^{2+} and Fe^{3+}	Clark (1999)
	868 nm	Hematite	Grove et al. (1992)
	~870 nm	ET band of Fe^{3+}	Baumgardner et al. (1985), Hunt and Salisbury (1970), Hunt et al. (1971)
	884 nm	ET band of hematite	Viscarra Rossel and Behrens (2010)
	~900 nm	Transition bands of Fe^{2+} and Fe^{3+}	Clark (1999), Hunt et al. (1971), van der Meer and de Jong (2001)
	900-930 nm	ET band of goethite	Morris et al. (1985)
	920 nm	ET band of goethite	Viscarra Rossel and Behrens (2010)
	~930 nm	Goethite	Hunt (1982)
	930 nm	3υ of O–H stretch	Viscarra Rossel and Behrens (2010)
	932 nm	Goethite	Grove et al. (1992)
	940 nm	υ+υ of O–H stretch	Viscarra Rossel and Behrens (2010)
	~1000-1100 nm	ET band of Fe^{2+}	Baumgardner et al. (1985), Hunt and Salisbury (1970), Hunt et al. (1971), van der Meer and de Jong (2001)
	1025 nm	ET band of Fe^{3+}	Ben-Dor and Banin (1994)
	1075 nm	ET band of Fe^{3+}	Ben-Dor and Banin (1994)
• **550–590 nm (CF):**	Decrease in reflectance in the blue wavelength range towards the ultraviolet light		e.g. Clark (1999), Hunt et al. (1971), Morris et al. (1985)
• **450–750 nm (HF):**	Decrease in reflectance in the visible range		e.g. Clark (1999), Hunt et al. (1971), Morris et al. (1985)

Abbreviations: AF: absorption features, CF: curve features, HF: hull features, λ_{max}: wavelength of maximal absorption predominantly found in literature, CR: range where continuum removal is performed, ET: electronic transition bands, υ: overtone bands, υ+υ: combination bands of fundamentals and overtones.

Clay minerals

The most prominent diagnostic spectral feature of clays are the Al-OH combination bend around 2200 nm and a minor hydroxyl overtone feature around 2340 nm. Slight shifts of the characteristic wavelength position, especially of the 2200 nm feature, are due to the specific type of clay mineral mixture and its chemical composition and the alteration level of the individual clay minerals (see section 2.3). This leads to a certain spread of the wavelength position given in literature. Illite and montmorillonite are the dominant clay minerals in the study area (Mills 2009) and as such, 2206 nm is chosen as center wavelength of this feature as it is the most cited band position of illite and montmorillonite. The kaolinite doublet feature appears only subsidiary in some field spectra. As the strong absorptions at 1400 and 1900 nm which are due to bound water and hydroxyl in some clay minerals are masked by atmospheric water, they are not usable for the quantification of clay content from field or airborne data. Effects of an increasing content of finer grain sizes lead to an increase in reflectance (e.g. Ben-Dor et al. 2003, Clark 1999, Stoner and Baumgardner 1981, van der Meer and de Jong 2001). They are included as shape of the spectral curve in the VIS and also in the NIR/SWIR range (HF features). This results in two AF and two HF to be used for the determination of soil clay content. Thereby the 2340 nm feature is not unique to clay minerals. Figure 5-4 gives an example of laboratory spectra of samples showing different clay contents.

Discussion and handling of overlapping and ambiguous spectral features

Some of the spectral features described in this section do not fulfill the requirement of being unique to one material. They can be divided into two groups. (1) Two or more spectral absorption features each diagnostic for one material overlap in a specific wavelength region and cannot be resolved as separate features. (2) The reflectance in a broad wavelength range may be caused by multiple properties, for instance the trend of the reflectance in the VIS. Thus, those spectral features cannot be seen as really diagnostic. The decomposition of those spectral characteristics into the fractions of the properties contributing to it, for instance by the application of a recursive function, could be helpful to solve the problem of ambiguous features. However, this could not be implemented in this study. Nonetheless, tests on the inclusion of those spectral features showed that these features are still valuable and their inclusion should be ranked higher than the ambiguities resulting from them. The individual feature ambiguities and their implications are discussed in detail in the following.

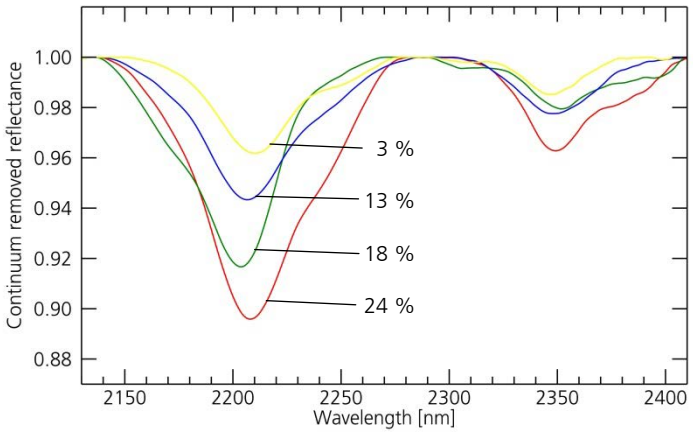


Figure 5-4: Detailed view of the clay mineral absorptions at 2200 nm and 2340 nm in continuum removed laboratory spectra. Corresponding clay contents are given. For the absorption around 2340 nm a non-linearity between chemical content and the depth of the absorption is obvious, which may be a result from its overlap with the organic carbon absorption around 2330 nm (see below for further details).

Table 5-3: The spectral characteristics that are used for the delineation of soil organic carbon are given with details used for parameterization. Their wavelength position and assignment as they were described in previous studies are given.

Feature and type	Wavelength	Assignment	Reference
• **2200 nm (AF):** CR: 2100–2290 nm λ_{max}: 2206 nm	2160/2168 + ≈ 2200 nm	Kaolinite	Grove et al. (1992)
	2163+2209 nm	δ of AlOH bend of kaolinite doublet and υ+υ of OH stretch	Clark et al. (1990)
	2160+2208 nm	δ of AlOH bend of kaolinite doublet and υ+υ of OH stretch	van der Meer and de Jong (2001), Viscarra Rossel and Behrens (2010)
	≈ 2200 nm	δ of AlOH bend and υ+υ of OH stretch of montmorillonite and illite	Baumgardner et al. (1985) Clark et al. (1990), Hunt (1977), Hunt (1982), van der Meer and de Jong (2001)
	2200/2204 nm	Montmorillonite	Grove et al. (1992)
	2206 nm	υ of OH stretch of montmorillonite	Lagacherie et al. (2008), Viscarra Rossel and Behrens (2010)
	2206 nm	υ of OH stretch of illite	Viscarra Rossel and Behrens (2010)
	2208 nm	δ of AlOH bend of kaolinite doublet	Viscarra Rossel and Behrens (2010)
	2216 nm	illite	Grove et al. (1992)
	2230 nm	δ of AlOH bend of smectites	Viscarra Rossel and Behrens (2010)
• **2340 nm (AF):** CR: 2270–2410 nm λ_{max}: 2340 nm	2308/2312 nm	Kaolinite	Grove et al. (1992)
	2336 nm	Illite	Grove et al. (1992)
	2340 nm	υ of OH stretch of illite	Chabrillat et al. (2002), Clark et al. (1990), van der Meer and de Jong (2001), Viscarra Rossel and Behrens (2010)
	2372/2376 nm	Kaolinite	Grove et al. (1992)
• **450–700 nm (HF):**	Overall increase in reflectance in the visible range		Ben-Dor et al. (2003), Clark (1999), Stoner and Baumgardner (1981), van der Meer and de Jong (2001)
• **1460–1750 nm (HF):**	Overall increase in reflectance in the near to shortwave infrared range		Ben-Dor et al. (2003), Clark (1999), Stoner and Baumgardner (1981), van der Meer and de Jong (2001)

Abbreviations: AF: absorption features, HF: hull features, λ_{max}: wavelength of maximal absorption predominantly found in literature, CR: range where continuum removal is performed, δ: fundamental bands, υ: overtone bands.

In the wavelength range 2330 to 2340 nm there are several overlapping absorptions: organic materials around 2330 nm, certain clay minerals around 2340 nm and a prominent absorption of carbonate minerals around 2345 nm (compare section 2.3). However, the presence of the spectral signature of carbonate minerals can be neglected as result of the local conditions present in the study area, where there is no carbonate bedrock and carbonate precipitation sediments are only found very locally and mostly in deeper soil layers. Where these are revealed at the surface, they potentially can influence the derivation of soil organic carbon and clay content by an increased absorption at this wavelength range and additionally by their brighter reflectance over nearly the entire VIS to SWIR range. However, the occurrence of carbonates does limit the transferability of the feature-based quantification approach to carbonate-rich areas. The overlap of the organic carbon and clay feature around 2330 to 2340 nm cannot be decomposed but is included as descriptive for both constituents in the same way.

Another ambiguity in regard to soil organic carbon results of the spectral features of dry plant materials and soil organic matter. It is not possible to resolve the origin of a spectral feature only from analysis of the individual feature since soil organic carbon is built of dry plant materials and their derivation products, and as such, spectral features are similar. The spectral influence of dry plants is removed in the further processing of the HyMap data by a combination of spectral unmixing and a subsequent approximation of the soil signature contained in a mixed pixel which is presented in section 5.2. Within this analysis, dry plant

materials are identified by spectral unmixing algorithms because they are based on the joint occurrence of several distinct features in the entire available spectrum (e.g. the flat shape of the spectral curve in the VIS/NIR combined with specific NIR/SWIR features). If a vegetation signature remains in the image spectra after this procedure, it may bias the determination of soil organic carbon due to increased carbon related absorptions.

As is shown in section 2.3 the overall reflectance of a soil over the entire spectrum is the result of all the spectral properties of a soil's chemical constituents and its physical properties. This means that HF, describing the shape of the continuum of a spectrum, are likely not unique to one parameter. The shape of the spectral curve in the VIS (450-700 nm) is included as feature for all three constituents and the shape of the curve in the SWIR (1460-1750 nm) is included for the determination of soil organic carbon and also clay content. The effect of each of the considered soil constituents on these spectral regions is physically caused and well-documented (see above and section 2.3). Also for these spectral characteristics the fraction of each constituent and additional influencing factors to the measured reflectance cannot be determined. Additionally factors are known that have a significant effect on the overall reflectance of a soil, with soil moisture being the most prominent one (see section 2.3). However, an influence of soil moisture can be neglected, because all soils were dry at the time of the field measurements and the airborne survey as it is typical for the local climate conditions and it is shown by local weather data (ARC-ISCW 2011).

5.1.2 Feature parameterization

In parameterization, the spectral datasets are analyzed for the selected spectral characteristics of the three introduced types of features (AF, CF and HF see above) and are transferred to numerical parameters describing the shape of the spectral features. These derived parameters are called "spectral feature variables" and are used for subsequent regression analysis. The following section and figure 5-5 describe the determination of spectral variables for every type of spectral features.

AFs are parameterized from continuum removed reflectance spectra. Continuum removal, as is presented in section 2.1 is calculated individually for an interval around each feature being mapped (CR in table 5-1). The wavelength location of the actual feature shoulders ($S_{left/right}$) is calculated individually for each spectrum as the points where the spectral curve first drifts away and then joins back the continuum. AF are described by the following six variables, similar to spectral absorption feature analysis described by other authors (e.g. Clark et al. 2003, Grove et al. 1992, van der Meer 2004): depth (d_{max}) and wavelength (λ_{dmax}) of maximal absorption, absorption at supposed characteristic wavelength according to respective literature (λ_{lit}), feature width as distance between the two shoulders (w), the area between normalized continuum and spectral curve (A) and asymmetry of the feature (A_{left}/A_{right}) as shown in figure 5-5b. The maximal depth of an absorption relative to the continuum is calculated based on equation 2.7 in section 2.1. It has to be mentioned that the absolute absorption depth of features near the spectrum's borders, e.g. the iron feature around 550 nm, is controlled very much by the extent of the spectrum as this determines the starting point of the continuum. The range of all spectra in this study is fixed from 450 to 2450 nm as was defined in section 3.2.

To describe the shape of the spectral curve in a defined wavelength range, two characteristics are introduced. CFs describe changes in reflectance occurring in a specific wavelength range that are characteristic for the surveyed parameter. This is relevant to describe the gradient of the spectral curve in

the visible that occurs in consequence to the strong absorption of iron oxides towards the ultraviolet, of which only the right slope is visible in the wavelength ranges surveyed by VIS to SWIR spectroscopy. A CF is characterized only by the mean slope of the spectral curve (sl_{ρ}), calculated from a line fitted in the given wavelength range. Second, HF are introduced to describe spectral shape without the influence of local absorptions. These features capture a soil constituent's effect on a broader spectral region that is not expressed as absorption feature. For example, this can be a change in the overall reflectance caused by a decrease in the average particle size, an increased presence of organic carbon or water in the soil. As important wavelength regions where those changes occur, one region in the VIS (450 to 740 nm) and a second in the SWIR range (1460 to 1750 nm) are chosen (see section 5.1.1). HF features are parameterized by mean slope ($sl_{\rho c}$) and mean reflectance ($r_{\rho c}$) of the convex hull of a spectrum in a defined wavelength range. Both are calculated also by using a line fit.

The parameterization of the three types of spectral features from a sample spectrum results in a group of spectral variables for each spectrum and soil parameter (e.g. d_{max} of AF1, λ_{dmax} of AF1, etc.). Table 5-4 shows

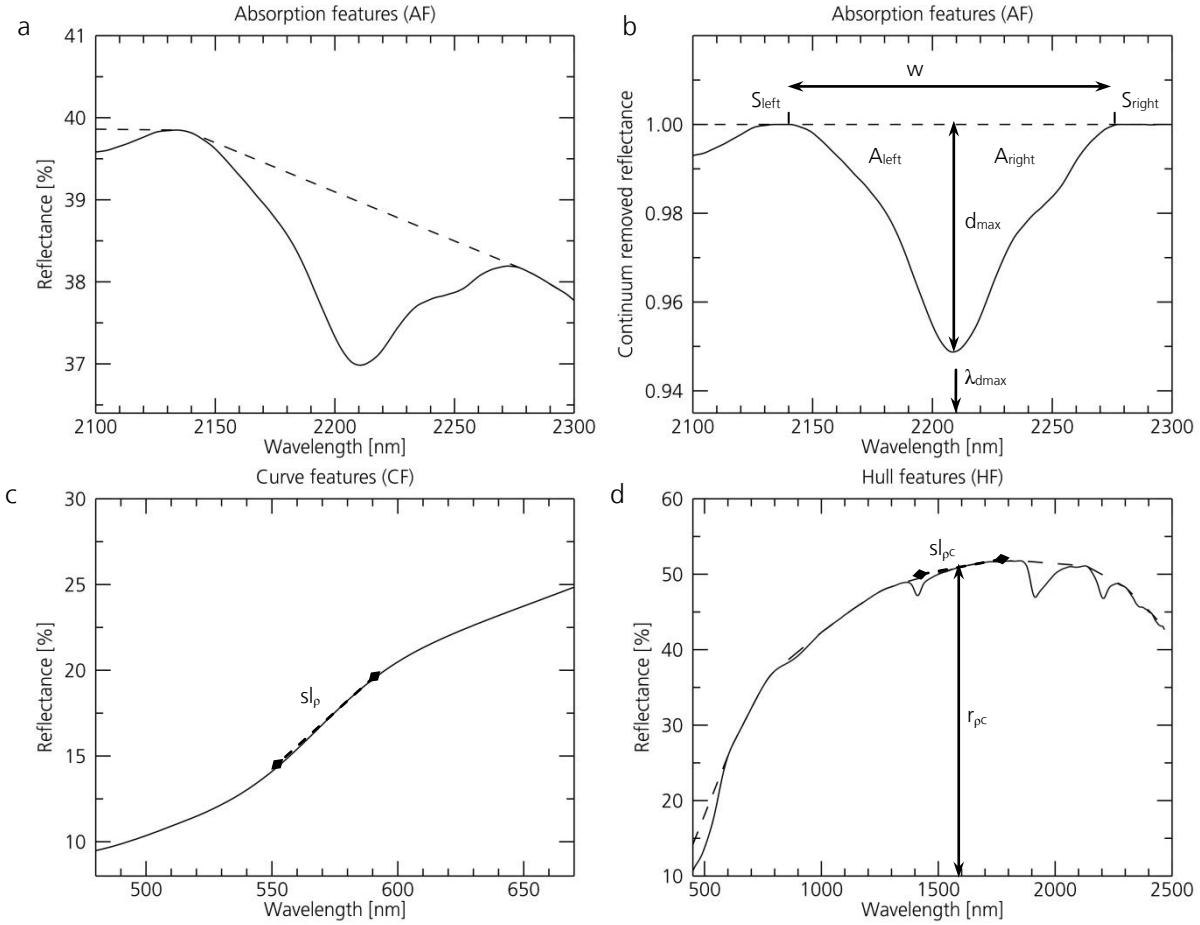


Figure 5-5: Parameterization of spectral features by the calculation of spectral feature variables for the assigned three types of spectral features is shown for an exemplary reflectance spectrum measured in the laboratory.
⁓ Reflectance spectrum, ·– spectral continuum, ◆→ line fitted in given interval.
a, b: reflectance spectrum and continuum removed reflectance of the clay absorption feature (AF) near 2200 nm and determination of six spectral variables describing the shape of the feature, c: parameterization of curve features (CF) as changes in shape of the spectral curve in a given interval, d: feature variables calculated to describe the shape of the continuum in a given interval (HF). See text for used symbols and abbreviations.

the number of spectral features analyzed for the determination of soil organic carbon, iron oxides and clay and the resulting number of variables that describe the spectral shape and are used in regression analysis to determine parameter contents.

Table 5-4: Number of spectral features used and total number of variables for determination of soil constituents.

	Absorption features (6 variables)	**Curve features** (1 variable)	**Features of convex hull** (2 variables)	**Total number of variables**
Organic carbon	2[1]	-	2	15[1]
Iron oxides	3	1	1	21
Clay	2	-	2	16

[1]For C_{org} AF near 1730 nm only 5 variables are calculated because λ_{lit} is not defined (see section 5.1.1).

5.1.3 Model calibration by MLR

Multiple linear regression techniques are used to establish a functional regression relationship between the chemical contents of a soil constituent and the corresponding spectral signal. Ideally, the spectra used for calibration should be from the same sensor as the spectra the calibration finally is applied to, or the sensor characteristics should be as similar as possible. Due to the small scale heterogeneity of the South African study area, nearly every HyMap pixel is comprised of a mixture of several surface materials (see also section 5.2). Because of this, HyMap spectra recorded at the sample locations cannot be used for calibration directly. Even though in section 5.2.2 a method is proposed to approximate the soil signature of image pixels by the use of spectral unmixing, it is decided not to include the large influence of spectral unmixing in the calibration procedure. Thus, calibration is performed based on spectral field measurements. Throughout literature this approach is accepted as common practice for a transfer of the developed functional relationship to airborne image data (e.g. Schwanghart and Jarmer 2011 used laboratory spectra for calibration of C_{org} prediction models applied to HyMap data). Spectral datasets are pre-processed (see section 3.2.1 for details) and resampled to meet HyMap's spectral resolution prior to parameterization and regression analysis. A functional regression relationship is determined for soil organic carbon, iron oxides and clay content separately and independent from the other soil constituents. For this the same fully automated workflow as implemented in the IDL software is used. In each case the selected spectral features describing contents of the particular soil constituent are considered in parameterization. The calibration range for each model is defined by the chemical range of the reference samples collected in the field ("calibration limits"). Constituents' contents exceeding these ranges need to be extrapolated by the models and may therefore not be correct. Because the considered soil constituents all are at a comparably low level, it is assumed that the relationship between the chemical contents and the corresponding diagnostic features follows a linear trend and is not affected by a saturation of absorption features at high constituent contents as it is discussed in section 2.4. The following section explains every single step of the calibration workflow leading to the functional regression relationship. Figure 5-6 provides an overview of the sequence of steps and of important final and additional side outputs. Details regarding the technical implementation are included in appendix D.

Determination of spectral variables and quality checks

As a first step in the model calibration workflow, the spectral characteristics of the selected soil constituent to be modeled are parameterized from pre-processed field spectra as it is described in section 5.1.2. This

results in a group of variables for each spectrum that first undergo three independent quality checks. This control is needed to take into account the dataset's characteristics resulting e.g. from sensor signal-to-noise ratio or band width. It aims for an identification of problematic spectral variables prior to regression analysis. First, the values calculated for each spectral variable are analyzed for their variance. The normalized standard deviation (σ_{jn}, see table 5-5) gives the magnitude in variance in relation to the magnitude of values calculated, where the division by the mean includes the actual range of a variable's values and makes the standard deviation comparable. This is necessary to compare the variance for instance of absorption depths which show values below 1 [-] and of the wavelength position of an absorption with values between 450 and 2450 [nm]. For a further inclusion of a variable, σ_{nj} is required to be higher than 0.001. This excludes very small spectral variables which would only cause an undesirable statistical adaptation during regression analysis (e.g. the depth of the 2100 nm C_{org} absorption that is very weakly established).

Next, the list of spectral variables is tested for redundancies - two variables for which exactly the same values are determined from the spectra. This can happen for example due to HyMap's spectral resolution of 13 to 17 nm. For instance, the wavelength position of the maximal absorption depth (λ_{dmax}) and the absorption depth at the wavelength position according to literature (λ_{lit}) can result in the same band position for all spectra resulting in identical values calculated for the two variables. Duplicate columns must be eliminated prior to regression analysis because the method cannot deal with ambiguities. Considering the importance of the variables, the procedure is set to remove the variable λ_{lit} instead of λ_{dmax} in this case.

In a final step, the percentage of invariant values is determined for each variable. If identical values are determined e.g. for more than 75 % of a variable, it is assumed that calculated values are controlled by a superordinate factor (e.g. sensor band positions), and thus are not suitable for regression analysis. Specific spectral variables that fail to have the desired characteristics are excluded from the analysis. These tests were also used to identify the suitability of additional specific spectral features for model development, such as the organic carbon feature around 2100 nm. Table 5-5 sums up the thresholds set in various quality control steps of the calibration procedure.

Table 5-5: Summary of thresholds set in individual steps controlling the quality of derived spectral variables within the calibration workflow.

	Calculations and threshold	Reason
Initial quality check of calculated spectral variables	$\sigma_{jn} = \frac{\sigma_j}{j_{mean}}$	Exclude too small spectral variables to prevent them to provide undesired statistical adaptation
	$\sigma_{jn} > 0.001$	
	$x_j \neq x_k$	Exclusion of redundant columns because regression analysis cannot deal with ambiguity
	< 75 % identical values	Identify variables controlled by a superordinate factor and not suitable for regression analysis
Outlier samples and insignificant variables during cross-validation	$\lvert R^2_i - R^2_{i\,mean} \rvert > 2 \cdot \sigma_{R^2}$	Identification of outlier samples that have a high influence on established regression relationship
	$c_{j\,mean} > 2 \cdot \sigma_{cj}$	Significance criterion for spectral variables

Abbreviations: j, k: spectral variables, $x_{j,\,k}$: values calculated for variables j and k, σ_j: standard deviation of all x_j, j_{mean}: mean of all values x_j, σ_{nj}: normalized standard deviation of spectral variable j, R^2_i: determination coefficient calculated in cross-validation when sample i was excluded, R^2_{mean}: average of all R^2_i, σ_{R^2}: standard deviation of R^2, c_{jmean}: mean of correlation coefficients (c_j) of spectral variables j, σ_{cj}: standard deviation of c_j

The calculated feature variables are subsequently standardized by subtraction of their mean and division by their standard deviation which results in all variables having a mean of 0 and a standard deviation of 1. This allows an unbiased comparison of regression coefficients.

First MLR analysis

Based on the standardized spectral variables and the chemical reference data, an initial multiple regression analysis is performed according to the method described above including all samples and the spectral variables that passed the quality checks. The accuracy of the established functional relationship is given by common accuracy measures like determination coefficient (R^2), root mean squared error (RMSE), ratio of performance to deviation (RPD), etc. The measures used to assess the accuracy of a multivariate relationship are given in appendix C. This results an initial functional relationship that gives a first correlation without any further interaction.

Leave-one-out analysis is conducted to analyze the sample population for outliers and spectral variables for significance. Outliers are defined by Haaland and Thomas (1988) as "samples that are not representative of the calibration samples, and therefore, their estimated concentrations must be treated with caution". In this cross-validation (CV) in case of n samples, subsequently one sample (i) is removed from the pool and the remaining n-1 samples are used to set up a regression relationship with the determination coefficient R^2_i that is assigned to the sample that is left out. Comparison of R^2_i for all n samples enables the identification of outlier samples that, when included in the sample population, have a high influence on the regression relationship established, and thus its accuracy. A sample i is treated as outlier if the absolute deviation of R^2_i from the mean of all R^2_i is higher than twice the standard deviation of the R^2_i (see table 5-5). Samples identified as outliers can be removed from the population if a reason for this is eminent.

Spectral variables are considered insignificant for regression analysis if the absolute value of a regression coefficient's mean is smaller than twice its standard deviation determined in leave-one-out analysis (see table 5-5). This threshold is used as significance test with a 95 % confidence (see Kessler 2007). The influence of a variable identified as insignificant is random and thus negligible, even though its inclusion can slightly improve the functional relationship. Such variables should be removed to improve the robustness of the final functional relationship. In contrast to the removal of outlier samples, insignificant variables are removed by default in the calibration workflow, however the removal can also be manually prohibited (see appendix D for details).

Final MLR regression modeling and cross-validation

A final multiple linear regression model is established based on the significant feature variables and chemical reference values with possibly selected outlier samples removed. The model is then calibrated to predict contents between the lower and upper calibration limit that is defined by the range of the chemical reference values.

The calibration model is validated by an evaluation with cross-validation (CV), to assess the predictability of the calibrated model (Kessler 2007). This internal validation method uses the same dataset, the calibration is established with. The advantage of it is that all available samples can be used for calibration of the regression model (125 samples for field and 163 for laboratory datasets). Very often, validation of regression models is done using an "independent" test set that was previously separated for the entire dataset (though, in a technically sense also test sets out of one sample population are not really independent). Although this seems to be the preferred validation method in past and recent studies, the accuracies of test-set and cross-validation are usually in the same range, given that both are done at the correct level (e.g. both across different sampling sites). This is confirmed for the models based on the available datasets. Because of this, cross-validation is selected to take advantage of the entire sample

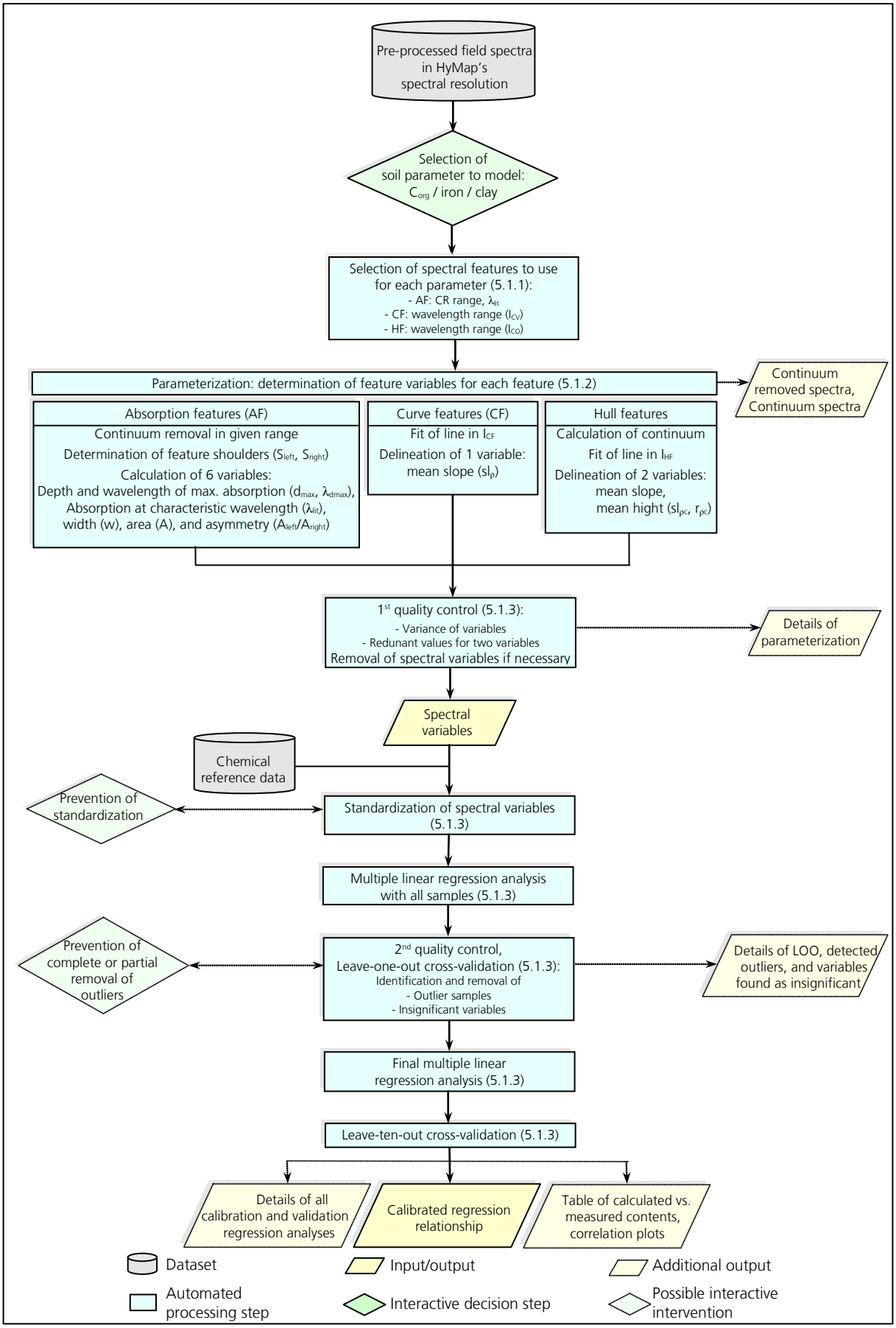


Figure 5-6: Overview of the sequence of steps and of important final and additional side outputs during the calibration of the soil constituents prediction models. The sections where the individual steps are explained in detail are given in parenthesis.

population leading to more robust calibration models in regard to the prediction of soil constituents from airborne imagery, than would be achieved by using only e.g. 2/3 of the sample population.

In the cross-validation, in subsequent steps, a part of the samples is excluded and a regression model is established using the remaining sample population (see above). The excluded samples then are predicted with the model calibrated with the other samples. The predictability of the regression model is determined based on the predictions for all samples. Here, 10 randomly determined samples were excluded for CV. Leave-ten-out cross-validation results are calculated as an average of 100 calculations since the result of a single run is influenced by the random selection of sample groups for the test sets.

Outputs of the calibration workflow

By the application of the described workflow, a functional regression relationship can be established based on a spectral library and related chemical reference data. The calibrated regression relationship also includes the determination coefficient (R^2_{Cal}) of the final calibration model and of the cross-validation (R^2_{CV}) the constant regression coefficient (b_0), the regression coefficients ($b_{1...m}$) for all m applied spectral variables, and their mean and standard deviation. This file is used to apply the established regression model to other datasets (point spectral data or image data). Enclosed metadata (e.g. calibration limits, number of bands, first and last band positions, etc.) ensure that the calibrated model is only applied to data of the same spectral resolution and band position. Figure 5-6 illustrates the steps during the calibration procedure, while appendix D shows details of the technical implementation with an overview of all generated files (Table D-11) and examples of the most crucial outputs (Figure D-10 and Figure D-11).

5.2 Extract spectral information from mixed soil pixels

The patchiness of natural vegetation that is inherent to most semiarid landscapes causes a large spatial heterogeneity of the study area. This is reflected in hyperspectral imagery in terms of mostly non-pure spectral signatures in image pixels depending on the imagery's geometric resolution of several meters (3.3 by 3.3 m for this HyMap imagery). Pixels that contain more than one cover type result in spectral mixtures. In semiarid environments this is often a mixture of bare soil and changing fractions of green and dry vegetation (Bachmann 2007).

The partial vegetation coverage limits the usage of Imaging Spectroscopy for soil applications, because the relationships between soil constituents and reflectance are disturbed by the presence of other materials, mostly vegetation (e.g. Mulder et al. 2011). A direct application of feature-based regression approaches is challenging since the spectral features are confounded if there is vegetation coverage (refer to section 2.3.4 and 2.4). This applies especially for the iron features in the VIS that are masked by the intense chlorophyll absorption of vegetation (Mulder et al. 2011, Stevens et al. 2008).

Results of a quantitative prediction of soil constituents thus can only be meaningful if the spectra the analysis is based upon are not influenced by other materials, such as vegetation. Thus, analysis has to be limited to pure soil signatures prior to the prediction of soil constituents. There are two possible approaches to address this – (1) pixels that are not dominated by bare soil are excluded for further processing, or (2) adequate methods that spectrally decompose the mixture are used to extract the spectral signal of the soil

that is of interest. Soil constituents for the excluded areas are interpolated afterwards from the non-vegetated areas (Ben-Dor et al. 2002).

An identification of soil dominated image information can be achieved by applying simple thresholds, for instance using vegetation indices, such as the normalized difference vegetation index (NDVI). Many vegetation indices are straightforward and usually based on a ratio of spectrally distinct features in the VIS/NIR bands to indicate whether the pixel contains green vegetation or not. They only perform well for the differentiation between green vegetation and bare soil, but are less suited to discriminate between soil and dry vegetation, since these materials show similar spectral shape in the VIS/NIR region. This is applicable in agricultural areas where bare soil surfaces are often pure and can be well discriminated from their vegetated surrounding. The application of a NDVI threshold to extract soil dominated pixels in semiarid areas where bare soil usually coexists with green and also dry vegetation, would lead to rough results where pixels classified as dominated by soil may still be influenced by dry vegetation, which would especially bias a spectral feature based delineation of soil organic carbon.

Spectral mixture analysis (SMA) is a more computationally intensive approach, but is able to provide an assessment of the spectral mixture with respect to the cover types since it applies the full spectral range hyperspectral imagery provides. In SMA, the spectra of pure surface materials (the endmembers, EM) can directly be derived from imagery and used to calculate the fraction of each endmember contributing to the spectral mixture of a pixel (see e.g. Settle and Drake 1993). In previous studies dealing with the prediction of soil constituents, spectral unmixing was often applied and further analysis was limited to soil dominated areas (e.g. Selige et al. 2006, Schwanghart and Jarmer 2011). For instance Schwanghart and Jarmer (2011) allowed vegetation coverage up to 10% to determine soil dominated pixels that were further used for the prediction of soil organic carbon. Taking into account that the large spatial heterogeneity of the South African study area mostly prevents pure bare soil pixels, even the consideration of soil pixels including a certain amount of vegetation such as 10 %, would result in very few pixels that can be analyzed. This would provide no spatially continuous coverage. An increase of the included vegetation coverage would, however, increasingly bias the predictions.

An approach where unmixing ground cover fractions were used to remove the influence of maize from the spectra of partly vegetated agricultural fields in Belgium was recently described by Bartholomeus et al. (2011). They showed that the error in soil organic carbon prediction rises with vegetation coverage and that this can partly be overcome by the application of what they called Residual Spectral Unmixing (RSU). However, their application was limited to agricultural areas where spectral mixtures comprised of two endmembers and could be well described and modeled. So far this was not tested in natural areas, where spectral mixtures are more complex due to many spatially variant surface materials.

To allow for the prediction of soil related information even in a natural ecosystem with intense spectral mixing of surface materials, in this study an approach is developed to approximate the soil signature contained in a mixed pixel. In a first step, spectral mixture analysis is used to deconvolve pixels that contain a mixture of soil and vegetation into their constituent components (section 5.2.1). In a second step, the resulting unmixing information is used to approximate the soil signature included in a mixed soil pixel (section 5.2.2). The procedure applied here prepends the feature-based delineation of soil information. It allows the methodology to be applicable to image data with a ground resolution of several meters and the application in areas where soil is the dominant land cover but coexists with vegetation.

5.2.1 Spectral mixture analysis

Typically, the first step in SMA is to find the endmembers as spectral signatures of pure ground materials comprising to the mixture and the second step to unmix the mixed pixel spectra as combinations of endmember spectra. This results in estimates of ground cover fractions (van der Meer et al. 2012). Successful spectral unmixing requires a good estimation of the number of endmembers present in a hyperspectral scene. The spectra of endmembers as pure surface materials are the pixel spectra of an image that lie at the vertices in the n-dimensional space, with n being the number of spectral bands. As such the endmember spectra are often selected directly from image data. The usage of direct field measurements as EM's would involve pre-knowledge about the different materials occurring in the scene to ensure the completeness of the EM set and also disadvantages in terms of identical measurement conditions (Rogge et al. 2006). Semi-automated endmember extraction tools apply different approaches to determine the endmembers of an image, such as the N-FINDR, vertex component analysis (VCA) and orthogonal subspace projection (OSP). Some of them include a manual post-processing and selection of the final endmember set to be used in spectral unmixing. Selected endmembers may be grouped to land cover classes. As suggested by several studies (e.g. Bachmann 2007, Asner and Lobell 2000), the spectral mixture for semiarid environments can be modeled by a combination of endmembers of the three classes "photosynthetic active vegetation" (PV), "non-photosynthetic active vegetation" (NPV), and "bare soil" (BS). This separation is also applied for a classification of the EMs and proper spectral unmixing in this study.

Subsequent spectral mixture analysis (SMA) applies the knowledge of endmember spectra to classify mixed pixels by deconvolving each pixel spectrum into fractional abundances of its endmembers (Adams et al. 1986 and 1993). Recent algorithms apply a linear approach, in which the reflectance of an image pixel is assumed to be a linear combination of the fraction of each endmember and the reflectance of an endmember. The following basic equation of the linear mixture model can be solved by a least squares approach:

$$\rho_S = \sum_{1}^{n} \rho_{EMn} \cdot f_{EMn} + E \qquad [-] \qquad (5\text{-}1)$$

with the mixed pixel spectrum ρ_S, the endmember spectra ρ_{EMn}, their fractional cover estimates f_{EMn}, n as the number of endmembers comprising the mixture of the pixel and the error term E. Fractional abundances are most accurate in SMA, when only the endmembers that comprise the spectral mixture of a pixel are used. Linear mixture assumes a spatial segregation of the materials comprising the mixture, with multiple scattering on a small-scale mixture of the surface materials (non-linear mixture) is neglected (Keshava and Mustard 2002). In the following, the algorithms used for EM extraction and SMA and their workflow are presented.

Extraction of image endmembers

In this study, potential spectral endmembers are derived using the spatial-spectral endmember extraction tool (SSEE) described by Rogge et al. (2007). In addition to detect its endmembers based on spectral characteristics alone, the algorithm includes spatial information. The SSEE method works by analyzing a scene in subsets and comprises three main steps: (1) the application of singular value decomposition determines a set of basis vectors that describe most of the spectral variance for the image subsets, (2) the full image data are projected onto these locally defined eigenvectors to determine a set of candidate endmember pixels, and (3) spatial constraints are applied to average spectrally similar endmembers (Rogge

et al. 2007). Using this semi-automatic approach, no a priori knowledge of the area is required and the EMs are only selected from the imagery. The application of image subsets within SSEE allows for an identification of endmembers that have a low spectral contrast with each other, but occur spatially independent. Figure 5-7 illustrates an example, where two land cover types (i, j) are spectrally similar but spatially independent. Due to their low spectral contrast the EMs may not be discriminated by using an exclusively on spectral information based approach that is directly applied to the full image. However, they do show up as two separate EMs because of their spatial separation when the image is analyzed in subsets like within the SSEE approach.

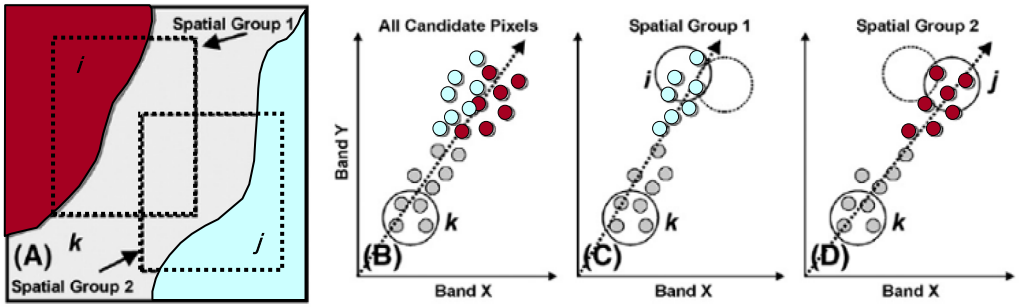
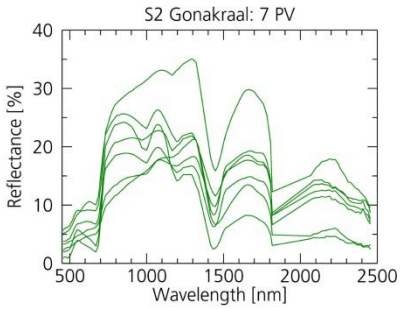
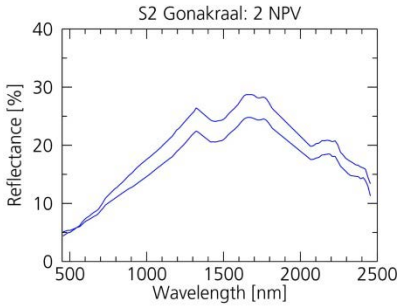
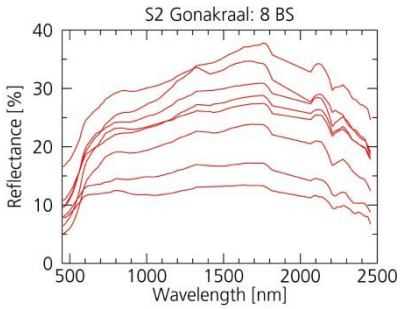


Figure 5-7: The spatial-spectral endmember extraction algorithm (SSEE) used for the derivation of image endmembers is able to detect local image endmembers even if they have a low spectral contrast to each other but occur spatially separated (adapted from Rogge et al. 2007).

Table 5-6: Classification criteria for derived image endmembers. The EMs of scale 2 subset of farm Gonakraal (see section 5.4) are given as example.

Class	Spectral characteristics	Sample EM set
Photosynthetic active vegetation **- PV -**	✓ Low to medium reflection in VIS ✓ High reflection in NIR ✓ Low reflection in SWIR ✓ Different degree of chlorophyll absorptions in VIS ✓ Strong water absorptions in NIR and SWIR ✓ Absorptions of plant materials in NIR and SWIR ✗ No clay absorption in SWIR	
Non-photosynthetic active vegetation **- NPV -**	✓ Spectrally "flat" in VIS/NIR ✓ Reflection increases continuously in VIS/NIR ✓ Absorptions of dry plant materials and organic matter in NIR and SWIR ✗ No distinctive absorption features in VIS/NIR ✗ No clay absorption in SWIR ✗ No water absorptions in NIR/SWIR	
Bare soil **- BS -**	✓ High reflection in VIS, NIR and SWIR ✓ Typical clay feature in SWIR observable ✓ Absorptions of iron oxides in VIS/NIR, organic matter and clay in the SWIR ✗ No chlorophyll absorptions in VIS/NIR	

The SSEE algorithm is used to derive image endmembers from the pre-processed HyMap imagery. During pre-processing, artificial objects are masked (see section 3.2.2) and thus the SSEE endmember sets only include natural materials. Settings for the SSEE processing are given in Table D-13 in the appendix (see Rogge et al. 2007 and references therein for details). From the pool of endmembers determined by the SSEE algorithm, a representative number of endmembers is manually selected depending on the present variance of the endmembers and the size of the input imagery. One EM set is prepared for each subset. For the entire northern and southern transect, a set of candidate EMs is derived by SSEE for each 1000 lines, which once more are projected onto the eigenvectors and the major endmembers are extracted for the final set of candidate EM of the entire regions. The selected EMs are further assigned to the classes PV, NPV and BS to facilitate the evaluation and allow for validation of the unmixing results. This is done based on their spectral features, with the classification criteria summarized in table 5-6.

Iterative spectral mixture analysis

Spectral unmixing is performed using an iterative spectral mixture approach (ISMA) which is described in detail in Rogge et al. (2006). ISMA applies an iterative approach to solve the linear mixture problem given in equation 5-1 and to determine which endmembers are best suited to resolve the mixture of the pixel. ISMA applies unconstrained unmixing based on the assumption that if the perfect EM set is used it is not necessary to use a constrained approach. A uniform reflectance shade endmember is also included.

For each image pixel, ISMA includes two main parts. First, unconstrained linear unmixing is applied using the pixel EM set. No further constraints apply, such as the usage of only one endmember per class, preference of mixtures of a similar RMSE according to spatial neighborhood, etc. The RMSE between the pixel spectrum and the modeled spectrum is calculated. From the EM set, the one EM with the lowest abundance (positive or negative) is removed (excluding the shade EM). If the number of pixel endmembers is >1 a new unmixing solution is calculated with the remaining EM. The process is repeated until only one EM and the shade EM remain. The root mean squared error (RMSE) then can be given as function of iteration as it is depicted in figure 5-8. In the second part of ISMA the change of the RMSE as function of iteration is analyzed to locate the critical point, where the optimal number of endmembers (n) is applied (figure 5-8). RMS will be low until a point, where actual EMs comprising the pixels mixture are removed in the iterative process. This is the critical iteration determining the optimal per pixel EM set, which is used to calculate the pixel fractional abundances.

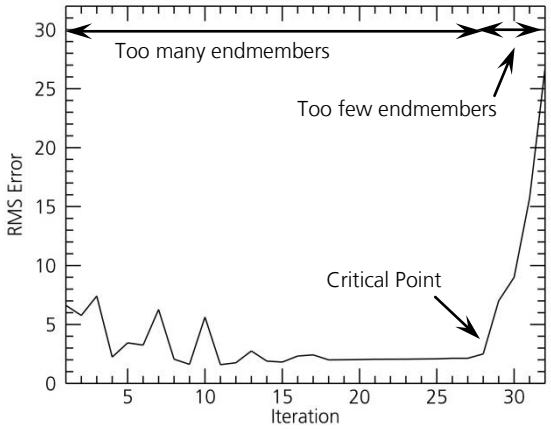


Figure 5-8: Graphical representation of the RMSE as a function of the number of iterations within ISMA spectral unmixing. The critical point marks the iteration between having too many and too few endmembers and thus the optimal endmember set to resolve the spectral mixture of the given pixel (after Rogge et al. 2006, redrawn based on hyperspectral imagery of the South African study area).

With the determination and application of the optimal set of endmembers to be used for the unmixing of each pixel, ISMA ensures a realistic unmixing resulting in low RMSE, non- negative abundances and the sum of the abundances close to 1. The settings used for processing the hyperspectral imagery with the SSEE endmember sets and the ISMA algorithm are given in Table D-13 in the appendix. The maximum number of endmembers to use for each pixel was set to 5.

5.2.2 Approximation of soil spectral signature from mixed pixels

The second step to prepare the hyperspectral imagery of the South African study area for a successful prediction of soil constituents includes the approximation of the soil signature contained in a mixed pixel. Thereby the unmixing fractional cover estimates are used to remove the spectral influence of vegetation on a non-pure soil spectrum. In the following the soil signature approximated from a mixed pixel is referred to as soil residuum (SR, ρ_{SR}). SR is described by the inversion of the linear mixture model by applying the endmember spectra of the classes PV and NPV contributing to the spectral mixture of the pixel and their cover fractions:

$$\rho_{SR} = \rho_S - \rho_{EM1} \cdot f_{EM1} - \rho_{EM2} \cdot f_{EM2} - \ldots - \rho_{EMn} \cdot f_{EMn} \quad [-] \quad (5\text{-}2)$$

with the reflectance ρ_{SR} of the soil residuum, the mixed pixel spectrum ρ_S recorded by the HyMap sensor, the spectrum of the individual endmembers of green and dry vegetation used to unmix the specific pixel ρ_{EM} and their cover fractions f_{EM}. SR then is supposed to include the spectral signal coming from the soil fraction and any unexplained signals which were not explained within spectral mixture analysis. To approximate the full "endmember" spectrum, the extracted residual soil spectrum is scaled by the summarized fractions of the m non-soil endmembers identified in the pixel.

$$\rho_{SR\,scl} = \rho_{SR} \cdot \left(1 - \sum_{1}^{m} f_{EMm}\right) \quad [-] \quad (5\text{-}3)$$

The spectral mixture in natural environments of a size of several square kilometers can be locally very complex due to land cover changes coupled with topography, land use, etc. This results in an endmember set which is selected for imagery of a given size not being perfect for the unmixing of every pixel. Thus, additional criteria are introduced which ensure the quality of the approximated soil residuum. The following quality requirements are set up to filter out pixels where SR might be not approximated properly due to the predominance of a land cover class that is not bare soil or insufficient unmixing. In the following, also typical values for each threshold are given, though they are scene dependent.

- A minimal soil fractional cover (e.g. 0.35),
- A maximal fractional cover of PV and NPV (e.g. 0.25),
- A maximum unmixing root mean squared error (e.g. 5),
- A range for the sum of the cover fractions around 1 (e.g. 0.4-1.4),
- A minimum mean reflectance of the approximated soil residuum, to filter out dark pixels where eventually unexplained signal highly influences the calculated soil residuum (e.g. 8%).

The threshold values need to be set manually according to the characteristics of the specific environment and unmixing quality to achieve good results for the particular imagery. They can be adapted to achieve better results for regional subsets if desired. It is assumed that for a pixel fulfilling these requirements, the soil signature included in the mixture can be approximated with a high accuracy. For the definition of the

thresholds, the unmixing abundances are not scaled. Tests if the scaling of the cover fractions allows a more accurate discrimination of unbiased and influences approximated soil residues did not provide proof that the scaling lead to significant advantages. Thus, unmixing ground cover fractions are applied as they are calculated by the unconstrained ISMA unmixing. Pixels failing the thresholds are masked for further analysis. In one of the last steps of prediction, the soil information of excluded pixels is interpolated to a certain extent by a selective filter operation (see section 5.3). This procedure results in a reflectance image of the bare soil surface, which is used to predict the spatial distribution of soil organic carbon, iron oxides and clay content in the Albany Thicket Biome, South Africa. Figure 5-9 gives an overview of the applied workflow.

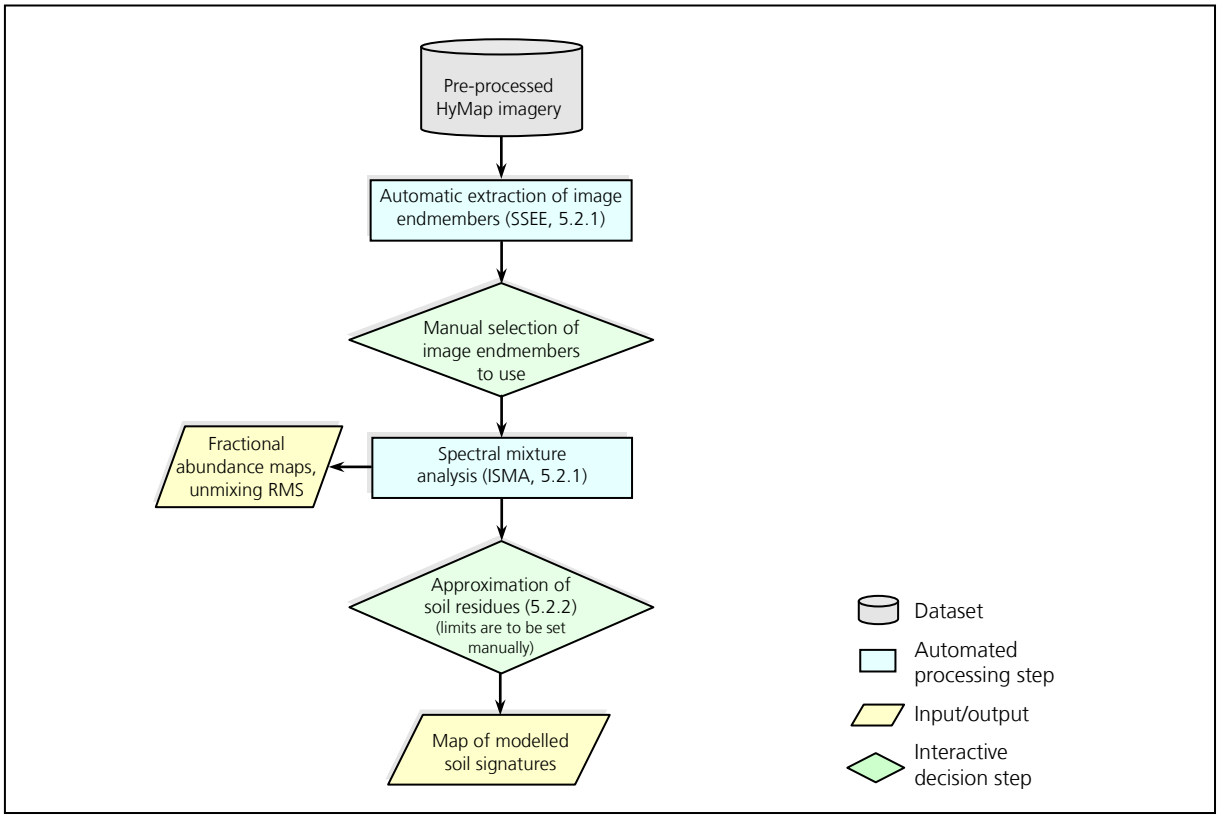


Figure 5-9: Processing workflow of pre-processed HyMap imagery to approximate the soil spectral information contained in mixed pixels that can be used to predict soil constituents. The sections where the individual steps are explained in detail are given in parenthesis.

5.3 Prediction of soil constituents from hyperspectral imagery

After calibration, the established functional relationship can be applied to predict soil constituents from spectra. The procedure is outlined in the following for the prediction of soil parameters from hyperspectral imagery in consistency with the objective of the study. The methodology can further be used for the prediction of soil constituents from spectral libraries of unknown or known corresponding chemistry (i.e. test set validation). These procedures are outlined in appendix D.2. The prediction applies the hyperspectral image data of approximated soil signatures which are produced according to the procedure described in section 5.2.2. To predict soil constituents in a first step, the image spectra are parameterized in the same way as during calibration. The spectral variables found significant in calibration are extracted and

standardized by the same parameters as it was done in calibration. Subsequently, the calibrated regression relationship is applied to the set of spectral variables of each pixel, which leads to a predicted constituent content for each image pixel, a soil signature is approximated.

Generation of the final map product of predicted soil constituents

The resulting initial map of predicted soil constituents is processed further to generate smooth map products. Pixel values exceeding the range of the chemical reference of the samples used for calibration (calibration limits) are first detected and masked. This results in the valid range of predicted soil constituents. As an additional step, a selective spatial filter is applied to the image of valid calculated contents to take account of neighborhood effects. The traditional way to account for this is the application of a smoothing filter, mostly a low pass filter with a kernel size of 3 or 5 (e.g. Bartholomeus et al. 2011). The filter operation here is slightly more complex and connected with an interpolation step due to the non-continuous pattern of pixels, for which soil constituents are predicted which results of the exclusion of individual pixels within the approximation of the soil residual signal. Spatial averaging is carried out with a low pass filter in a 3 x 3 neighborhood including only pixels for which a value is calculated and with a double weight given to the center pixel. Single pixels, where no parameter value is predicted, in particular pixels that are excluded during the approximation of the pixel soil residuum, are further interpolated in this step from the surrounding. Both the interpolation and filtering operation are only carried out if no more than 4 pixels are missing in the defined neighborhood. This additional constraint ensures that only the pixels at the border of larger data gaps are interpolated. If existent, larger spots where no soil information can be predicted remain unchanged in the data. To allow for applications that require a higher degree of generalization, the same filter is also applied to the dataset using a larger kernel size of 5 pixels. This product focuses on spatially larger variances in the spatial distribution of a soil constituent and removes a part of the small scale variability. Interpolation then is applied only when less than 6 pixels are missing, allowing for larger and less restricted spatial filtering.

Following this procedure, the prediction of soil constituents from hyperspectral image data results in a five band image file. In each band background pixels, for which no soil constituents are predicted, show a value of 0.

- Band 1: The contents of the predicted soil constituent calculated for each pixel are stored in this band without changes.
- Band 2: Values exceeding the calibration limits are masked, which results in this band showing the valid range of predicted soil constituents.
- Band 3: Result of the selective spatial filter and interpolation operation applied to band 2 with a kernel size of 3 x 3 pixels.
- Band 4: Result of the selective spatial filter and interpolation operation applied to band 2 with a kernel size of 5 x 5 pixels.
- Band 5: Pixel mask flagging pixels outside the range of the chemical reference data (operation producing band 2) with a value of 1 and pixels that were interpolated by the 3 x 3 selective filter operation leading to band 3 with a value of 2. This band is only used as input for the confidence assessment (section 5.3.1).

Band 3 is seen as the final result for the prediction of soil parameters. It represents the calculated parameter contents within the calibration limits and after a selective spatial filtering that provides the consideration of neighborhood effects but keeps the filter size small and thus retains small scale variability.

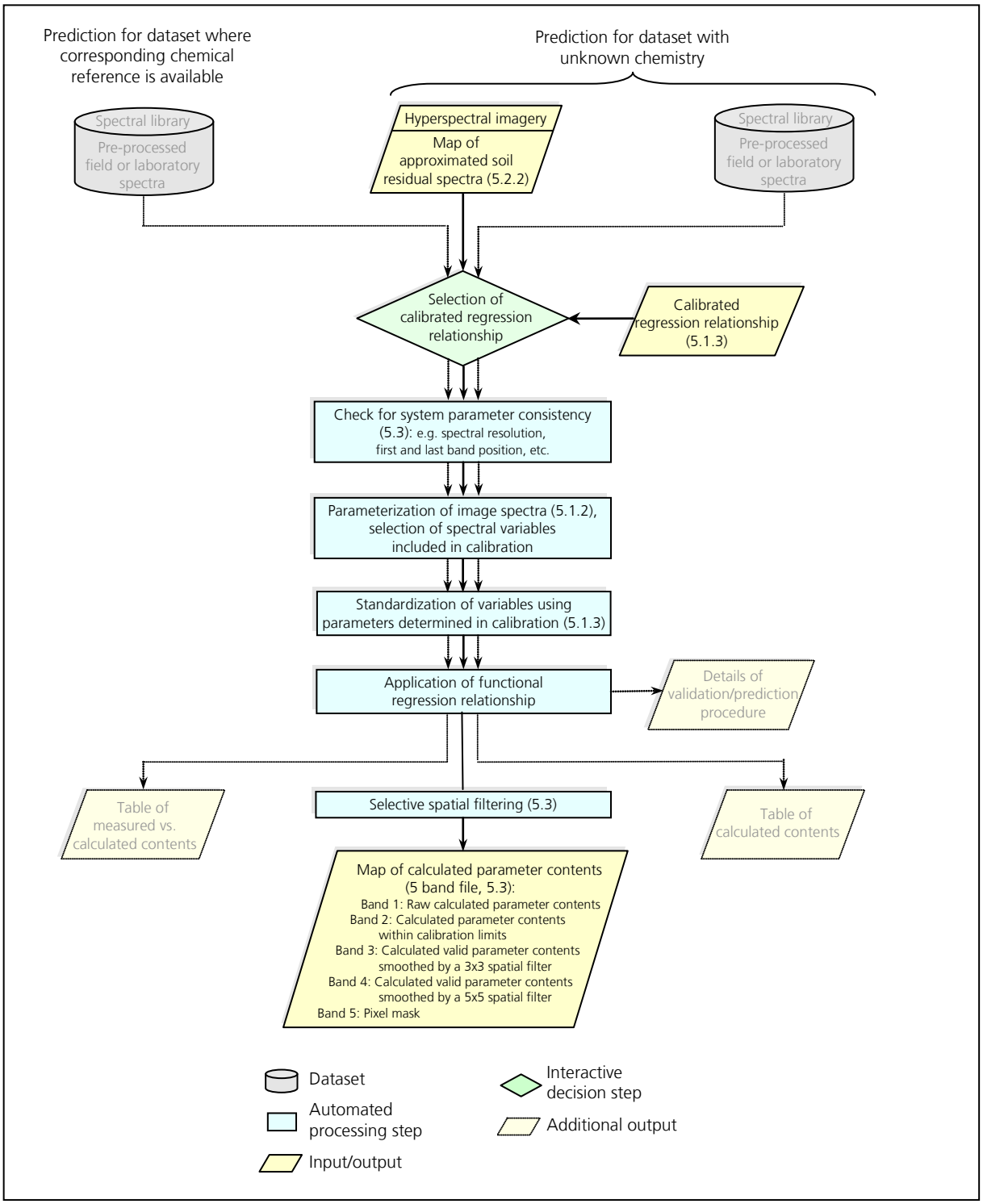


Figure 5-10: Procedure for the prediction of soil constituents from hyperspectral imagery. The prediction of soil constituents based on spectral libraries of known (validation) or unknown chemistry using a calibrated functional relationship is also indicated. This procedure is once applied for the prediction of each soil constituent. The sections where the individual steps are explained in detail are given in parenthesis.

Processing of entire dataset

The processing of the entire dataset of 320 km^2 comprising 52 Mio. pixels for the northern (2 979 x 17 451) and about 14 Mio pixels for the southern transect area (1 925 x 7 032) is realized in parts of 1 000 lines, each subset with a lateral extension between 1 600 and 1 750 samples. Input parameters for each step, e.g. the approximation of soil signatures, are tested in advance on a number of the subsets to meet the entire dataset and subsequently applied to the entire dataset. Appendix D shows details on the technical implementation of the workflow. Some processing times giving rough indicators of the expenditure of time are given in appendix D.6.

5.3.1 Additional performance and quality indicators of the computational approach

To assess the quality of each image pixel, for which soil constituents are predicted, a reliability measure is flagged in an additional data layer. The confidence layer comprises all data layers included in the workflow, like spectral unmixing results, approximated soil signatures and maps of calculated soil constituents. It starts with the calibrated and pre-processed HyMap imagery and ends with the calculation of quantitative information for soil dominated image pixels. For instance, the thresholds set for the approximation of soil residues (e.g. a minimal soil fraction, maximum vegetation influence, etc.) are treated as boundary conditions for the inclusion of pixels, for which soil information can be predicted (see section 5.2.2). If parameters approach these limits, intermediate and final products can show reasonable results but may as well be influenced. This is expressed in a reduced confidence.

The quality of initial image data is not included in this consideration. It is examined e.g. by the EUFAR quality indicators based on the raw radiance data (Bachmann et al. 2010). Local incidence angle is included as the only flight dependent parameter as the proportion of shadows influencing the quality of both the spectral unmixing and the approximation of the pixel soil signature rise with increasing viewing angle. For each contributing dataset one confidence layer is calculated where the dataset is classified based on its data range and presumed weight. The quality levels range from 10 for best conditions to 1 for reduced accuracy, close to the applied boundary conditions. A superior layer of the overall confidence is calculated as the average of five sub-layers with the additional constraint that the assigned final value may at most be two levels higher than the minimal single layer value. The confidence of maps of calculated clay content are assessed as side product but since the derivation of the clay content is problematic and has some limitations (see section 6.1.2) it is not included in the assessment of the prediction confidence. The superior confidence layer is source for the final three quality flags “good”, “medium” and “reduced”. A pixel flagged as “reduced” may show reasonable contents but indicates that there are several factors indicating limited validity. The designed confidence index is to be seen as a summarized indicator for a possible reduction in quality of the derived parameter maps which is established based on the involved data and to facilitate the identification of critical pixel properties. Figure 5-11 depicts the contributing layers and the applied parameters. A full overview file of the confidence index calculated for a specific dataset and soil parameter analysis showing the detailed set-up and listing applied increments can be found in the appendix D.3.

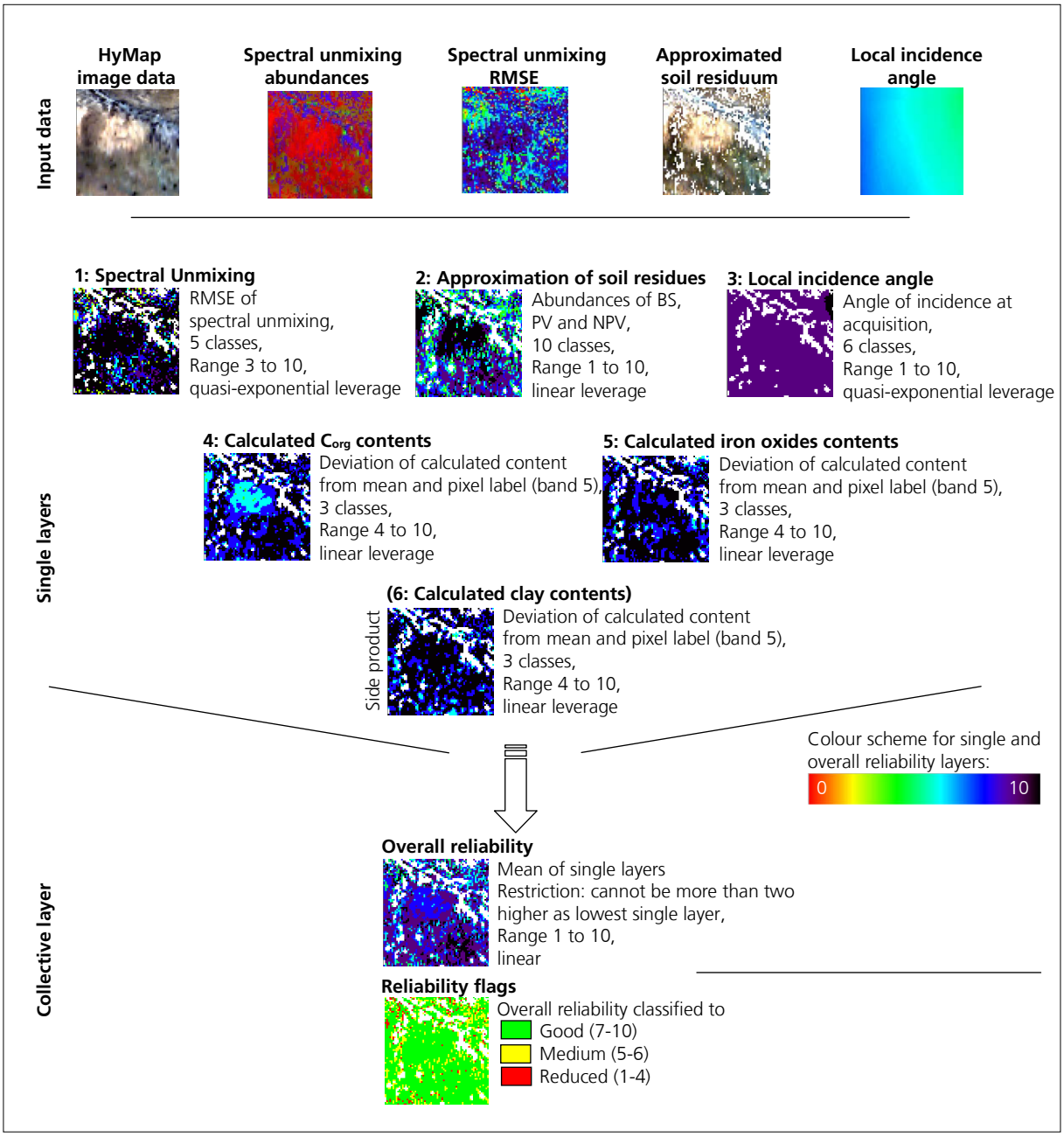


Figure 5-11: Principle of the calculation of single confidence layer of all data products produced in the workflow and their aggregation to an overall confidence layer showing three category flags.
An example for a 60 x 60 pixels subset is given from a portion of the farm Gonakraal (340534 E, 6358159 S) to illustrate typically occurring distributions in the given range of values and number of classes of each layer. Note the linear stretch of the single and overall quality layers with dark colors indicating highest and red indicating lowest confidence levels. A final confidence flag reduces all layers to a traffic light like three class coloring. White pixels were excluded within the approximation of the residual soil signature, e.g. because of a low soil fractional cover, and no soil constituents were calculated for them.

5.3.2 Validation methodology

A number of approaches exist to compare model derived soil parameters to in-situ field measurements, differing in the initial localization of the measured site within the hyperspectral imagery, the potential inclusion of the neighborhood of the pixel, that is determined as the one comprising the sample site, and

finally the consideration of additional background information, such as accuracy assessments. For this purpose in this study the ground reference points that are covered by the given imagery (might be a subset of the entire dataset) are first determined based on their UTM coordinates. The assignment of a certain pixel as location of each sample site is critical since uncertainties result from the orthorectification of the imagery (geometric accuracy ± 2 pixels, about 7 m) and the GPS position of the sample sites determined in the field (averaged precision ± 5 m). This leaves some tolerance for the determination of the pixel location of the sampled site. Thus, the pixel including the determined location of the sample site is used as center pixel but in addition its surrounding of 3 by 3 pixels is considered indirectly in the analysis. Based on the map of predicted soil constituents and the 3-class confidence assessment, the samples included in a given imagery are assigned to three groups:

- "Valid sample points": A constituent content can be predicted for the center pixel and no significant influences of the center point or its neighborhood can be detected in the confidence analysis. The confidence for the most of the pixels in the direct surrounding is flagged as "good".

- "Influenced sample points": A constituent content can be predicted for the center pixel but some influencing factors are detected in the confidence analysis, which means that the confidence of the center pixel or a number of pixels of its surrounding is lower than "good". In these cases, a negative influence on the accuracy of the center pixel is expected and thus it is flagged accordingly and treated carefully.

- "No value calculated": The center pixel is excluded during either the approximation of the soil residual spectra (see section 5.2.2) or the control of the predicted values if they match the calibration limits (see section 5.3) and thus no prediction is available for the center spot. The location of the sample sites is selected to include a certain fraction of bare soil but also to represent the characteristic land cover conditions of the surrounding area. This is done to ensure a calibration dataset including the entire spectral and chemical variance present in the study area. Nonetheless it results in not all sites showing the fractional cover of bare soil that is necessary to properly approximate the residual soil spectral signature from hyperspectral imagery.

Only sample points classified as valid are used for a comparison of measured and predicted contents of the soil constituents. Spatially predicted soil constituents can be compared to field reference measurements in several ways, for instance (1) by comparing the measured contents to the content that is predicted for the center pixel, (2) to a content calculated as average from a given area around the center spot or (3) to the content predicted for one of the pixels in a given neighborhood around the center pixel that matches best to the measured content. Here it is selected to compare the measured reference directly to the value predicted for the center pixel within the band, which is filtered and interpolated with the 3 x 3 filter kernel (band 3 see previous section). The selective spatial filtering operation that is already applied to the calculated soil information during the prediction procedure is adjusted to take account of the neighborhood, though with a very light impact. Details of the technical realization of the comparison of predicted soil constituents to measured ones can be found in appendix D.

5.3.3 Up-scaling of the approach to EnMAP hyperspectral satellite data

The compiled workflow is also applied to predict soil constituents from two simulated datasets that match the characteristics of the future EnMAP satellite (see section 3.2.3). The delineation of soil constituents from this data needed a re-calibration of the regression models due to the different number of bands of the EnMAP sensor. Therefore the bare soil field spectra in full ASD resolution are resampled to the spectral range and band positions of the simulated EnMAP data. Endmembers are derived from the simulated data also using the SSEE algorithm, though due to the lower spatial resolution of the data in combination with the small scale mixture within the study area it is expected that the derivation of pure endmembers may be incomplete. The problematic to derive endmembers from low spatial resolution data is not in the focus of this study and as such not followed up further. Thus, the EM sets of the corresponding HyMap imagery were re-sampled to match the spectral resolution of the EnMAP data and were included for selection of the final EM sets. All following steps are identical to the processing of the HyMap imagery.

5.4 Investigation of scaling issues

To visualize the dependency of the multi-stage workflow presented in the previous sections from the size and characteristics of the imagery they are applied to, three spatial scales were selected from the 320 km^2 of HyMap imagery of the study area. Each scale includes subsets of the imagery of a certain extension, going from small subsets showing a local area in scale 1 to the entire imagery of a regional extent in scale 3.

- Scale 1 (S1) comprises small subsets of the imagery with an area of some km^2. Six of these subsets are selected within the areas of the scale 2 subsets.
- Scale 2 (S2) comprises subsets of an area of 10 to 20 km^2. The five subsets selected in this scale are named according to the farm portion they predominantly show. One of these subsets is located in the smaller southern transect.
- Scale 3 (S3) comprises the entire dataset of approximately 320 km^2 containing both transect parts.

For selecting the lower scale' subsets, the focus is set to areas where bare soil areas exist in combination with morphological features that potentially determine significant changes in soil constituents (e.g. a valley, drainage systems, increasing vegetation coverage, etc.) and local reference measurements are available. Based on this criteria the five subsets of scale 2 are selected within the two transect parts, each being representative for the land coverage pattern of a certain part. Scale 1 subsets are selected within the scale 2 subsets to include a certain surface feature such as a plain, areas close to a river, etc. All scale 2 and scale 1 subsets are selected so that they include all land cover classes to ensure the possibility to find proper endmembers for all classes. Table 5-7 gives an overview of the dimension and characteristics of the subsets of all three levels of scaling.

Figure 5-12 and figure 5-13 show the location of the small- and medium-sized subsets within the imagery of the entire study area comprising scale 3. The data of the lower scales are extracted as subsets from the pre-processed HyMap imagery comprising scale 3. Pre-processing of all imagery therefore is identical for all

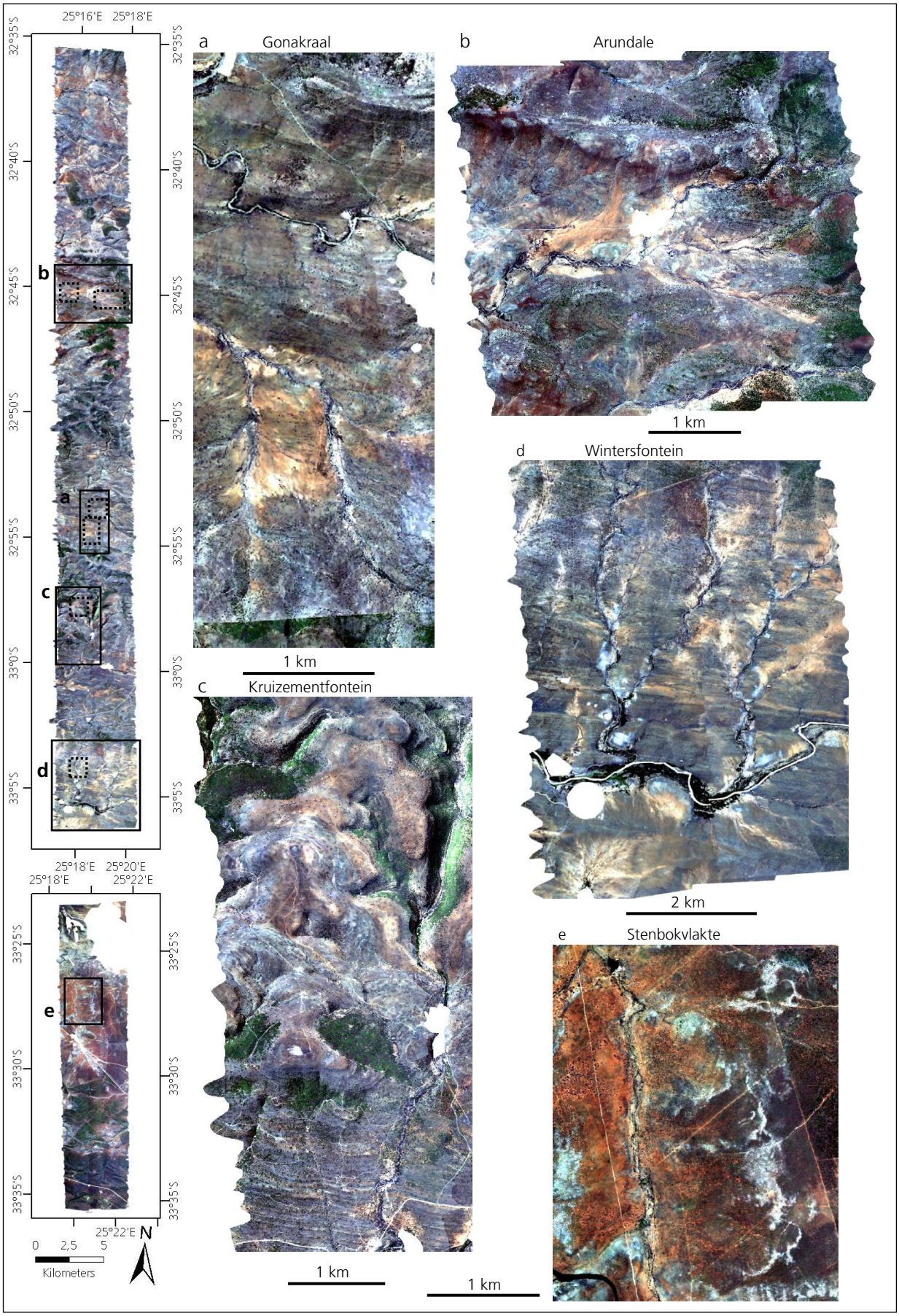


Figure 5-12: Overview of the five medium-sized subsets that are processed in scale 2 (true color with R: 640 nm, G: 549 nm, B: 456 nm). Their location in indicated in the overview map, together with the location of the small subsets of scale 1 (dotted lines).

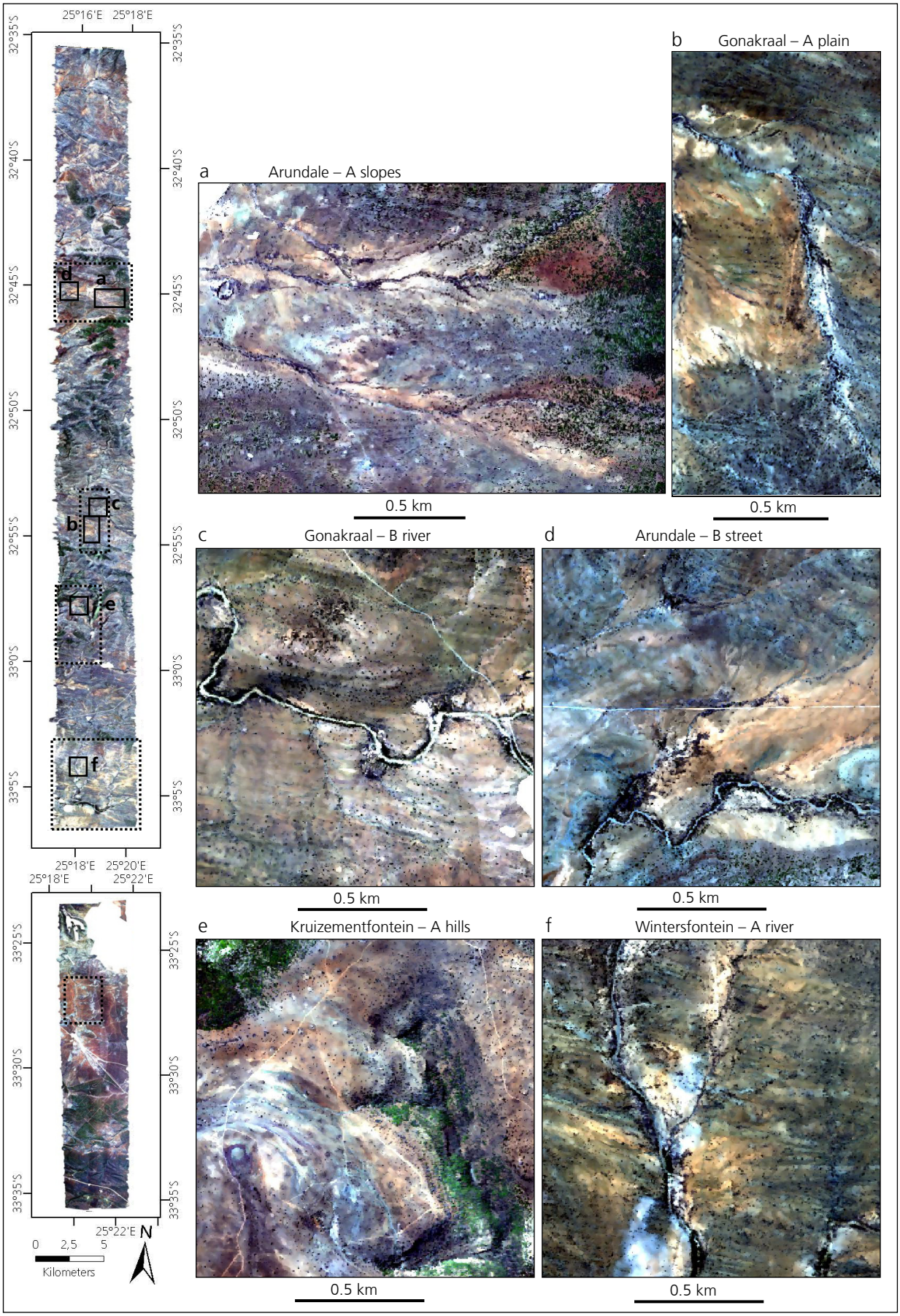


Figure 5-13: Overview of the six small subsets that are processed in scale 1 (true color with R: 640 nm, G: 549 nm, B: 456 nm). Their location in indicated in the overview map, together with the location of the medium-sized subsets of scale 2 (dotted lines).

Table 5-7: Overview of details of the subsets attributed to the three spatial scales.

	Subsets	Dimension			Flight lines	Characteristics
		km	km²	Pixel		
Scale 1	**Gonakraal - A plain**	0.9 x 1.9	1.7	270 x 585	1-2	Plain with reduced vegetation coverage and large bare soil surfaces, divided by erosive channels
	Gonakraal - B river	1.3 x 1.3	1.7	405 x 405	1-3	Area around major drainage channel, positive C_{org} anomaly around hill in NE
	Arundale - A slopes	2.0 x 1.3	2.6	600 x 400	2-3	Erosive slopes with drainage systems, decreasing vegetation coverage, doleritic bedrock in E
	Arundale - B street	1.3 x 1.3	1.7	400 x 400	1	Accumulation areas at foothills, drainage system, without influence of dolerite
	Wintersfontein - A river	1.3 x 1.3	1.7	400 x 400	1	Areas with reduced vegetation coverage and drainage systems
	Kruizementfontein - A hills	1.3 x 1.3	1.7	400 x 400	1-2	NE facing slope with prominent layering of sandstones and mudstones and locally steep valleys
Scale 2	**Gonakraal**	1.8 x 4.5	8.10	546 x 1352	1-3	Vegetation gradient from bare plains to fence line contrast in S
	Arundale	4.9 x 4.1	20.2	1480 x 1250	1-3	Sparsely vegetated valley surrounded by dolerite capped hills with increased vegetation coverage
	Wintersfontein	5.5 x 6.6	36.2	1662 x 2000	1-3	Large bare soil areas and prominent layering of mudstones and sandstones, drainage systems
	Kruizementfontein	2.9 x 5.5	14.3	866 x 1650	1-2	Highly vegetated hilly terrain, in particular high coverage of dry vegetation
	Stenbokvlakte	2.6 x 3.3	8.7	800 x 1000	4-6	Prominent soil area in southern transect, high vegetation coverage
Scale 3	**Northern section**	4.3-4.8 x 57.2	257.6	2978 x 17451	1-3	Entire northern transect area, undulating terrain with large soil dominated areas in valleys and usually vegetated hill slopes
	Southern section	4.5 x 22.5	101.3	1925 x 7032	4-6	Entire southern transect area showing high vegetation coverage, agricultural and artificial areas in N and NE

imagery. The scale 2 subsets are used for testing the methodology since the processing of the entire imagery for testing issues is computationally too intensive. The scale 2 subset of farm Gonakraal in the presentation of the results is often times used as example, because this subset includes land cover pattern typical for the Albany Thicket Biome on a small area. In addition it is not influenced by e.g. high topography, artificial areas or other factors such as the doleritic intrusions, leading to unusual or non-significant pattern in analysis results.

6 Results and discussion

This section presents the results of the multi-stage workflow proposed in section 5. The established regression models (section 6.1.1) and the analysis of the spectral mixtures within the hyperspectral imagery (section 6.2) thereby present intermediate results towards the predicted maps of soil constituents for the Albany Thicket Biome. These maps are interpreted in section 6.3. The overall accuracy of the predictions and their consistency is evaluated (section 6.4) and the transfer of the methodology to simulated EnMAP satellite data is shown (section 6.5). Finally, the benefit of the methodology to support restoration efforts in the degraded study area are shown in section 6.6.

6.1 Performance of feature-based prediction models

For soil organic carbon, iron oxides and clay content, a model is established following the feature-based regression method described in section 5.1. An evaluation of the consistency and significance of the proposed method is based on the three available soil spectral datasets (in-situ field, bare soil field, laboratory) both in full ASD spectral resolution and resampled to the pre-processed HyMap imagery (115 bands). This procedure results in 12 models per soil constituent. Model performance is evaluated based on the models' correlation coefficient (R^2) for predicted versus measured compositions, root mean square error (RMSE) and ratio of performance to deviation (RPD) for both calibration and cross-validation. Predictions are evaluated following the qualitative classifications of Chang et al. (2001) which was also used to evaluate PLS prediction models in section 4.3. The established models are subsequently compared to models that were established based on the same spectral datasets and using partial least squares regression analysis (PLS).

6.1.1 Calibration and validation accuracies

Table 6-1 gives the detailed accuracy measures for all established regression models. All models were built on the entire sample population, no outliers were removed. The numbers of spectral variables calculated in parameterization and finally applied for regression analysis are given. The number of spectral variables found as insignificant and excluded for the final model development depends on the variables' variation during cross-validation. Thus, the number of insignificant variables varies slightly for all models established for a certain soil constituent. However, the influence of several of the applied spectral variables in a regression equation is minor compared to the most important variables in the regression relationship. Table 6-2 lists the five spectral variables, which show the highest coefficients in the regression equation of every feature-based regression model, and thus, are the most important ones for model development. Their influence is determined based on the regression coefficient and given in % of the summed absolute values

of all regression coefficients. Negative signed influences indicate negative regression coefficients. The absolute values of the regression coefficients, and thus the influences of each specific variable, strongly

Table 6-1: Calibration and validation accuracies for the prediction models developed with the feature based regression approach for the three soil constituents and each of the three available spectral datasets in two spectral resolutions. The models that are applied for a prediction of soil constituents from hyperspectral imagery are highlighted.

		Spectral dataset	No. spectral variables	Calibration			Cross-Validation		
				R^2_{Cal}	$RMSE_{Cal}$	RPD_{Cal}	R^2_{Val}	$RMSE_{Val}$	RPD_{Val}
ASD spectral resolution	C_{org}	1 - In-situ field	14 of 16	0.73	0.47	1.92	0.55	0.52	1.42
		2 - Bare soil field	13 of 16	0.77	0.43	2.10	0.63	0.49	1.64
		3 - Laboratory	15 of 16	0.75	0.44	2.01	0.66	0.50	1.73
	Iron oxides	1 - In-situ field	19 of 21	0.58	0.80	1.55	0.24	1.09	0.99
		2 - Bare soil field	17 of 21	0.58	0.80	1.55	0.27	0.91	1.18
		3 - Laboratory	21 of 21	0.66	0.84	1.73	0.46	0.99	1.38
	Clay	1 - In-situ field	15 of 16	0.23	4.24	1.14	0.12	4.87	0.95
		2 - Bare soil field	12 of 16	0.17	4.39	1.10	0.10	4.83	0.95
		3 - Laboratory	16 of 16	0.20	4.55	1.12	0.11	4.98	0.98
HyMap spectral resolution	C_{org}	1 - In-situ field	13 of 16	0.75	0.46	1.99	0.61	0.52	1.71
		2 - Bare soil field	13 of 16	0.79	0.42	2.17	0.66	0.46	1.77
		3 - Laboratory	15 of 16	0.77	0.43	2.08	0.72	0.45	1.95
	Iron oxides	1 - In-situ field	18 of 21	0.56	0.82	1.51	0.26	0.97	1.09
		2 - Bare soil field	19 of 21	0.61	0.77	1.61	0.30	0.93	1.19
		3 - Laboratory	19 of 21	0.64	0.86	1.67	0.45	1.01	1.35
	Clay	1 - In-situ field	15 of 16	0.23	4.25	1.14	0.14	4.70	0.97
		2 - Bare soil field	13 of 16	0.14	4.47	1.08	0.09	4.92	0.93
		3 - Laboratory	13 of 16	0.22	4.49	1.13	0.17	4.83	1.02

Table 6-2: Influence (I) of the five most important spectral variables in the regression relation of each of the feature-based regression models for three spectral libraries and two spectral resolutions. The models that are applied for a prediction of soil constituents from hyperspectral imagery are highlighted.

			Rank 1		Rank 2		Rank 3		Rank 4		Rank 5	
		Spectral dataset	I [%]	Variable	I [%]	Variable	I [%]	Variable	I [%]	Variable	I [%]	Variabl
ASD spectral resolution	C_{org}	1 - In-situ field	38.0	$r_{HF\ SWIR1}$	-23.8	$S_{HF\ VIS}$	-11.2	$r_{HF\ VIS}$	-7.4	$S_{HF\ SWIR1}$	3.2	AS_{AF2330}
		2 - Bare soil field	31.2	$r_{HF\ SWIR1}$	-21.4	$S_{HF\ VIS}$	-9.2	$r_{HF\ VIS}$	9.1	$d_{max\ AF1730}$	-8.4	$S_{HF\ SWIR1}$
		3 - Laboratory	23.2	A_{AF1730}	14.0	$r_{HF\ SWIR1}$	-11.3	$d_{max\ AF1730}$	-9.7	$r_{HF\ VIS}$	-8.2	$S_{HF\ VIS}$
	Iron oxides	1 - In-situ field	-15.7	A_{AF900}	12.4	$d_{max\ AF550}$	-11.6	$d_{\lambda lit\ AF550}$	11.4	$d_{max\ AF900}$	-6.7	A_{AF550}
		2 - Bare soil field	-17.5	A_{AF900}	13.0	$d_{max\ AF900}$	10.3	$d_{max\ AF700}$	-9.7	$d_{\lambda lit\ AF550}$	8.5	A_{AF550}
		3 - Laboratory	26.5	$d_{max\ AF900}$	-11.4	$d_{\lambda lit\ AF900}$	-10.7	A_{AF900}	9.9	A_{AF550}	-7.1	$d_{\lambda lit\ AF550}$
	Clay	1 - In-situ field	-22.2	$d_{\lambda lit\ AF2200}$	22.0	$A_{AF\ 2200}$	-7.0	$S_{HF\ SWIR1}$	7.0	$r_{HF\ SWIR1}$	6.3	A_{AF2340}
		2 - Bare soil field	-19.8	$d_{\lambda lit\ AF2200}$	14.4	A_{AF2200}	12.0	$d_{max\ AF2200}$	-7.8	$d_{\lambda lit\ AF2340}$	-7.8	$S_{HF\ SWIR1}$
		3 - Laboratory	-33.2	$d_{max\ AF2200}$	30.7	$d_{\lambda lit\ AF2200}$	-8.6	$d_{\lambda lit\ AF2340}$	5.0	A_{AF2200}	5.0	d_{max}
HyMap spectral resolution	C_{org}	1 - In-situ field	36.5	$r_{HF\ SWIR1}$	-20.3	$S_{HF\ VIS}$	-14.1	$r_{HF\ VIS}$	-6.6	$S_{HF\ SWIR1}$	6.3	AS_{AF2330}
		2 - Bare soil field	27.8	$r_{HF\ SWIR1}$	-17.1	$S_{HF\ VIS}$	-11.3	$r_{HF\ VIS}$	10.0	$d_{max\ AF1730}$	7.8	AS_{AF2330}
		3 - Laboratory	18.7	A_{AF1730}	-11.8	$d_{max\ AF1730}$	-8.8	$\lambda_{dmax\ AF2330}$	8.5	$r_{HF\ SWIR1}$	-8.4	d_{max}
	Iron oxides	1 - In-situ field	-20.3	A_{AF900}	17.4	$d_{max\ AF900}$	11.4	$d_{max\ AF700}$	-8.6	$d_{\lambda lit\ AF550}$	-8.2	A_{AF700}
		2 - Bare soil field	16.3	A_{AF550}	-12.0	A_{AF900}	10.9	$d_{max\ AF900}$	-10.6	$d_{max\ AF550}$	10.1	$d_{max\ AF700}$
		3 - Laboratory	27.5	$d_{max\ AF900}$	-13.3	$d_{\lambda lit\ AF900}$	-9.9	A_{AF900}	-8.5	A_{AF700}	7.2	$S_{HF\ VIS}$
	Clay	1 - In-situ field	-19.9	$d_{\lambda lit\ AF2200}$	14.4	A_{AF2200}	9.3	$d_{max\ AF2200}$	9.3	$r_{HF\ SWIR1}$	-9.1	$S_{HF\ SWIR1}$
		2 - Bare soil field	-18.0	$d_{\lambda lit\ AF2200}$	12.5	A_{AF2200}	11.3	$d_{max\ AF2200}$	-9.9	$d_{\lambda lit\ AF2340}$	9.5	λ_{dmax}
		3 - Laboratory	-23.4	$d_{\lambda lit\ AF2340}$	17.7	A_{AF2200}	-12.4	$d_{max\ AF2200}$	10.4	$d_{max\ AF2340}$	9.9	A_{AF2340}

Symbols: absorption features (AF): A_{AF}: area, $d_{max\ AF}$: maximum depth, $\lambda_{dmax\ AF}$: wavelength of $d_{max\ AF}$, $d_{\lambda lit\ AF}$: depth at wavelength position given in literature, AS_{AF}: asymmetry factor. Hull features (HF): r_{HF}: mean reflectance in interval, S_{HF}: slope in interval (see section 5.1.2 for definition).

depend on the number and character of spectral variables considered in the regression models. Thus, their absolute values cannot expect to be identical for different models, but are still considered valuable clues to identify the most important spectral variables.

Soil organic carbon prediction models

All prediction models for soil organic carbon provide good calibration accuracies with R^2_{Cal} between 0.73 and 0.79, RPD_{Cal} between 1.92 and 2.17 and $RMSE_{Cal}$ around 0.45 %. The correlation is shown in scatter plots of measured vs. calculated soil organic carbon concentrations, with an example shown in figure 6-1 for the model built upon bare soil field spectra in HyMap's spectral resolution. Cross-validation accuracy is slightly lower but still within the accuracy of medium class models (R^2_{Val} between 0.55 and 0.72, RPD_{Val} between 1.42 and 1.95, $RMSE_{Val}$ around 0.50 %). The models apply 13 to 15 of 16 calculated spectral variables. In the models developed based on field spectra, the influences concentrate on a few variables, with both spectral variables describing the two hull features of great importance (refer to section 5.1.2). In all these models, $r_{HF\ SWIR1}$, $s_{HF\ VIS}$ and $r_{HF\ VIS}$ are the dominant factors in the regression relationships (see table 6-2), and together account for an influence of around 60 %. The two last variables show a negative influence as expected, since increasing contents of soil organic carbon reduce reflectance in the visible range and effect the slope and mean reflectance. Although the same effect was expected for $r_{HF\ SWIR1}$, its coefficient is positive. In the models developed based on the laboratory dataset, the influences are distributed on more spectral variables and, in particular, the variables describing the 1730 nm absorption are predominant. A_{AF1730}, $r_{HF\ SWIR1}$ and $d_{max\ AF1730}$ show up as important variables.

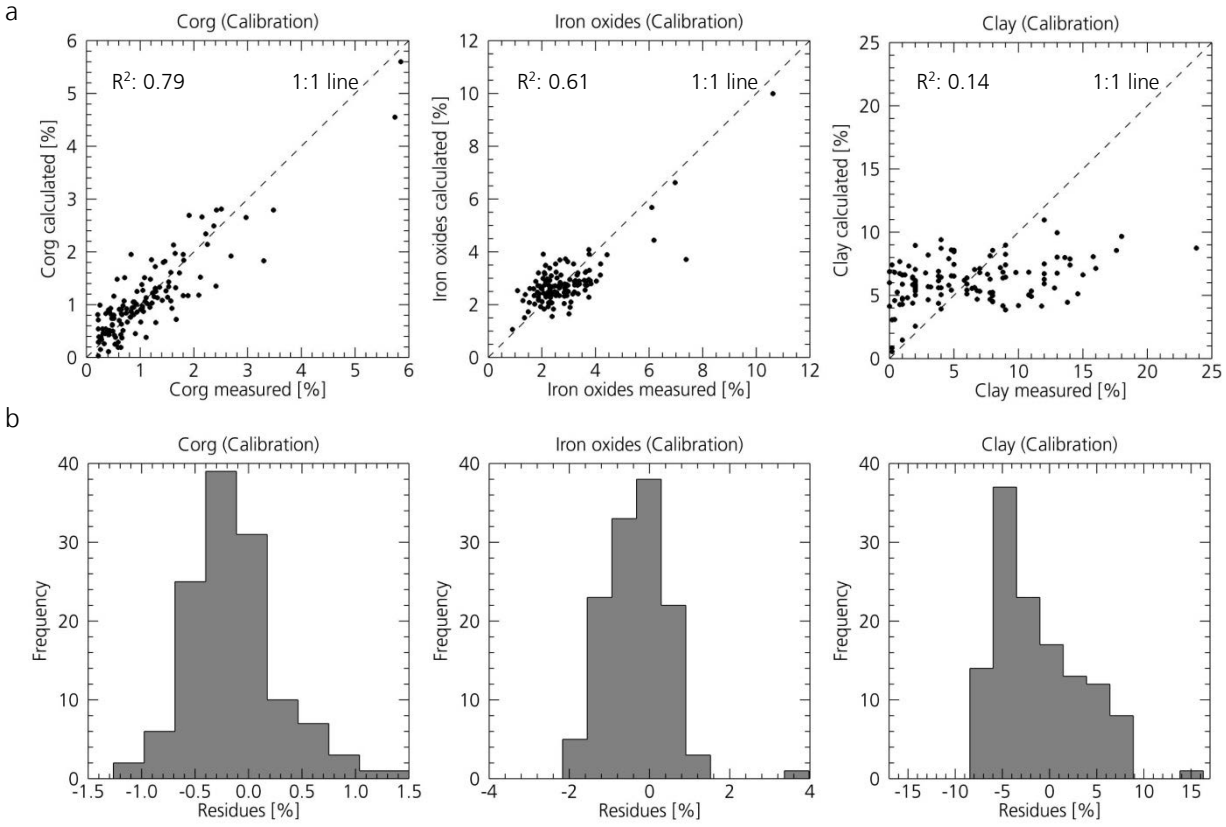


Figure 6-1: Scatter plots of measured versus calculated contents (a) and histograms of modeling residues (b) of predictions using the feature-based regression approach. Models were developed based on bare soil field spectra in HyMap's spectral resolution.

Iron oxides prediction models

Calibrated iron oxide models show R^2_{Cal} between 0.58 and 0.66, RPD_{Cal} between 1.51 and 1.73 and $RMSE_{Cal}$ around 0.80 %. Among these models, the ones based on laboratory spectra perform significantly better than the models based on the two field datasets. Even if R^2_{Cal} values for all iron oxides prediction models are reasonable, lower RPD_{Cal} values already indicate reduced predictability. Cross-validation results confirm this with R^2_{Val} all below 0.5 and corresponding RPD_{Val} below 1.4, indicating models with no predictive power. Only the models built upon the laboratory soil spectra are close to the range of medium prediction models for both spectral resolutions (R^2_{Val} of 0.46, RPD_{Val} of 1.35 and 1.38, $RMSE_{Val}$ of around 1.0 %). The models for the prediction of iron oxides are built with 17 to 19 or all of the 21 spectral variables describing spectral feature properties. In the models developed based on field spectra, A_{AF900}, $d_{max\ AF900}$, $d_{max\ AF700}$, $d_{max\ AF550}$, $d_{\lambda lit\ AF550}$ and A_{AF550} are of prominent importance (see table 6-2). Maximum absorption depths are usually positively signed, as expected. It is observed that the area of the 900 nm absorption feature is consequently signed negatively. Negative coefficients also appear for the absorption depth at 550 nm ($d_{\lambda lit\ AF550}$), which was identified as the wavelength that was predominantly given in the respective literature. This directly links a decreasing absorption depth at this wavelength with increasing iron content, which may indicate that with increasing iron content the wavelength of maximum absorption of the iron oxides present in the South African study area trigger shifts away from 550 nm. In fact, the wavelength of maximal absorption determined in the spectra was found to average around 505 nm.

For the models developed based on laboratory spectra, $d_{max\ AF900}$, $d_{\lambda lit\ AF900}$ and A_{AF900} appear as the most important spectral variables. Negative coefficients of $d_{\lambda lit\ AF900}$ and A_{AF900} are observed which may suggest that with increasing iron oxides content, the wavelength position of the absorption shifts away from 900 nm and further more pronounced and steep features instead of broad ones occur.

Clay prediction models

Calibration of feature-based clay prediction models leads to low correlations between 0.14 and 0.23 with RPD_{Cal} around 1.10 and $RMSE_{Cal}$ above 4.0 % for both spectral resolutions. Cross-validation results are also poor with R^2_{Val} mostly below 0.15, RPD_{Val} mostly below 1.0 and high $RMSE_{Val}$ close to 5.0 %. Soil clay content prediction models are built based on 12 to 16 of 16 derived spectral variables. The influences observed for spectral variables are highly variable and it appears that no significant order of spectral variables can be determined. However, $d_{\lambda lit\ AF2200}$, A_{AF2200}, $d_{max\ AF2200}$ and $d_{\lambda lit\ AF2340}$ frequently appear among the most important spectral variables of all six models (see table 6-2). There seems to be no connection between the selection of specific variables and the field or laboratory origin of the spectral base data. Some variables (e.g. A_{AF2200}, $r_{HF\ SWIR1}$) appear consequently positively signed, whereas for instance $d_{\lambda lit\ AF2200}$ and $d_{max\ AF2200}$ occur positively and negatively signed.

Outlier samples within the MLR modeling

In the chemical reference data, two samples for soil organic carbon and one sample for iron oxides attract attention because they exceed the overall distribution of the other samples (compare figure 3-14). These samples appear in the calibration scatter plots of figure 6-1a close to the 1:1 line, indicating that they may significantly influence the model calibration. However, when testing regression models built on the presented spectral libraries excluding these samples, similar models and results were achieved as they are presented here.

6.1.2 Discussion and placement of the approach

Of all functional relationships established with feature-based MLR, spectra measured in a laboratory lead to the best models, as expected and also observed in previous studies (e.g. Lagacherie et al. 2008), whereas in-situ field spectral surveys usually had the lowest performance. This is very likely the result of small stones lying on the bare soil surface and influencing the correlation between spectral signature and chemical concentrations. As the detailed results show, the prediction models for soil organic carbon are of good quality close to the accuracy of excellent prediction models according to the applied classification (average R^2_{Cal} 0.76/$RMSE_{Cal}$ 0.44/RPD_{Cal} 2.05, and R^2_{Val} 0.64/$RMSE_{Val}$ 0.49/RPD_{Val} 1.70). The selection of spectral variables that account for the most influence on regression relationships is very constant and supports the significance of the established models. Different variables are selected for field and laboratory data. It seems to be reasonable especially for field data, where the variables of the spectral continuum in the VIS and SWIR range show up as the most influential ones.

Accuracies of iron oxides prediction models are of medium quality, but lower than that for the soil organic carbon prediction models (average R^2_{Cal} 0.61/$RMSE_{Cal}$ 0.82/RPD_{Cal} 1.60, and R^2_{Val} 0.33/$RMSE_{Val}$ 0.98/ RPD_{Val} 1.20). Significantly lower validation accuracies indicate a reduced predictive power. The selection of spectral variables shows some consistency in the selection of the most important variables, although the influences are distributed on more spectral variables. The area and maximum depth of the 900 nm absorption are of prominent importance. Also here, different variables are selected if field or laboratory data are used as input.

All regression models established with the feature-based regression method do not adequately predict soil clay content (average R^2_{Cal} 0.20/$RMSE_{Cal}$ 4.40/RPD_{Cal} 1.12, and R^2_{Val} 0.12/$RMSE_{Val}$ 4.86/RPD_{Val} 0.97). No significant pattern of the selection of important variables indicates a high degree of statistical adaptation and finally a low significance of each model. Scatter plots of measured vs. calculated indicate for all models a systematic overestimation and underestimation of low and high clay contents, respectively (figure 6-1).

For soil organic carbon and iron oxides, the modeling residues as deviations between measured and calculated contents are close to a normal distribution (figure 6-1b). They are not autocorrelated and the residuals' mean is not significantly differing from zero. Thus, it is expected that the models represent well the original distribution of each constituent, although these are not normally distributed (see figure 3-14). However, this does not apply for clay predictions where modeling residues are skewed to negative values. Figure 6-1b gives an example of the distribution of the residues for the bare soil field spectra in HyMap's spectral resolution.

Influence of the spectral resolution on model development

The spectral resolution of the input spectral libraries affects the size, shape and location of absorption features. Thus, the lower spectral resolution of the HyMap sensor (13-17 nm) compared to the ASD field spectrometer (1 nm) may influence the correlation of the derived spectral variables with chemical reference. For the analyzed field and laboratory spectral libraries, the resampling of the ASD spectral resolution to the HyMap resolution had a low influence on model performances, which is consistent with that reported by other authors (e.g. Ben-Dor and Banin 1994, Lagacherie et al. 2008).

Comparison of feature-based to PLS models

For the available spectral libraries both approaches provide similar capabilities to set up significant regression models, particularly for soil organic carbon and iron oxides and accuracies of MLR models are in the same range that statistical PLS approaches provide (see section 4.3). Also the same trends were observed, such as reduced validation accuracies compared to calibration for iron oxides models and generally low predictabilities of clay models. However, the feature-based MLR technique in general performs slightly lower compared to conventional PLS techniques. The difference is small for soil organic carbon models (comparable validation performance of MLR and PLS models), but more apparent for the prediction of iron oxides. In the same way as the feature-based approach, also PLS fails to establish significant clay content prediction models. A full comparison of the two approaches using the available field and laboratory spectral libraries is given in Bayer et al. (2012).

Influences on clay prediction models

Three factors are possible explanations for the observed poor correlation of spectral and chemical information regarding clay content – (1) the chemical determination of soil clay content, (2) the presence of soil crusts and (3) the variable characteristics in the study area, i.e. geology. In chemical analysis, soil clay content was determined by pipette analysis and as such based on particle size only. In general the clay particle size content is a good measure for the content of clay minerals and thus was applied in a number of studies (e.g. Ben-Dor and Banin 1995b, Waiser et al. 2007). However, this does not apply if a significant amount of clay-sized particles are not clay minerals. Therefore in addition to grain size information, the examination of the mineralogy of the clay fraction would be necessary, e.g. by x-ray diffraction analysis (XRD).

In a field environment, the presence of soil physical crusts is an additional factor. Soil physical crusts in the study area can locally be very well developed and reach over 1 cm in thickness, as discussed in section 3.1.2. Soil crusts are problematic for spectral analyses, since the chemistry of the surface layer which is measured by a field spectrometer does not correspond to the chemistry of the bulk sample of the upper 1 cm analyzed in the laboratory, thus leading to a decorrelation of spectral and chemical information.

Finally, the changing characteristics within the large study area describe another factor influencing in particular clay content predictions. Because soil development highly depends on the source material, soils which developed based on sandstones very likely differ in their chemical and spectral properties from ones which developed based on mudstones. A common way to deal with this is to split the dataset in groups according to their characteristics and to generate separate regression models for each group (see e.g. Richter et al. 2009, Stevens et al. 2010). However, this requires the groups to be well separable from each other based on the determined characteristics (e.g. two groups in the texture triangle) or the availability of additional information that would allow a separation (e.g. spatial information on geology or soil types with a high degree of detail). This is not applicable for the data of the South African study area because the basement rock type was not identifiable at every sample site, and also based on the texture distribution of the reference samples, a discrimination of two or more groups would be vague (see figure 3-15). Further, there is no additional information of soil types or geology available that matches the degree of detail that would be needed for such analysis.

Placement of the approach

Compared to other studies working in agricultural environments (e.g. Stevens et al. 2008 and 2010), the accuracy of developed prediction models for the feature-based MLR and also the PLS approach are slightly lower. This is likely a result of the large size and the highly variant characteristics of the South African study area caused by the non-agricultural environment and differences in geology and soil types. The difficulty to achieve prediction models for large areas with changing conditions (referred to as global calibrations) was addressed in previous studies (e.g. Stevens et al. 2010). Based on AHS-160 data Stevens et al. (2010) predicted soil organic carbon using among others PLS techniques for an agronomic Luxembourgian region. An RPD_{Val} of 1.47 was obtained for a global calibration, which is in the same range as the prediction models for soil organic carbon established here with both the feature-based MLR and PLS approaches. Stevens et al. could improve this RPD_{Val} to 2.76 using local calibrations based on agro-geological regions and soil types. In this study the lack of detailed spatially continuous information matching the small-scale variation of geology and soils prevented the investigation of the performance of local calibrations (see e.g. figure 3-15).

A further factor lowering prediction RPD is the small variability in measured contents of organic carbon in our study area (see figure 3-14a). The ground reference sites were mainly selected to have a significant bare soil component so that these sites can also be used as validation targets for airborne hyperspectral imagery. Thus, no soil samples were taken in densely vegetated regions characterized by increased input of C_{org} and thus higher concentrations that would have increased the variability in the ground reference which is directly related to the modeling RPD. This corresponds with findings reported in e.g. Stevens et al. (2010), Udelhoven et al. (2003).

Conclusions for the prediction of soil constituents

The results show that it is possible to establish regression relationships on a physical basis that reach the predictability of PLS derived regression models. This can be achieved by the application of a set of spectral characteristics for each of the investigated soil constituents and the inclusion of various properties of these spectral features. Statistical adaptation within regression analysis is reduced to a minimum. It was shown that the established models for soil organic carbon and iron oxides are stable over different spectral libraries and spectral resolutions. With the combination of several spectral properties of each soil constituent, this approach is suggested to be more robust compared to similar physical approaches that are based on the analysis of only one diagnostic spectral feature or band ratio. Additionally, the established models are straightforward and have minimal computational burden owing to a limited number of variables in the regression relationship.

The developed feature-based approach is tested using hyperspectral imagery of the South African study area. In this case, the models developed based on bare soil field spectra resampled to the resolution of the HyMap imagery are applied. The bare soil field spectra are used, because the field conditions match the airborne measurements better than measurements in a controlled laboratory environment, but without the influence of overlying stones as for the in-situ field measurements. Calibration and validation accuracies of these models are highlighted in Table 6-1, while figure 6-1a shows their scatter plots of measured vs. calculated contents. The coefficients of all variables used in the regression models are given in appendix E.1.

6.2 Modeling of soil spectral information

The processing of the hyperspectral imagery for a quantitative prediction of soil constituents included spectral mixture analysis, which comprises of the extraction of image endmembers and the spectral unmixing itself. Subsequently based on these results an approximation of the soil signature contained in mixed pixels. The results are presented here with visual examples given for the Gonakraal farm subset. Results for other subsets and additional statistics can be found in appendix E.3 to E.4.

6.2.1 Spectral mixture analysis

Sets of image endmembers

From the sets of EM candidate spectra automatically derived using the SSEE approach, the final EM sets are selected manually. Representative endmembers are found for all materials. A pre-classification of the imagery supporting the extraction of a certain class, e.g. masking of vegetative areas based on NDVI to support the extraction of bare soil endmembers, was not necessary. One independent EM set is prepared for each subset of the six subsets in scale 1 and the five selected subsets of scale 2 (see section 5.4 for description of the three scales). In scale 3, one EM set was established for the northern section (235 km^2) and one for the southern section of the study area (92 km^2) since the two areas differ in their characteristics. From EMs with similar spectral shape but different overall albedo, the ones with higher reflectance are selected to minimize the noise fraction. Table 6-3 lists the selected EM sets for each image subset according to their classification as green and dry vegetation and bare soil. They comprise of 13 to 31 endmember spectra with more EMs extracted with increasing size of the dataset. Usually, more EMs classified as green vegetation and bare soil are needed, because the spectral variance within these two classes is significantly higher compared to the dry vegetation class. The EM sets comprise no spectra of artificial materials or water. A visualization of the SSEE endmember sets of all subsets can be found in appendix E.3.

Table 6-3: Overview of EM sets derived from the hyperspectral imagery using the SSEE approach for the three scales.

	Subsets	Size in pixels	EMs extracted by SSEE	EMs selected to be used for unmixing			
				PV	NPV	BS	all
Scale 1	**Gonakraal - A plain**	270 x 585	73	5	3	6	14
	Gonakraal - B river	405 x 405	56	6	2	5	13
	Arundale - A slopes	600 x 400	60	7	4	6	17
	Arundale - B street	400 x 400	71	6	4	7	17
	Wintersfontein - A river	400 x 400	70	5	3	6	14
	Kruizementfontein - A hills	400 x 400	69	6	3	6	15
Scale 2	**Gonakraal**	546 x 1352	105	7	2	8	17
	Arundale	1480 x 1250	151	10	5	9	24
	Wintersfontein	1662 x 2000	264	7	5	9	21
	Kruizementfontein	866 x 1650	116	9	4	9	22
	Stenbokvlakte	800 x 1000	134	6	3	7	16
Scale 3	**Northern section**	2978 x 17451	155	13	4	12	31
	Southern section	1925 x 7032	197	7	4	10	21

Spectral mixture analysis

The ISMA algorithm used on average 3 to 4 endmembers to unmix each pixel (maximum was set to 5, each excluding the shade EM). Figure 6-2 shows an example unmixing result for S2 subset Gonakraal. The map of ground cover fractions represents the pattern observed in the field. The spatial distribution of the summed ground cover fractions are close to 1 where spectral mixtures of the image spectra match well within the endmember spectra. For example, on the flat plains where bright soils occur and are mostly pure (see A in figure 6-2c). The sums are low (e.g. in inclined areas that are partly shadowed), because the shadow EM used in the ISMA unmixing accounts for a significant fraction. For example in the south, where a valley cuts into the slopes and at a small hill north of the river (see B and C in figure 6-2c). Here, the brightness of image spectra is usually lower than of the endmember spectra. Unmixing RMSE are very low overall, with an average of 1.6 % for this subset (standard deviation of 0.8 %). RMSE are generally higher in vegetated areas as result of the high spectral variance of vegetation that cannot be fully represented by the endmembers (see D and E in figure 6-2d). Small-scale patterns in the RMSE are the result of local mixtures.

Some general features that are distinctive in the scale 2 and 3 subsets and a brief comparison of the unmixing results of all scales are discussed here. Statistics of the unmixing results for the subsets of all scales can be found in appendix E.3. Among the scale 2 subsets, the Wintersfontein area shows the largest bare soil ground cover fractions (average bare soil cover fraction of 0.57), directly followed by the Gonakraal area which in this study often serves as example area. The subsets Stenbokvlakte (only S2 subset in southern part), Kruizementfontein and also Arundale show significant amounts of green and dry vegetation with no

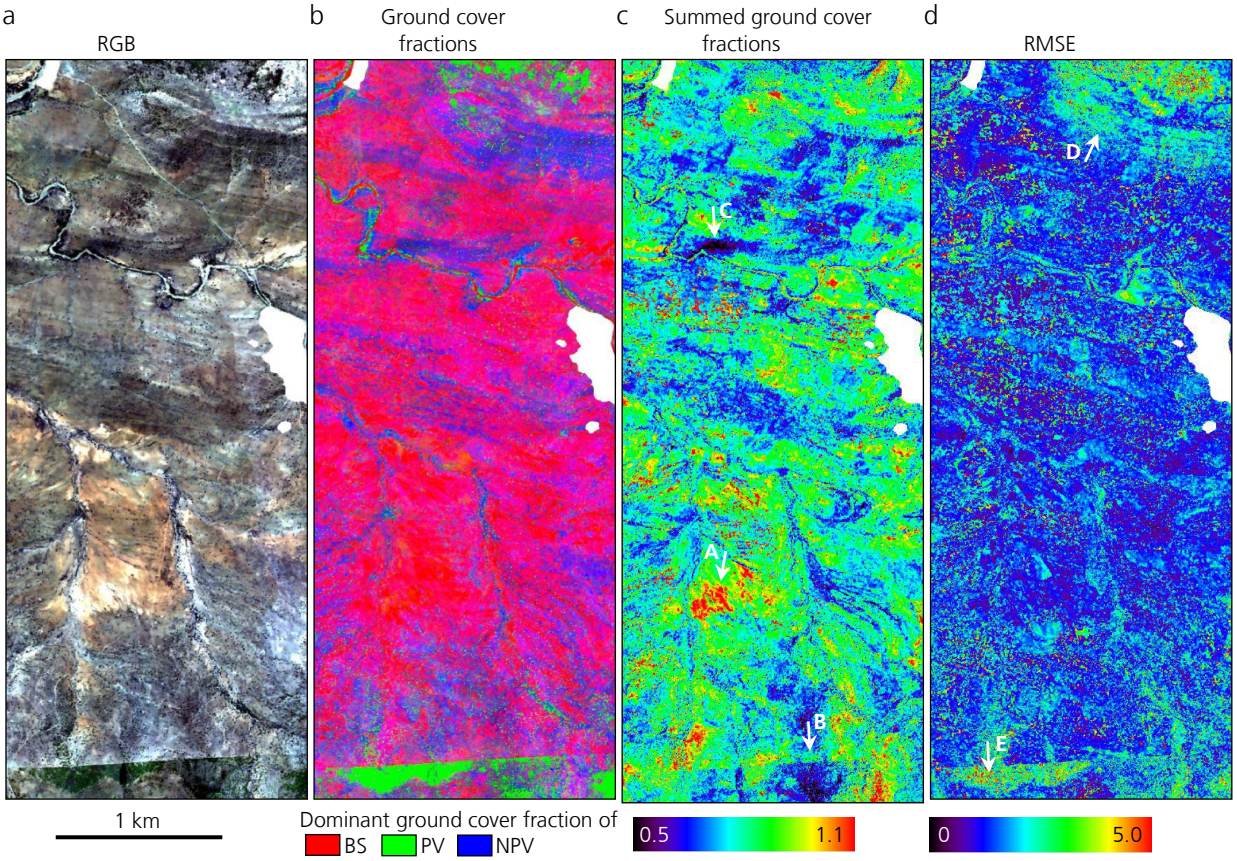


Figure 6-2: Spectral unmixing results for scale 2 subset of farm Gonakraal (a, RGB with R: 640 nm, G: 549 nm, B: 456 nm). With the unmixing cover fractions grouped to the classes PV, NPV and BS (b), summed ground cover fractions (c) and the unmixing RMSE (d).

Table 6-4: Statistics of ISMA spectral unmixing results of the three scales. Distinctive features are highlighted.

	Subsets	Ground cover fractions PV	Ground cover fractions NPV	Ground cover fractions BS	Summed cover fractions	RMSE
Scale 1	6 subsets	Mean: 0.16 Stdev: 0.19 Min: 0.00 Max: 1.93	Mean: 0.28 Stdev: 0.16 Min: 0.00 Max: 1.62	Mean: 0.54 Stdev: 0.21 Min: 0.00 Max: 2.75	Mean: 0.85 Stdev: 0.16 Min: 0.07 Max: 3.27	Mean: 1.56 Stdev: 0.82 Min: 0.00 Max: 22.92
Scale 2	5 subsets	Mean: 0.23 Stdev: 0.23 Min: 0.00 Max: 2.19	Mean: 0.28 Stdev: 0.17 Min: 0.00 Max: 4.12	Mean: 0.43 Stdev: 0.23 Min: 0.00 Max: 2.80	Mean: 0.75 Stdev: 0.18 Min: 0.00 Max: 4.27	Mean: 1.56 Stdev: 0.84 Min: 0.00 Max: 36.91
Scale 3	2 subsets	Mean: 0.22 Stdev: 0.19 Min: 0.00 Max: 1.51	Mean: 0.27 Stdev: 0.13 Min: 0.00 Max: 2.12	Mean: 0.39 Stdev: 0.21 Min: 0.00 Max: 4.39	Mean: 0.69 Stdev: 0.17 Min: 0.00 Max: 4.96	Mean: 2.20 Stdev: 1.32 Min: 0.00 Max: 190.14

larger soil areas. The averages of summed ground cover fractions for all subsets are above 0.75, except for the Kruizementfontein area, where cover fractions may be lower due to increased amounts of shadow resulting of the highly undulating and partly hilly terrain. The overall low unmixing RMSE are slightly increased for the vegetation rich subsets Stenbokvlakte and Kruizementfontein. Singular pixels showing high unmixing RMSE result from pixels having either an erroneous signal or show materials that cannot be modeled with the non-artificial materials comprised in the EM sets. Occasionally small water ponds of the size of a few pixels show up here. Within the entire dataset (scale 3), the smaller southern section shows significantly more vegetation than the northern part.

The RMSE of the entire imagery given for the subsets of all scales is sufficient and the unmixing results are of good quality. Table 6-4 lists the spectral unmixing results as statistics of all subsets within each scaling level, with the statistics of the scale 3 subsets giving the statistics of the entire dataset. Ground cover fractions depend on the selection of the image subsets. Thus, bare soil fractions increase from scaling level 3 to the scale 1 subsets, since the smaller subsets focus on soil dominated regions. Trends in the sums of the cover fractions and the RMSE are indicators for the unmixing accuracy. The cover fractions sums decrease from the detailed scale 1 subsets to the scale 3 data that comprise the entire imagery, while RMSE are comparable for scale 1 and 2 subsets and higher for the scale 3 data. Both factors are a result of the endmember sets. The scale 3 endmember sets are selected to comprise the predominant endmembers, which best explain the dominant mixtures of the large area. Because these EMs are not perfectly adapted to all local mixtures, unmixing RMSE are usually higher here. Despite this, the endmember sets of the small scale 1 subsets are well adapted to the imagery and also local mixtures and unmixing performs better with lower RMSE.

Validation of unmixing results

The ISMA spectral unmixing results can be validated relatively using another well established unmixing approach and further absolutely using the ground cover fractions assessed in the field (compare section 3.2.1). It is to be noted that a comparison of unmixing ground cover fraction to field assessments is limited by a number of factors: (1) If unmixing ground cover fractions are unscaled (sum not equal 1), they vary depending on the brightness of the selected endmembers as was discussed above. If mainly dark endmembers are used for the unmixing, the ground cover fractions will be significantly higher than if endmembers with a similar shape but comparably high overall albedo were used. (2) The field assessments are flawed, which is also influenced by the subjectiveness of personal assessments. Based on publications,

the accuracy of these field assessments can be assumed to be around ± 10 % (see Bachmann 2007). (3) A time difference exists between the field assessments done in June and September/October 2009 and the HyMap imagery taken on 14th October 2009. This difference in time might bias the comparison of the results, although this is expected to be minimal since weather conditions were constantly dry during this period which limits the variability of Thicket vegetation. (4) Finally, the accuracy of georectification is assumed to be around ± 2 to 3 pixels. In this highly variant environment, where ground cover fractions of adjacent pixels show significant differences, the presence of such a shift introduces further inaccuracies.

For the validation based on field reference data, the ground cover fractions assessed according to life form are grouped to the three cover classes. Vegetation is classified according to its state (green, dry, mixed), whereas mixed vegetation was split in half and assigned to PV and NPV. Spekboom shrubs are assigned to PV since they are evergreen. The ground cover percentage of larger stones present at the surface is assigned to the class BS. ISMA unmixing ground cover fractions calculated for a 3 x 3 pixel (corresponds to 9.9 x 9.9 m) kernel around the center pixel matching the GPS location of the sample site are averaged. This matches very well the 10 x 10 m for which the ground cover fractions were assessed in the field. The ground cover fractions of the unconstrained ISMA unmixing are further scaled (sum-to-1) to remove the effect of differently scaled endmembers and the influence of the shade endmember and allow for a comparison to the field measurements. In total 9 of the 163 sample sites are excluded due to incomplete field information, or because these sites lie in or are very close to artificial areas that are not suited for a representative comparison.

Correlation coefficients between field assessments and ISMA derived ground cover fractions are low with around 0.25 for the classes NPV and BS, and below 0.10 for PV (figure 6-3). Cover fractions of dry vegetation (NPV) seem to be underestimated in the field or overestimated by the spectral unmixing approach. The opposite is true for bare soil fractions, which seem to be overestimated in the field measurements. The severe differences in the correlation for green vegetation (PV) are most likely a result of uncertainties remaining from geocorrection of the imagery. In the South African study area, during extremely dry years as it occurred 2009 when the hyperspectral imagery was taken, green vegetation is only present in patches of isolated shrubs or trees but not continuously over larger areas. Thus, a shift of few pixels can significantly change the calculated cover fractions, the field assessments are compared to. This effect of course is in the same way present for the other two classes, but appears minor since their cover fractions are generally larger than the ones of green vegetation. In summary, the validation of the unmixing results using field reference data appears insufficient, may be due to the given influences.

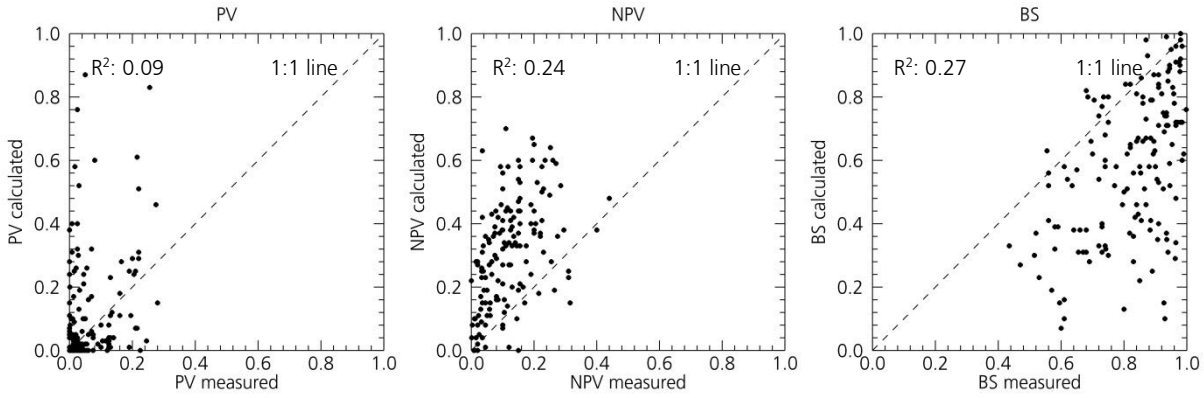


Figure 6-3: Comparison of ground cover fractions assessed in the field and calculated by ISMA spectral unmixing for 154 sample sites.

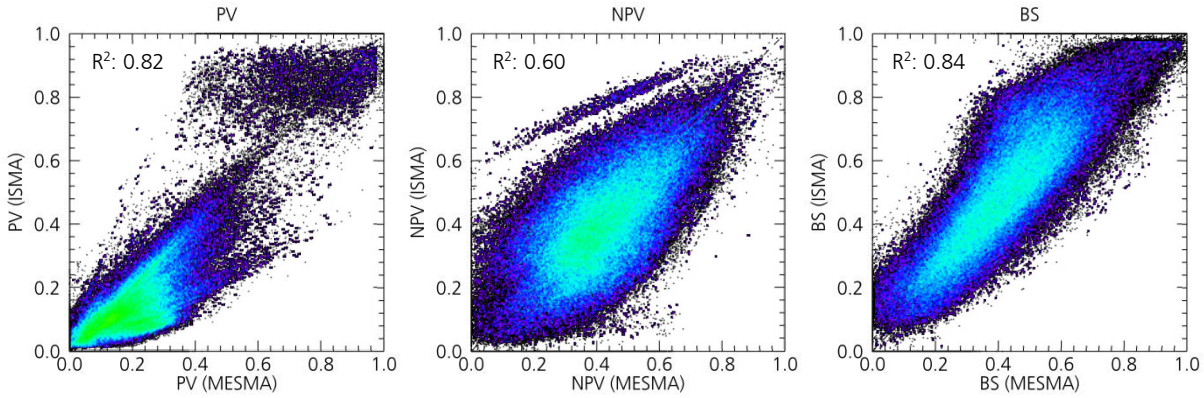


Figure 6-4: Pixel-based comparison of ground cover fractions of two unmixing approaches ISMA and µMESMA calculated for scale 2 subset of farm Gonakraal (PV: photosynthetic active vegetation, NPV: non-photosynthetic active vegetation, BS: bare soil).

For a second, relative validation, the ISMA unmixing results are compared to another well established spectral unmixing approach. The µMESMA approach was selected, which was developed at DLR especially for the unmixing of semiarid regions. A full description can be found in Bachmann (2007). In relation to the processing of the spatially large hyperspectral dataset of the South African study area, the ISMA approach was preferred to be used for spectral unmixing, since µMESMA is computationally much more intensive. In comparison to ISMA, the µMESMA algorithm uses only one endmember for each of the classes green and dry vegetation and bare soil for a constrained unmixing of each pixel. µMESMA further uses the information of diagnostic absorption features and adjacency relations for the unmixing process. The ISMA cover fractions were scaled sum-to-1 to be comparable to the µMESMA cover fractions. Figure 6-4 shows scatter plots of the two independent unmixing approaches calculated for the Gonakraal subset using the same endmember set. Correlations are good and the majority of pixels lies on or close to the 1:1 line. Noticeable differences may be the result of the different methodical set-up of the two unmixing approaches (e.g. µMESMA being constrained to using only one endmember for each of the three classes per pixel). For example, this causes the development of local subgroups apart from the 1:1-line in the unmixing of green and also dry vegetation (see figure 6-4).

Conclusions

A well established and tested method (see Rogge et al. 2006) was successfully applied to derive ground cover fractions for the study area. The results overall show a reasonable range and unmixing RMSE are low. The derived ground cover fractions are consistent to another independent spectral unmixing approach (Bachmann 2007). Thus, it is assumed, that the applied unmixing results in physically realistic ground cover fractions, although they cannot be verified with the available ground truth data to complete satisfaction.

6.2.2 Approximation of per pixel soil residual signal

In section 5.2.2 the spectral mixtures within image pixels, resulting of the small scale heterogeneity of the South African study area, were resolved to prepare the hyperspectral imagery for an unbiased and quantitative assessment of soil constituents based on spectra. For soil dominated image pixels, the influence of vegetation on the spectral signal was removed and the soil signature contained in mixed pixels was approximated. This was done based on the pre-processed HyMap imagery and spectral unmixing results.

Within this procedure, a number of thresholds (e.g. bare soil and vegetation fraction, unmixing RMSE, mean reflectance) ensured the quality of the approximated soil spectra. In the following, these thresholds are discussed first, since they vary depending on the scene characteristics. A map of approximated soil signatures is given for the Gonakraal subset. Example soil signatures are compared to the original pixel spectra and their quality is evaluated.

Processing thresholds

Thresholds within the approximation of the soil residual signal ensured the quality of the retained soil signatures. The threshold values did not change the modeling of the soil signature itself, but they determine if the calculated soil signature is assigned to the pixel or the pixel is masked because the probability is too low such that a soil signature could be accurately determined. Thus, pixels that are not dominated by soil but show a high influence of vegetation or where the unmixing RMSE is high, were excluded. In addition, pixels are excluded, where the approximated soil residuum has a very dark mean reflectance in order to filter out pixels where signals not resolved during the unmixing and noise highly influence the calculated soil residuum. Although default values can be defined for each threshold, they can be adapted for each scene according to its characteristics. Table 6-5 gives the thresholds applied for the processing of each subset of the three scales.

For small image subsets of local extent (e.g. subsets of scales 1 and 2 see section 5.3.3), unmixing results are generally more accurate since the endmembers fit well the local surface materials and the sum of the cover fractions are closer to 1 (see discussion of unmixing accuracies in section 6.2.1). The thresholds may be slightly reduced, to still allow for an accurate derivation of the soil residuum for larger areas. For

Table 6-5: Overview of thresholds and results of the approximation of soil signatures from hyperspectral imagery for the subsets of each scale. Background or masked pixels (artificial objects) are excluded for the calculation of the percentages.

		Thresholds						% of pixels					
	Subsets	Maximal RMSE	Minimal fraction BS	Sum of cover fractions	Maximal fraction PV	Maximal fraction NPV	Minimal mean reflectance [%]	Valid Pixels	High unmixing RMSE	Low fraction BS	Sum of cover fractions very low or high	High fraction of vegetation	Low mean reflectance
	Default values	Factor 5	0.30	0.4-1.4	0.20	0.35	10						
Scale 1	**Gonakraal - A plain**	4.26[1]	0.30	0.4-1.4	0.20	0.35	10	52.0	0.2	8.6	0.8	36.9	1.6
	Gonakraal - B river	4.97[1]	0.30	0.4-1.4	0.20	0.35	10	78.1	0.3	12.2	0.3	4.2	5.0
	Arundale - A slopes	5.03[1]	0.30	0.4-1.4	0.20	0.35	10	69.7	0.2	17.0	0.1	11.6	1.4
	Arundale - B street	3.26[1]	0.30	0.4-1.4	0.20	0.35	10	58.4	0.3	16.5	0.2	21.9	2.7
	Wintersfontein - A river	4.25[1]	0.35	0.4-1.4	0.30	0.35	9	69.0	0.4	16.4	0.2	11.6	2.4
	Kruizementfontein - A hills	8.09[1]	0.30	0.4-1.4	0.20	0.35	10	20.7	0.1	44.3	0.7	33.3	0.9
Scale 2	**Gonakraal**	5.81[1]	0.35	0.4-1.4	0.20	0.30	10	40.8	0.2	37.1	0.3	19.3	2.3
	Arundale	4.56[1]	0.25	0.4-1.4	0.20	0.40	10	28.3	0.3	39.6	0.1	30.3	1.3
	Wintersfontein	5.59[1]	0.35	0.4-1.4	0.30	0.35	9	72.7	0.4	20.5	0.1	2.7	3.8
	Kruizementfontein	5.60[1]	0.30	0.4-1.4	0.20	0.35	10	11.2	0.2	81.8	0.4	6.2	0.1
	Stenbokvlakte	7.69[1]	0.30	0.4-1.4	0.30	0.35	10	13.0	0.6	64.2	0.1	22.2	0.0
Scale 3	**Northern section**	8	0.35	0.4-1.4	0.20	0.25	10	32.5	0.7	47.5	0.4	13.6	5.4
	Southern section	8	0.35	0.4-1.4	0.25	0.30	10	10.1	0.5	79.5	0.0	9.9	0.0
	Entire dataset							26.8	0.6	55.6	0.3	12.7	4.0

[1]Factor 5 of RMSE

example, the minimum soil fraction then must be higher than only 30 % and the pixels are allowed to include up to 20 % of green and 35 % of dry vegetation. The thresholds may also be adapted due to other specific characteristics in the scene. For instance, in the Wintersfontein subsets, dark mudstones occur close to the surface and reduce the surface reflectance over large areas. Thus, the minimal mean reflectance of the approximated soil signatures was reduced (9 % instead of 10 %) because in these areas soil signatures were approximated with good quality but may have a mean reflectance below 10 %. For the S1 and S2 subsets, the threshold for the unmixing RMSE (T_{RMSE}) was calculated based on the average of the RMSE ($RMSE_{mean}$) over the entire subset and its standard deviation (σ_{RMSE}).

$$T_{RMSE} = RMSE_{mean} + f \cdot \sigma_{RMSE} \quad [-] \quad (6\text{-}1)$$

The factor f for all subsets is set to 5, which was done based on various tests. Another threshold derived from the two thresholds set for green and dry vegetation (T_{PV}, T_{NPV}) limits the total vegetation fraction regardless if green or dry. It was set in between the threshold of each vegetation status and the sum of both.

$$T_{PV+NPV} = \frac{T_{PV} + T_{NPV}}{2} \cdot 1.5 \quad [-] \quad (6\text{-}2)$$

The unmixing of larger datasets of regional extent, such as the scale 3 imagery of the entire northern section of the study area (235 km^2), cannot provide comparable accuracy because of the comparably small EM set that is used for the unmixing. If the same threshold values would be applied to this imagery, soil signatures would be approximated for more pixels but a significant number of pixels would result in erroneous soil signatures. Thus, the thresholds need to be adapted to ensure a good overall quality of the approximated soil signature. This resulted in a minimum soil fraction that must be higher than 35 % and maximum fractional covers of 20 % for green and 25 % for dry vegetation for the northern transect, whereas for the south 25 and 30 % were used. For the approximation of the soil signature of the entire datasets, the RMSE threshold is set to 8. No factor is used.

In the right half of table 6-5, the total percentage of image pixels is given, where a soil residuum could be approximated (valid pixels), and in addition the percentages of pixels that are excluded for each factor. For the entire imagery, for 26.8 % of the pixels, a soil signature could be approximated with the proposed method. Pixels that fail the given thresholds are masked for further processing. To a certain extent they are interpolated by a selective filter operation in one of the last steps of the prediction of soil constituents (see section 5.3).

Spatial distribution of approximated soil signatures

Figure 6-5 shows the map of approximated soil signatures for the Gonakraal subset (scale 2). For pixels that fail the thresholds, a removal code flags the reason. If more than one reason is eminent to remove the pixel, the pixel is flagged according to the first threshold that is failed. Both maps directly give an impression of the spatial distribution of factors limiting the approximation of proper soil signatures based on the original hyperspectral imagery. In the Gonakraal area soil signatures could be approximated nearly spatially continuously for the bare plains between the small erosive channels. In the surrounding areas, where soil surfaces are covered by mostly dry vegetation at different intensity, soil signatures are only approximated for local patches, depending on the intensity of the vegetation coverage. The same applies for the slopes in the south that rise up to the prominent fence line contrast. On the slopes, only pixels where the influence of vegetation is low and does not prevent the accurate approximation of soil signatures are applied. Across the fence line, the influence of vegetation is too high and pixels are masked.

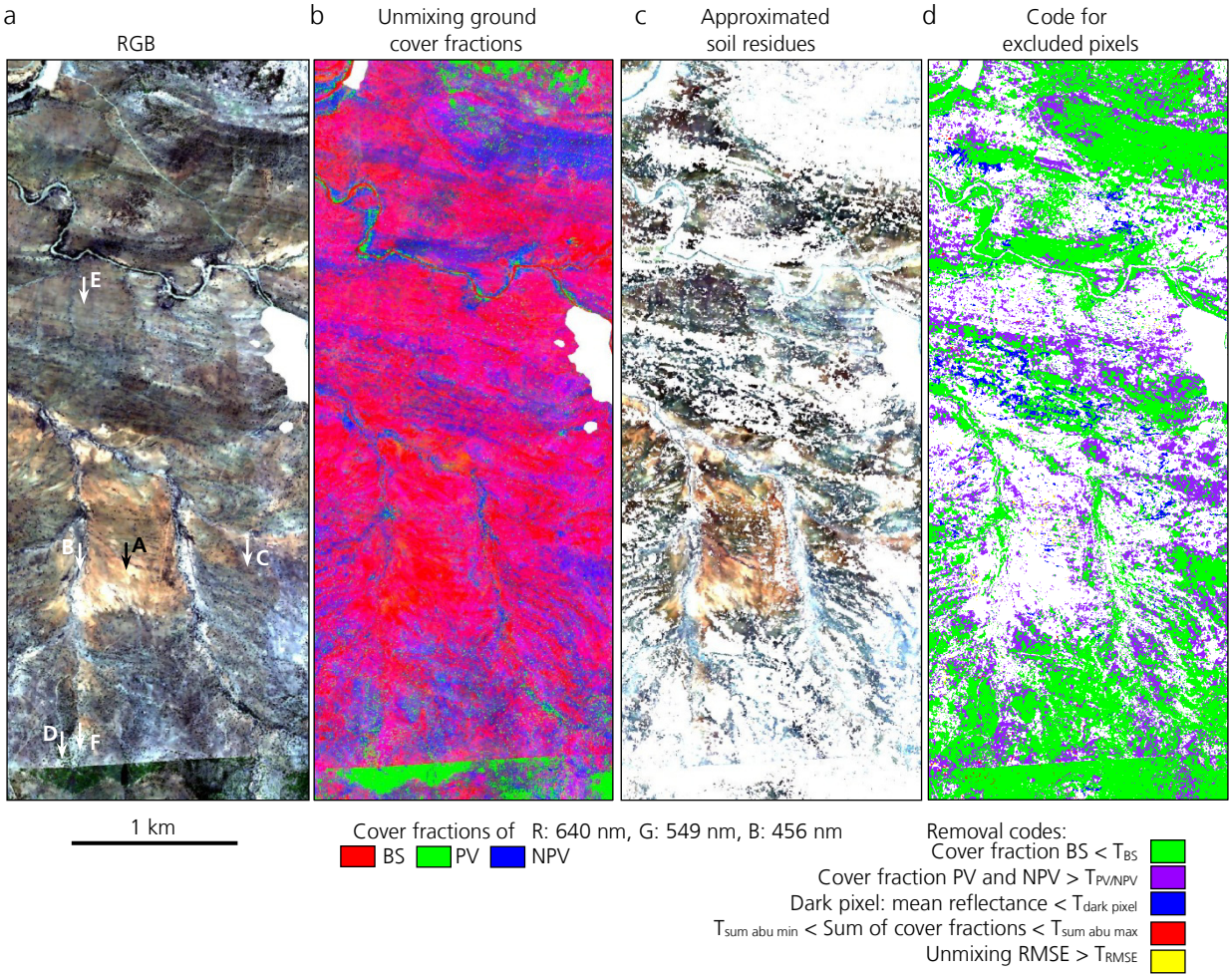


Figure 6-5: Approximated soil residues for the farm Gonakraal (farm area is masked). A true color image (a, with R: 640 nm, G: 549 nm, B: 456 nm) and unmixing fractional covers grouped to the classes PV, NPV and BS (b) are given to visualize the selective removal of pixels during the procedure. Approximated soil residues are shown as RGB (c) and the reason due to which pixels were removed during this step is given for each pixel (d). The different thresholds are termed with T. Labeled points in image a refer to the locations of spectra of figure 6-6.

Spectra of approximated soil signatures

Figure 6-6 gives examples of approximated soil signatures of the Gonakraal area for different ground cover fractions. The locations of the sample pixels are indicated in figure 6-5. Approximated soil signatures are equal to the original pixel spectra if no influence of vegetation was determined in spectral unmixing and ground cover fractions of PV and NPV are 0. An example is given on the bare plains (figure 6-6a and A in figure 6-5a). In areas with a predominance of soil but together with mixed fractions of vegetation (figure 6-6b-e and B-E in figure 6-5a), the vegetation signal is filtered well by the approach, and in the VIS range a significant reduction of the red edge, which is characteristic for green vegetation, is apparent. Only where the unmixing does not identify green vegetation, the red edge remains as artifact in the spectra of approximated soil residues (figure 6-6f and F in figure 6-5a).

For the 26.8 % of the image pixels of the entire study area, for which a soil residual spectra could be approximated, the spectra are overall of good quality and show reasonable mean reflectance and also characteristic soil spectral features. Only singular spectra passing the thresholds are biased by remaining vegetation signal (e.g. red edge) or are erroneous for other reasons (e.g. due to a poor quality of the original pixel spectra).

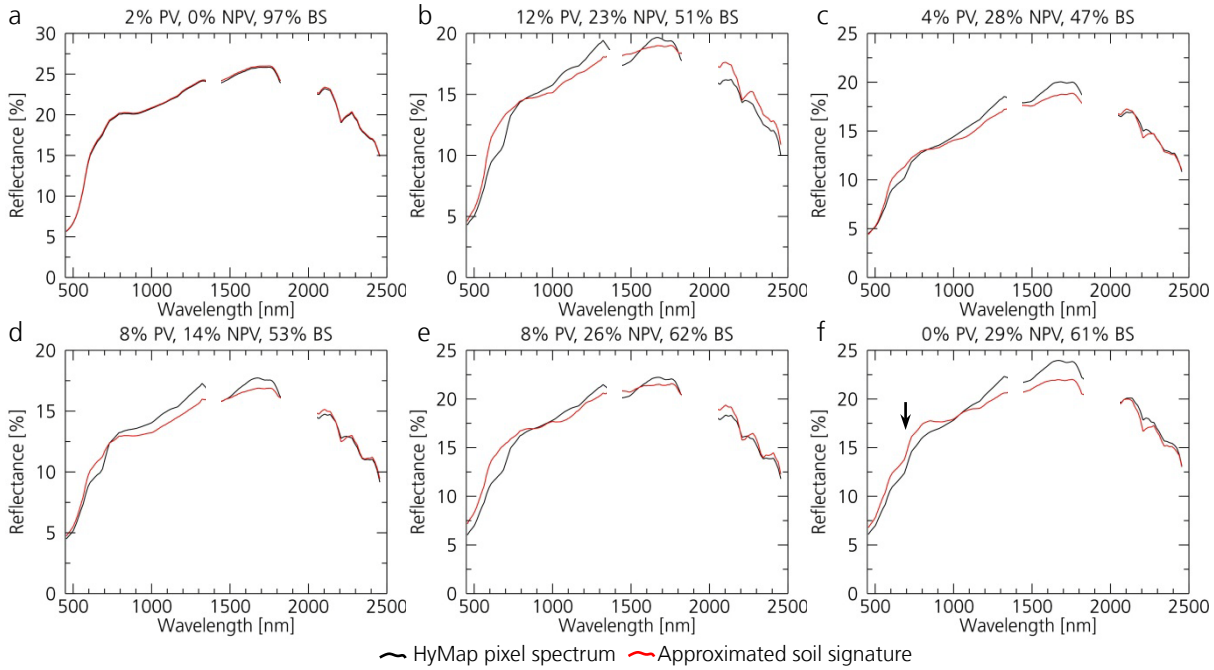


Figure 6-6: Examples for pixel soil signatures approximated based on the hyperspectral imagery and derived ground cover information.

Conclusions

Using this methodology, soil signatures could be extracted from image pixels having a vegetation coverage up to about 40 % (green and dry vegetation combined). The applied thresholds provide a separation of pixels which are suitable for this procedure based on their spectral unmixing parameters. In the predominant cases reliable soil signatures can be approximated. The procedure generated a reflectance image of the bare soil surface, in most cases without an influence of any vegetation. It provides a large spatial coverage despite spectral mixing within image pixels.

6.3 Discussion of the spatial distribution of soil constituents

Soil constituents are predicted for the South African study area by the application of the developed regression relationships to the maps of approximated soil signatures (refer to section 5.3). Maps of topsoil organic carbon, iron oxides and clay content are evaluated first for the entire study area. The spatial distribution patterns of the three soil constituents, linked to geomorphic features and processes within the ecosystem, are discussed in detail based on local examples of the different scales in the following sections. A comparison of the predicted soil constituents with contents measured in-situ in the field is conducted in section 6.4.1.

Description of predicted maps

The prediction of soil constituents for hyperspectral image data results in a five band image file (see section 5.3). The individual bands allow an overview of all levels of detail, from the unchanged calculated soil constituents to more generalized maps that allow a good overview of larger regions. An example for the

three bands of processed calculated values is given in figure 6-7, showing the soil organic carbon prediction for the Gonakraal subset. From the initial map of soil constituents predicted using the proposed methodology, first pixels are detected which values exceed the range of the chemical analysis of the reference samples and then masked (calibration limits). This only applies to singular pixels, thus statistics and the actual map have minimal change. Figure 6-7b shows the map of valid and unfiltered pixels. The further inclusion of neighborhood effects (selective smoothing and interpolation of missing pixels) results in two filtered bands, one smoothed with a small filter kernel size of 3 x 3 and the second with a larger kernel of 5 x 5 pixels. During this step, pixels where no soil residuum could be approximated are interpolated to a certain extent. A stronger smoothing and interpolation of more missing pixels is done in the filter operation with the larger kernel. The resulting image of figure 6-7d shows the predicted soil information more continuously, focusing on larger spatial variances, while figure 6-7b gives the information unfiltered and figure 6-7c slightly filtered. The histograms of the three bands (figure 6-8) reflect the influence of the filter and interpolation operations. A subtle skew of the distributions of soil organic carbon to higher values and of clay content to lower values is observed. With further smoothing, the distributions narrow their range and the number of extreme values declines (indicated by the two downward pointing arrows in figure 6-8). The increase of the total amount of values around the average is a result of the interpolation of individual missing pixels (indicated by the upward pointing arrow in figure 6-8). The cut off of calculated C_{org} contents below the lower calibration limit of 0.21 % C_{org} is shown in figure 6-8.

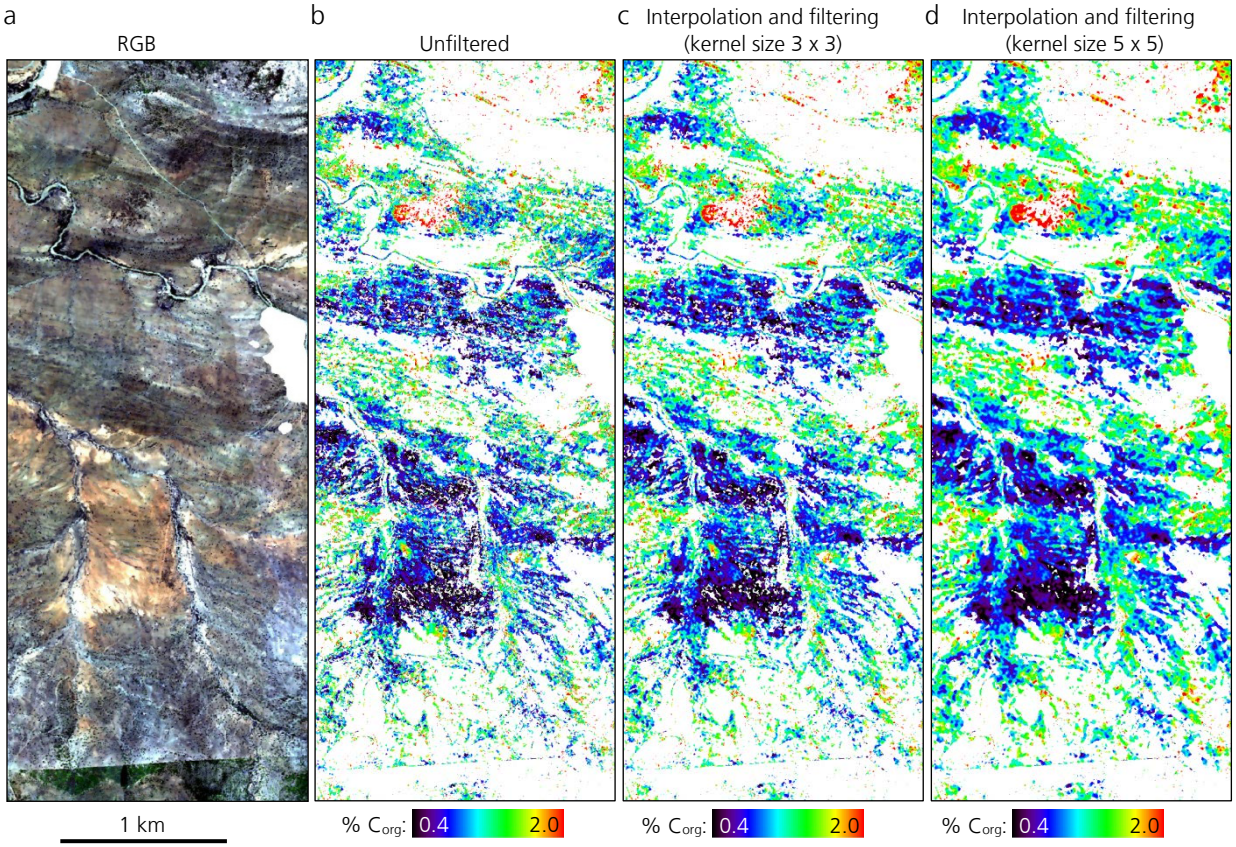


Figure 6-7: Maps of predicted topsoil organic carbon contents are produced unfiltered (b) and filtered using two different filter kernel sizes (c and d). They are shown here for the Gonakraal area (scale 2 data). True color composite with R: 640 nm, G: 549 nm, B: 456 nm.

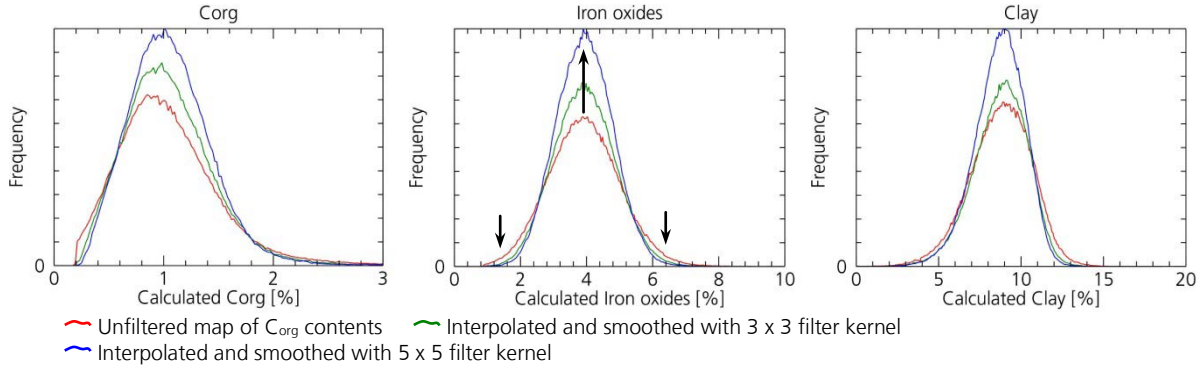


Figure 6-8: Histograms of the different output bands for the prediction of the three soil constituents reflect the influence of the filter and interpolation operations. They are exemplarily shown for the Gonakraal subset.

Through the interpolation of individual pixels that are excluded in previous processing steps from the surrounding, the amount of pixels within the entire study area for which soil constituents are predicted, can be increased from about 27 % to about 42 % in the map smoothed with the 5 x 5 kernel. In the following sections, for clearness in the figures, the map smoothed with the 5 x 5 kernel (band 4 in section 5.3) is always shown, because this map is the most spatially continuous. White areas in the maps mark pixels where no soil constituents are predicted. These are (1) pixels where no soil residuum is approximated (see section 5.2.2), and (2) pixels including artificial objects and excluded within pre-processing (section 3.2.2).

Overview of predicted soil constituents maps

Table 6-6 gives an overview of the range of soil constituents predicted for the about 8.4 Million pixels within the entire study area. For each constituent, the calculated values of about 98 % of the pixels are within the calibration range, which is defined by the chemical range of the reference samples collected in the field. For the prediction of soil organic carbon and iron oxides, negative values are calculated for less than 0.5 % of the pixels. With 2.4 % this fraction is significantly higher for the prediction of clay content, which is probably a result of the inaccurate calibration of clay prediction models (see section 6.1). For about 1.5 % of the pixels, contents outside the calibration limits are calculated. These contents might be reasonable, though the pixels are excluded for the final maps of predicted soil constituents because they need to be extrapolated by the model. Extrapolations would only be accurate for total linear behavior across the entire range of possible values, which does not apply since especially for high concentrations there exists an intensity saturation of absorption bands. For iron oxides this was for instance reported by Torrent et al. (1983).

Table 6-6: Overview of statistics of predicted soil constituents for the entire study area.

% of pixels		**C_{org}**	**Iron oxides**	**Clay**
Negative values calculated	($c_{calc} < 0$)	0.30	0.15	2.40
Calculated contents below lower calibration limit	($0 < c_{calc} < c_{min}$)	1.30	0.70	-
Calculated contents within calibration range	($c_{min} \leq c_{calc} \geq c_{max}$)	98.40	98.10	97.60
Calculated contents exceed calibration range	($c_{calc} > c_{max}$)	0.0001	1.07	0.00

% of particular soil constituent		**C_{org}**	**Iron oxides**	**Clay**
Calibration range [%]	(c_{min}, c_{max})	0.21-5.85	0.90-10.62	0.00-23.80
Statistics [%] (values of band filtered and interpolated with kernel size 5 x 5)		Mean: 1.23 Stdev: 0.49 Min: 0.21 Max: 5.84	Mean: 4.14 Stdev: 1.32 Min: 0.90 Max: 10.62	Mean: 7.06 Stdev: 2.10 Min: 0.00 Max: 22.30

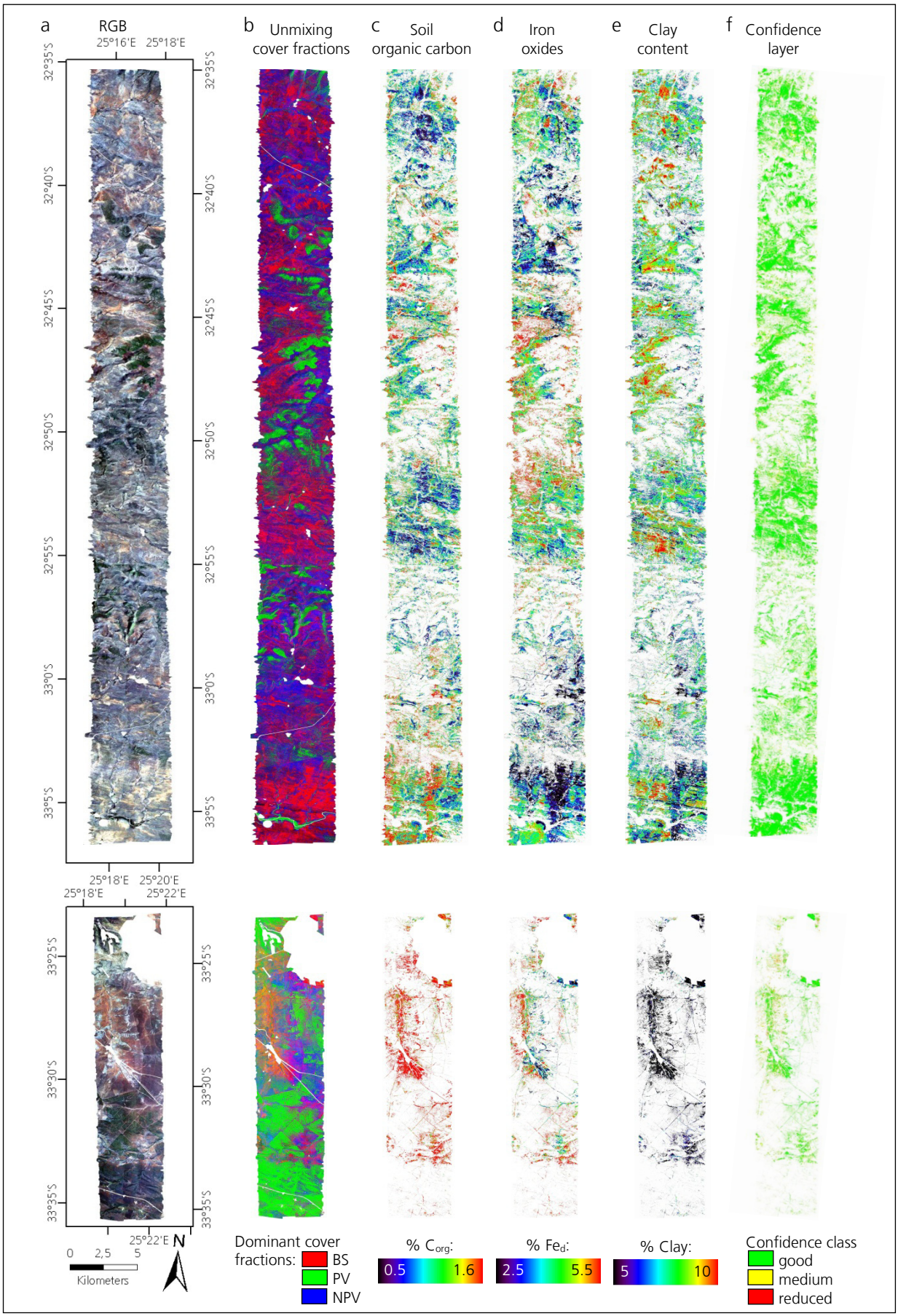


Figure 6-9: Overview of the results of the multi-stage methodology for the entire South African study area with ground cover fractions (b), maps of predicted topsoil constituent quantities (c-e) and per pixel confidence classification layer (f). For areas masked white, no soil constituents can be predicted.

The good match of the range of predicted soil constituents with the field measurements showed that the field sampling represented well the variability of all three investigated parameters in areas for which soil constituents are predicted and that the prediction model is well adapted to the present conditions. It is expected that soil organic carbon contents are nonetheless higher in densely vegetated areas. Those areas were not sampled, since they are not in the scope of a prediction of soil constituents from hyperspectral imagery.

An overview of the maps of soil organic carbon, iron oxides and clay contents for the entire study area predicted by the proposed methodology is depicted in figure 6-9. These maps provide snapshots for the spatial distribution of these parameters in October 2009. The distribution of areas for which soil constituents are predicted is well defined by the distribution of vegetation. Where soil is the dominant cover fraction over large regions, soil constituents are predicted nearly spatially continuous. In regions where the original dense Thicket vegetation is retained, no soil constituents are predicted due to the high cover of green vegetation. This applies particularly to the slopes of the mountainous regions crossing the central part of the northern transect and large areas within the southern transect. Dry vegetation occurs throughout the entire region. Its presence is also increased around the slopes, and therefore further limits the areas where soil constituents can be predicted. The layer classifying the confidence of the predictions for each pixel (see section 5.3.1) flags a good quality for 78.1 % of the pixels. A medium (19.8 %) or reduced confidence (2.1 %) is flagged predominantly in areas close to vegetation patches, in particular in the in the southern transect part. A regional example for the confidence classification is given in section 6.3.4.

6.3.1 Soil organic carbon

For 98.4 % of the image pixels, soil organic carbon contents within the calibration range (0.21-5.85 % C_{org}) are calculated. This fraction would increase to 99.7 % of the pixels if also non-negative contents outside this range would be considered (i.e. 0-0.21 %). Resulting histograms of the modeled C_{org} concentrations show a realistic range when compared to the field measurements (figure 6-8a), only very low and high values seem underestimated. This is likely an effect of the spatial resolution of the HyMap sensor. Since surfaces with extremely low or high contents usually have limited spatial extents, an up-scaling to the HyMap spatial resolution of 3.3 by 3.3 m results in a smoothing of C_{org} minima and maxima. The average content of approximated soil organic carbon content for the entire imagery is 1.23 %, with a standard deviation of 0.49 % C_{org}. For the 163 reference samples collected at random points within the study area, an average C_{org} content of 1.21 % with a standard deviation of 0.89 % C_{org} is determined in chemical analysis.

The spatial distribution of soil organic carbon is a result of the interdependence of the duration and intensity of the enrichment in C_{org} by the accumulation of plant litter on one side and the reduction of C_{org} by soil erosion on the other. This relationship in general is affected by topography. Based on the spatial variability of soil organic carbon, an ecosystem can then be divided in two zones. Soil development is progressive, where under relative stability, soils become more developed and the topsoil layer is enriched in organic carbon. In areas of regressive soil development, soil erosion takes place and reduces the amount of newly built soil material and accumulated organic matter and thus C_{org} contents are low (after Birkeland 1990). Areas with diverging flow directions are mostly active erosion zones (e.g. hill tops and slopes) while sediment is accumulated where water flow converges, such as in concave valleys and plain slope segments. The described direct link is most pronounced in ecosystems with a more or less evenly distributed vegetation

coverage, providing continuous carbon enrichment of soils. Due to the fact that the dense pristine Thicket vegetation is not distributed evenly within the study area but mostly on the hill slopes, this relationship is modified by the non-homogenous carbon input resulting from the distribution of vegetation. Figure 6-10 and figure 6-11 show local examples for the prediction of soil organic carbon contents with the proposed method.

Areas of high C_{org} concentrations

In the South African study area high soil organic carbon contents occur (1) within highly vegetated areas i.e. in the central part of the northern transect and the entire southern transect, (2) on partly vegetated gently dipping slopes and (3) within partly vegetated flat river plains. In all these situations organic carbon contents are usually high primarily because the present vegetation causes an increased carbon input into the soils. However, this is only observed if the soils are prevented from being eroded, either by low topography (e.g. in river plains), or under gently dipping conditions by the holding capacity of vegetation patches. The isolated vegetation patches then produce discontinuous runoff patterns, and soil organic carbon patterns vary depending on the distance to the vegetation (e.g. figure 6-10a-c, figure 6-11a). Within densely vegetated areas, predicted contents are limited to isolated spots and soil constituents are not predicted spatially continuously. Increased concentrations of organic carbon for soils provide positive effects on their structure, infiltration and increased nutrient availability (Schachtschabel et al. 2002), which are all factors promoting plant growth and litter fall and making these soils more resistant to soil erosion. Thus, there is a positive feedback mechanism driven by organic matter in soils (Hill and Schütt 2000) and a correlation between increased soil organic carbon contents and vegetation (see also end of this section). Especially in semiarid regions, those mechanisms are important to sustain soil fertility, increase soil moisture storage and mitigate droughts (Tiessen et al. 1994).

Low C_{org} concentrations

Soil organic carbon contents are usually low in areas where the input of soil organic carbon is minor due to a reduced vegetation coverage and/or soil erosion is predominant and outweighs organic matter input for longer periods (regressive soil development after Birkeland 1990). This generally applies for (1) bare areas where almost no vegetation is present, (2) erosive slopes with medium to high gradients, where a reduced vegetation cover is unable to prevent the erosion of the organic topsoil layer, (3) elevated ridges where erosion is usually high because of their topographic exposition and Thicket vegetation coverage is low and (4) runoff trenches that are not overgrown by vegetation. An example for erosive slopes is given in figure 6-11b for an area in the northern part of the study area. The terrain drops off towards the small river on the fan to the southeast. Both parts of the slope are mostly bare, except for some vegetation present along small erosion channels. The soil organic carbon map reveals that in the eastern part erosion is dominant, while in the western part litter accumulated from the plants is retained on the soils, decomposes and increases the soil organic carbon contents. Soil erosion of partly vegetated areas frequently occurs at inclined hill slopes where the original vegetation coverage is significantly constricted by intensive grazing. The carbon maps of figure 6-10c and figure 6-11a show examples of such a set-up within the study area. The remaining vegetation fixes the attached sediment and prevents its erosion. Figure 6-12 shows two examples for the holding capacity of patches of small herbs on a slightly inclined surface. The fourth scenario referred to above applies to erosion trenches where rain water immediately runs off (e.g. figure 6-10c). Thus, they are often bare and topsoil C_{org} is eroded during the periodically occurring rainfalls.

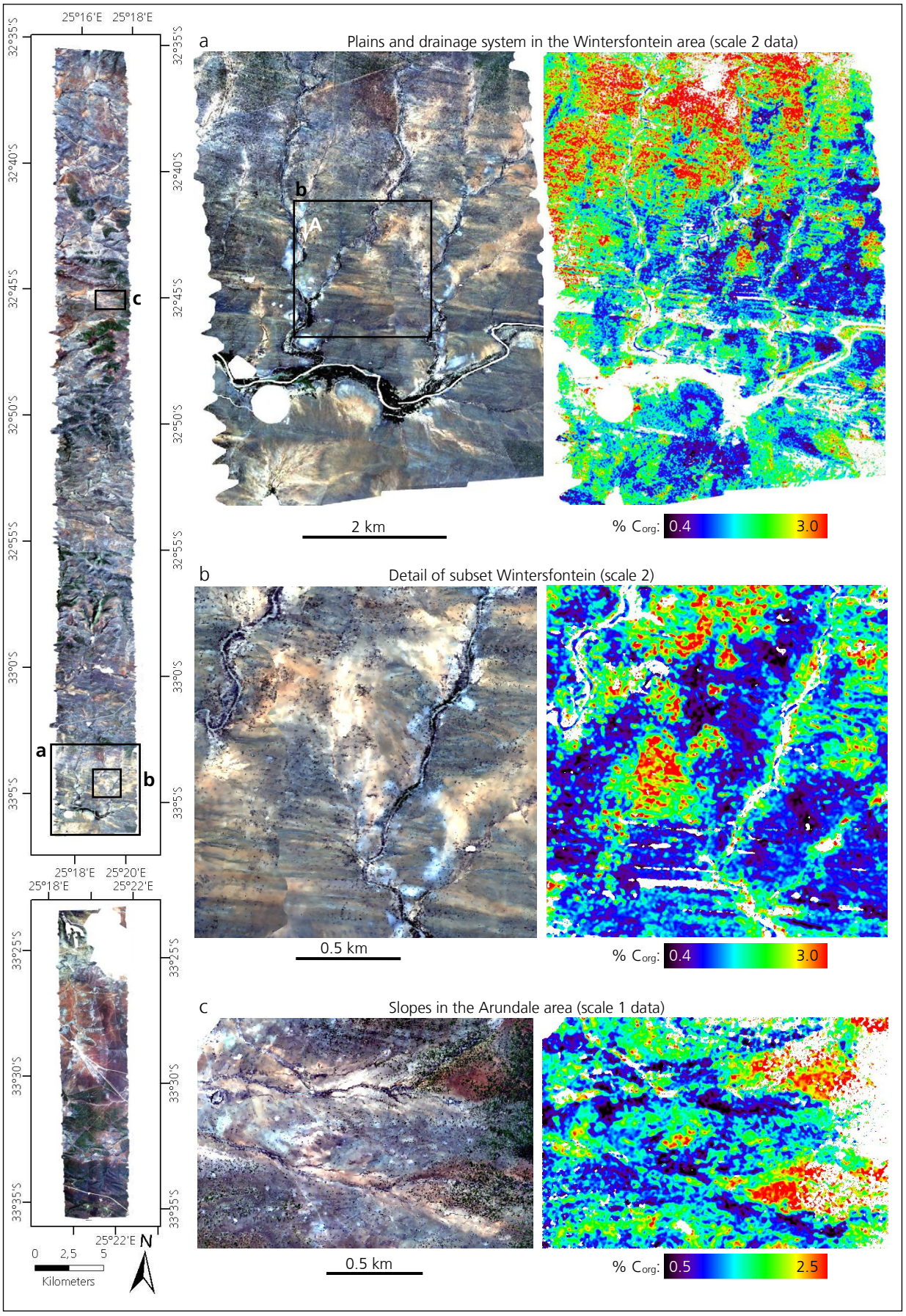


Figure 6-10: Local examples for predicted maps showing the spatial distribution of topsoil organic carbon contents for different scales (band 4 see section 5.3) - Part I. For areas masked white, no soil constituents can be predicted. True color composites with R: 640 nm, G: 549 nm, B: 456 nm.

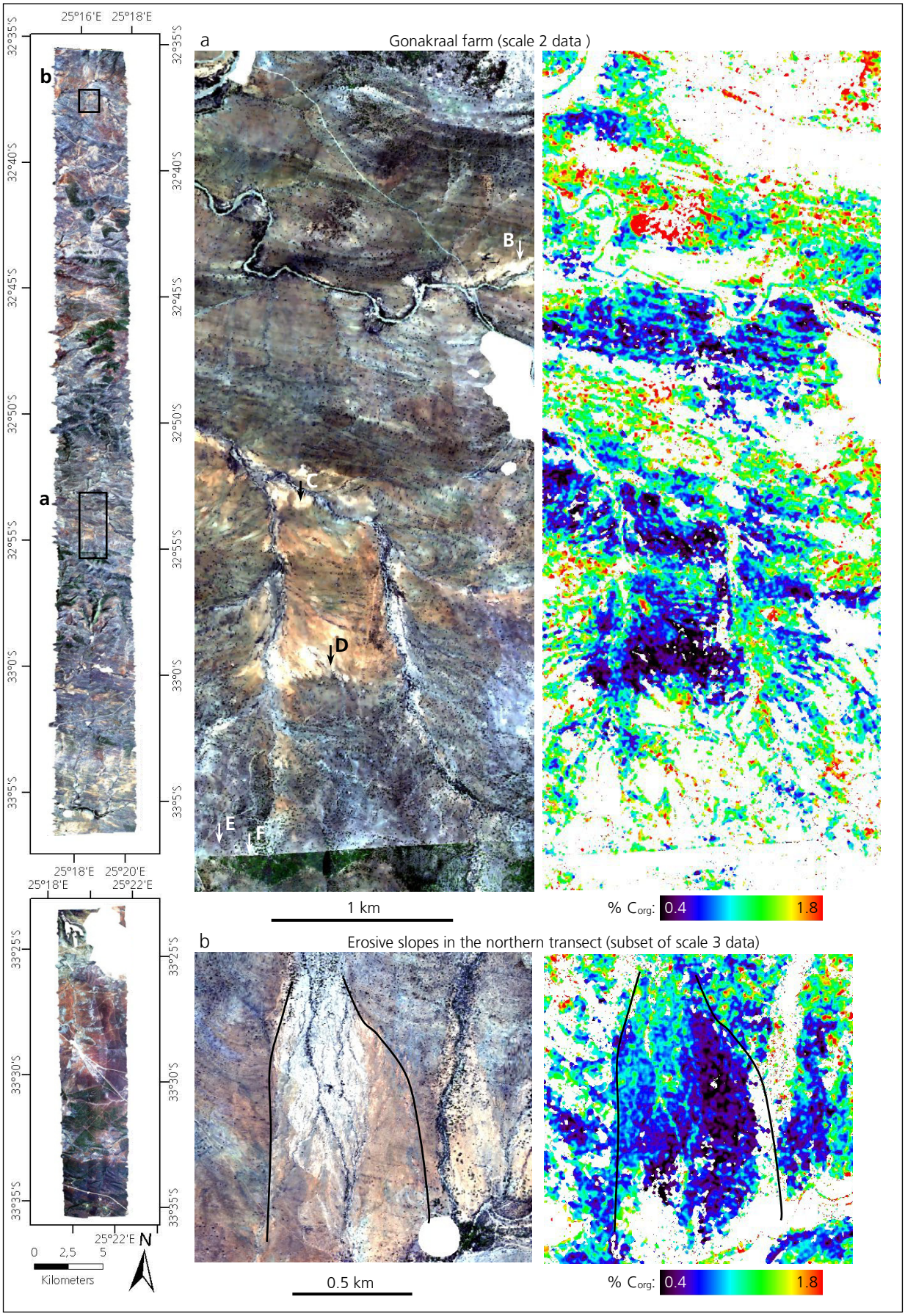


Figure 6-11: Local examples for predicted maps showing the spatial distribution of topsoil organic carbon contents for different scales (band 4 see section 5.3) - Part II. For areas masked white, no soil constituents can be predicted. True color composites with R: 640 nm, G: 549 nm, B: 456 nm.

Analogue to the positive feedback mechanism driven by high C_{org} contents, low organic carbon contents cause negative influences on soil and ecosystem development, i.e. plant growth is limited while soils become prone to erosion due to a lowered structure stability (compare section 3.1.2). Since low C_{org} contents are an indicator for present erosion and soil net loss, there is consequently a high probability that soil profiles are truncated in these areas. Long-term erosion leads to an increase of bedrock exposure at the soil surface in source areas, which can be observed frequently in the study area (e.g. figure 6-13b).

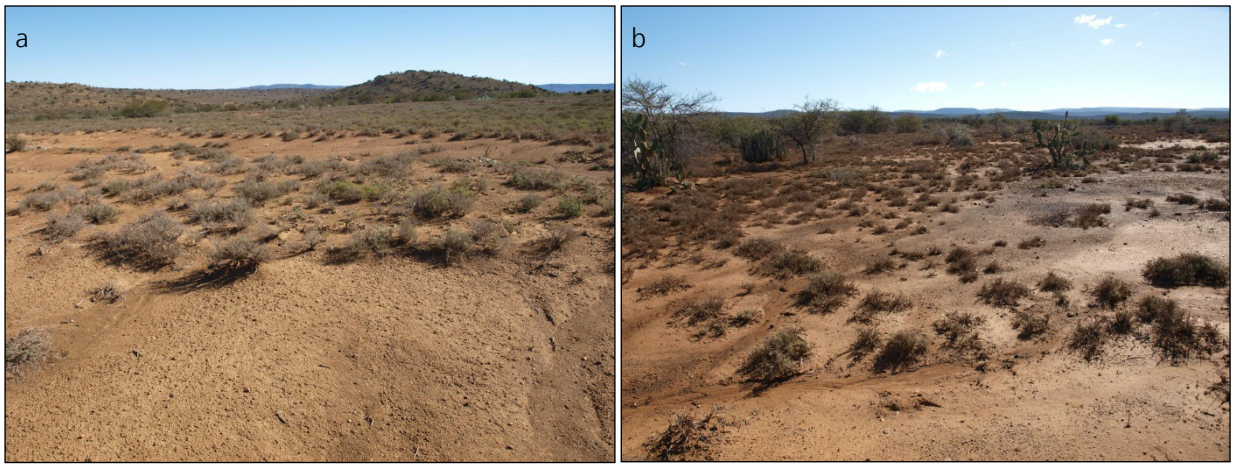


Figure 6-12: Directly around patches of isolated herbs sediment is held back from being eroded on a slightly inclined surface (a, picture taken at location B in figure 6-11a). Discontinuous runoff and erosion patterns form between the vegetation patches (b, picture taken at location A in figure 6-10a). Images were taken in 2009.

Examples from the Gonakraal area

The Gonakraal subset depicted in figure 6-11a includes examples for nearly all described situations. A drainage pattern is obvious originating from elevated rangelands in the South and heading towards the larger river systems in the North. Soil organic carbon contents are mostly below 0.60 % on the bare plains in the center of the subset. Soil erosion is a prominent factor here and it is likely that the organic topsoil layer is eroded across large areas (see figure 6-13a). A soil profile taken in 2011 on this plain revealed that organic carbon contents are very low in the surface layer (0.19 % C_{org}), yet rise with increasing depth to contents about 0.80 % C_{org} (see Kompter 2012). Contents are lowest in bare regions, e.g. locations C and D in figure 6-11a. At location D soil organic carbon contents around 0.35 % are predicted. The field survey in 2011, two years after the acquisition of field reference information, showed that at this location locally the entire soil layer is eroded and a deeper cemented horizon is exposed, shown in figure 6-13b. A soil organic carbon content of 0.33 % is measured, which matches well the predicted contents. This emphasizes that the developed soil organic carbon maps can be used to detect and monitor topsoil removal by erosion. In the middle of the plain, where some isolated shrubs and also dry vegetation is present, organic carbon contents are higher and about 1 %. Close to the erosive channels, C_{org} contents increase due to the higher vegetation coverage as a result of the improved water availability.

A rather progressive soil development with higher C_{org} concentrations in top soils is observed on the flat areas surrounding the plain in the center that are partially covered by vegetation. To the south of the Gonakraal area the terrain rises, vegetation cover increases and organic carbon contents increase to about 1.5 %. In this region, a spatially complete prediction of soil organic carbon can no longer be provided due to the increased vegetation coverage. The photo of figure 6-13c is taken close to the fence line (location 4

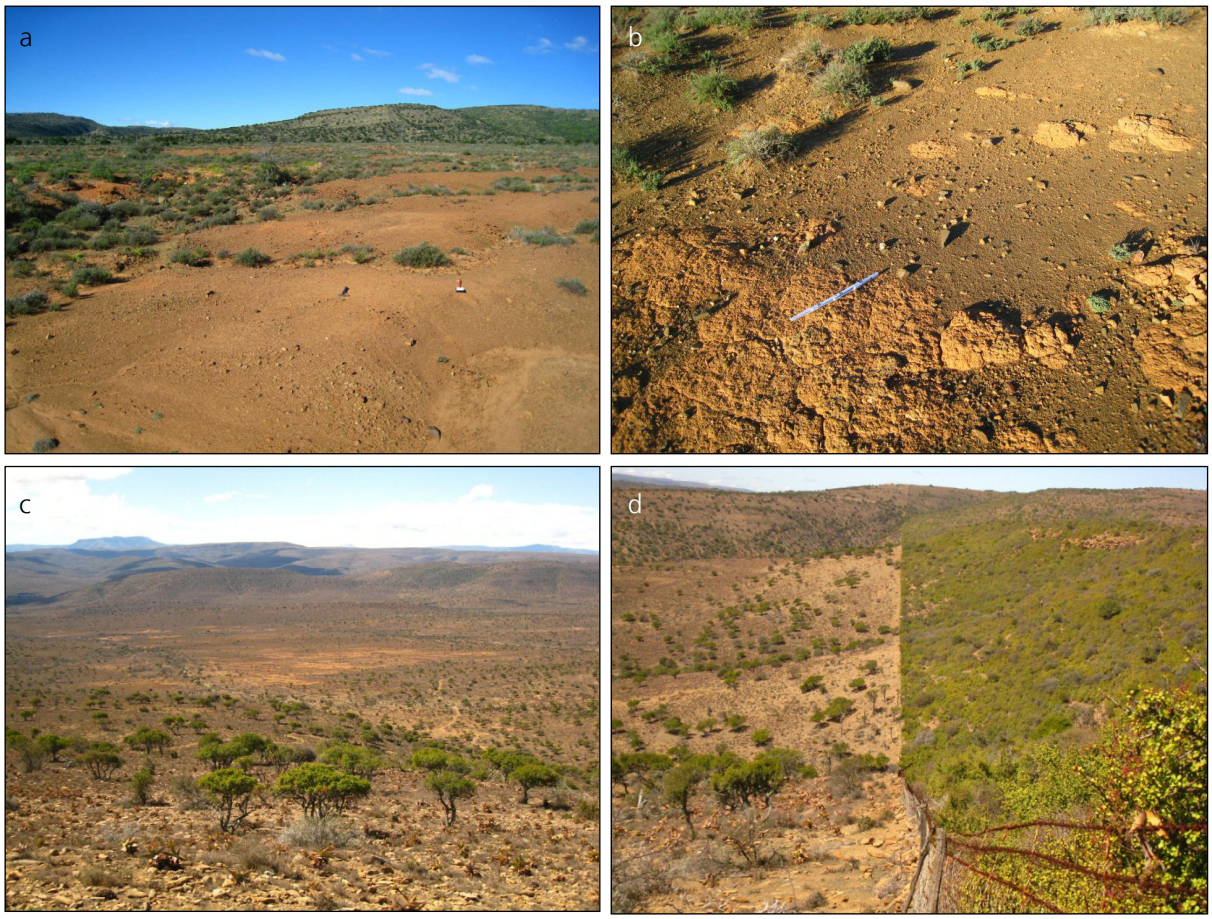


Figure 6-13: Field photos of distinct areas within the Gonakraal area.
A region on the flat plain in the center of the Gonakraal area where soils are highly eroded (a, taken at location D in figure 6-11a). In the north of the plain the organic topsoil layer is locally eroded and deeper cemented soil horizons are exposed (b, taken at location C in figure 6-11a). A partially vegetation coverage prevents a spatially continuous prediction of soil constituents on the slopes in the south (c, taken at location E in figure 6-11a). Where dense pristine vegetation is present, soil constituents can only be predicted locally (d, taken at location F in figure 6-11a). Images a and b were taken in 2011, c and d in 2009.

in figure 6-11a). The occurrence of green vegetation as isolated trees and the irregular patterns of dry vegetation explain the patchiness of this region in the maps of predicted soil constituents. On the pristine side of the fence line, C_{org} can only be predicted for small areas within the dense shrub matrix (figure 6-13d).

Link between C_{org} and vegetation

In the soil organic carbon maps derived for the study area, there seems to be a relationship between the occurrence of vegetation and higher soil organic carbon contents. Scatter plots shown in figure 6-14 reveal a present, though weak correlation between soil organic carbon and vegetation. Two reasons arise for this correlation. The first is a result from the relationship and positive feedback mechanism between plants and soil organic carbon concentrations as evaluated above. The presence of plants initially determines an increased carbon input to the soils and in addition prevents the erosion of the organic rich topsoil layer. This basic relationship makes a correlation between vegetation coverage necessary as it was also observed.

A second reason for the observed correlation results of the applied workflow: the correlation of vegetation coverage and carbon contents can be the result of a vegetation signature remaining in the spectral signature of the approximated soil signatures. This may be observed if the selected vegetation endmembers cannot explain the vegetation signature present in the mixed pixel. A vegetation signature remaining in the spectra used for a prediction of soil organic carbon would by all means lead to higher soil organic carbon contents, since the material and thus spectral features of some plant components are identical to soil organic matter. Nonetheless, over the entire study area low organic carbon contents also occur close to vegetation patches where the erosion of the organic rich upper soil layer is expected and predicted soil organic carbon contents are low (e.g. on slopes). Because of this and the overall reasonable spatial distribution and chemical range of predicted soil organic carbon contents, it is assumed that the second source of organic carbon does not generally apply and that the application of the thresholds within the approximation of the soil residual signatures is adequate for an identification of not accurately determined soil signatures. However, it may apply locally that a vegetation signature remains in the approximated soil residues and generate local artifacts in the spatial distribution of C_{org} (e.g. see red spot in figure 6-11a).

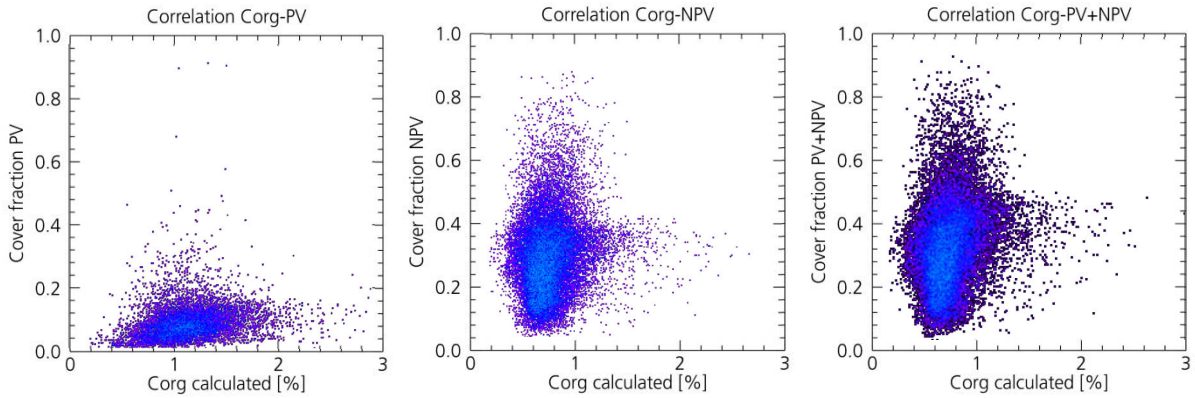


Figure 6-14: Scatter plots showing the correlation of green and dry vegetation and derived soil organic carbon quantities for the entire study area.

Factors influencing C_{org} estimations

Artifacts within the maps of predicted soil organic carbon contents may be generated by several factors: (1) from the proposed methodical workflow and (2) from the influence of other surface materials. An example for each of these factors is given in the following. Artifacts in the C_{org} maps resulting from erroneous approximated soil signatures may be best identified from the maps of the individual spectral variables. Figure E-18 and Figure E-19 in appendix E.4 give an example for the spectral variables within the C_{org} determination for the Gonakraal subset. At location 5 marked in figure 6-11a, high organic carbon contents are predicted within an area of low overall concentrations. Here, the original pixel spectra show a vegetation influence which is not represented in the spectral unmixing results. The approximated soil residual spectra, which are the basis for the delineation of organic carbon contents, thus are biased by a remaining vegetation signature. This results in a depth of the 1730 nm absorption above the ordinary and comparable high C_{org} contents (see Figure E-18c in appendix).

Surface materials that are significantly different from bare soil and vegetation may disturb the correlation between soil spectral and chemical properties, and thus, bias the delineation of soil organic carbon. An example can be found in the Arundale area, which is a valley surrounded by dolerite capped hills. The dolerites occur as iron rich magmatic intrusions in the sandstones and mudstones of the northern section (see section 3.1). Soils that form upon doleritic bedrock are of dark reddish color owing to the oxidation of

iron-rich minerals. The darkness in the VIS/NIR region causes the prediction of higher soil organic carbon contents than field measurements taken in these regions. The dolerites change the spectral properties of the soils in a way that is non-significant for the mechanism of carbon input through litter. Because the model is not calibrated for this, predicted contents are biased in these regions. Figure 6-15 shows an example of a strong spatial correlation between high organic carbon contents and the doleritic intrusions, which show up in the RGB image with a dark reddish soil color.

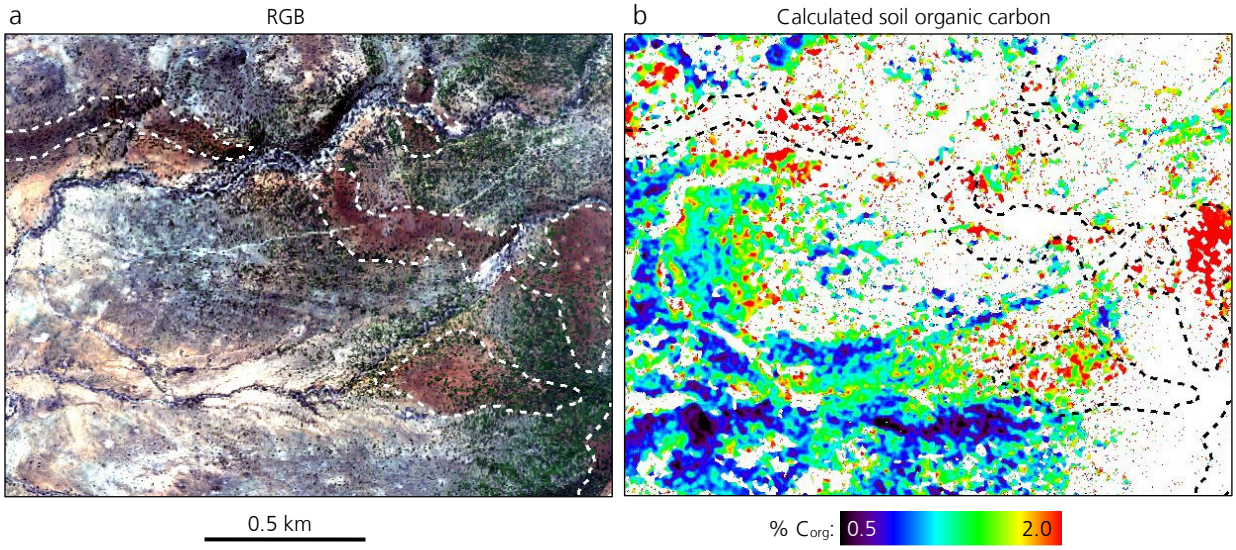


Figure 6-15: Example for doleritic intrusions influencing predicted soil organic carbon quantities from the Arundale farm area (data calculated within scale 3). In the true color composite (a), the dolerites show up with an intense dark red color and in the soil organic carbon maps (b) with high contents. For areas masked white, no soil constituents can be predicted.

Conclusions

The range of predicted soil organic carbon contents matches well within the chemical reference data and C_{org} contents within the calibration limits of the models are calculated for 98.4 % of the pixels. Spatial soil organic carbon patterns can be linked to progressive soil development with C_{org} enrichment in the topsoil and areas where active soil erosion outweighs the carbon enrichment and C_{org} contents are low. The observed patterns match well with previous studies focusing on links between soil organic carbon and topography in other semiarid regions (e.g. Hill and Schütt 2000, Schwanghart und Jarmer 2011). This relationship in the Albany Thicket Biome is significantly modified by the non-homogenous carbon input resulting from the remaining dense pristine Thicket vegetation on hill slopes.

6.3.2 Iron oxides

Iron oxides contents within the calibration range (0.90-10.62 % Fe_d) are calculated for 98.1 % of the image pixels. Modeled iron oxides concentrations have an average of 4.14 % Fe_d with a standard deviation of 1.32 % Fe_d. In comparison, the average content of the ground reference samples is lower and about 3.06 %, although with a comparable standard deviation of 1.44 % Fe_d. This indicates a low overall variability of the pedogenic iron oxides in the South African study area. In general, the spatial distribution of iron oxides in soils is a result of the parent rock material and the duration and intensity of pedogenesis (Qafoku and Amonette 2006, Schachtschabel et al. 2002). Pedogenic iron oxides primarily form as

weathering products of iron bearing mineral rock components (e.g. biotite, olivine, magnetite) and accumulate over time in the soil profile (Schachtschabel et al. 2002). They precipitate as clay-sized fractions, often as coatings on other soil materials such as sands and clays. This accounts for a high specific surface (Kuntze et al. 1988, Qafoku and Amonette 2006), making iron oxides one of the main pigmenting agents of soils and substantially contributing to the physical and chemical properties of a soil (Qafoku and Amonette 2006). Due to their water-insolubility in the normal pH range and their comparably high weight, iron oxides are very stable and remain close to the site of their production (Birkeland 1990). Thus, iron oxides are one of the end products of the weathering process in soils and accumulate as erosion residues (Qafoku and Amonette 2006). The dependence of iron oxides concentration on the parent rock material is discussed below.

Regions enriched in iron oxides

Resulting from the relationships described above, an increased occurrence of iron oxides is an indication for more developed older soils, assuming similar geologic conditions (e.g. Qafoku and Amonette 2006, Schachtschabel et al. 2002, Torrent et al. 1983). Within the South African study area this indicates a higher stage of soil development in the southern transect compared to the northern transect. Soil profiles taken in 2011 at 32 sites confirm this conclusion with deeper average profile depths in the southern part (average depth of 32 cm in northern transect and 52 cm in the southern transect, see Kompter 2012). The higher concentrations of iron oxides in soils of the southern transect are also reflected in soil color, where more reddish Munsell color hues were found (Kompter 2012). Also the climatic constraints providing the region where the southern transect is situated with slightly higher precipitation, and the higher vegetation coverage favor a more intense pedogenesis, and thus, higher stage of soil development in the southern part. Apart from that, higher contents of iron oxides are found in different morphologic set-ups, ranging from plateaus to slopes stabilized by vegetation. They are enriched, when other soil components such as organic carbon are dissolved and eroded. In addition, high concentrations of iron oxides are found in areas that are influenced by doleritic intrusions, due to increased oxidation of iron rich minerals within these magmatic rocks. This irregularity is discussed below.

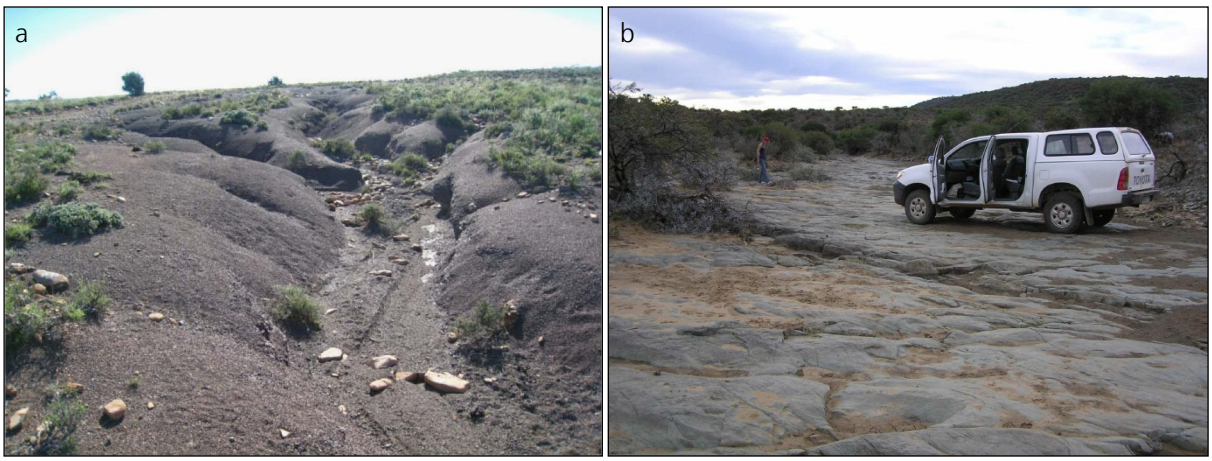


Figure 6-16: Field photos of erosive channels where bedrock is exposed within the South African study area. a: In this small runoff channel the entire shallow soil layer is eroded and the mudstones exposed (picture taken at location 1 in figure 6-17). b: For the same reason, in large river beds, the parent material are exposed which in this case are bright slightly metamorphosed sandstones (picture taken at location 2 in figure 6-17). Image a was taken in 2011 and b in 2009.

Reduced iron oxides concentrations

Low iron oxides concentrations occur, where (1) soils are young and soil formation is incomplete, (2) aeolian abrasion removes the iron-rich coatings by particle-by-particle friction driven by the wind energy (Ben-Dor et al. 2006) and (3) within and close to erosive channels. The third set-up is apparent within the entire study area, in both smaller and larger drainage channels (e.g. figure 6-17a and c, figure 6-18a and b). All drainage channels in the study area only bear water after heavy rains. This also applies to the largest channel in the south of the northern transect. Field photos show that potentially developed soils in these channels are completely washed away and the basement rock is exposed (figure 6-16). It is suggested that the soil material, and with it the iron oxides, is laminary eroded by the water flooding the channels during and after heavy rains.

Link between iron oxides and geology

Correlation between the distribution of iron oxides and the underlying geology is frequently observed within the study area. The sandstones and mudstones, which are the dominant geologic parent material in the study area, may both contain various amounts of iron oxides. The clay minerals forming the mudstones contain iron in their lattice structure. While the individual grains of the sandstones, which in this region are predominantly made of quartz, may be coated with iron oxides. Due to their sedimentary origin, the iron present in both materials itself was formed as weathering product from the original parent material. In these sediments, significantly higher Fe-contents are found within sandstones compared within the mudstones. Figure 6-17b and figure 6-18a show examples for areas where the underlying geology is clearly visible as changes in soil color. In the mountainous Kruizementfontein area (figure 6-17b) an alternation of almost horizontally lying layers of sandstones and mudstones are visible on the slopes. The hill tops are formed by more resistant sandstones. In this area an intense occurrence of dry herbs prevents a spatially continuous prediction of iron oxides. In the mostly flat Gonakraal area, the geologic basement is also visible on the surface by changes in soil color (figure 6-18a). A folded sequence of sandstones and interbedded mudstones occurs in the eastern part of the subset. Here, the same relationship between iron oxides contents and the two rock materials is also apparent.

Examples from the Gonakraal area

Iron oxides contents in the northern part of the Gonakraal area show little variation and range between 4 and 6 % Fe_d (green to red colors in figure 6-18a). Despite that, contents are mostly below 3 % Fe_d in the erosive channels cut in the flat plains and northward facing slopes. High variances also occur on the plain in the center of the subset. In the southern part of the plain (location A in figure 6-18a), which is devoid of any vegetation, contents are low (about 2-3.5 % Fe_d). This may be the result of laminar erosion of sediment. The same applies for location B (figure 6-18a and figure 6-13b), where the soil layer locally is removed completely as was already discussed above as an explanation for low organic carbon concentrations and is shown in figure 6-13b. Iron oxides contents are higher in the northern part of the plain, where laminar erosion of the entire soil layer is prevented by a minor vegetation layer. Here, iron oxides accumulate within pedogenesis and are further enriched when water-dissolvable contents such as certain organic materials are partly removed. With increasing occurrence of green vegetation towards the fence line contrast in the south, iron oxides contents do not significantly change from the overall trend. Lower contents are still predominantly close to channels (C and D in figure 6-13a), whereas the presence of iron oxides increases on the slopes in between. In the eastern part of the subset, iron contents are modified by bedrock geology.

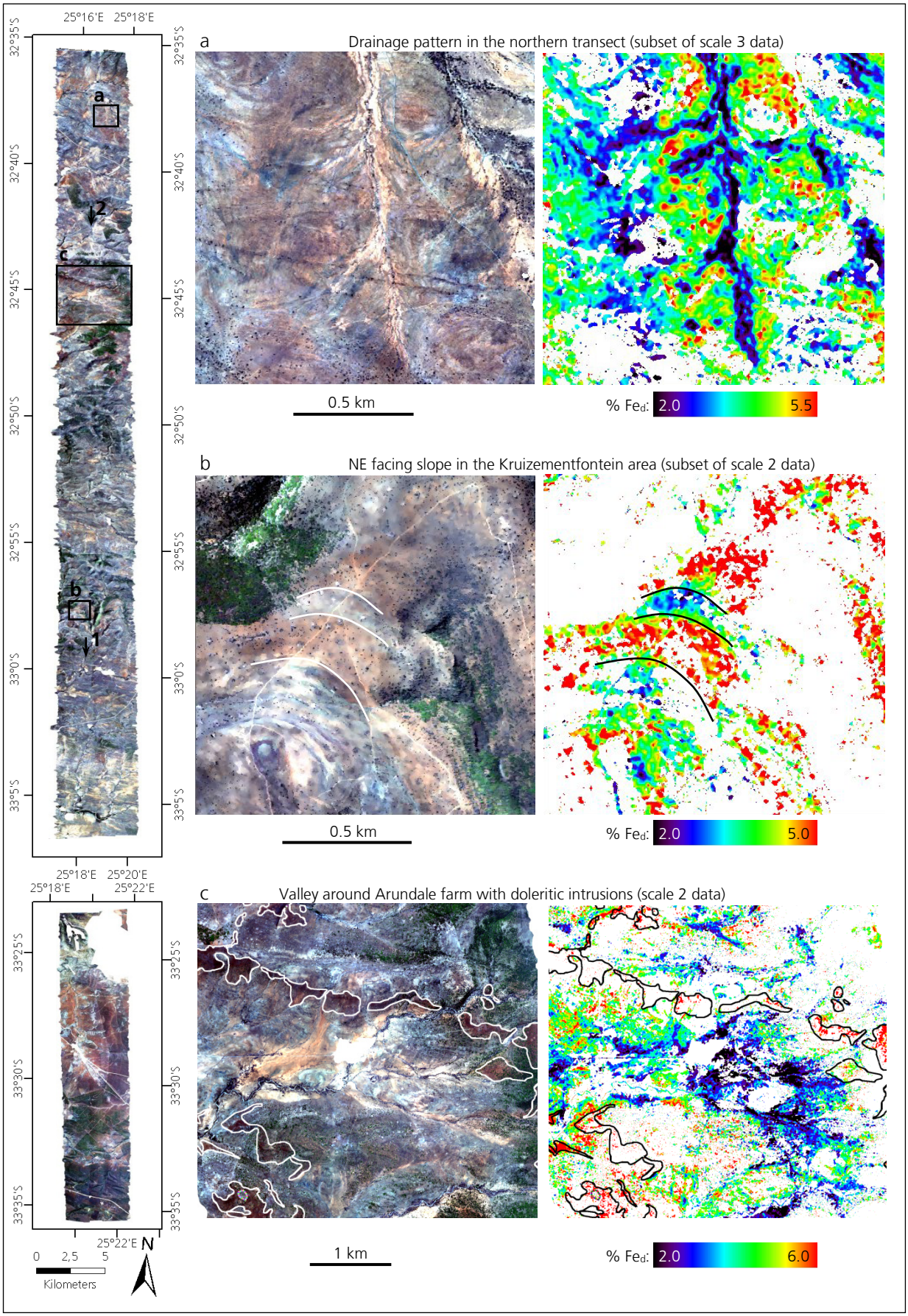


Figure 6-17: Local examples for predicted maps showing the spatial distribution of topsoil iron oxides contents for different scales (band 4 see section 5.3) - Part I. For areas masked white, no soil constituents can be predicted. True color composites with R: 640 nm, G: 549 nm, B: 456 nm.

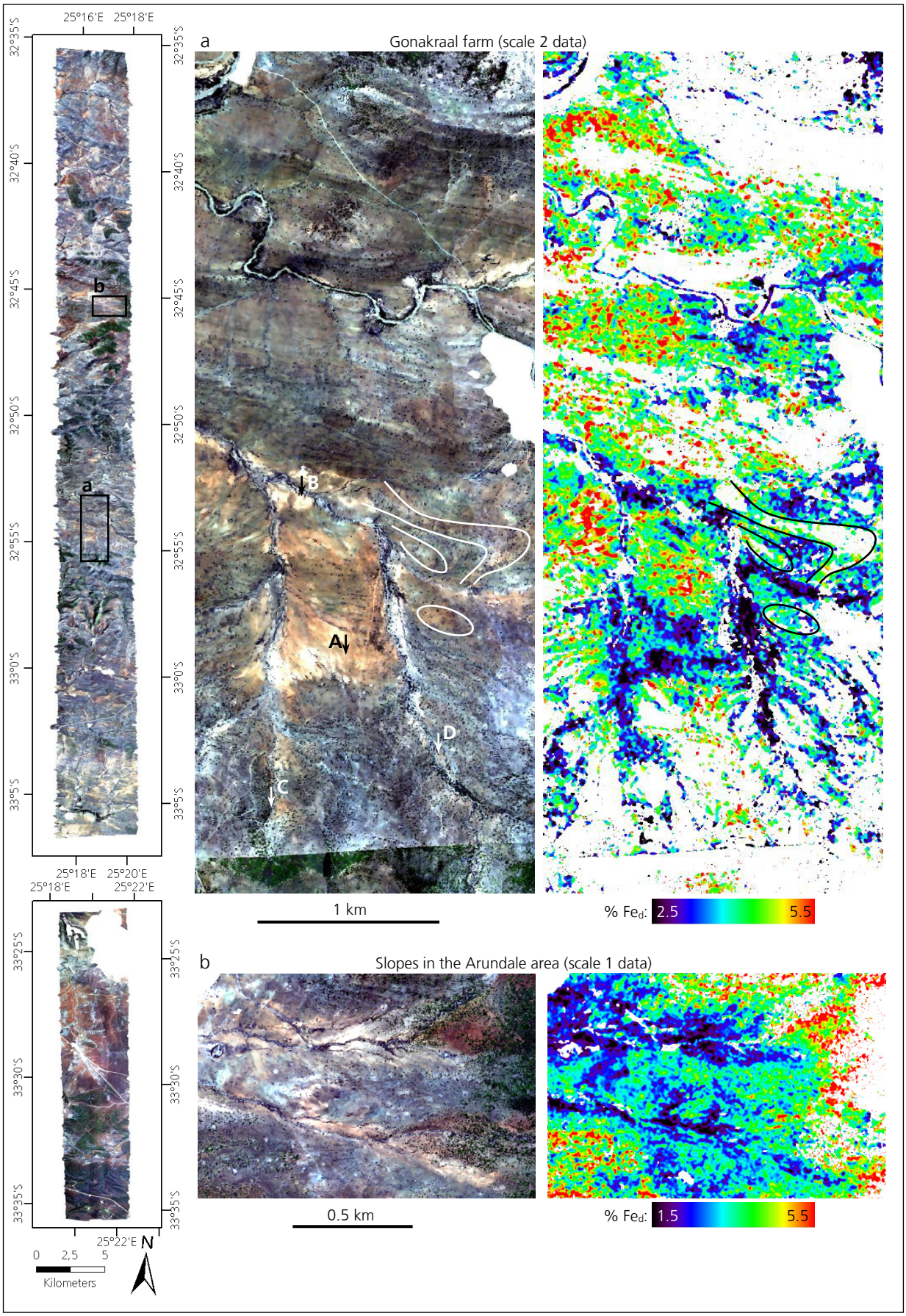


Figure 6-18: Local examples for predicted maps showing the spatial distribution of topsoil iron oxides contents for different scales (band 4 see section 5.3) - Part II. For areas masked white, no soil constituents can be predicted. True color composites with R: 640 nm, G: 549 nm, B: 456 nm.

Influences on iron oxides predictions

The presence of doleritic intrusions, e.g. in the Arundale area (figure 6-17c), increases the mapped iron oxides contents. This can be attributed to an increased oxidation caused by weathering of Fe-minerals in the bedrock. In figure 6-17c the outline of the main areas, where dolerites are exposed in this region are shown. They match very well with regions of high iron oxides concentrations. However, contents calculated for areas influenced by dolerites are biased, since the production mechanism and source of the iron oxides is significantly different from the one within the remaining area. Thus, the model is not calibrated for such an influence. Iron oxides predictions are impaired if a vegetation signal remains in the approximated soil signatures. In particular the chlorophyll absorption in the VIS range triggered by green vegetation overlaps with characteristic iron oxides absorption features and may induce biased iron oxides concentrations (see e.g. Richter 2010a).

Conclusions

Overall, the soils in the study area reveal differences in soil development by the spatial distribution of pedogenic iron oxides. Soils in the southern transect part are more developed. Hence, it remains open if this is a result of a longer or more intense soil development. By way of comparison, soils in the northern section reflect a lower soil development, because the characteristic intense red soil colors coupled with higher iron oxides contents are not observed over large areas. Subtle differences in iron oxides content result of weathering intensity, laminar erosion of soil material (river channels) and geologic parent material.

6.3.3 Clay

The low accuracies of the regression models developed for soil clay content do not allow a quantitative prediction of soil clay content (see discussion in section 6.1.2). Nevertheless, the regression model is applied for a qualitative assessment of the spatial distribution of soil clay, taking into account an increased error of the predicted clay contents. The calculated clay contents in the following are interpreted in relation to each other and rather than by means of their absolute values. Furthermore, the mapping of different clay mineral mixtures as an additional result of the spectral unmixing procedure is presented.

Spatial patterns in clay contents in the entire study area

Within the entire 320 km² of the study area, intermediate to low level clay contents are found on hill tops and in partially vegetated areas, such as the Kruizementfontein area or the hills surrounding the Arundale area. Clay contents are significantly higher on erosive slopes, fans and plains, partially vegetated or bare (e.g. fan in the northern section depicted in figure 6-11b, fans into the Arundale valley). Increased clay contents are as well predicted for some smaller roads and drainage channels.

A strong link exists between clay concentrations in soils and the underlying geology. Increased clay contents are expected to be found upon mudstones resulting of the weathering of the parent material, and thus, increased input of clay sized particles from the mudstones (e.g. figure 6-19b). In addition to this, locally increased clay concentrations may be attributed to the existence of physical crusts on the soil surfaces. When a reduced or removed vegetation layer exposes soil surfaces, the soils are highly vulnerable to form

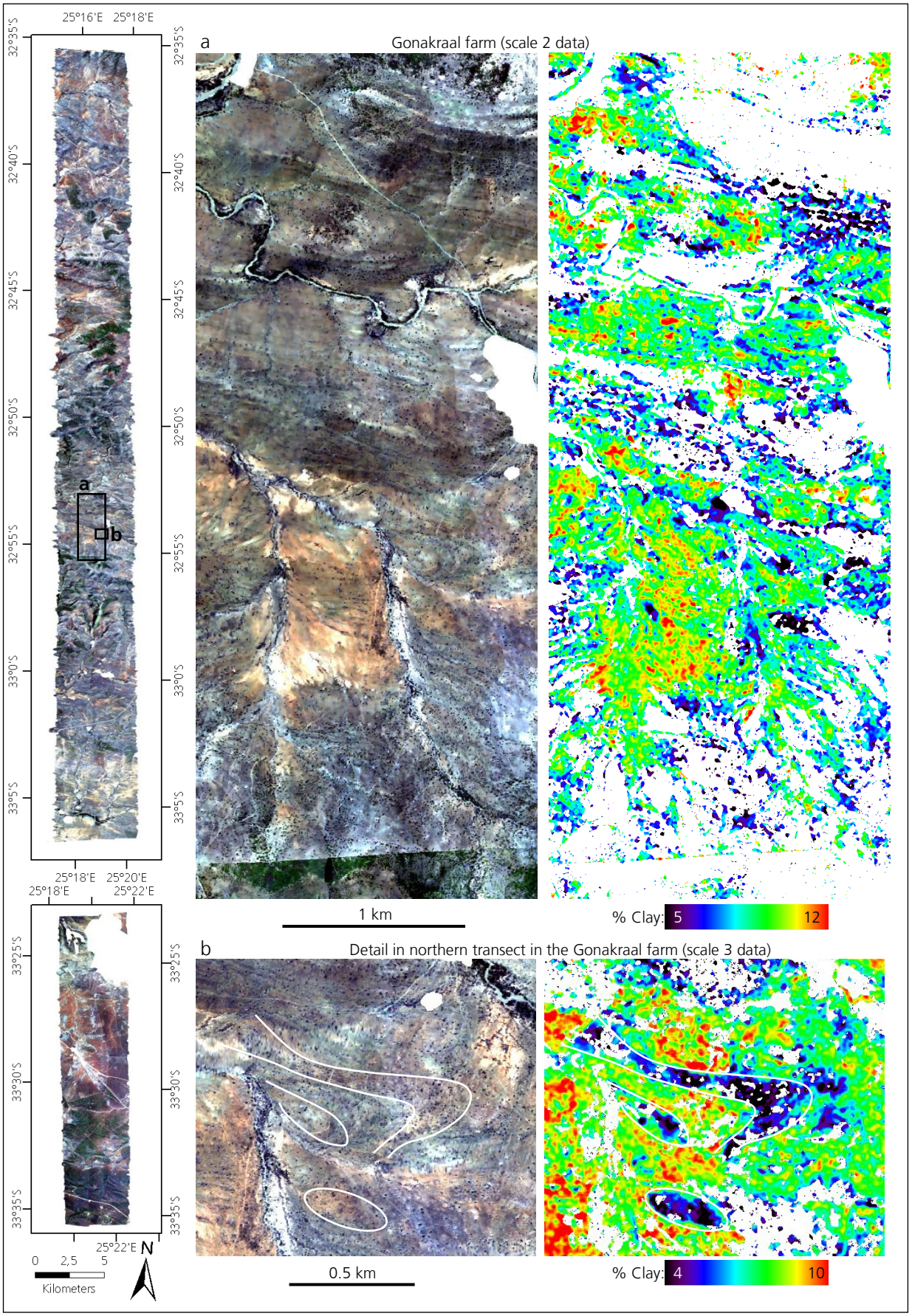


Figure 6-19: Local examples for predicted maps showing the spatial distribution of topsoil clay content (band 4 see section 5.3). Note that due to poor calibration accuracies of the developed regression models, the maps provide lower accuracy and are used only for qualitative interpretation. For areas masked white, no soil constituents can be predicted. True color composites with R: 640 nm, G: 549 nm, B: 456 nm.

crusts when rain drops impact on the unprotected surfaces (see section 3.1.2). A local increase of topsoil clay content in connection with bare soils can be used as an indicator for the presence of soil physical crusts as the crust formation process causes an enrichment of clay sized particles at the soil surface, assuming constant weathering conditions.

Examples from the Gonakraal area

Clay contents between 8 and 10 % are predicted for the soils in the Gonakraal area (figure 6-19a), with overall variances small. Clay contents above the average (between 10 to 12 % clay) are predicted for the bare plains. Considering the distribution of soil organic carbon and iron oxides, these concentrations cannot be explained by a progressive soil development promoting the enrichment of clay particles in the soils. The high concentrations are more likely attributed to the existence of soil physical crusts as was outlined above. Accordingly, intermediate to lower clay contents on the slopes rising up to the fence line do not indicate the existence of crusted soil surfaces, as a partial vegetation layer protects the soil surfaces from crusting and intense erosion. The link of clay contents and geology can also be observed at the sandstone-mudstone fold (figure 6-19b), where predicted clay contents are significantly higher above mudstones than sandstones. This variation is more apparent in the maps of scale 3, since in the scale 2 data the sandstone areas are mostly excluded within the approximation of soil residues.

Mapping of clay mineral alteration products

Based on the spectral characteristics of the selected endmembers and the spectral unmixing ground cover fractions, two different groups of clay mineral mixtures can be differentiated. Most of the bare soil endmembers show a 2200 nm clay absorption only comprising the HyMap band at 2208 nm. However, some EMs show a broad 2200 nm absorption that is comprised in the HyMap bands at 2208 and 2226 nm. For instance for the Gonakraal subset 2 of the 8 bare soil EMs show such a broad clay absorption. Additionally, the broad clay absorption at 2200 nm seems to be linked to a deeper absorption at 2340 nm (the described correlation is not revised sufficiently, so that besides the possibility presented in the following other reasons for the increased 2340 nm absorption also arises). Clark et al. (1990) observed a similar shift in the band position of montmorillonites and based on chemical composition attributed it to an increased Calcium content.

The dominant mixture of clay minerals in the study area is made of illite and montmorillonite (Mills 2009). According to Clark et al. (1993, and see references in table 5-3), illite and montmorillonite produce an absorption at about 2200 nm but no significant absorption at about 2340/2350 nm. The doublet feature typical for kaolinite appears only subsidiary in some field spectra but was not observed frequently in the HyMap spectra. In the study area, the broad absorption around 2200 nm might be caused by a presence of muscovite in the clay mineral mixture. In comparison to illite and montmorillonite, muscovite has a broader absorption between 2195 and 2225 nm and just like illite has an additional absorption band at 2340/2350 nm, as opposed to montmorillonite (Clark et al. 1990). Therefore, a clay mineral mixture dominated by illite and muscovite is imaginable, following on the one hand the broad 2200 nm absorption and on the other hand the observed joint occurrence of this absorption with an increased absorption at about 2340 nm. The occurrence of muscovite as primary weathering product suggests the presence of a source of these minerals in combination with an early stage of soil development. In contrast, the mixture of montmorillonites and illites, comprising the predominant clay mineral mixture within the study area, produces a more distinct absorption and a diminished absorption at 2340/2350 nm. Montmorillonite, as

one mineral of the smectite group, is the more advanced weathering product of both muscovites and also illites (Karathanasis 2006, Schachtschabel et al. 2002) and as such indicates intermediate weathering.

Figure 6-20 illustrates the ground cover fractions of the EMs with distinct and broad absorption for the Gonakraal farm area. Soils with distinct clay absorption dominate the entire area, especially where the soil fractional cover is high, such as on the bare plains, indicating well advanced weathering. Soils with a broad clay absorption and resulting intermediate weathering products are predominant in the topographic lows, especially the erosion channels and drainage systems, though there is not very much soil matrix present, as field photos showed (see section 6.3.1). Thus, it is likely that sediment is eroded before it reaches a higher alteration stage. In the sandstone-mudstone fold, the muscovite containing soils dominate the soil fraction upon the mudstone parts, though both groups occur. The presence of muscovite can here be attributed to the mudstones as source material, and less likely to a less advanced soil development. On the contrary on the sandstone areas, where clay contents are generally lower (see figure 6-19), illite-montmorillonite soils occur.

Conclusions

The maps of predicted soil clay concentrations reveal spatial patterns that can be linked to erosion processes and geologic parent material. Nevertheless, the maps are only suitable for qualitative evaluation because of the poor performance of the regression relationship established for soil clay content, resulting from the

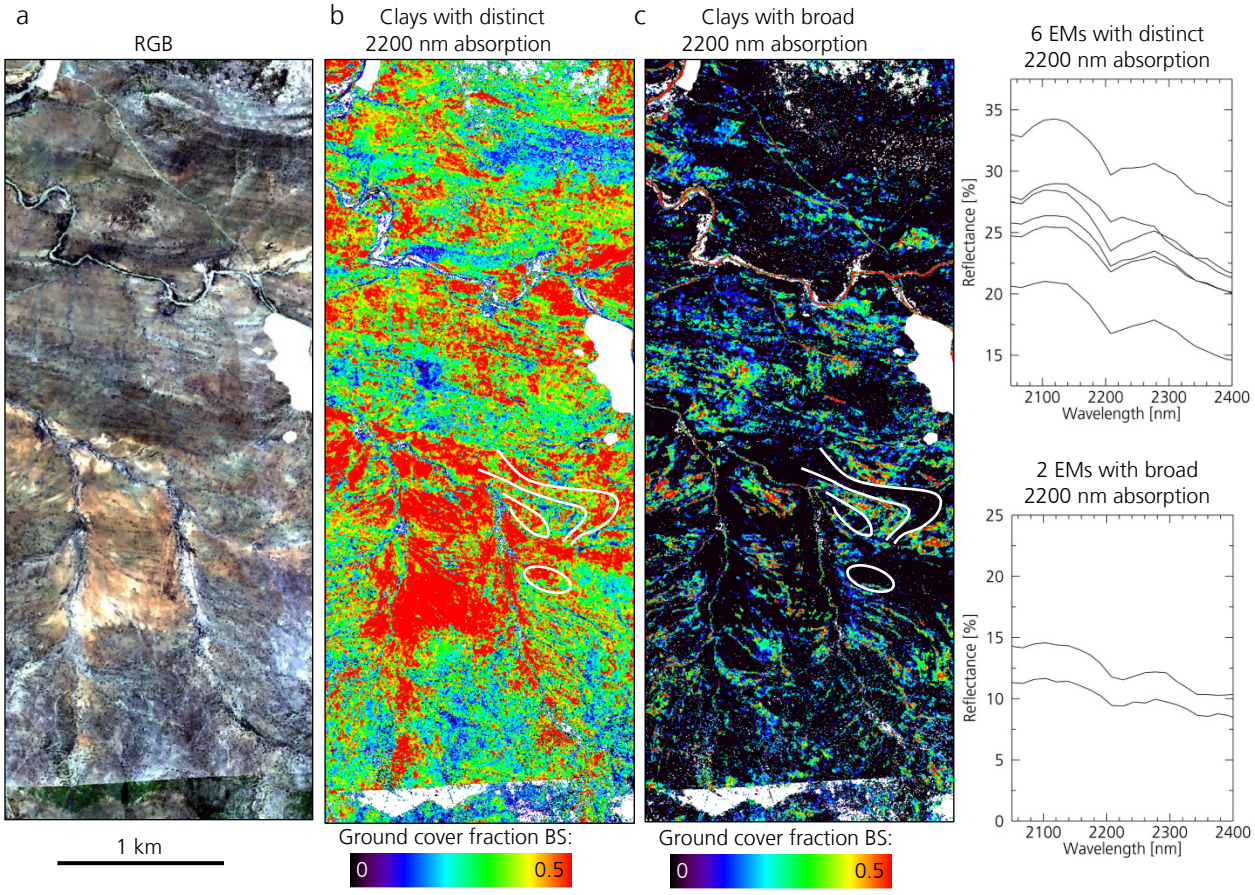


Figure 6-20: Spatial distribution of soils with distinct (b) and broad (c) clay mineral absorption at 2200 nm. It is shown by the cover fractions determined in spectral unmixing for the Gonakraal farm. White areas indicate artificial areas and pixels, where the determined bare soil fraction is 0. True color composite with R: 640 nm, G: 549 nm, B: 456 nm.

large variance in characteristics of the study area (e.g. in regard to geology, surface characteristics, soil crusting). Further spatial information can be derived from different clay related spectral features within the endmember spectra, by which the spatial distribution of clays in different weathering stages can be visualized.

6.3.4 Confidence assessment of soil constituents predictions

The confidence layer gives an assessment of the reliability of the predicted soil constituents. It is derived by combining all input layers contributing in the workflow to the final maps of soil constituents like spectral unmixing RMSE, approximated soil signatures and the range of predicted soil organic carbon and iron oxides contents (see section 5.3.1). The predicted maps of clay content were not included in this calculation.

A good confidence is flagged within the entire dataset comprising 320 km^2, for 78 % of the pixels, for which soil constituents are predicted. This indicates that no factors could be identified that could possibly limit the probability of a reliable prediction of soil constituents from hyperspectral imagery. The good confidence is mainly flagged for areas where soil is the dominant land cover fraction, for instance on the bare plains within the Gonakraal area (figure 6-21a). For 20 % of the predicted pixels minor influencing factors are identified, such as a vegetation coverage close to the limit up to which a soil signature can be approximated, an increased unmixing RMSE, etc. These pixels flagged with medium confidence are often found where vegetation coverage is present and spectral mixture within pixels is apparent. A reduced confidence is shown for only 2 % of the pixels. These are pixels, for which a soil constituent is calculated but there exist a number of factors influencing the soil constituents prediction. If thresholds (e.g. within the approximation of the soil signatures), would diminish (see section 6.2.2), the pixels identified here would have to be the first to be excluded.

In the Gonakraal area 76 % of the pixels show a good confidence, with 22 % showing medium and 2 % of the pixels reduced confidence. Medium confidence is identified in regions where vegetation coverage increases (e.g. on the slopes to the south) but also some areas in the northern section. The reason for the patterns of the three reliability classes can be identified based on the individual layers that contribute to the 3-class confidence layer. An example for the Gonakraal subset is given in the appendix (Figure E-20). The highest variance here is found in the confidence layer of the approximation of soil residues, which is also the reason for the medium confidence level in the northern areas. On the flat plains, predicted carbon contents that are very close to the lower calibration limit locally reduce the assessed confidence. Small groups of pixels showing a reduced confidence are spread nearly over the entire subset. They occur where several factors arise that the final predicted contents may be biased.

Conclusions

The derived confidence layer gives an assessment of the expected reliability of predicted soil organic carbon and iron oxides. This assessment of the prediction accuracy based on factors within the workflow that may influence the predictions replaces a map of modeling errors, that cannot be presented because no spatially continuous ground reference information is available.

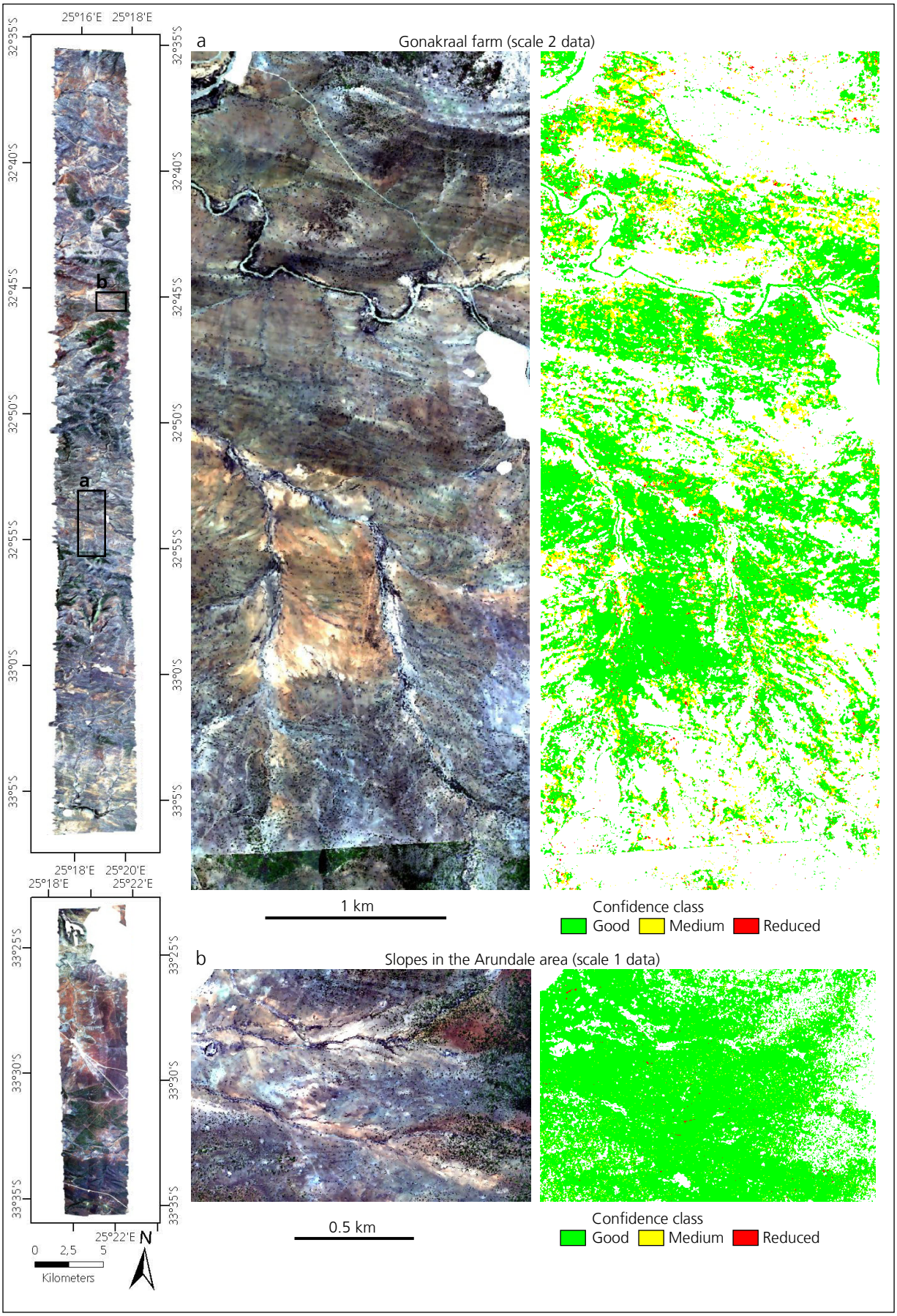


Figure 6-21: Maps giving a confidence assessment for the predicted soil constituent quantities established based on factors within the workflow that may influence the predictions. For areas masked white, no soil constituents can be predicted. True color composite with R: 640 nm, G: 549 nm, B: 456 nm.

6.4 Performance evaluation

The performance of the proposed multi-stage methodology is evaluated based on the agreement of predicted and measured soil constituents' contents and the consistency of the method for data of different size and characteristics. The benefit of the individual steps within the method is evaluated based on a comparison of their performance to existing approaches. Finally, the most profitable step is identified.

6.4.1 Regional validation accuracy

The in-situ measured contents are directly compared to the constituent contents predicted for the ground truth sites based on the HyMap imagery. The selected method for this comparison is given in section 5.3.2. No additional interpolation with the adjacent pixels is applied therein. This comparison of model derived with in-situ measured soil constituents provides a regional validation based on the hyperspectral imagery of the South African study area. The performance determined in this assessment is made up as a compound of all steps within the workflow. It not only rates the performance of the regression model itself, which was determined in cross-validation.

Regional validation of predicted soil constituents

From the initial 163 sites, soil samples were collected from, soil signatures are approximated and soil constituents are predicted for about 70 sites. From these 70 sites, a medium or reduced confidence is flagged for about 23 sites. These pixels are ranked as "influenced" and are only of restricted use for validation purposes (see section 5.3.2 for details on this classification). According to this, about 47 valid sites remain for comparison of measured and modeled soil constituents. Although this is a low number compared to the originally 163 sampled sites, the remaining samples still provide a good basis for validation of the multi-stage methodology proposed in this study. Table F-23 in the appendix gives an overview of the samples included for validation of each subset and of the entire dataset.

Figure 6-22a shows scatter plots of measured versus calculated contents for each soil constituent. Correlation coefficients (R^2) are low for all parameters. Significant spreads are observed for soil organic carbon and iron oxides. For clay content, a systematic overestimation of low and underestimation of high concentrations is observed, in the same way this is observed in the calibration scatter plots of the developed clay content regression model (see figure 6-1a). Corresponding scatter plots including both the 47 valid and the 23 sample sites classified as "influenced" are shown in Figure F-21 in the appendix. They show the same trends observed here, though point clouds are denser and regression coefficients lower. Modeling residues as the deviation between in-situ measured and modeled constituent contents are shown in figure 6-22b. For soil organic carbon and also iron oxides they are close to normally distributed, indicating that the proposed methodology does not introduce systematic errors. However, for soil organic carbon the average of the residues is shifted towards negative values since predicted are higher than observed C_{org} concentrations. Only the residues of the clay prediction show large deviations over the entire range, being the result of the poor model accuracy and thus significance. The same trends for the three soil constituents are observed in the residues histograms of all valid and influenced sample sites (Figure F-21 in appendix). For all three soil constituents, no correlation between modeling residues and the vegetation coverage can be detected. Scatter plots showing the deviation of measured and calculated contents versus the unmixing ground cover fraction of vegetation are given in appendix F.

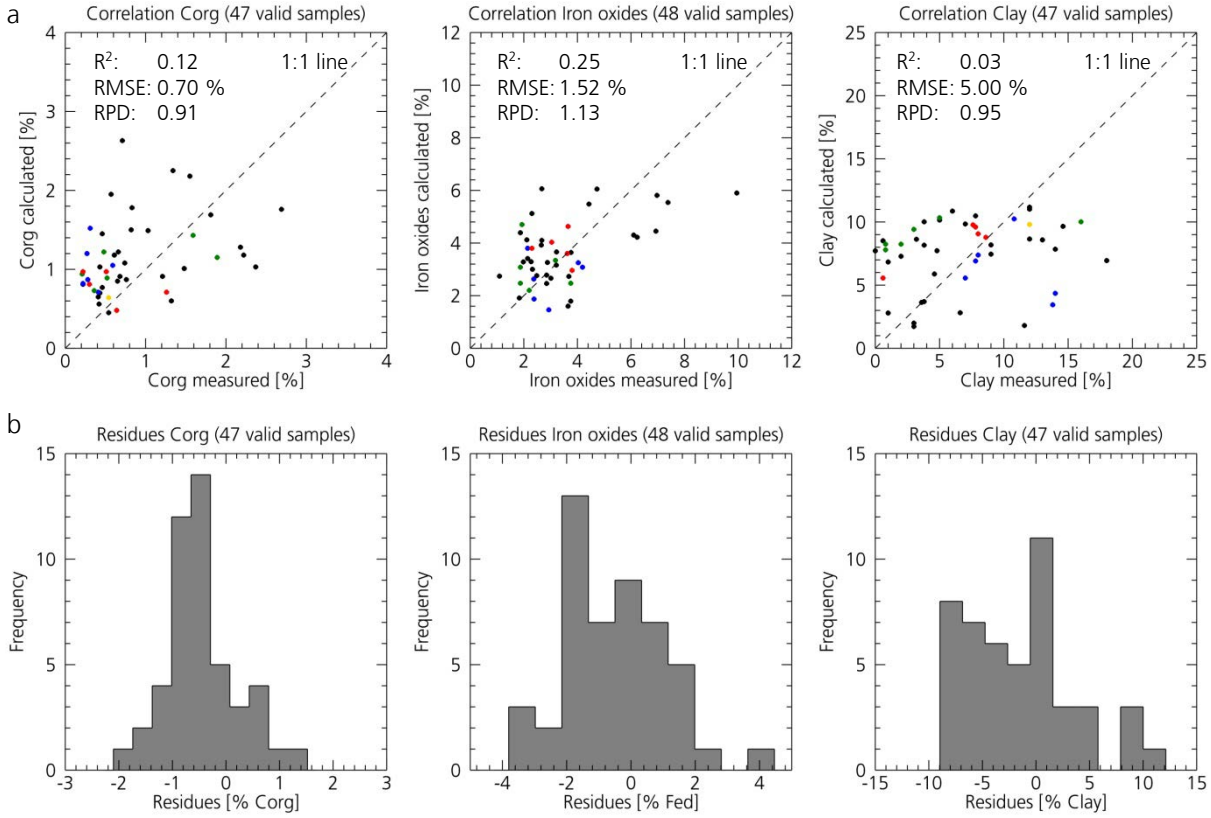


Figure 6-22: Scatter plots for in-situ measured versus predicted soil constituents as regional validation of the multi-stage methodology (a) and histograms of modeling residues (b).
Color coding in a: • Validation sites outside the scale 2 subsets, sites within • Gonakraal, • Arundale, • Wintersfontein, • Kruizementfontein and • Stenbokvlakte area.

Conclusions and placement

The comparison of model derived with measured soil constituents provides a regional validation of the proposed methodology based on the hyperspectral imagery of the South African study area. It serves as an assessment of all steps within the workflow from the initial regression model over the processing of the image spectra and approximation of the contained soil signatures towards the prediction of soil constituents. However, correlation coefficients of measured and modeled soil constituents are low and indicate remaining inaccuracies within the workflow. The systematic over-and underestimation of clay contents supports the rating of the clay predictions as only of limited use for the reasons given in section 6.1.2.

A similar approach for the prediction of soil organic carbon contents for agricultural maize fields with partial vegetation coverage in Belgium was applied by Bartholomeus et al. (2011). Using PLS regression techniques and a similar approach to approximate the per pixel soil signature, they predicted C_{org} contents based on AHS-160 data (spatial resolution of 2.6 m) and obtained an RPD_{Val} of 1.50. However, the homogenous mixture of only two endmembers within the agricultural areas and the availability of endmember spectra measured in the field made the spectral mixtures well identifiable and resolvable. Besides, Gomez et al. (2008b) used Hyperion satellite data for soil organic carbon predictions in Australian soils and achieved prediction models with an R^2 of 0.49, RPD of 1.42 and RMSE of 0.8 % (results of model cross-validation using Hyperion spectra). Low correlations are due to high noise present in Hyperion spectra and spectral mixing within pixels (30 m spatial resolution of Hyperion). These two examples show among others that the

accuracy of soil constituents predictions declines with the intensity of spectral mixture within pixels. Although the pixel size of the HyMap imagery used in this study with 3.3 m is small, the small-scale mixtures of surface materials within the study area and the complexity of mixtures of numerous different surface materials have a significant influence on prediction accuracy, although spectral mixtures are resolved and soil signatures approximated for each soil dominated image pixel.

6.4.2 Dependency of determination of soil constituents on scaling

To investigate differences in performance when the compiled methodology is applied to hyperspectral imagery of different size and characteristics, three spatial scales are introduced in section 5.4, coming from the data of the entire 320 km² imagery (scale 3), to subsets of 10 to 40 km² (scale 2) and finally the smallest subsets of about 2 km² (scale 1). Pre-processing and the steps within the further applied workflow are identical for all the data. Differences in the maps of predicted soil constituents are exclusively caused by differences within the performances of the individual methods that are caused by image characteristics and size such as endmember selection and spectral unmixing. It is expected that accuracies are higher for the smaller imagery (scale 1 and 2), for instance because the endmembers are better adapted to local mixtures and spectral unmixing quality is higher compared to the scale 3 processing, where one endmember set is used for each entire part of the study area. The following discussion focuses on differences in the predicted soil constituents resulting from the processing of each scale. Some distinctive features arising within the processing of the data of the three scales regarding the derivation of endmembers, performance of spectral unmixing and approximation of soil signatures, are evaluated in appendix F.2.

For the following analysis, the maps of soil constituents are compared at each spatial scale. Therefore, from the maps of scale 3 comprising the entire study area subsets matching the extension of the scale 2 imagery (5 medium sized subsets) and scale 1 imagery (6 small subsets) are clipped. In the same way from the maps resulting of the processing within scale 2, the extension of the five scale 1 subsets are clipped. In the following, for the medium-sized subsets the maps processed in scale 2 and 3 are compared and for the small subsets the maps processed in scales 1, 2 and 3 are compared.

First of all, the number of pixels for which soil constituents are predicted as fraction of the total number of pixels within the imagery differs within the scales (table 6-7). Within scale 2 and 3, soil constituents can be predicted for about 50 to 60 % of the pixels, depending on the individual subsets. Only within scale 1, soil constituents can be predicted for significantly more pixels (i.e. about 70 %). This is achieved because of

Table 6-7: Percentage of pixels for which soil constituents were predicted.

		Percentage of pixels for which soil constituents were predicted[1]
Medium subsets	All scale 2 subsets	50 %
	Scale 3 data with the extension of the medium sized subsets	46 %
Small subsets	All scale 1 subsets	70 %
	Scale 2 data with the extension of the small subsets	49 %
	Scale 3 data with the extension of the small subsets	60 %

[1]The number of pixels slightly differs for each soil constituent due to differences in pixels, that were removed because of unreliable values or being interpolated. Differences in all cases are below 1 %. Band 4 of the predicted soil constituents' maps was used to derive these numbers.

two factors. First, the quality of spectral unmixing is higher for smaller subsets because the endmembers are adapted to local mixtures, and thus, the dominant land cover classes can be well identified. This provides the basis for the second factor, which is the approximation of the soil residues. Well resolved spectral mixtures and identified land cover types enable a better adaptation of the thresholds used within the approximation of soil residues. The thresholds control the exclusion of unsuitable pixels where no soil signature can be approximated. Due to the small size of each scale 1 subset, the thresholds can very well be adapted, which supports the discrimination between suitable and unsuitable pixels that are to be excluded.

Comparison of soil constituents determined for scales 1, 2 and 3

Figure 6-24 shows examples for soil organic carbon and iron oxides contents predicted by applying the multi-stage methodology first to the entire imagery of the study area (scale 3) and second when applied only to the imagery of a medium sized subset (scale 2). The maps of predicted soil constituents for these two scales are compared for the extension of the Gonakraal area. In general, the same patterns for the spatial distribution of high and low contents are observed. Soil organic carbon contents are in an average about 0.25 %higher in scale 2 than in scale 3 (average of all scale 2 subsets, see table 6-8). Predicted iron oxides contents are well in the same range for the two processing scales, though slightly lower for the scale 3 processing (about 0.14 % higher Fe_d as average of all subsets, see table 6-8). Although predicted contents show minimal variation as was shown, the deviations are below the RMSE of the regression models determined in cross-validation, which is 0.46 % for C_{org} and 0.93 % for iron oxides (refer to Table 6-1 in section 6.1.1). Skews of histograms (e.g. a positive skew for soil organic carbon, see figure 6-24) do not change over the different scales. Figure 6-23 shows scatter plots of these data for all three soil constituents. The predicted contents match well, though the plots also show existing deviations.

Figure 6-25 and figure 6-26 give examples for soil constituents maps predicted for all three levels of scaling and the extension of the subsets in scale 1. Also here observed patterns resemble each other and the main structures are well identifiable in all maps. Subtle differences exist in local details and in the absolute values of the predicted contents (see table 6-8).

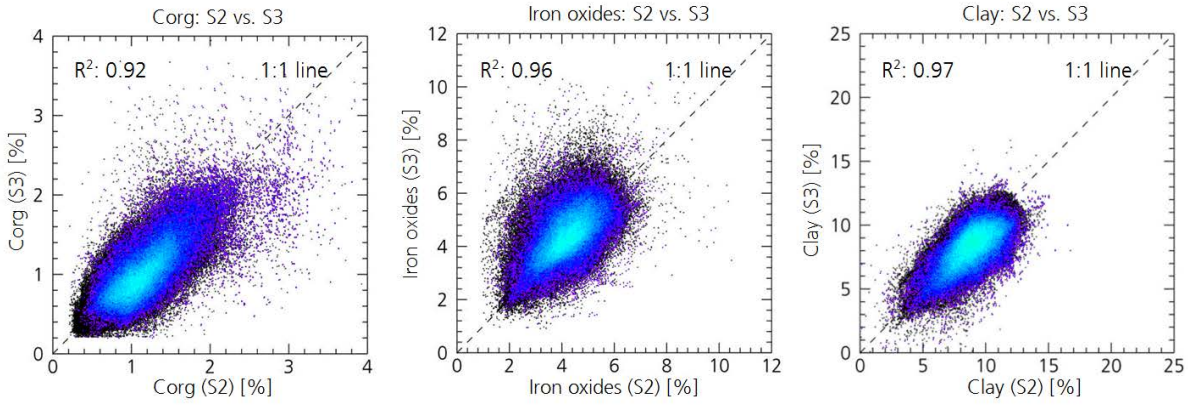


Figure 6-23: Comparison of predicted soil constituents calculated in scale 2 and scale 3 shown in scatter plots for the Gonakraal area.

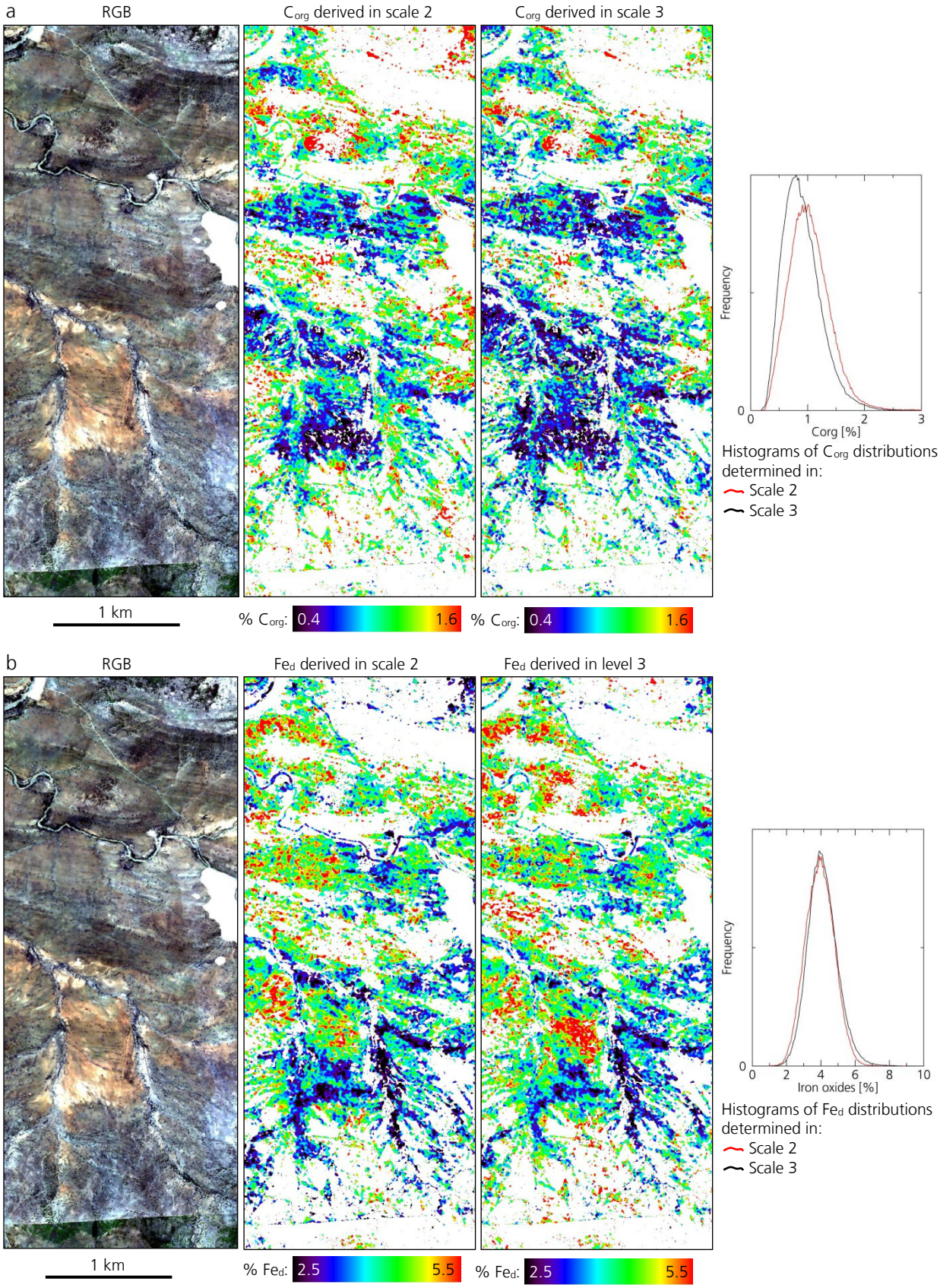


Figure 6-24: Comparison of topsoil organic carbon (a) and iron oxides contents (b) derived in scales 2 and 3 for the Gonakraal area as one example of a medium sized subset. Only pixels for which iron oxides are predicted in both scales are displayed and considered for calculation of histograms. For areas masked white, no soil constituents can be predicted. True color image with R: 640 nm, G: 549 nm, B: 456 nm.

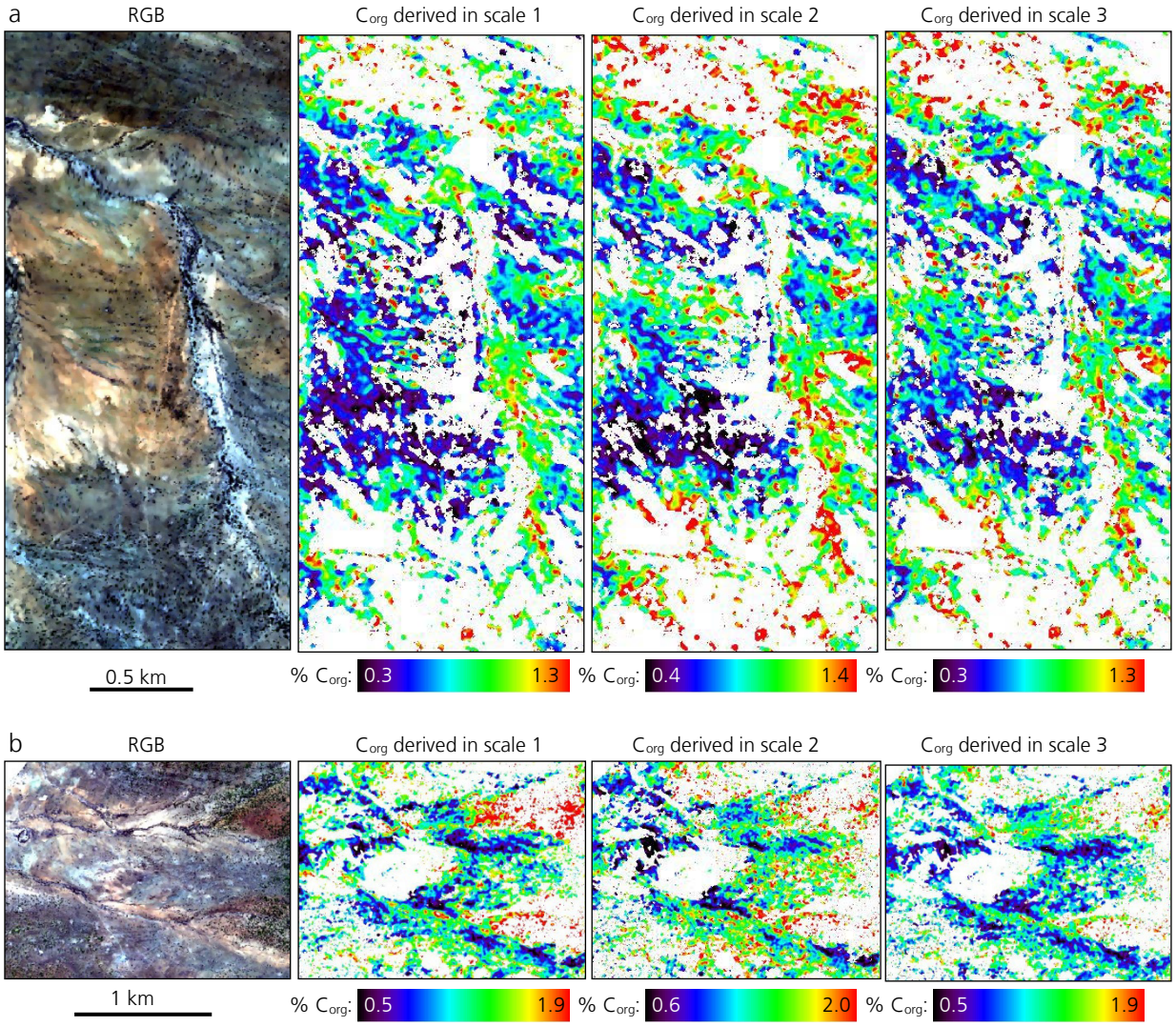


Figure 6-25: Comparison of topsoil organic carbon derived in scales 1, 2 and 3 for small sized subsets of a bare plain in the Gonakraal area (a) and slopes in the Arundale area (b, subset slightly cut in east and west). Only pixels for which C_{org} is predicted in all three scales are displayed. For areas masked white, no soil constituents can be predicted. True color image ith R: 640 nm, G: 549 nm, B: 456 nm.

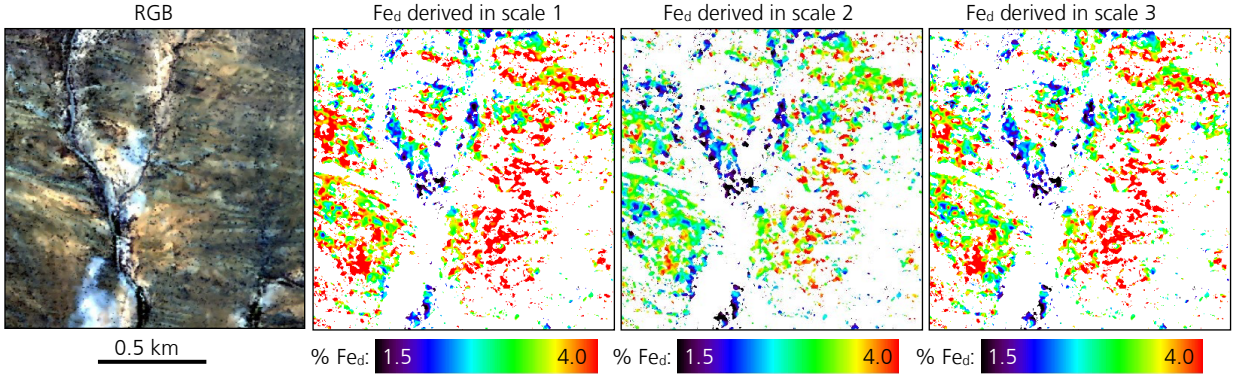


Figure 6-26: Comparison of topsoil iron oxides derived in scales 1, 2 and 3 for the small sized subset of a river in the Wintersfontein area. Only pixels for which Fe_d is predicted in all three scales are displayed. For areas masked white, no soil constituents can be predicted. True color image with R: 640 nm, G: 549 nm, B: 456 nm.

General trends in the statistics of predicted soil constituents for the three levels of scaling

A comparison of the statistics of the contents predicted within each scale reveals general trends (table 6-8). For iron oxides and clay content, there is a constant trend that contents predicted in the higher data scales (larger imagery) are higher. This trend is not observed for soil organic carbon, where predicted contents are mostly lower the larger the imagery (higher scales), though the trend is not altogether constant. The standard deviation of the contents predicted in the lower scale's processing is usually higher than of scale 3. This trend is more obvious for soil organic carbon and clay content, though in reverse for iron oxides. Higher standard deviations mean that the data of the smaller subsets comprise a variability in contents that is not fully covered by the scale 3 data. The reason may be more inaccurate predictions removing some of the present variance.

Validation accuracies over the three levels of scaling

Trends within the prediction of soil constituents for the three levels of scaling can also be visualized by comparing the prediction accuracies of the in-situ measured versus the soil constituents contents predicted within every scale. The scatter plots for ground truth samples included in the medium subsets which are predicted in the scales 2 and 3 are shown in Figure F-24 in the appendix and for the samples within the small subsets that are predicted in all three scales are shown in Figure F-25. Correlation coefficients (R^2) usually are higher in the lower scale subsets than in scale 3. For clay this trend is not observed, resulting from the systematic overestimation of low and underestimation of low contents due to the poor quality clay regression model (see section 6.1.2) leading to variations in the adjustment of the regression line, and thus, correlation. However, the correlations are only indicators for a better performance of the lower scale processing but not conclusively due to the low number of samples the correlations are built upon (17 samples within medium and 7 samples within small subsets).

Table 6-8: Statistics of soil constituents predicted in the different scales for the extension of the scale 2 and scale 1 subsets. For calculation of the statistics only pixels were included for which in the corresponding scales soil constituents were predicted for. The values present an average of all subsets with the extension of scale 1 and 2.

		C_{org}	Iron oxides	Clay
Medium subsets	Statistics of all **scale 2** subsets	Mean: 1.47 Stdev: 0.58 Min: 0.21 Max: 5.50	Mean: 3.53 Stdev: 1.05 Min: 0.90 Max: 10.62	Mean: 7.06 Stdev: 2.01 Min: 0.00 Max: 17.42
	Statistics of **scale 3** data with the extension of the medium sized subsets	Mean: 1.22 Stdev: 0.44 Min: 0.21 Max: 5.45	Mean: 3.67 Stdev: 1.15 Min: 0.90 Max: 10.62	Mean: 7.25 Stdev: 1.98 Min: 0.00 Max: 17.90
Small subsets	Statistics of all **scale 1** subsets	Mean: 1.08 Stdev: 0.42 Min: 0.21 Max: 4.35	Mean: 3.57 Stdev: 0.86 Min: 0.91 Max: 10.44	Mean: 7.47 Stdev: 1.90 Min: 0.07 Max: 14.21
	Statistics of **scale 2** data with the extension of the small subsets	Mean: 1.17 Stdev: 0.46 Min: 0.21 Max: 5.25	Mean: 3.62 Stdev: 0.99 Min: 0.90 Max: 10.58	Mean: 8.17 Stdev: 1.76 Min: 0.03 Max: 16.74
	Statistics of **scale 3** data with the extension of the small subsets	Mean: 0.99 Stdev: 0.34 Min: 0.21 Max: 4.76	Mean: 3.85 Stdev: 1.03 Min: 0.91 Max: 10.60	Mean: 8.43 Stdev: 1.39 Min: 0.02 Max: 16.09

Confidence assessment for the three scales

While for the entire dataset 78 % of the pixels are flagged with a good accuracy, this percentage is higher and usually above 90 % when soil constituents are predicted for only medium and small sized subsets (see Table F-25 in appendix). In these smaller subsets usually below 1 % of the pixels have a reduced accuracy as expected. In the same way, the confidence assessment of the small subsets is lower and less pixels are flagged with a good confidence when they were predicted in the processing of scale 2 and 3 (first lines in Table F-25). This clearly shows that the workflow produces more reliable results when applied to local subsets compared to the entire imagery. When comparing the expected reliability for local subsets of different characteristics, it is rated that within areas having for instance an increased vegetation coverage or higher topography such as the Stenbokvlakte and Kruizementfontein areas, the confidence of the predicted soil constituents is usually lower compared to areas without such factors.

Conclusions

The maps of predicted soil constituents are in general consistent over the three scales of scaling the methodology was applied to. Subtle differences in predicted contents occur, though they are below the cross-validated RMSE of the regression models. Trends exist, that in the higher scale processing higher contents are predicted and the standard deviation decreases. In particular, the trend observed for the standard deviation of predicted contents indicates that the present variability in soil constituents is better covered when soil constituents are predicted for small and medium subsets. Nevertheless, these trends are not altogether constant but also show deviations. The comparison of validation accuracies using samples predicted in multiple scales indicates a better performance when smaller subsets are processed. This is confirmed by the assessed confidence of predicted soil constituents which showed a good confidence for more pixels in the lower scales compared to scale 3. These aspects indicate that soil constituents predictions using the compiled multi-stage method are more accurate when applied to medium and small sized imagery (up to about 40 km^2). In contrast, the analyzed factors and confidence indicator suggest a lower performance when the methodology is applied to large datasets of several hundred square kilometers. This indicates that the quality of the spectral unmixing is the factor limiting the prediction accuracy of soil constituents.

6.4.3 Placement of the approach

Within the proposed multi-stage methodology, the individual steps are rated with different accuracy and benefit. The feature-based regression approach is shown to provide similar predictability to PLS regression techniques as the commonly applied method. It is shown that the good quality soil organic carbon prediction models allow a reliable quantification based on hyperspectral airborne data collected of the study area. This is also shown for the medium quality iron oxides predictions models although they reveal drawbacks in model validation. Spatial patterns of this parameters can be linked to soil development and degradation phenomena. Because the prediction models for soil clay content do not reach significant accuracies, they are not suited to be applied for a quantification over large areas. However, it is shown that they can be used for the prediction of relative concentrations to discriminate areas of high and low clay concentrations. Compared to other studies, model accuracies seem to be slightly reduced by the variant characteristics present within the large study area (“global calibrations”) and the small variability in

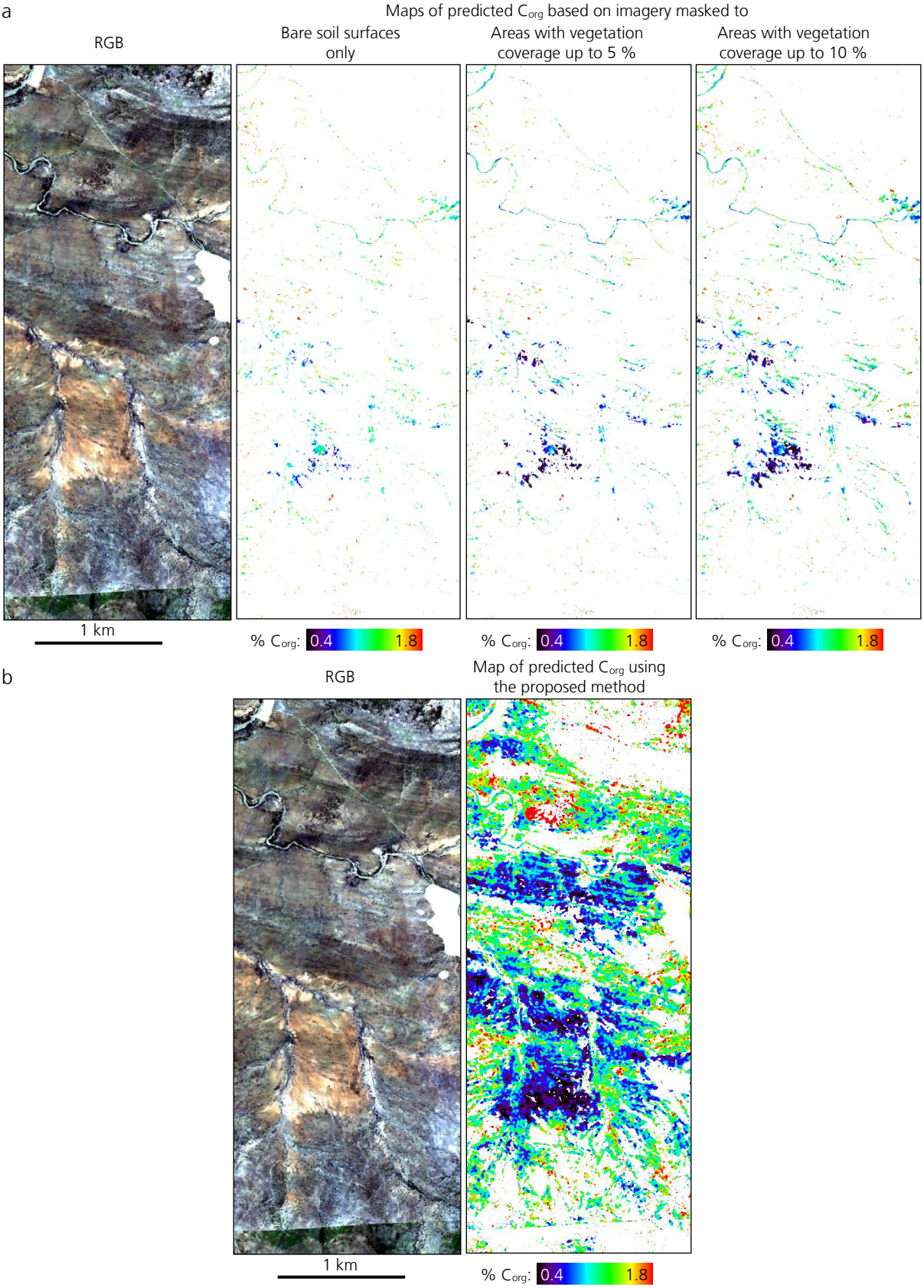


Figure 6-27: Comparison of the spatial extent of topsoil organic carbon contents predicted using the feature-based regression approach but with different input data for the Gonakraal area (scale 2 data). a: Using image data where pixels with a certain amount of vegetation (sum of green and dry vegetation) are masked without residual unmixing, b: C_{org} map using the proposed methodology. Maps presented in part a are interpolated and filtered by 3 x 3 filter kernel, while map presented in part b is interpolated and filtered by 5 x 5 filter kernel. For areas masked white, no soil constituents can be predicted.

measured contents especially for soil organic carbon (average content of 1.21 % C_{org}). Both factors have been discussed earlier as lowering model accuracies (e.g. Stevens et al. 2010, Udelhoven et al. 2003). The presented feature-based regression approach thus provides an alternative to statistical PLS techniques, providing advantages particularly in terms of its physical significance. It is also expected that the applied approach benefits in terms of transferability. However, a proof of this cannot be realized in this study but will be in the focus of further investigations.

The greatest challenge, but also the largest benefit of the methodology, is provided within the approximation of pure soil signatures from mixed image spectra. This is done prior to the prediction of soil constituents using Imaging Spectroscopy. So far these were often limited in pre-processing to bare soil areas. For example, using a threshold in maximum vegetation coverage either determined in spectral unmixing (e.g. Richter 2010a, Schwanghart and Jarmer 2011) or by using a vegetation index (e.g. Stevens et al. 2008), or using classification algorithms (e.g. Stevens et al. 2010). Figure 6-27a shows an example from the study area, where soil constituents are predicted only for soil dominated image pixels, having between 0 % and 10 % vegetation coverage. Predictions then comprise only between 2 (no vegetation allowed in pixels) and 4 % of the image pixels of the sample area (10 % vegetation coverage within pixels allowed). They cover only local regions such as the prominent bare soil areas. Little or no spatially continuous information can be provided. In addition, it should be considered that in particular soil organic carbon and iron oxides predictions from pixel spectra including any amounts of vegetation may be biased by this. Prediction errors increase with the permitted amount of vegetation included in the spectral signature.

Based on the above needs, an approach for removing the spectral influence of vegetation and approximating each pixel's soil signature is proposed in this study. This step within the workflow provides the basis for an successful and unbiased prediction of soil constituents from hyperspectral imagery. The approach is shown to work well in most of the cases. Only single spectra showing artifacts remain of this procedure, which may influence the prediction of soil constituents. The success and significance of soil constituent predictions from partially vegetated areas depend on the quality of the unmixing and the quality of the approximated soil residual spectra. Similar conclusions were drawn by Bartholomeus et al. (2011), one of few studies where a partially vegetation coverage was considered and resolved. They predicted soil organic carbon in agricultural fields where the spectral mixture between soil and vegetation is limited to one vegetation type, and thus, is well characterizable. Yet in this environment the importance of spectral mixtures was apparent.

Using the proposed methodology of resolving spectral mixtures and approximating soil spectral signatures before a prediction of soil constituents, the amount of pixels within our sample area for which soil constituents can be predicted could be increased to 52 % (figure 6-27b). Thus, the number of pixels for which soil constituents are predicted is 12 times as high as when limiting image information to pixels containing up to 10 % of vegetation and about 22 times as high as when allowing only bare soil pixels. Predicted soil constituents are further more accurate because the analyzed soil signatures are not biased by an influence of vegetation. This emphasizes the benefit of the soil signature approximation in areas with very inhomogeneous land cover as found in the South African study area.

Limitations: Evaluation of impact factors

The application of the proposed methodology and its transfer to other geographic regions is limited by a number of factors. Some of them represent general boundary conditions for model development, but others can be minimized by adequate treatment and pre-processing.

1. Models always represent a simplification of the real world. The developed regression models for each soil parameter describe the complex relationship between the chemical constituents and its spectral properties with few variables. As such, the models are not expected to characterize the processes in its entirety, but to describe them as accurate and robust as possible.

2. Fundamental basis for model development is that soil organic carbon, iron oxides and clay affect soil reflectance in a diagnostic way, which is described by a linear relationship following the Beer's Law model. For all three parameters this was extensively reported in previous studies (e.g. Ben-Dor et al. 2008, Clark 1999, Viscarra Rossel and Behrens 2010).

3. Other soil constituents influence the development of a prediction model for a single soil constituent. Although this is minimized by the application of several diagnostic spectral features and a number of their properties for each soil constituent.

4. Chemical reference data are treated as accurate values and suited for the given application. Determination methods are selected based on their proven suitability and successful application. However, local conditions may cause the elimination of an established method (e.g. determination of clay mineral content based on particle size only). In addition, the inaccuracy included in each reference measurement limits the accuracy of the presented methodology.

5. There are a number of impact factors related to the acquisition of in-situ and airborne data. In regard to the hyperspectral imagery, these are mainly sensor related factors, atmospheric correction and geometric considerations. The detectability and shape of spectral features depends on the signal-to-noise ratio and radiometric calibration of the sensor. A number of studies, however, proved the suitability of the HyMap sensor for quantitative soil analysis (e.g. Chabrillat et al. 2002, Richter 2010a, Schwanghart und Jarmer 2011). In the same way, HyMap imagery in general can well be corrected for atmospheric disturbances (Richter 2010b). Individual artifacts remaining from this procedure are minor and can be removed by spectral polishing in restricted wavelength regions. The spectral resolution of the HyMap sensor influences the size, shape and position of absorption features, though as is shown in section 6.1.1, this influence is minor. Different ways of locating ground truth sites introduces ambiguities in validation accuracies. They are further increased when local soil measurements of about 1 m^2 are compared to contents derived for an image pixel covering about 10 m^2. The last two mentioned factors are the result of the spatial resolution of the applied sensor and the applied orthorectification.

6. Changes in surface conditions over the large study area, as result of e.g. variances in geology, stones overlying the surface, surface roughness or the presence of soil crusts, introduce a further spectral variance that is not diagnostic for the three considered parameters. Model calibrations show that these factors are not decisive for soil organic carbon and iron oxides models though they are probably an additional factor for the systematic miscalibration of the clay prediction models.

7. Vegetation coverage in general limits the applicability of the established regression models. The proposed method to approximate the soil residual signal contained in a mixed pixel provides a possibility to retrieve information of soil constituents for partially vegetated natural areas. Nevertheless, this method introduces inaccuracies resulting of a possible incomplete resolution of the complex spectral mixtures. A vegetation signature remaining in the spectra used for the prediction of soil constituents has an effect in particular on soil organic carbon predictions in case of remaining SWIR features, and on iron oxides in case of remaining VIS/NIR features.

8. A transfer of the methodology to other geographic regions is possible for other semiarid areas. It is only restricted to carbonate free areas because the presence of the major carbonate absorption around 2345 nm will highly disturb the small organic carbon and clay absorptions around 2330/2340 nm, which are considered in model development.

6.5 Prediction of soil constituents from EnMAP hyperspectral satellite imagery

Soil constituents are also predicted from simulated data that match the characteristics of the future EnMAP satellite (see sections 3.2.3 and 5.3.3). Specific results from within the processing of the simulated EnMAP data towards the final maps of soil constituents are summarized here, while further details are found in appendix G. Results are compared to the ones of the medium sized HyMap subsets of the Gonakraal and Wintersfontein areas (scale 2) which match the extension of the simulated EnMAP data.

Spectral Mixture Analysis and approximation of soil residues

A challenge is the endmember extraction from the simulated EnMAP data. The small scale mixtures in the study area in relation to the large pixel size of the EnMAP sensor result in problems in extracting EMs from the imagery. Only incomplete EM sets can be derived that are insufficient for spectral unmixing. For this reason, and the fact that this problematic is not in the focus of this study, the final EM set to be used for spectral unmixing is selected out of a combination of the few EnMAP derived EMs and the EMs derived from the HyMap imagery of the selected subsets re-sampled to the spectral resolution of the EnMAP sensor. Table G-26 and Figure G-26 in the appendix give an overview of the EMs in each dataset and the EM sets that are finally selected to unmix the EnMAP data. Results of spectral unmixing are overall similar to the ones obtained from the processing of the HyMap imagery. A comparison of the ground cover fractions of HyMap and EnMAP for the two areas is shown in Figure G-27 in the appendix. Average and standard deviations of the cover fractions determined for each class are usually lower (Table G-27). Nevertheless, unmixing RMSE are lower compared to the HyMap scenes.

The settings applied for the approximation of the soil signatures from mixing pixels follow the trends of the corresponding HyMap subsets. The percentage of vegetation ground coverage is 5 % below the maximum fraction of PV and NPV that is allowed in scale 2 processing of the HyMap imagery (Table G-28). The percentage of pixels for which soil signatures can be approximated decreased from about 41 to 38 % for the Gonakraal and from about 73 to 58 % for the Wintersfontein area. This is a result of the increased influence of vegetation on soil dominant pixels due to the lower spatial resolution, hence increased spectral mixing within the simulated EnMAP data.

Delineation of soil constituents

The accuracies of the newly calibrated regression models matching the spectral resolution of the EnMAP sensor are compared to the HyMap models in Table G-29 in the appendix. Calibrations accuracies are in the same range as the models developed for HyMap data, though consistently lower. For instance, the calibration accuracy (R^2_{Cal}) of organic carbon models declines from 0.79 to 0.77 and for iron oxides models from 0.61 to 0.59 (R^2_{Val} 0.66/0.62 for C_{org} and 0.30/0.29 for Fe_d). $RMSE_{Cal/Val}$ shows minimal increases.

The soil constituents derived based on the simulated EnMAP data are compared to the HyMap derived maps of the two subsets (scale 2 data are used). The HyMap maps are not resampled to match the spatial resolution of the EnMAP data, because the pattern of pixels where no values are calculated and the parameters of a thereby necessary interpolation operation would strongly influence resulting maps and comparisons. Thus, the comparison is limited to a visual comparison of observed patterns and statistics. Table 6-9 gives a summary of the statistics of the three soil constituents prediction from the EnMAP data and their comparison to the corresponding HyMap predictions. Detailed results are found in appendix G.

Figure 6-28 and figure 6-29 show the maps of topsoil organic carbon and iron oxides contents for simulated EnMAP and HyMap data. Spatial patterns observed in the EnMAP derived maps of soil organic carbon resemble the HyMap derived maps (figure 6-28). Although, the EnMAP spatial resolution has the effect of a low pass filter and strongly reduces small-scale variations. Low contents are also found in particular on the bare plain in the center of the Gonakraal area and within the bare parts in the east of the Wintersfontein area. The range of predicted contents is comparable to the HyMap derived contents within the Wintersfontein area, though about 0.3 % C_{org} higher within the Gonakraal area. An increase of soil organic carbon contents within partly vegetated areas seems to be apparent, but the observations are only based on singular pixels because no spatially continuous prediction is provided in these areas. Topsoil iron oxides concentrations predicted based on EnMAP data are about 1 % Fe_d lower than they are predicted from HyMap data (figure 6-29). In both datasets lower iron oxides concentrations are observed close to the erosive channels, though the spatial resolution prevents the detection of small structures in the EnMAP derived maps. In the Gonakraal area an enrichment of iron oxides in the center of the plain and in the western parts shows up. The 30 m spatial resolution of the EnMAP data is usually too low to identify a correlation between soil iron oxides contents and the small geologic structures as is frequently observed in the HyMap derived iron oxides maps. The close layering in the study area (dm to tens of m) mostly prevents these observations within the EnMAP derived maps. Higher iron oxides contents are only observed in the soils overlying a large sandstone unit trending east-west in the center of the Wintersfontein area (figure 6-29b). Maps of predicted soil clay contents are found in Figure G-28. Observed patterns are similar, though the lower spatial resolution complicates the correlation of clay contents with distinct geologic layers or crusting intensity.

Table 6-9: Summary of statistics for the prediction of topsoil organic carbon for the simulated EnMAP data in comparison with the corresponding HyMap data (scale 2). Statistics are given of smoothed with 3 x 3 filter kernel for the EnMAP and 5 x 5 filter kernel for the HyMap data.

	Subsets	C_{org}	Iron oxides	Clay
Gonakraal	**EnMAP**	Mean: 1.53 Stdev: 0.46 Min: 0.43 Max: 4.58	Mean: 2.81 Stdev: 0.82 Min: 0.92 Max: 8.19	Mean: 7.60 Stdev: 1.31 Min: 0.47 Max: 14.55
	Medium HyMap subsets	Mean: 1.07 Stdev: 0.38 Min: 0.21 Max: 4.62	Mean: 3.96 Stdev: 0.85 Min: 0.90 Max: 10.47	Mean: 8.73 Stdev: 1.51 Min: 0.02 Max: 16.74
Wintersfontein	**EnMAP**	Mean: 1.50 Stdev: 0.48 Min: 0.23 Max: 5.73	Mean: 2.63 Stdev: 0.89 Min: 0.91 Max: 10.38	Mean: 6.09 Stdev: 1.52 Min: 0.02 Max: 15.59
	Medium HyMap subsets	Mean: 1.80 Stdev: 0.71 Min: 0.21 Max: 5.76	Mean: 3.46 Stdev: 1.00 Min: 0.90 Max: 10.62	Mean: 6.34 Stdev: 1.91 Min: 0.00 Max: 16.74

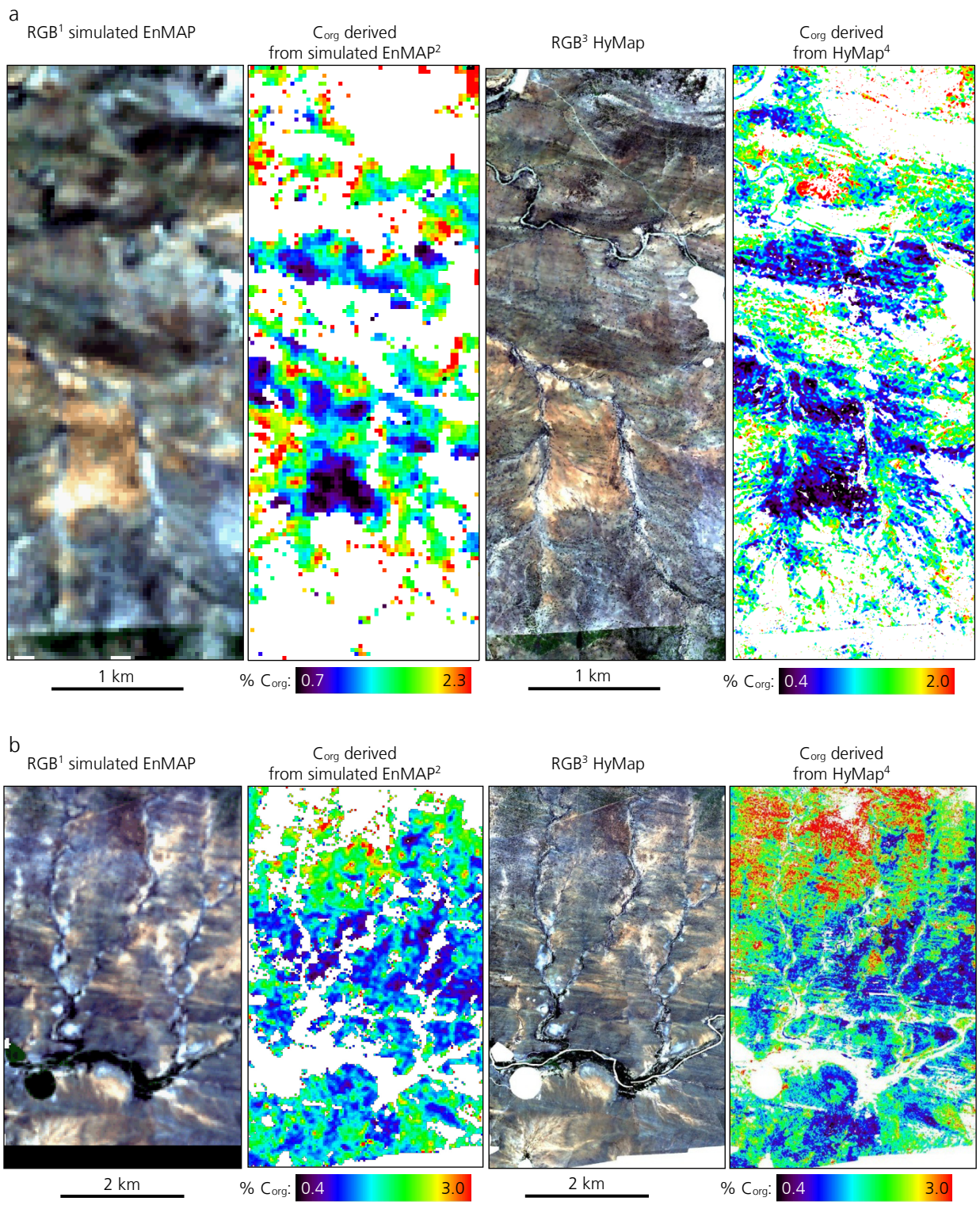


Figure 6-28: Comparison of the spatial distribution of topsoil organic carbon derived from simulated EnMAP data and HyMap data (scale 2) for the Gonakraal area (a) and the Wintersfontein area (b). For areas masked white, no soil constituents can be predicted.
[1]True color EnMAP images with R: 640 nm, G: 549 nm, B: 460 nm, [2]Predicted contents after interpolation and filtering with kernel size 3 x 3, [3]True color HyMap images with R: 640 nm, G: 549 nm, B: 456 nm, [4]Predicted contents after interpolation and filtering with kernel size 5 x 5.

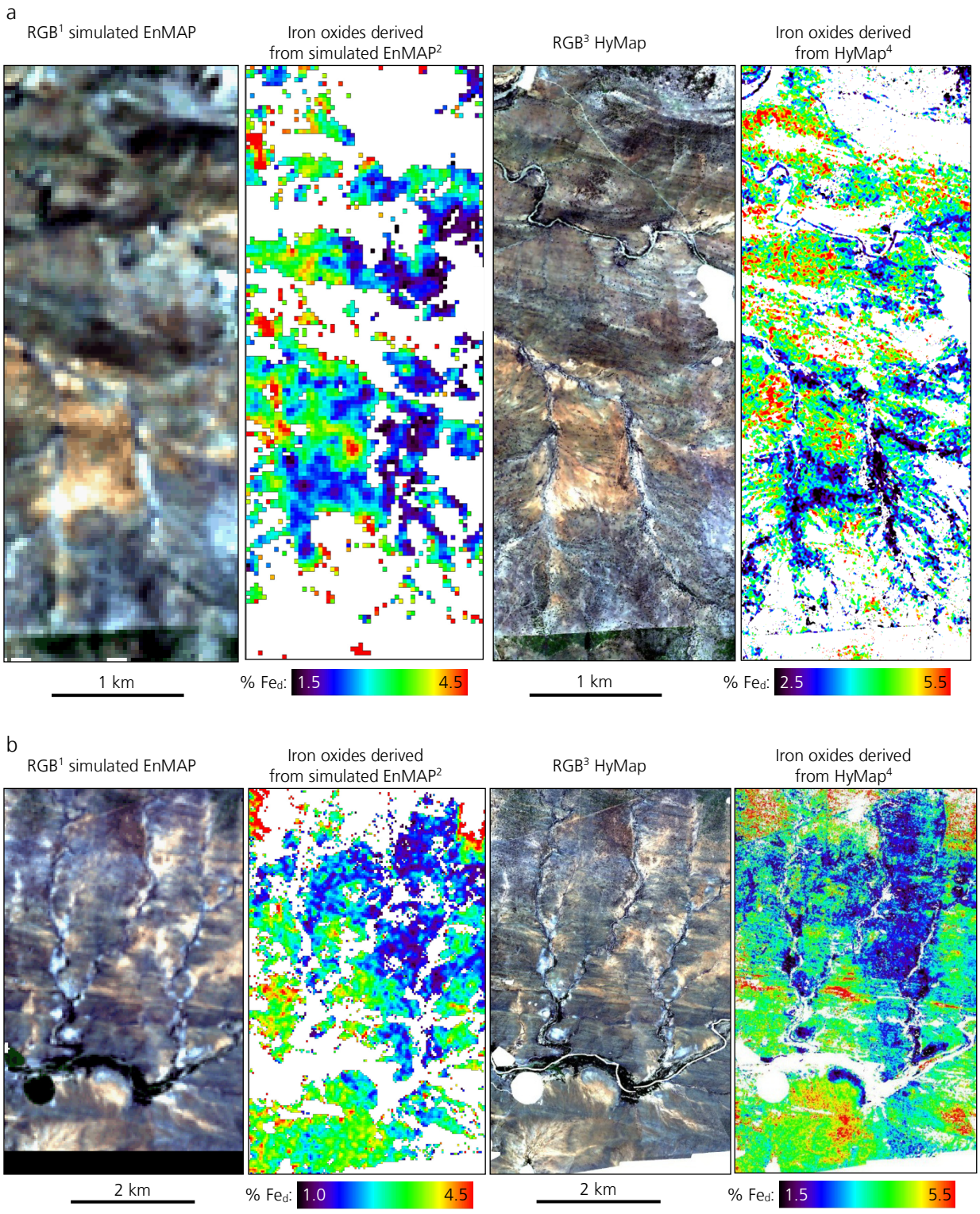


Figure 6-29: Comparison of the spatial distribution of topsoil iron oxides contents derived from simulated EnMAP data and HyMap data (scale 2) for the Gonakraal area (a) and the Wintersfontein area (b).
[1]True color EnMAP images with R: 640 nm, G: 549 nm, B: 460 nm, [2]Predicted contents after interpolation and filtering with kernel size 3 x 3, [3]True color HyMap images with R: 640 nm, G: 549 nm, B: 456 nm, [4]Predicted contents after interpolation and filtering with kernel size 5 x 5.

Conclusions

The observations show that soil constituents can be predicted based on simulated EnMAP satellite data. Within the processing workflow the greatest challenge seems to be the derivation of proper image endmembers. This is a result of the highly variant land cover in the study area and the 9-fold increased pixel size of the EnMAP imagery in comparison to the HyMap imagery. The lower spatial resolution further causes more pixels to be influenced by vegetation, and thus, a reduction of the percentage of pixels for which a soil residuum can be approximated and soil constituents can be predicted. Nevertheless, the spatial patterns observed for soil organic carbon and iron oxides are similar to the ones determined based on the HyMap imagery, although the lower spatial resolution prevents the detection of smaller structures. Ranges of calculated contents slightly differ from the HyMap derived contents. Predicted soil organic carbon contents based on EnMAP data are significantly higher for one subset, while iron oxides contents are generally lower in both regions. Both may be a result of either the EnMAP sensor's characteristics, artifacts produced within the data simulation or are a result of the applied pre-processing. A comparison of in-situ measured with contents predicted from the EnMAP imagery can not be realized due to the small extent of the simulated EnMAP scenes, though it is expected that correlations are below the ones determined for the HyMap imagery.

6.6 Interpretation of results for degradation assessment and carbon related land surface processes

The presented maps show patterns of the spatial distribution of key soil constituents over an area of 320 km^2 within the Albany Thicket Biome as they existed in October 2009. Soil constituents' information can be derived nearly spatially continuous where bare soil is the dominant land cover fraction and vegetation coverage is below about 40 %. Section 6.3 shows that the soil constituents can be linked to geomorphic features and active processes directly related to land degradation.

Identified land surface processes

Based on the established maps of soil constituents and known mechanisms of enrichment or reduction of constituents in soils, areas with progressive and regressive soil development are identified (after Birkeland 1990). In areas with progressive soil development, the soil profiles in general become deeper and better differentiated by horizons over time. In this process of relative stability, topsoil organic carbon content increases due to decomposition of litter and clay-sized particles, and iron oxides accumulate in the weathering process (see Ben-Dor et al. 2008). Nevertheless, pedogenesis is regressive when soil forming processes are interrupted by erosion, which reduces soil profile depth. This is linked to a reduced topsoil organic carbon content, an enrichment of iron oxides as erosion residues and the increased formation of a surface crust rich in clay-sized particles. In the study area this frequently occurred on partly vegetated slopes, where erosion is supported by the existing gradient, but also on flat areas with reduced vegetation coverage (refer to figure 6-10 and figure 6-11). Continued erosion produces truncated profiles. Along linear flow structures in the study area, erosion is so heavy that it locally removed the entire sediment (figure 6-13b).

This shows that undisturbed and developed soils can be discriminated from eroded soils by the constituent components of their uppermost topsoils, thus soil constituents can be used as indicators for degradation

(Ben-Dor et al. 2008 and 2009, Viscarra Rossel et al. 2011). The analysis highlights the importance of even small vegetation patches, first for carbon input through littering and second for sediment stabilization and as sediment traps. The benefit of increased carbon concentrations in the soil, such as an increased stability and nutrient availability, can only be derived indirectly from the presented maps.

In addition, to the soil constituents' maps, other map products (either generated within the workflow or products that can be generated using one of the presented methods) are valuable indicators to identify the status of Thicket degradation. Maps of the dominant land cover classes resulting from spectral unmixing are an example of this (see figure 6-9 for an overview), where they show the spatial distribution of dense and patchy vegetation and where uncovered soil is predominant. Where vegetation coverage is low soils are prone to degradation and erosion.

Vegetation coverage - a limiting factor when analyzing Thicket degradation?

A major factor within the generation and interpretation of soil maps is the missing soil information wherever soil is masked by vegetation. The indirect analysis of soil properties based on relations between vegetation and soil attributes, e.g. by using plant functional types (PFT) or Ellenberg indicator values, was previously described (Mulder et al. 2011). For instance, the concept of PFT can be used to map typical groups of species that grow on typical soils. However, such analysis is unable to provide soil related information with the same degree of detail as is shown in this study. The traditional way to interpolate soil properties for vegetated areas from the surrounding non-vegetated areas (e.g. Ben-Dor et al. 2002) is not applicable in the Albany Thicket Biome where soil properties, in particular soil organic carbon, highly vary from open to areas beyond a canopy (see e.g. Mchunu 2012, Mills and Cowling 2006). The described method as such is not able to directly visualize differences in soil constituents between pristine and open Thicket sites. Though as shown above and in section 6.3 the presented maps allow for analyzing a number of processes directly related to degradation. Thus, they indirectly deliver indicators visualizing the effects of degradation in the Albany Thicket Biome without a direct comparison of soil constituents' contents below a vegetation canopy and in the open.

Identification of degradation hotspots

Based on the maps and the derived information, hotspot regions for degradation within the study area can be identified. These are areas that face the most challenges of soil erosion as a result of advanced land degradation. They should be treated most urgently and selected first for restoration. However, the integration of all available information and development of an "index of land degradation", as was done by Hill and Schütt 2000, will be subject to future research. In such a classification additional information quantifying existing surface crusting can be included. Goldshleger et al. (2001) and Ben-Dor et al. (2002) mapped crusted soils based on their infiltration and runoff properties and thereby identified soils with an increased erosion potential. A direct mapping of parent material in soils to describe soil development and erosional state of soils as done by Hill et al. (1995) is not feasible in the study area because the sandstones and mudstones as the predominant parent material are not sufficiently spectrally distinct from the soil material.

Contribution to an assessment of soil related carbon pools and their changes

The ability of soils to sequester carbon is of increasing interest as a potential way to mitigate the effect of greenhouse gases in the atmosphere (Lal 2004). In this context, the number of studies dealing with

quantitative mapping of soil carbon from remotely sensed data rapidly increased in the last decades (e.g. Bartholomeus et al. 2008 and 2011, Ben-Dor et al. 2002, Gomez et al. 2008b, Stevens et al. 2008 and 2010). The presented maps of topsoil organic carbon show the high spatial variability in soil carbon stocks in the Albany Thicket Biome. The effect of human induced changes in land coverage can be identified and visualized. The maps present a direct assessment of the current inventory and status of carbon stocks in soils. A soil carbon stock balance of the entire area is prevented by densely vegetated areas for which no soil carbon stocks can be derived. In addition, the generated maps and relationships observed allow for a better characterization of processes determining carbon accumulation and losses on a regional scale. This contributes to a reduction of the uncertainty in assessments of terrestrial carbon pool sizes and their variations. The available soil organic carbon maps in combination with maps derived from possible future surveys (ideally using the same scanner system and methodology) would allow to visualize potential changes of local carbon pools. These could either result from further advanced degradation and soil erosion or from rehabilitation initiatives leading to a long-term increase in soil carbon stocks (e.g. Mills and Cowling 2006).

Application of results for Thicket restoration

The soil constituents maps and especially the knowledge of degradation hotspots can directly be used as basis for local restoration efforts. Besides the adoption of recommended management practices, restoration includes an extensive restoration of degraded landscapes because transformed Thicket vegetation does not recover spontaneously (Vlok et al. 2003). The extensive restoration of degraded landscapes using in particular cuttings from the species *P. afra* was suggested as the most cost effective method for Thicket restoration (Mills and Cowling 2006, Swart et al. 1994) as was shown in several pilot projects. Restoration of degraded landscapes has a positive effect not only on carbon sequestration and biodiversity but also on enhanced ecosystem services such as water quality, food security and not least the value of an intact ecosystem for tourism.

7 Main conclusions and outlook

The quantification of soil carbon stocks and the assessment of ecosystem degradation are of major importance in context of necessary carbon sequestration and ongoing land degradation. This study showed an innovative approach to map the fine-scale heterogeneity of key soil constituents related to degradation and including organic carbon by a combination of field and Imaging Spectroscopy. The considered landscape in the semiarid Albany Thicket Biome, South Africa, made great demands in terms of its regional extent with variable land cover, surface and soil conditions. This required a specific treatment that is provided within the entire methodology. The major findings are in particular:

- The combination of several diagnostic spectral features and their properties allowed to overcome non-significant influences on spectral features and to establish physically based regression models. Models that accurately predicted soil organic carbon and iron oxides contents in the Thicket Biome were established. Despite that, no correlation could be established for soil clay content.
- Robust subpixel soil signatures could be approximated from image pixels of a natural semiarid environment despite a high level of spectral mixing of the various surface materials.
- Soil constituents could be predicted based on spectral signatures for the Albany Thicket Biome. Predictions were more accurate for medium and small subsets (below about 20 km^2). However, the agreement of field reference and predicted contents was low, although no systematic errors were identified for soil organic carbon and iron oxides.
- A confidence layer provided a pixel-wise assessment of the reliability of predicted soil constituents.
- The spatial coverage of soil constituents predictions could be significantly enlarged compared to traditional approaches using the proposed multi-stage methodology.
- The maps of quantified soil organic carbon contents directly provided an assessment of carbon stocks in the study area in their status of October 2009.
- Areas of stable progressive soil development could be differentiated from erosion affected areas. Thereby, degradation hot spots could be identified within the Albany Thicket Biome where land changes occur due to intense overgrazing.
- Soil constituents could be predicted from simulated EnMAP satellite data. The greatest challenge within this application was to resolve the spectral mixtures within the low spatial resolution data.

Evaluation of the results in detail

Within the proposed multi-stage methodology, the individual steps were rated with different characteristics, accuracy and benefit. It was shown that based on chemical ground reference data and field spectra of 125 sampled sites, physically based models for the prediction of soil organic carbon, iron oxides and clay content

from spectra could be established. The approach benefited from combining several spectral features (distinct absorptions and properties of the spectral curve and continuum), each of them diagnostic but potentially influenced by other features or non-significant disturbances (see section 4). The approach provided a physical basis, was considered straightforward and required minimal computation. Statistical adaptation within regression analysis was reduced to a minimum. The developed regression models showed good correlations in particular for soil organic carbon with cross-validated R^2 of 0.66 and RMSE of 0.46 % C_{org}. The prediction models of organic carbon and iron oxides exhibited high model stability and came close to the predictability of PLS regression analysis. Thus, the feature-based regression approach was shown to provide an alternative to common statistical techniques, providing advantages in particular in terms of its physical significance and its transferability (section 6.1.1), in which the latter could not be proven in the extent of this study. Clay content models did not reach significant accuracies – neither in feature-based MLR nor in PLS. The reason for this was identified in the chemical determination of the clay content based on particle size only, in combination with the local presence of soil physical crusts and the highly variable characteristics in the study area (e.g. geology). Clay models thus were only applied for qualitative predictions.

The hyperspectral imagery received adequate treatment to resolve the intensive spectral mixing within image pixels resulting of small scale changes in land cover. For soil dominated pixels, the influence of vegetation on the spectral signature could be removed to a large extent and the pure soil signature contained in mixed pixels was approximated by a residual spectral mixture analysis. Processing thresholds assured the quality of the soil residues (e.g. low unmixing RMSE, minimal bare soil fraction and maximum fraction of vegetation). The greatest challenge but also the largest benefit of the methodology was provided within this step, which built the basis for a successful and unbiased prediction of soil constituents from hyperspectral imagery. This procedure allowed the successful approximation of pure soil signatures despite a vegetation coverage of up to about 40 % depending on the image characteristics (section 6.2.2). The spectra were of good quality and showed characteristic soil spectral features. Only in individual spectra a residual vegetation signature in the VIS/NIR remained. The amount of pixels for which soil constituents could be predicted could be increased significantly compared to common approaches, where analyzed pixels were limited to ones with no or a low vegetation coverage (e.g. 10 % vegetation coverage). For an example area within the study area, the number of pixels for which soil constituents were predicted was between 12 and 22 times as high using the proposed approach (section 6.4.3). Thus, with this approach it was possible (1) to enlarge the spatial extent of derived information of soil constituents so that it was nearly spatially continuous in soil dominated areas and (2) to reduce the influence of materials such as vegetation on the quantification of soil constituents to a minimum. This accounted for a significant benefit in particular in areas with very inhomogeneous land cover as found in the semiarid South African study area.

Soil constituents were predicted for the South African study site by applying the established regression relationships to the maps of locally approximated soil signatures. This analysis included a selective interpolation of individual pixels for which no soil signature was approximated. Thereby, the amount of pixels for which soil constituents were predicted could be increased from 27 to about 42 %. Where soil is the dominant land cover class, soil constituents could be predicted nearly spatially continuous. For each soil constituent about 98 % of the predicted values were within the calibration range of the prediction models. This showed, that the field sampling represented well the variability of all three investigated parameters in areas for which soil constituents could be predicted based on hyperspectral imagery and that the prediction models were well adapted to the present conditions.

Maps were presented that showed patterns of the spatial distribution of key soil constituents as they existed in October 2009. A confidence layer showed a pixel-wise assessment of the reliability of predicted soil constituents which was derived based on factors within the workflow that may influence the predictions. For soil organic carbon and iron oxides the observed patterns could be linked to geomorphic features and land surface processes. Areas of relative stability and enrichment of soil organic carbon and iron oxides were discriminated from areas of partial or even intense erosion. These findings were in agreement with earlier studies focusing on links between soil organic carbon and iron oxides concentrations and their input or erosion mechanisms (see section 6.3). It was concluded that in the Albany Thicket Biome in particular the relationship between topography and soil organic carbon content is significantly modified by the non-homogenous pattern of remaining vegetation. Areas where degradation is most severe were identified as hot spots. This is possible although the maps did not show soil constituents in vegetation dominated areas. These findings supported the adoption of soil constituents as proxies for degradation.

The comparison of in-situ measured to the constituents' contents predicted for the ground truth sites based on the HyMap imagery provided a regional validation of the entire multi-stage methodology for the South African study area. Scatter plots provided evidence of existing correlations, though correlation coefficients were low. Modeling residues for soil organic carbon and iron oxides nevertheless showed that the proposed methodology did not introduce systematic errors. Exclusively for clay, systematic mispredictions were observed in accordance with the calibrated clay prediction models. No correlation between modeling residues and the determined vegetation coverage could be detected for all three soil constituents. The low correlations of measured and predicted soil constituents was seen as an indicator that significant factors remain that disturb the correlation of spectral characteristics and soil constituents.

The consistency of the compiled methodology was proven for imagery of different size and characteristics, including small and medium-sized subsets (below about 20 km^2) and finally the imagery of the entire study area. The same patterns were observed and subtle differences occurring in predicted contents were below the cross-validated RMSE of the regression models. Nevertheless trends in the predicted contents, local validation accuracies and confidence assessments over the three scales indicated a more accurate prediction of soil constituents for medium and small subsets, which was likely a result of the more accurate unmixing of smaller subsets because the endmembers represented local materials and mixtures. In conclusion the spectral unmixing was identified as the factor limiting the prediction accuracy of soil constituents.

An automation of the proposed methodology was given for its crucial steps. Manual intervention was only needed in two steps - first to select a final endmember set out of a pre-selected set, and second to eventually adjust the default values for the thresholds that ensure a proper approximation of soil residues according to the characteristics of the given scene. A manual control of these results was recommended to ensure proper handling and high quality base data for an analysis of soil constituents. The procedures for calibrating a feature-based regression relationship and its application to the image data of approximated soil signatures were fully automated. This included the prediction and the selective filtering and interpolation of the soil constituents maps. They provided maps for two generalization levels resulting of a small and a larger filter size.

Soil constituents were successfully predicted also from simulated data that matched the characteristics of the future EnMAP satellite. The resolution of the spectral mixtures in the 30 by 30 m EnMAP pixels was identified as the greatest challenge within the workflow. It was shown that spatial patterns observed for soil organic carbon and iron oxides are similar to the ones determined based on the HyMap imagery, although the lower spatial resolution prevented the detection of smaller structures and prevented a direct comparison of predicted contents.

Applications in general and within the Albany Thicket Biome

The established maps of soil constituents helped to discriminate between areas with progressive soil development under relative stability on the one hand and erosion affected areas on the other hand. This allowed the identification of degradation hot spots within the Albany Thicket Biome as areas, that face the most challenges of soil erosion as a result of advanced land degradation and should be selected first for restoration. The maps further presented a direct assessment of the current inventory and status of soil carbon stocks in the Albany Thicket Biome.

The information derived from Imaging Spectroscopy could be the base for the formulation of appropriate land management practices and for local restoration efforts including site specific management; both parts are needed for the long-term rehabilitation of degraded landscapes (section 6.6). The development of an ecosystem's soil carbon pool during restoration transactions could be monitored using the proposed approach. This could be valuable to quantify the increase of carbon stocks during this process, which would be needed when trading the sequestered carbon, for instance in the free market as it is currently done within some projects restoring selected land portions. This would be, however, limited by the point the vegetation coverage gets too pronounced.

Prospects of possible further developments and applications

Further developments and extensions of the presented approach might improve the accuracy of predictions and the applicability of the methodology. For instance a recursive calculation of multiple soil parameters in dependency to each other and realized by an iterative procedure could be valuable to resolve overlapping absorption features (addressed in section 5.1.1). Extensive laboratory measurements of artificial mixtures with known chemistry would be necessary to prove the feasibility and benefit of such analysis in advance. The implementation of a degradation index, as joint product of multiple layers such as soil constituents' maps, land cover fractions, etc. would be valuable to visualize the heterogeneous degradation patterns in the Albany Thicket Biome. As discussed in section 6.6 this was not realized within the study but will be part of future research.

The method was developed for semiarid areas in general. It is not adapted to specific conditions in the study area, though prediction models were regionally validated in a South African study area. It is envisioned that the compiled methodology could be applied for soil constituents mapping in other semiarid regions with partial vegetation coverage and considering amounts of bare soil within the land surface. This transfer to other regions of similar environmental conditions will further be investigated.

In consideration of spectral mixing within image pixels and approximation of undisturbed soil signatures as the critical factor identified within the workflow, the focus of further research should be drawn to this matter. Its importance rises in particular when envisioning the application of future hyperspectral satellite sensor systems such as EnMAP, HISUI and HyspIRI. These sensors provide the capacity and technical requirements for determining the spatial distribution of key soil constituents over large areas and in multi-temporal surveys. The large ground coverage and repeatability of data acquisitions of hyperspectral satellite sensors support change detection research by the mapping of past and future land use effects on selected soil constituents, which can contribute to and monitor long-term sustainable and site adapted soil management. However, the link between a fine-scale soil heterogeneity, as it was found in the study area, and the provided reduced spatial resolution of such sensor systems even increase problems resulting of spectral mixing and the connection of the data with ground truth information, as they were identified within the analysis of HyMap and simulated EnMAP data.

8 Bibliography

ACOCKS, J.P.H.: Veld Types of South Africa. Second edition, Memoir of the Botanical Survey of South Africa, 1975, No 40.

ADAMS, J., SMITH, M. & JOHNSON, P.: Spectral mixture modeling: A new analysis of rock and soil types at the Viking Lander 1 site. *Journal of Geophysical research*, 1986, Vol. 91, pp. 8098–8112.

ADAMS, J., SMITH, M. & GILLESPIE, A.: Imaging spectroscopy: Interpretation based on spectral mixture analysis. In: PIETERS, C. M. & ENGLERT, P. A. (eds.). Remote Geochemical Analysis: Elemental and Mineralogical Composition. Cambridge, U.K.: Cambridge Univ. Press, 1993, pp. 145–166.

Agricultural Research Council, Institute for Soil Climate and Water (ARC-ISCW), provision of climate station data, 2011.

ARDUINO, E., BARBERIS, E., AJMONE MARSAN, F., ZANINI, E. & FRANCHINI, M.: Iron oxides and clay minerals within profiles as indicators of soil age in Northern Italy. *Geoderma*, 1986, Vol. 37, pp. 45-55.

ASNER, G. & LOBELL, D.: A Biogeophysical Approach for Automated SWIR Unmixing of Soils and Vegetation. *Remote Sensing of Environment*, 2000, Vol. 74, pp. 99-112.

BACHMANN, M.: Automatisierte Ableitung von Bodenbedeckungsgraden durch MESMA-Entmischung. PhD Thesis, Julius-Maximilian-Universität Würzburg, 2007.

BACHMANN, M., S. ADAR, E. BEN-DOR, J. BIESEMANS, M. GRANT, J. HANUS, G.B.M. HEUVELINK, S. HOLZWARTH, A. HUENI, E. KNAEPS, M. KNEUBUEHLER, K. MEULEMANN, E. DE MIGUEL, A. PIMSTEIN, E. PRADO, I. REUSEN, T. RUHTZ, M. SCHAALE: Generic Quality Indicators and Quality Layers. EUFAR FP7, JRA2-HYQUAPRO, Report DJ2.2.1, 2010.

BARTHOLOMEUS, H., SCHAEPMAN, M., KOOISTRA, L., STEVENS, A., HOOGMOED, W. & SPAARGAREN, O.: Spectral reflectance based indices for soil organic carbon quantification. *Geoderma*, 2008, Vol. 145, pp. 28-36.

BARTHOLOMEUS, H., KOOISTRA, L., STEVENS, A., VAN LEEUWEN, M., VAN WESEMAEL, B., BEN-DOR, E. & TYCHON, B.: Soil Organic Carbon mapping of partially vegetated agricultural fields with imaging spectroscopy. *International Journal of Applied Earth Observation and Geoinformation*, 2011, Vol. 13, pp. 81-88.

BARTON, C. & KARATHANASIS, A.: Clay Minerals. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition, CRC Press, 2006, pp. 276-280.

BAUMGARDNER, M., SILVA, L., BIEHL, L. & STONER, E.: Reflectance Properties of Soils. *Advances in Agronomy*, 1985, Vol. 38, pp. 1-44.

BAYER, A., BACHMANN, M., MÜLLER, A. & KAUFMANN, H.: A comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African ecosystem. *Applied and Environmental Soil Science*, 2012, Vol. 2012, Article ID 971252, 20 p.

BELLON-MAUREL, V. & MCBRATNEY, A.: Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils - Critical review and research perspectives. *Soil Biology & Biochemistry*, 2011, Vol. 43, pp. 1398-1410.

BEN-DOR, E. & BANIN, A.: Visible and near-infrared (0.4-1.1 µm) analysis of arid and semiarid soils. *Remote Sensing of Environment*, 1994, Vol. 48, pp. 261-274.

BEN-DOR, E. & BANIN, A.: Near infrared analysis (NIRA) as a simultaneously method to evaluate spectral featureless constituents in soils. *Soil Science*, 1995a, Vol. 159, pp. 259-269.

BEN-DOR, E. & BANIN, A.: Near-Infrared Analysis as a rapid method to Simultaneously Evaluate Several Soil Properties. *Soil Science Society of America Journal*, 1995b, Vol. 59, pp. 364-372.

BEN-DOR, E., INBAR, Y. & CHEN, Y.: The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400-2500 nm) during a controlled decomposition process. *Remote Sensing of Environment*, 1997, Vol. 61, pp. 1-15.

BEN-DOR, E., IRONS, J. & EPEMA, G.: Soil Reflectance. In: RENCZ, A. (ed.): Remote Sensing for the Earth Sciences, Vol. 3: Manual of Remote Sensing. John Wiley and Sons, 1999, Vol. 3, pp. 111-188.

BEN-DOR, E., PATKIN, K., BANIN, A. & KARNIELI, A.: Mapping of Several Soil Properties Using DAIS-7915 Hyperspectral Scanner Data - A Case Study Over clayey Soils in Israel. *International Journal of Remote Sensing*, 2002, Vol. 23, pp. 1043-1062.

BEN-DOR, E., GOLDLSHLEGER, N., BENYAMINI, Y., AGASSI, M. & BLUMBERG, D.: The Spectral Reflectance Properties of Soil Structural Crusts in the 1.2 to 2.5 µm Spectral Region. *Soil Science Society of America Journal*, 2003, Vol. 67, pp. 289-299.

BEN-DOR, E., LEVIN, N., SINGER, A., KARNIELI, A., BRAUN, O. & KIDRON, G.: Quantitative mapping of the soil rubification process on sand dunes using an airborne hyperspectral sensor. *Geoderma*, 2006, Vol. 131, pp. 1-21.

BEN-DOR, E., TAYLOR, R., HILL, J., DEMATTÊ, J., WHITING, M., CHABRILLAT, S. & SOMMER, S.: Imaging Spectrometry for Soil Applications. *Advances in Agronomy*, 2008, Vol. 97, pp. 321-392.

BEN-DOR, E., CHABRILLAT, S., DEMATTÊ, J., TAYLOR, G., HILL, J., WHITING, M. & SOMMER, S.: Using Imaging Spectroscopy to study soil properties. *Remote Sensing of Environment*, 2009, Vol. 113, pp. 38-55.

BIRKELAND, P.: Soil-geomorphic research - a selective overview. *Geomorphology*, 1990, Vol. 3, pp. 207-224.

BISHOP, J., LANE, M., DYAR, M. & BROWN, A.: Reflectance and emission spectroscopy study of four groups of phyllosilicates: smectites, kaolinite-serpentines, chlorites and micas. *Clay Minerals*, 1994, Vol. 43, pp. 35-54.

BRADFORD, J. & HUANG, C.: Mechanisms of crust formation: physical components. In: SUMNER, M. & STEWART, B. (eds.): Soil crusting: chemical and physical processes. Lewis Publishers, Boca Raton, 1992, pp. 55-72.

BROWN, D., SHEPHERD, K., WALSH, M., MAYS, M. & REINSCH, T.: Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, 2006, Vol. 132, pp. 273-290.

CHABRILLAT, S., GOETZ, A., KROSLEY, L. & OLSEN, H.: Use of hyperspectral images in the identification and mapping of expansive clay soils and the role of spatial resolution. *Remote Sensing of Environment*, 2002, Vol. 82, pp. 431-445.

CHANG, C.-W. & LAIRD, D.: Near-infrared reflectance spectroscopic analysis of soil C and N. *Soil Science*, 2002, Vol. 167, pp. 110-116.

CHANG, C.-W., LAIRD, D., MAUSBACH, M. & HURBURGH JR., C.R.: Near-infrared reflectance spectroscopy - Principal components regression analyses of soil properties. *Soil Science Society of America Journal*, 2001, Vol. 65, pp. 480-490.

CLARK, R.: Chapter 1: Spectroscopy of Rocks and Minerals and Principles of Spectroscopy. In: RENCZ, A. (ed.): Remote Sensing for the Earth Sciences, Vol. 3: Manual of Remote Sensing. John Wiley and Sons, 1999, pp. 3-58.

CLARK, R. & ROUSH, T.: Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*, 1984, Vol. 89, pp. 6329-6340.

CLARK, R., KING, T.V.V., K.M. & SWAYZE, G.: High Spectral Resolution Reflectance Spectroscopy of Minerals. *Journal of Geophysical Research*, 1990, Vol. 95, pp. 12653-12680.

CLARK, R., SWAYZE, G., GALLAGHER, A., KING, T. & CALVIN, W.: The U. S. Geological Survey, Digital Spectral Library: Version 1: 0.2 to 3.0 microns. U.S. Geological Survey Open File Report 93-592, 1993, 1340 pages.

CLARK, R.N., SWAYZE, G.A., LIVO, K.E., KOKALY, R.F., SUTLEY, S.J., DALTON, J.B., MCDOUGAL, R.R. & GENT, C.A.: Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems. *Journal of Geophysical Research*, 2003, Vol. 108, pp. 1-44.

CLOUTIS, E., HAWTHORNE, F., MERTZMAN, S., KRENN, K., CRAIG, M., MARCINO, D., METHOT, M., STRONG, J., MUSTARD, J., BLANEY, D., III, J.B. & VILAS, F.: Detection and discrimination of sulfate minerals using reflectance spectroscopy. *Icarus*, 2006, Vol. 184, pp. 121-157.

COCKS, T., JENSSEN, R., STEWART, A., WILSON, I. & SHIELDS, T.: The HyMap airborne hyperspectral sensor: the system, calibration and performance. 1st EARSEL Workshop on Imaging Spectroscopy, Zurich, 1998.

Council for Geoscience (CGS): Atlas on geological storage of carbon dioxide in South Africa. 2010, compiled by CLOETE, M., Council for Geoscience, Pretoria, South Africa.

COWLING, R. & VLOK, J.: On the origin of southern African subtropical thicket vegetation. *South African Journal of Botany*, 2005, Vol. 71, pp. 1-23.

COWLING, R., LOMBARD, A., ROUGET, M., KERLEY, G., WOLF, T., SIMS-CASTLEY, R., KNIGHT, A., VLOK, J., PIERCE, S., BOSHOFF, A. & WILSON, S.: A conservation assessment for the Subtropical Thicket Biome. TERU Reprt 43, Terrestrial Ecology Research Unit, Zoology Department, University of Port Elizabeth, South Africa, 2003.

CROWLEY, J.K., SWAYZE, G.A.: Mapping minerals, amorphous materials, environmental materials, vegetation, water, ice, and other materials: the USGS Tricorder Algorithm. In: Summaries of the Fifth Annual JPL Airborne Earth Science Workshop, JPL Publication 95-1, pp. 39–40, 1995.

CROWLEY, J., BRICKEY, D. & LAWRENCE, C.: Airborne Imaging Spectrometer Data of the Ruby Mountains, Montana: Mineral Discrimination Using Relative Absorption Band-Depth Images. *Remote Sensing of Environment*, 1989, Vol. 29, pp. 121-134.

DALAL, R. & HENRY, R.: Simultaneous determination of moisture or organic carbon, and total nitrogen by near infrared reflectance spectroscopy. *Soil Science Society of America Journal*, 1986, Vol. 50, pp. 120-123.

DE JONG, S., ADDINK, E. & VAN BEEK, L.P.H., D.D.: Physical characterization, spectral response and remotely sensed mapping of Mediterranean soil surface crusts. *Catena*, 2011, Vol. 86, pp. 24-35.

DEHAAN, R. & TAYLOR, G.: Field-derived spectra of salinized soils and vegetation as indicators of irrigation-induced soil salinization. *Remote Sensing of Environment*, 2002, Vol. 80, pp. 406-417.

DEMATTÊ, J. & GRACIA, G.: Alteration of soil properties trough a weathering sequence as evaluated by spectral reflectance. *Soil Science Society of America Journal*, 1999, Vol. 63, pp. 327-342.

DUNCAN, A. & MARSH, J.: The Karroo igneous Province. In: JOHNSON, M., ANHAEUSSER, C. & THOMAS, R. (eds.): The Geology of South Africa. Geological Society of South Africa, Johannesburg, and Council for Geoscience, Pretoria, 2006, pp. 501-520.

ELVIDGE, C.: Visible and near infrared reflectance characteristics of dry plant materials. *International Journal of Remote Sensing*, 1990, Vol. 11, pp. 1775-1795.

ESHEL, G., LEVY, G.J. & SINGER, M.J.: Spectral Reflectance Properties of Crusted Soils under Solar Illumination. *Soil Science Society of America Journal*, 2004, Vol. 68, pp. 1982-1991.

ESWARAN, H., R. LAL AND P.F. REICH: Land degradation: an overview. In: BRIDGES, E.M., HANNAM, I.D., OLDEMAN, L.R., PENING DE VRIES, F.W.T., SCHERR, S.J. AND SOMPATPANIT, S. (eds.). Responses to Land Degradation. Proceedings of 2nd International Conference on Land Degradation and Desertification, Khon Kaen, Thailand. Oxford Press, New Delhi, India, 2001.

Food and Agriculture Organization of the United Nations: Guidelines for soil description. 4th edition, Rome, 2006, FAO.

GALVAO, L., PIZARRO, M. & EPIPHANIO, J.: Variations in Reflectance of Tropical Soils: Spectral-Chemical Composition Relationships from AVIRIS data. *Remote Sensing of Environment*, 1998, Vol. 63, pp. 166-181.

GEE, G.W. & BAUDER, J.W.: Particle-size analysis. In: A. KLUTE (ed.): Methods of soil analysis. Part 1. 2nd edition, Agron. Monogr. 9., ASA and SSSA, Madison, USA, p. 383-411, 1986.

GEERING, H. & BING SO, H.: Texture. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition, CRC Press, 2006, pp. 1759-1764.

GENU, A. & DEMATTÊ, J.: Determination of soil attribute contents by means of reflected electromagnetic energy. *International Journal of Remote Sensing*, 2006, Vol. 27, pp. 4807-4818.

GOLDSHLEGER, N., BEN-DOR, E., BENYAMINI, Y., AGASSI, M. & BLUMBERG, D.G.: Characterization of soil's structural crust by spectral reflectance in the SWIR region (1.2-2.5 µm). *Terra Nova - Oxford*, 2001, Vol. 13, pp. 12-17.

GOLDSHLEGER, N., BEN-DOR, E., BENYAMINI, Y., BLUMBERG, D. & AGASSI, M.: Spectral properties and hydraulic conductance of soil crusts formed by raindrop impact. *International Journal of Remote Sensing*, 2002, Vol. 23, pp. 3909-3920.

GOMEZ, C., LAGACHERIE, P. & COULOUMA, G.: Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma*, 2008a, Vol. 148, pp. 141-148.

GOMEZ, C., VISCARRA ROSSEL, R. & MCBRATNEY, A.: Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma*, 2008b, Vol. 146, pp. 403-411.

GREENLAND, D.: Soil: The Essence of Life and Its Interaction with Other Natural Resources. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition, CRC Press, 2006, pp. 1642-1643.

GROVE, C., HOOK, S. & PAYLOR II, E.: Laboratory reflectance spectra of 160 minerals, 0.4 to 2.5 micrometers. Jet Propulsion Laboratory, National Aeronautics and Space Administration, 1992, JPL Publication 92-2.

GUANTER, L., SEGL, K. & KAUFMANN, H.: Simulation of Optical Remote-Sensing Scenes with Application to the EnMAP Hyperspectral Mission. *IEEE Transactions on Geoscience and Remote Sensing*, 2009, Vol. 47, pp. 2340-2351.

GURALNICK, L., RORABAUGH, P. & HANSCOM, I.Z.: Seasonal Shifts of Photosynthesis in Portulacaria afra (L.) Jacq. *Plant Physiology*, 1984, Vol. 76, pp. 643-646.

HAALAND, D. & THOMAS, E.: Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, 1988, Vol. 60, pp. 1193-1202.

HAUBROCK, S.: Quantification of soil properties for analyzing surface processes using spectroscopy and laser scanning. PhD Thesis, Humboldt University Berlin, Germany, 2009.

HAUBROCK, S., CHABRILLAT, S., LEMMNITZ, C. & KAUFMANN, H.: Surface soil moisture quantification models from reflectance data under field conditions. *International Journal of Remote Sensing*, 2008, Vol. 29, pp. 3-29.

HILL, J. & SCHÜTT, B.: Mapping Complex Patterns of Erosion and Stability in Dry Mediterranean Ecosystems. *Remote Sensing of Environment*, 2000, Vol. 74, pp. 557-569.

HILL, J., MÉGIER, J. & MEHL, W.: Land degradation, soil erosion and desertification monitoring in Mediterranean ecosystems. *Remote Sensing Reviews*, 1995, Vol. 12, pp. 107-130.

HOARE, D., MUCINA, L., RUTHERFORD, M., VLOK, J., EUSTON-BROWN, D., PALMER, A., POWRIE, L., LECHMERE-OERTEL, R., PROCHES, S.M., DOLD, A. & WARD, R.: Albany Thicket Biome. In: MUCINA, L. & RUTHERFORD, M.C.: The Vegetation of South Africa, Lesotho and Swaziland. South African National Biodiversity Institute, Pretoria, South Africa, 2006, pp. 541-567.

HOOK, S.J., GREEN, R.O., MIDDLETON, E.M., BRUCE, C. & FOOTE, M.: Summary of HyspIRI Mission Concept and Instruments. 2012 HyspIRI Symposium, 16/17 May 2012.

HUMMEL, J., SUDDUTH, K. & HOLLINGER, S.: Soil moisture and organic matter prediction of surface and subsurface soils using an NIR soil sensor. *Computers and electronics in Agriculture*, 2001, Vol. 32, pp. 149-165.

HUNT, G.: Spectral signatures of particulate minerals in the visible and near infrared. *Geophysics*, 1977, Vol. 42, pp. 501-513.

HUNT, G.: Spectroscopic properties of rocks and minerals. In: CARMICHAEL, R. (ed.): Handbook of Physical Properties of Rocks. CRC Press, Boca Raton, 1982, Vol. 1, pp. 295-385.

Hunt, G. & Salisbury, J.: Visible and near-infrared spectra of minerals and rocks: 1. Silicate Minerals. *Modern Geology*, 1970, Vol. 1, pp. 283-300.

HUNT, G., SALIBURY, J. & LENHOFF, C.: Visible and near-infrared spectra of minerals and rocks: III. Oxides and Hydroxides. *Modern Geology*, 1971, Vol. 2, pp. 195-205.

HUNTINGTON, T.: Available Water Capacity and Soil Organic Matter. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition, CRC Press, 2006, pp. 139-143.

IPCC (INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE): Land use, land use change, and forestry. A Special Report of the IPCC, 2000.

IPCC (INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE): Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC. Cambridge University Press, Cambridge, United Kingdom, 2007a.

IPCC (INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE): Climate Change 2007 – Synthesis Report. Report at IPCC Plenary XXVII, Valencia, Spain, 2007b.

ISLAM, K., SINGH, B. & MCBRATNEY, A.: Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Australian Journal of Soil Research*, 2003, Vol. 41, pp. 1101-1114.

IUSS Working Group (WRB): World Reference Base for Soil Resources 2006. First update 2007, World Soil Resources Reports No. 10, 2007, FAO, Rome, Italy.

JOHNSON, M.R., V.V.C., VISSER, J., COLE, D., DE V. WICKENS, H., CHRISTIE, A., ROBERTS, D. & BRANDL, G.: Sedimentary rocks of the Karroo Supergroup. In: JOHNSON, M., ANHAEUSSER, C. & THOMAS, R. (eds.): The Geology of South Africa. Geological Society of South Africa, Johannesburg, and Council for Geoscience, Pretoria, 2006, pp. 461-500.

KARATHANASIS, A.: Clay Minerals: Weathering and Alteration of. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition,CRC Press, 2006, pp. 281-286.

KESHAVA, N. & MUSTARD, J.: Spectral Unmixing. *IEEE Signal Processing Magazine*, 2002, Vol. 19, pp. 44-57.

KESSLER, W.: Multivariate Datenanalyse. Wiley-VCH, 2007, pp. 325.

KOMPTER, E.: Pedological site investigation and characterisation using spectroscopic methods - Examples of a South African ecosystem endangered by degradation. Master thesis, Friedrich Schiller University Jena, Germany, 2012.

KOOISTRA, L., WEHRENS, R., LEUVEN, R.S.E.W. & BUYDENS, L.M.C.: Possibilities of visible-near-infrared spectroscopy for the assessment of soil contamination in river floodplains. *Analytica Chimica Acta*, 2001, Vol. 446, pp. 97-105.

KOOISTRA, L., WANDERS, J., EPEMA, G., LEUVEN, R., WEHRENS, R. & BUYDENS, L.: The potential of field spectroscopy for the assessment of sediment properties in river floodplains. *Analytica Chimica Acta*, 2003, Vol. 484, pp. 189-200.

KRONBERG, P.: Fernerkundung der Erde - Grundlagen und Methoden des Remote Sensing in der Geologie. Enke, 1985, pp. 394.

KUNTZE, H., ROESCHMANN, G. & SCHWERDTFEGER, G.: Bodenkunde. UTB Ulmer, 1988, pp. 568.

LAGACHERIE, P., BARET, F., FERET, J.-B., MADEIRA NETTO, J. & ROBBEZ-MASSON, J.: Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. *Remote Sensing of Environment*, 2008, Vol. 112, pp. 825-835.

LAL, R.: Soil carbon dynamics in cropland and rangeland. *Environmental Pollution*, 2002, Vol. 116, pp. 353-362.

LAL, R.: Soil carbon sequestration to mitigate climate change. *Geoderma*, 2004, Vol. 123, pp. 1-22.

LAL, R.: Preface to the Second Edition. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition, CRC Press, 2006.

LECHMERE-OERTEL, R., KERLEY, G. & COWLING, R.: Patterns and implications of transformation in semi-arid succulent thicket, South Africa. *Journal of Arid Environments*, 2005, Vol. 62, pp. 459-474.

LLOYD, J., VAN DEN BERG, E. & PALMER, A.: Patterns of Transformation and Degradation in the Thicket Biome, South Africa. Terrestrial Ecology Research Unit, University of Port Elizabeth, 2002, Report No. 39.

LOBELL, D. & ASNER, G.: Moisture Effects on Soil Reflectance. Soil Science Society of America Journal, 2002, Vol. 66, pp. 722-727.

LOW, A.B. & REBELO, A.G. (eds): The Vegetation of South Africa, Lesotho and Swaziland. Department of Environmental Affairs and Tourism, Pretoria, 1996.

MARAIS, C., COWLING, R.M., POWELL, M., MILLS, A.: Establishing the platform for a carbon sequestration market in South Africa: The Working for Woodlands Subtropical Thicket Restoration Programme. XIII World Forestry Congress Buenos Aires, Argentina, 2009.

MARTENS, H. & NAES, T.: Multivariate Calibration. John Wiley & Sons, 1989, pp. 419.

MCHUNU, S.E.: Distribution and Stability of Soil Carbon in Spekboom Thicket, Eastern Cape, South Africa. Master Thesis, University of Stellenbosch, South Africa.

MCINTYRE, D.: Permeability measurements of soil crusts formed by raindrop impact. *Soil Science*, 1958, Vol. 85, pp. 185-189.

MEHRA, O. & JACKSON, M.: Iron oxide removal from soils and clays by a dithionite-citrate system buffered with sodium bicarbonate. *Clays and clay minerals*, 1958, Vol. 7, pp. 317-327.

METTERNICHT, G. & ZINCK, J.: Remote sensing of soil salinity: potentials and constraints. *Remote Sensing of Environment*, 2003, Vol. 85, pp. 1-20.

METZ, B., DAVIDSON, O.R., BOSCH, P.R., DAVE, R. & MEYER, L.A. (eds.): Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, USA, 2007.

MILLS, A., Personal communication, 2009.

MILLS, A.J. & COWLING, R.M.: Rate of Carbon Sequestration at Two Thicket Restoration Sites in the Eastern Cape, South Africa. *Restoration Ecology*, 2006, Vol. 14, pp. 38-49.

MILLS, A.J. & COWLING, R.M.: Below-ground carbon stocks in intact and transformed subtropical thicket landscapes in semi-arid South Africa. *Journal of Arid Environments*, 2010, Vol. 74, pp. 93-100.

MILLS, A. & FEY, M.: Declining soil quality in South Africa: effects of land use on soil organic matter and surface crusting. *South African Journal of Science*, 2003, Vol. 99, pp. 429-436.

MILLS, A. & FREY, M.: Transformation of thicket to savanna reduces soil quality in the Eastern Cape, South Africa. *Plant and Soil*, 2004a, Vol. 265, pp. 153-163.

MILLS, A. & FEY, M.: Effects of vegetation cover on the tendency of soil to crust in South Africa. *Soil Use and Management*, 2004b, Vol. 20, pp. 308-317.

MILLS, A., O'CONNOR, T., DONALDSON, J., FEY, M., SKOWNO, A., SIGWELA, A., LECHMERE-OERTEL, R. & BOSENBERG, J.: Ecosystem carbon storage under different land uses in three semi-arid shrublands and a mesic grassland in South Africa. *South African Journal of Plant and Soil*, 2005a, Vol. 22, pp. 183-190.

MILLS, A.J., COWLING, R., FEY, M., KERLEY, G., DONALDSON, J., LECHMERE-OERTEL, R., SIGWELA, A., SKOWNO, A. & RUNDEL, P.: Effects of goat pastoralism on ecosystem carbon storage in semiarid thicket, Eastern Cape, South Africa. *Austral. Ecology*, 2005b, Vol. 30, pp. 797-804.

MILLS, A.J., TURPIE, J., COWLING, R.M., MARAIS, C., KERLEY, G.I.H., LECHMERE-OERTEL, R.G., SIGWELA, A.M. & POWELL, M.: Assessing costs, benefits, and feasibility of restoring natural capital in Subtropical Thicket in South Africa. In: ARONSON, J., MILTON, S. & BLIGNAUT, J. (eds.): Restoring Natural Capital: Science, Business and Practice. Society for Ecological Restoration. Island Press. Washington, 2007, pp. 179–187.

MING, D.W.: Carbonates. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition,CRC Press, 2006, pp. 281-286.

MORRIS, R., LAUER, H., LAWSON, C., GIBSON JR., E., NACE, G. & STEWART, C.: Spectral and other physicochemical Properties of submicron Powders of Hematite (α-Fe2O3), Maghemite (γ-Fe2O3), Magnetite (Fe3O4), Goethite (α-FeOOH), and Lepidocrocite (γ-FeOOH). *Journal of Geophysical Research*, 1985, Vol. 90, pp. 3126-3144.

MUCINA, L. & RUTHERFORD, M.C.: The vegetation of South Africa, Lesotho and Swaziland. South African National Biodiversity Institute, Pretoria, South Africa, 2006, pp. 808.

MULDER, V., DE BRUIN, S., SCHAEPMAN, M. & MAYR, T.: The use of remote sensing in soil and terrain mapping - A review. *Geoderma*, 2011, Vol. 162, pp. 1-19.

MURPHY, R. & WADGE, G.: The effects of vegetation on the ability to map soils using imaging spectrometer data. *International Journal of Remote Sensing*, 1994, Vol. 15, pp. 63-86.

NEWTON, A., SHONE, R. & BOOTH, P.: The Cape Fold Belt. In: JOHNSON, M., ANHAEUSSER, C. & THOMAS, R. (eds.): The Geology of South Africa. Geological Society of South Africa, Johannesburg, and Council for Geoscience, Pretoria, 2006, pp. 521-530.

OHGI, N., IWASAKI, A., KAWASHIMA, T. & INADA, H.: Japanese Hyper-multi spectral Mission. 2010, Proceedings of IEEE IGARSS, July 2010, pp. 3756-3759.

PALACIOS-ORUETA, A. & USTIN, S.: Remote Sensing of Soil Properties in the Santa Monica Mountains I. Spectral Analysis. *Remote Sensing of Environment*, 1998, Vol. 65, pp. 170-183.

POWELL, M.: Restoration of degraded subtropical thickets in the Baviaanskloof Megareserve, South Africa: the role of carbon stocks and Portulacaria afra survivorship. Master Thesis, Rhodes University, Grahamstown, South Africa, 2009.

POWELL, M., Personal communication, 2011.

POWELL, M., SKOWNO,A., MILLS, A. & KNOWLES, T.: Providing a context for the rehabilitation of Subtropical Thicket. In: WILSON, S. (ed.): Proceedings of the 2004 Thicket Forum. Centre for African Conservation Ecology, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa, 2006a, Report No. 54.

POWELL, M., MILLS, A. & MARAIS, C.: Carbon sequestration and restoration: challenges and opportunities in subtropical Thicket. Proceedings of Natural Forests and Woodlands Savanna Symposium, IV. Port Elizabeth, 2006b, Pg. 214-223.

QAFOKU, N. & AMONETTE, J.: Iron Oxides. In: LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition, CRC Press, 2006, pp. 920-925.

REEVES III, J., MCCARTY, G. & MIMMO, T.: The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soils. *Environmental Pollution*, 2002, Vol. 116, pp. 277-284.

RICE, C.: Organic Matter and Nutrient Dynamics. LAL, R. (ed.): Encyclopedia of Soil Science. 2nd Edition, CRC Press, 2006, pp. 1180-1183.

RICHTER, N.: Pedogenic iron oxide determination of soil surfaces from laboratory spectroscopy and HyMap image data. PhD Thesis, Humboldt University Berlin, Germany, 2010a.

RICHTER, R.: Atmospheric/Topographic Correction for Airborne Imagery. German Aerospace Center, Oberpfaffenhofen, Germany, Report No. DLR-IB 565-02/10, 2010b.

RICHTER, N., JARMER, T., CHABRILLAT, S., OYONARTE, C., HOSTERT, P. & KAUFMANN, H.: Free Iron Oxide Determination in Mediterranean Soils using Diffuse Reflectance Spectroscopy. *Soil Science Society of America Journal*, 2009, Vol. 73, pp. 72-81.

ROGGE, D. & RIVARD, B.: Iterative spatial filtering for reducing intra-class variability and noise. Proceedings of 2nd WHISPERS, Reykjavik, Iceland, 2010.

ROGGE, D., RIVARD, B., ZHANG, J. & FENG, J.: Iterative Spectral Unmixing for Optimizing Per-Pixel Endmember Sets. *IEEE Transactions on Geoscience and Remote Sensing*, 2006, Vol. 44, pp. 3725-3736.

ROGGE, D., RIVARD, B., ZHANG, J., SANCHEZ, A., HARRIS, J. & FENG, J.: Integration of spatial–spectral information for the improved extraction of endmembers. *Remote Sensing of Environment*, 2007, Vol. 110, pp. 287-303.

ROGGE, D., BACHMANN, M., RIVARD, B. & FENG, J.: Hyperspectral flight-line leveling and scattering correction for image mosaics. Proceedings of IGARS, Munich, 2012.

SABINE, C., FEELY, R., GRUBER, N., KEY, R., LEE, K., BULLISTER, J., WANNINKHOF, R., WONG, C., WALLACE, D., TILBROOK, B., MILLERO, F., PENG, T.-H., KOZYR, A., ONO, T. & RIOS, A.: The oceanic sink for anthropogenic CO_2. *Science*, 2004, Vol. 305, pp. 367-371.

SAVITZKY, A. & GOLAY, M.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 1964, Vol. 36, pp. 1627-1639.

SCHACHTSCHABEL, P., H.-P., B., BRÜMMER, G., HARTGE, K. & SCHWERTMANN, U.: Lehrbuch der Bodenkunde. Spektrum akademischer Verlag, Heidelberg, 2002.

SCHLÄPFER, D. & RICHTER, R.: Geo-atmospheric processing of airborne imaging spectrometry data. Part 1: Parametric orthorectification. *International Journal of Remote Sensing*, 2002, Vol. 23, pp. 2609-2630.

SCHWANGHART, W. & JARMER, T.: Linking spatial patterns of soil organic carbon to topography — A case study from south-eastern Spain. *Geomorphology*, 2011, Vol. 126, pp. 252-263.

SELIGE, T., BÖHNER, J. & SCHMIDHALTER, U.: High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modelling procedures. *Geoderma*, 2006, Vol. 136, pp. 235-244.

SETTLE, J. & DRAKE, N.: Linear mixing and the estimation of ground cover proportions. *International Journal of Remote Sensing*, 1993, Vol. 14, pp. 1159–1177.

SHEPHERD, K. & WALSH, M.: Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*, 2002, Vol. 66, pp. 988-998.

SHONE, R.: Onshore post-Karroo Mesozoic Deposits. In: JOHNSON, M., ANHAEUSSER, C. & THOMAS, R. (eds.): The Geology of South Africa. Geological Society of South Africa, Johannesburg, and Council for Geoscience, Pretoria, 2006, pp. 541-552.

STENBERG, B., VISCARRA ROSSEL, R., MOUAZEN, A. & WETTERLIND, J.: Visible and Near Infrared Spectroscopy in Soil Science. *Advances in Agronomy*, 2010, Vol. 107, pp. 163-215.

STEVENS, A., VAN WESEMAEL, B., BARTHOLOMEUS, H., ROSILLON, D., TYCHON, B. & BEN-DOR, E.: Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma*, 2008, Vol. 144, pp. 395-404.

STEVENS, A., UDELHOVEN, T., DENIS, A., TYCHON, B., LIOY, R., HOFFMANN, L. & VAN WESEMAEL, B.: Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma*, 2010, Vol. 158, pp. 32-45.

STONER, E. & BAUMGARDNER, M.: Characteristic Variations in Reflectance of Surface Soils. *Soil Science Society of America Journal*, 1981, Vol. 45, pp. 1161-1165.

STONER, E.R., BAUMGARDNER, M.F., BIEHL, L.L. & ROINSON, B.F.: Atlas of soil reflectance properties. Research Bulletin 962. Agricultural Experiment Station, Prudue University, United States, 75 p., 1980.

STROBL, P., MÜLLER, A. SCHLÄPFER, D. & SCHAEPMAN, M.: Laboratory Calibration and Inflight Validation of the Digital Airborne Imaging Spectrometer DAIS 7915 for the 1996 Flight Season. 1997, Proceedings of SPIE, Vol. 3071 pp. 225-235.

STUFFLER, T., KAUFMANN, C., HOFER, S., FÖRSTER, K., SCHREIER, G., MUELLER, A., ECKARDT, A., BACH, H., PENNÉ, B., BENZ, U. & HAYDN, R.: The EnMAP hyperspectral imager—An advanced optical payload for future applications in Earth observation programmes. *Acta Astronautica*, 2007, Vol. 61, pp. 115-120.

SUDDUTH, K. & HUMMEL, J.: Soil Organic Matter, CEC and moisture sensing with a portable NIR spectrophotometer. *Transactions of the ASAE*, 1993, Vol. 36, pp. 1571-1582.

SWART, M., HOBSON, F. & STURAT-HILL, G.: Establishment of Spekboom. *Dohne Bulletin*, 1994, Vol. 3, pp. 10-13.

THAMM, A. & JOHNSON, M.: The Cape Supergroup. In: JOHNSON, M., ANHAEUSSER, C. & THOMAS, R. (eds.): The Geology of South Africa. Geological Society of South Africa, Johannesburg, and Council for Geoscience, Pretoria, 2006, pp. 443-460.

TIESSEN, H., CUEVAS, E. & CHACON, P.: The role of soil organic matter in sustaining soil fertility. *Nature*, 1994, Vol. 371, pp. 783-785.

TORRENT, J., SCHWERTMANN, U., FECHTER, H. & ALFEREZ, F.: Quantitative Relationships between Soil Color and Hematite Content. *Soil Science*, 1983, Vol. 136, pp. 354-358.

TURPIE, J.: An Ecological-economic appraisal of Conservation on Commercial Farm Land in Four Areas of South Africa. Final Report of the Ecological-Economics Research Component of the Conservation Farming Project, submitted to National Botanical Institute, 2003.

UDELHOVEN, T., EMMERLING, C. & JARMER, T.: Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study. *Plant and Soil*, 2003, Vol. 251, pp. 319-329.

VAN DER MEER, F.: Analysis of spectral absorption features in hyperspectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, 2004, Vol. 5, pp. 55-68.

VAN DER MEER, F. & DE JONG, S. (eds.): Imaging Spectrometry - Basic Principles and Prospective Applications. Kluwer Academic Publishers, 2001, Vol. 4, pp. 403.

VAN DER MEER, F., VAN DER WERFF, H., VAN RUITENBEEK, F., HECKER, C., BAKKER, W., NOOMEN, M., VAN DER MEIJDE, M., CARRANZA, E., BOUDEWIJN DE SMETH, J. & WOLDAI, T.: Multi- and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, 2012, pp. 112-128.

VELASCO-FORERO, S. & MANIAN, V.: Improving Hyperspectral Image Classification Using Spatial Preprocessing. *IEEE Geoscience and Remote Sensing Letters*, 2009, Vol. 6.

VISCARRA ROSSEL, R.: Robust modelling of soil diffuse reflectance spectra by bagging-partial least squares regression. *Near Infrared Spectroscopy*, 2007, Vol. 15, pp. 39-47.

VISCARRA ROSSEL, R.: ParLeS: Software for chemometric analysis of spectrometric data. *Chemometrics and Intelligent Laboratory Systems*, 2008a, Vol. 90, pp. 72-83.

VISCARRA ROSSEL, R.: The Soil Spectroscopy Group and the development of a global soil spectral library. In: MINASNY, B. (ed.): Pedometron, Newsletter of the Pedometrics Comission of the IUSS, Vol. 25, 2008b.

VISCARRA ROSSEL, R. & BEHRENS, T.: Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 2010, Vol. 158, pp. 46-54.

VISCARRA ROSSEL, R. & CHEN, C.: Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. *Remote Sensing of Environment*, 2011, Vol. 115, pp. 1443-1455.

VISCARRA ROSSEL, R., WALVOORT, D., MCBRATNEY, A., JANIK, L. & SKJEMSTAD, J.: Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 2006, Vol. 131, pp. 59-75.

VISCARRA ROSSEL, R., ADAMCHUK, V., SUDDUTH, K., MCKENZIE, N. & LOBSEY, C.: Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. *Advances in Agronomy*, 2011, Vol. 113, pp. 237-282.

VLOK, J. & EUSTON-BROWN, D.: The patterns within, and the ecological progresses that sustain, the Subtropical Thicket vegetation in the planning domain for the Subtropical Thicket Ecosystem Planning (STEP) project. Terrestrial Ecology Research Unit, University of Port Elizabeth, Port Elizabeth 6031, 2002, Report No. 40.

VLOK, J.H.J. AND EUSTON-BROWN, D.I.W. AND COWLING, R.M.: Acocks' Valley Bushveld 50 years on: new perspectives on the delimitation, characterisation and origin of subtropicsl thicket vegetation. *South African Journal of Botany*, 2003, Vol. 69, pp. 27-51.

VOHLAND, M., BESOLD, J., HILL, J. & FRÜND, H.-C.: Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma*, 2011, Vol. 166, pp. 198-205.

WAISER, T., MORGAN, C., BROWN, D. & HALLMARK, C.: In Situ Characterisation of soil clay content with visible near-infrared diffuse reflectance spectroscopy. *Soil Science Society of America Journal*, 2007, Vol. 71, pp. 389-396.

WALKLEY, A. & BLACK, I.: An Examination of the Degtjareff Method for Determining Soil Organic Matter, and A Proposed Modification of the Chromic Acid Titration Method. *Soil Science*, 1934, Vol. 37, pp. 29-38.

WANG, Y., NIU, R. & YU, X.: Anisotropic Diffusion for Hyperspectral Imagery Enhancement. *IEEE Sensors Journal*, 2010, Vol. 10, pp. 469-477.

WHITING, M., LI, L. & USTIN, S.: Predicting water content using Gaussian model on soil spectra. *Remote Sensing of Environment*, 2004, Vol. 89, pp. 535-552.

WOLD, S. AND MARTENS, H. AND WOLD, H.: The multivariate calibration problem in chemistry solved by the PLS method. In: RUHE, A. & KAGSTROM, B. (eds.): *Proceedings of the Conference on Matrix Pencils, Lecture Notes in Mathematics*, Springer-Verlag, Heidelberg, Germany, pp. 286-293, 1983.

WOLD, S., SJÖSTRÖM, M. & ERIKSSON, L.: PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 2001, Vol. 58, pp. 109-130.

Online references:

Note that any references to online available literature have been last checked for availability on 24th October 2012.

www.ercap.co.za

www.wikipedia.de

www.biota-africa.org

Appendix

A Technical details and statistics of available datasets

The following includes an overview of the collected field reference information (A.1) and the results of the chemical analysis for each sample site (A.2). Table A-3 also gives statistics of the chemical analysis of the two sampling trips and of the entire sample population. Appendix A.3 gives an overview of the soil survey conducted in 2011.

A.1 Field data collection

Table A-1: Site specific information collected during the field surveys in June and September/October 2009.

Location information & Site description	Sample site no. Location, farm portion, GPS coordinates: latitude, longitude, elevation, error (GARMIN ForeTrex 201) Date and time of spectrometric measurements, weather conditions Terrain slope: undulating (minor unevenness), flat (0-10°), medium (10-30°), hilly (> 30°) Vegetation type/pattern Land use: grazing (goats, sheep, cows), game, unknown Comments, e.g. erosion channels, floodplains, etc.
Soil and surface description	Moisture conditions: surface/crust: dry, slightly moist, moist, very moist under surface layer/crust: dry, slightly moist, moist Surface conditions (1 x 1 m): bare soil, with stones Crusting: none, slight (present but crushes immediately), medium, strong (very hard, with spatula not easily breakable) depth of crust [cm] Quantity of small stones (1 to 5 mm) in 1 x 1 m raster: none / with stones quantity: low (< 2 %), medium (2 – 30 %), high (> 30 %) cohesion: freely supported/overlying, baked-in (stones are covered with soil material), both occurring size [mm] type: mudstone, sandstone, mixed, … Quantity of boulders (stones > 5 cm) in 10 x 10 m raster: none / with stones quantity: % of surface coverage size [cm] type: mudstone, sandstone, mixed, …
Spectroscopic measurements	Spectra no. (ASD FieldSpec Pro 2) Description: bare soil (bs), soil with stones (ss), soil with vegetation (sv, i.e. for homogenous grass coverage), vegetation sample (veg), stone sample (st) Photo no.: shot on measured center spot from above, overview of sampling site with neighborhood, shots in every direction (N, E, S, W) Height of measurement: 1.15 m (normal height), 0.5 m (alternative height for small center spots)
Fractional cover assessment	Assessment of cover fraction in % of measured center spot (1x1 m) and neighborhood (10x10 m around measured spot) Discriminated units: bare soil, stones (see above stone coverage 10x10 m), litter, grass (< 0.3 m): green/mixed/dry herbs (0.3 to 0.8 m): green/mixed/dry shrubs: < 0.5 m / 0.5 to 3 m, green/mixed/dry trees (3 to 7 m): green/mixed/dry spekboom (always green) other regionally frequently occurring plants: aloe, different succulents, … With: green: > 66 % photosynthetic active plant stage, mixed: 33 to 66 % green/dry vegetation dry: > 66 % non-photosynthetic active vegetation stage

A.2 Laboratory chemical analysis of collected soil samples

Chemical concentrations Table A-2 and Table A-3 are given in mass %. They can be converted to g/kg by multiplication by 10, e.g. 1 % C_{org} = 10 g/kg.

Table A-2: Position and chemical analysis of samples taken in July and September/October 2009 (analysis provided by Central Analytical Facility, University of Stellenbosch, South Africa). Coordinates are given in UTM projection (Zone 35 South, Datum: WGS-84).

July 2009						September/October 2009					
Sample	x	y	C_{org} [%]	Fe_d [%]	Clay [%]	Sample	x	y	C_{org} [%]	Fe_d [%]	Clay [%]
4	347120.93	6290255.24	1.81	3.76	3.80	25	344894.08	6300253.09	1.80	2.63	5.80
6	345722.44	6290270.66	1.91	3.62	0.40	26	345367.96	6300454.78	1.34	6.10	14.80
11	347054.52	6292625.55	3.30	2.83	0.80	30	344950.69	6302000.90	3.01	1.08	3.00
12	346926.03	6292700.05	2.15	2.15	1.00	50	341048.92	6346854.05	2.41	3.15	11.00
13	346801.58	6292753.53	2.42	2.00	0.20	51	341403.15	6346908.57	0.87	2.98	10.80
14	346494.47	6293131.30	1.04	2.10	0.20	106	338581.68	6385638.09	1.32	2.48	3.80
15	346232.13	6293264.66	0.42	2.05	1.50	200	No GPS position available		0.45	2.35	6.40
18	345245.30	6295920.83	1.03	3.58	5.80	201	341825.77	6354219.80	1.57	1.35	5.00
19	345377.09	6295934.03	2.97	2.07	1.80	202	341091.44	6354637.19	0.75	3.75	6.80
20	346371.26	6296551.06	1.23	2.09	1.80	203	340716.54	6354397.12	1.13	3.68	11.80
21	346300.72	6296660.86	1.80	2.48	6.80	204	340477.92	6346739.43	0.66	2.85	3.20
22	345730.09	6298802.39	5.74	2.00	0.20	205	342239.18	6347629.63	0.66	2.88	1.20
40	341500.16	6338998.95	0.28	2.38	10.80	206	340632.51	6347636.97	0.94	2.08	3.80
41	341461.64	6339648.26	0.22	2.14	7.80	207	342979.48	6347038.15	0.41	3.75	4.80
42	341505.59	6340280.05	0.31	2.37	13.80	208	343397.00	6346582.33	0.83	3.33	12.80
43	341461.20	6341061.25	0.25	1.53	8.80	209	342000.64	6363332.47	1.24	3.03	0.80
44	341647.06	6342102.37	1.67	1.95	3.80	210	341520.29	6363633.10	1.20	2.68	2.80
45	341989.90	6342795.54	1.41	1.55	3.80	211	341446.70	6364602.35	1.03	1.98	0.60
46	341680.79	6343195.37	1.28	1.38	1.80	212	341069.19	6365431.41	1.57	3.08	0.60
47	340859.46	6344359.95	0.43	1.83	7.80	213	340607.81	6366406.61	1.47	2.23	0.80
53	341441.59	6348348.79	1.18	1.73	7.80	214	340201.49	6366170.47	3.17	4.15	12.80
54	341569.07	6349021.84	1.29	2.84	9.80	215	340195.20	6365285.34	1.67	3.75	10.80
55	341483.88	6349725.85	0.82	2.29	0.20	216	341292.37	6364010.96	0.68	2.88	5.00
56	340972.54	6350131.28	1.06	1.71	1.80	217	341251.30	6363129.70	0.54	2.30	3.80
58	340875.64	6352828.12	0.85	3.77	15.80	218	341385.43	6362413.18	0.93	2.30	0.60
59	340669.25	6350156.32	1.28	2.55	23.80	219	341418.88	6371774.15	2.41	4.60	6.80
60	340799.87	6350114.07	1.05	2.56	8.80	220	341267.37	6373454.15	1.89	1.93	0.80
63	341531.51	6359144.89	1.01	2.28	4.80	221	340766.42	6373664.60	0.36	3.75	5.00
64	341361.19	6359437.17	0.22	3.04	7.80	222	340156.24	6374139.44	1.29	2.25	8.80
65	340824.61	6359811.16	0.61	2.96	3.80	223	339809.88	6375277.28	0.48	2.20	0.80
72	339698.38	6362552.31	1.04	2.42	2.80	224	338753.18	6382003.20	3.10	2.25	0.80
86	339253.91	6373381.78	0.65	3.25	8.00	225	339409.35	6382197.92	0.56	3.68	13.20
87	339268.39	6373817.88	0.93	1.85	1.00	226	339944.29	6382467.17	0.66	3.63	6.80
88	339390.55	6374375.49	0.53	2.33	15.00	227	339811.49	6383083.88	0.87	3.20	8.80
89	339185.70	6375045.37	0.68	2.55	8.00	228	339320.14	6383017.17	0.46	3.65	14.60
90	338979.97	6375363.66	0.21	3.18	16.00	229	338675.61	6382911.36	2.96	3.98	19.60
92	339184.60	6376269.75	1.59	1.87	3.00	230	339337.30	6384045.53	1.33	3.05	11.60
93	338852.52	6378885.06	0.22	2.32	12.00	231	340419.87	6383749.09	1.00	6.25	19.60
94	338953.70	6379005.37	1.11	2.30	6.00	232	340232.56	6384490.25	0.58	3.00	23.60
95	339172.40	6379453.64	1.48	2.27	7.00	233	341480.09	6358385.46	0.51	3.65	0.60
96	338822.99	6380243.16	1.32	2.22	5.00	234	340913.65	6358351.94	0.47	6.18	17.60
97	338777.79	6380254.63	0.94	2.65	9.00	235	340246.91	6358387.75	0.59	2.43	4.60
98	338748.25	6380341.77	0.51	2.69	13.00	236	340912.13	6357692.02	0.64	3.80	7.60
99	338374.24	6380621.83	0.65	2.65	13.00	237	341135.91	6356751.81	0.89	2.30	0.40
100	338363.17	6381417.94	1.09	2.14	5.00	238	340486.18	6356899.91	1.26	2.30	8.60
101	339177.25	6381355.73	0.91	3.34	4.00	239	340002.51	6356383.00	1.43	2.50	4.80
102	340531.04	6380093.28	2.22	1.87	1.00	240	341822.76	6359770.65	0.83	4.18	13.60

July 2009						September/October 2009					
Sample	x	y	C_{org} [%]	Fe_d [%]	Clay [%]	Sample	x	y	C_{org} [%]	Fe_d [%]	Clay [%]
103	338483.85	6384497.51	1.08	3.21	9.00	241	341146.83	6366718.05	1.66	3.40	6.60
104	338293.86	6384585.37	0.36	3.22	12.00	242	341717.19	6366946.79	1.46	2.85	4.60
108	337927.13	6390060.29	0.49	2.51	4.00	243	341487.60	6367424.44	1.19	3.35	6.60
109	338357.05	6390347.86	0.88	2.25	1.40	244	340953.22	6367461.34	1.80	3.83	14.60
110	338561.83	6390686.10	1.65	2.43	9.00	245	340528.22	6367567.62	3.48	2.93	0.40
111	339125.91	6390632.00	0.88	2.41	7.00	246	347936.35	6286569.13	0.83	4.73	6.60
112	339469.96	6390484.50	0.76	2.85	9.00	247	347165.85	6286256.34	0.82	6.93	11.60
120	347321.07	6283440.43	2.69	6.97	9.00	248	346852.08	6286539.73	1.55	9.95	4.60
121	347320.73	6283403.82	1.62	10.62	13.00	249	346365.15	6286092.74	0.98	5.65	6.60
122	347562.22	6283514.13	1.53	6.10	11.00	250	345809.55	6285529.27	1.29	5.15	8.60
123	346573.82	6292934.03	0.71	1.09	1.00	251	345626.63	6286570.06	0.46	4.08	0.60
124	346280.61	6293196.66	0.46	4.43	3.00	252	346147.64	6289127.26	1.95	4.33	11.60
125	346202.61	6293250.88	0.57	2.67	3.00	253	346637.48	6287826.27	2.48	6.10	10.60
126	347463.11	6292820.58	1.73	3.05	2.00	254	347823.40	6287251.69	0.72	5.63	10.60
127	347207.27	6292838.71	0.23	0.90	1.00	255	347689.95	6288759.14	2.18	6.23	3.60
128	345329.86	6294357.17	0.65	1.29	2.00	256	346287.65	6290715.60	1.93	2.23	5.60
129	345990.55	6293959.57	0.64	2.88	4.00	257	346867.97	6291499.02	0.84	2.38	2.60
130	346461.71	6294371.92	2.09	2.83	2.00	258	345178.55	6301248.11	1.75	4.38	8.60
131	345608.22	6294270.67	1.11	1.97	1.00	259	344548.80	6300904.16	1.15	3.78	6.60
133	338273.46	6380649.02	0.42	3.01	12.00	260	345630.74	6301530.41	0.87	2.63	7.60
134	345524.53	6297468.13	1.22	1.77	2.00	261	344520.21	6302103.80	2.43	2.83	6.60
135	337158.85	6383777.21	1.53	2.04	2.00						
136	337282.46	6382356.32	0.61	2.14	2.00						
137	339801.62	6380844.56	1.51	1.90	2.00						
138	340183.08	6380682.13	0.48	3.31	6.00						
139	340590.31	6380139.70	2.37	7.38	18.00						
140	341519.58	6340281.39	0.59	4.03	14.00						
141	343000.77	6338821.26	0.27	4.19	8.00						
142	343632.33	6338974.44	0.41	2.93	7.00						
143	341458.77	6345258.03	0.68	2.23	4.00						
144	340031.06	6351574.47	1.01	3.61	12.00						
145	340887.22	6352863.80	0.54	3.68	12.00						
147	340572.14	6362465.52	0.51	2.29	2.00						
148	339552.83	6362638.67	1.21	2.11	0.00						
149	339352.93	6362210.64	2.25	2.00	0.00						
150	340102.74	6360529.26	0.26	2.22	2.00						
151	341711.76	6359457.22	0.30	3.62	8.00						
152	338976.36	6389546.07	0.22	2.53	2.00						
153	339479.33	6388095.80	0.35	3.39	6.00						
154	340255.57	6386973.74	0.33	2.71	4.00						
155	339892.16	6387633.32	0.43	3.21	14.00						
156	339010.88	6369654.71	2.12	2.52	0.00						
157	338694.36	6368097.97	0.74	2.67	6.00						
158	339287.38	6367740.52	1.36	1.69	2.00						
159	339148.08	6367610.71	2.51	2.68	10.00						
160	340523.20	6376320.19	0.70	2.13	4.00						
161	339913.51	6376289.30	1.28	2.97	8.00						
162	339293.68	6376074.11	0.52	1.87	2.00						
180	345377.09	6295934.03	5.85	1.49	2.20						

Table A-3: Statistics of chemical analysis of samples taken in July and September/October 2009 and of all samples (analysis provided by Central Analytical Facility, University of Stellenbosch, South Africa).

		C_{org}	Fe_d	Clay
July 2009	**Mean**	1.14	2.72	5.96
	Stdev	0.97	1.31	4.90
	Min	0.21	0.90	0.00
	Max	5.85	10.62	23.80
September/October 2009	**Mean**	1.29	3.54	7.17
	Stdev	0.76	1.50	5.29
	Min	0.36	1.08	0.40
	Max	3.48	9.95	23.60
All chemical reference	**Mean**	1.21	3.07	6.46
	Stdev	0.89	1.44	5.08
	Min	0.21	0.90	0.00
	Max	5.85	10.62	23.80

A.3 Soil profile information collected in 2011

The following sums up results from the soil survey in 2011 that are relevant within this study. The results at length are given in Kompter (2012). Munsell soil color classifications revealed the tendency of predominantly reddish soil colors in the southern part of the study area. pH measurements showed the local presence of calcium carbonate in 30 % of the samples in the southern transect. The analysis of soil constituents' distribution along soil profiles revealed characteristic trends and allowed the classification of soil types. Soil organic carbon contents were generally highest in the upper 10 cm and decreased with depth as was expected due to the C_{org} input through plant litter (see also Mills and Cowling 2010). Some profiles showed a reduction of C_{org} contents in the uppermost topsoil indicating that soil erosion outweighs litter accumulation at these locations. Clay contents were in general higher in the northern transect as result of the predominant geology of sandstones and mudstones in the northern, and sandstones in the southern transect. Within the soil profile, clay sized particles often accumulated between 5 and 25 cm (see Kompter 2012). Table A-4 gives statistics of samples taken from the topsoil layer collected in 2011 from 32 locations in the study area where the soil profiles were measured. Within chemical analysis of these samples, soil iron oxides was determined using EDTA extraction, determining the plant available iron oxides and not the total amount of the free iron oxides as it is done using the CBD extraction. Thus, the determined iron oxides concentrations are not comparable to the ones of the samples collected in 2009 that are used as basis for the model development in this study. Details on the sampling, characteristics of collected samples and the following analysis including the classification of soil types is found in Kompter (2012).

Table A-4: Statistics of the samples collected in 2011 from 32 locations within the study area where soil profiles were measured. Statistics of chemical reference includes surface samples only.

	Soil depth		C_{org}	Clay
	North (21 profiles)	**South (11 profiles)**	**[%]**	**[%]**
Min	5.00	25.00	0.29	0.00
Max	65.00	65.00	3.55	19.20
Mean	32.19	52.27	0.87	4.88
Stdev	19.22	15.42	0.71	4.74

B Pre-processing of datasets

Details of the applied pre-processing steps for the different datasets are shown in the following appendix. This includes the spectral libraries measured in field and laboratory (B.1), the airborne HyMap data (B.2) and the simulated EnMAP satellite data (B.3).

B.1 Field and laboratory spectra

Figure B-1 shows the reflectance curve of the Spectralon™ panel, that is used as brightness reference for spectroradiometric measurements in field and laboratory. Its reflectance is very close to 100 % and only in the SWIR shows larger deviations from it. The measurements are corrected for the self-reflectance of the panel during pre-processing (see section 3.2.1).

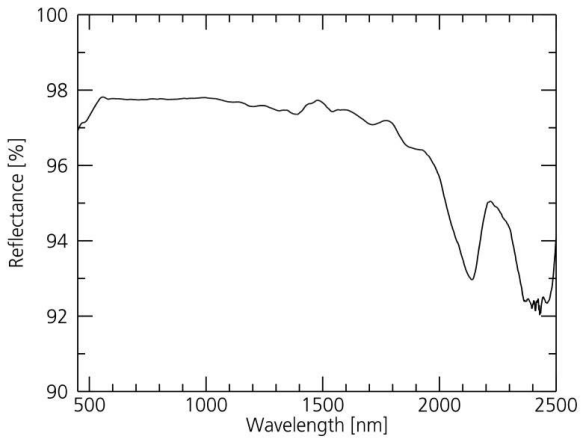


Figure B-1: Reflectance of Spectralon™ panel which is applied in the pre-processing of laboratory and field spectra. The spectrum was filtered to remove the ASD's increasing noise fraction in the SWIR.

A reverse spectral behavior than the white reference is desired for black-out materials, which are used in order to minimize influences on the spectroscopic measurements due to the reflection of the surrounding in a laboratory environment. For this reason, the walls and the slab of the spectroscopy laboratory are covered with a black pond foil and a special varnish is used for the tray, the samples are measured upon. Both materials are well suited for this purpose since they have a very uniform reflectance over the entire VIS to SWIR range with a reflection mostly lower than 3 %. Figure B-2a shows the reflectance of the pond foil and Figure B-2b the reflectance of the 3M varnish, the sample tray is covered with. The varnish produces a slightly rough surface and thus acts very closely to a Lambertian surface which reflects incident light uniformly in each direction. In both spectra, the ASD's typical increase of noise towards the longer wavelength range of each of the three sensors of the spectrometer and especially the low SNR in the SWIR are visible.

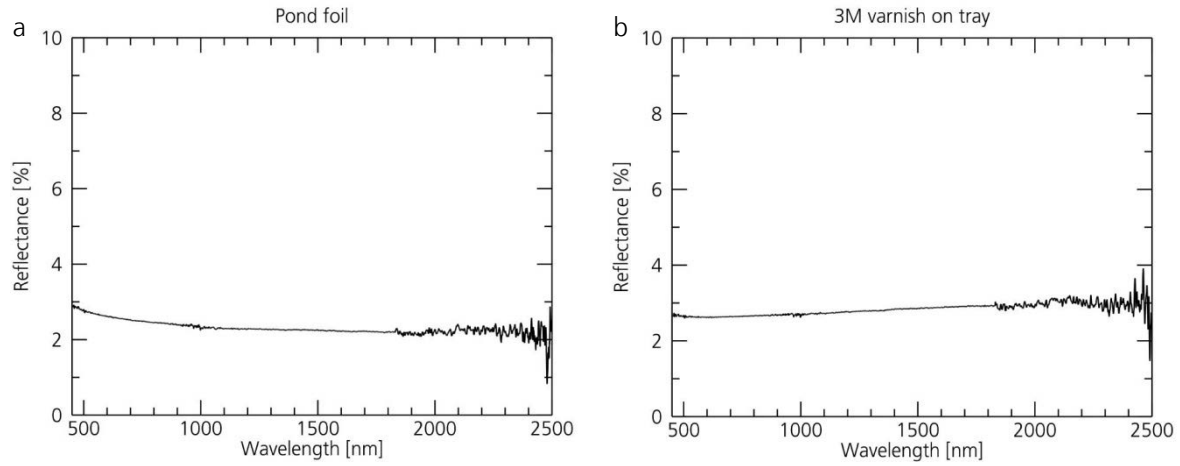


Figure B-2: Spectral curves of blackout materials used in the laboratory setup. Note the scaling between 0 and 10 %). Both curves were measured in the laboratory as mean of five singular measurements and no smoothing was applied.

B.2 HyMap imagery

Table B-5 gives an overview of flight parameters for the acquisition of the HyMap imagery at 14th October 2009. The flight parameters are prepared from gps files and quality control reports. The at-sensor-radiance HyMap imagery was processed with the following steps according to DLR's standard processing chain.

- Orthorectification, using flight parameters, SRTM-DEM (ground resolution of 30 m) and ground control points.
- Atmospheric correction, including spectral calibration and atmospheric correction with inflight calibration.

A digital elevation model available from the SRTM mission executed in 2000 is further used in processing. Its original spatial resolution of 30 m was resampled to the spatial resolution of the HyMap imagery of 3.3 m. In this process a low-pass filter was applied for smoothing of steps within the DEM. Slope and aspect information were created as further products based on the DEM using the appropriate ENVI functions.

Table B-5: Flight parameters of HyMap imagery and image sizes after orthorectification.

		Northern part (W→E)			Southern part (W→E)		
	Flight line	**1**	**2**	**3**	**4**	**5**	**6**
Flight parameters	Start of acquisition	12:33	12:13	11:54	11:18	11:28	11:39
	End of acquisition	12:49	12:29	12:10	11:25	11:35	11:46
	Mean heading	355°	176°	3°	11°	172°	7°
	Ground sampling distance	4.39	4.26	4.33	3.6	3.58	3.63
	No. of scan lines	15400	15096	14962	6751	6014	6784
	Altitude above sea level	2271	2257	2264	1776	1766	1787
	Altitude above ground	2098	2032	2069	1716	1711	1731
	% cloud pixels	0	0	0	0	0	0
Image size	Samples	2241	2054	1999	1349	1128	1364
	Lines	17402	17301	17300	7080	6901	7273

Orthorectification

The individual flight lines of the at-sensor-radiance HyMap imagery are orthorectified using ground control points (GCP) selected in the ENVI software and the resampled DEM in the PARGE software (Schläpfer and Richter 2002). The imagery was orthorectified using the projection UTM (Zone 35 South) with WGS-84 datum. All additional remote sensing data are re-projected to match this projection. The scan angle file in .bsq format is generated as additional output.

After the PARGE processing some artifacts remain in the data of the northern study area. Smaller artifacts comprise around ten lines each and are most visible in non-homogenous surface areas. One larger artifact remains in flight line 3 around the line 12500. These artifacts may result from large distortions in the raw imagery that originate from the strong movements of the plane that are directly transferred to the un-stabilized platform on which the HyMap sensor was mounted for the flight. They could not be resolved entirely during orthorectification with PARGE and thus remain as artifacts in the pre-processed imagery. Since the distortions were stronger in the raw imagery of the northern study area, the artifacts are limited to the flight lines 1 to 3. Figure B-3 shows a detailed view of the mentioned strong artifact and some of the smaller ones visible around it.

Atmospheric correction

Atmospheric correction is performed using DLR's ATCOR procedure developed by Richter (2010b). Spectral calibration is performed with the appropriate ATCOR module to detect and remove possible wavelength calibration errors. This is done based on the sensor definition, the range of bands per spectrometer, the solar geometry, atmospheric parameters and a certain number of target spectra of natural surfaces. They are selected from the imagery and used in an optimization procedure that minimizes the spikes of the derived surface reflectance spectra in the atmospheric absorption regions (see Richter 2010b for details). This leads to a file containing the spectral channels center wavelength shifts per spectrometer and a new wavelength file containing the updated wavelength for each channel.

Figure B-3: One large and some smaller artifacts remain in the orthorectified imagery. Detailed view of flight line 3 as true color image (with R: 640 nm, G: 549 nm, B: 456 nm).

Inflight radiometric calibration was applied to correct the laboratory calibration of the HyMap sensor. Ideally, multiple ground reflectance measurements of bright artificial surfaces are used to correct the imagery and calculate bandwise calibration coefficients (gain and offsets). Due to the lack of homogenous artificial surfaces of a certain size (at least one pure pixel would be required) the inflight calibration of the imagery of the South African study area was performed using a soil spectrum taken at a site dominated by bare soil (sample site 221). The field spectrum and photography of the center spot of this sample site is given in Figure B-4. A single target is not perfect but computationally sufficient to determine the coefficients.

The final processing of atmospheric correction was done using the option for rugged terrain as part of the ATCOR software. Table B-6 gives parameters of the individual flight lines used within the processing of atmospheric correction. Two different visibilities are applied as average of estimations for the single lines of every part of the study area. The visibility was kept constant during processing. Bands were interpolated around the atmospheric water vapor absorptions at 940, 1130, 1400 and 1900 nm (Table B-7). Each flight line was processed using the tiling command of ATCOR, since the flight lines 1 to 3, with around 17400 lines each, are too large to be processed as a whole.

Table B-6: Parameters within processing of atmospheric correction for each flight line.

	Northern part			**Southern part**		
Flight line	**1**	**2**	**3**	**4**	**5**	**6**
Solar zenith angle	43.7	40.0	36.7	30.9	32.4	34.0
Solar azimuth angle	295.4	300.3	305.6	320.7	316.3	312.3
Average visibility	25	30	30	25	25	25
Average water vapour column [cm]	1.44	1.42	1.46	1.70	1.71	1.74
Average aerosol optical thickness (550 nm)	0.284	0.247	0.247	0.301	0.298	0.194

Table B-7: Additional settings within processing of atmospheric correction.

Parameter	**Setting**
Visibility	No variable visibility because tiling of flight lines necessary due to length
Interpolations	No band interpolation in 760 nm oxygen region No band interpolation in 725 and 825 nm water vapor region Bands interpolated in 940 and 1130 nm water vapor region Interpolated bands in n 940 nm region: 905, 921, 937, 953, 969, 984 nm Interpolated bands in 1130 nm region: 1092, 1107, 1122, 1137, 1151, 1166, 1180 nm Bands interpolated in 1400 and 1900 nm water vapor region Interpolated bands in 1400 nm region: 1336, 1415, 1431, 1445, 1459, 1473, 1487 nm Interpolated bands in 1900 nm region: 1782, 1794, 1805, 1817, 1954, 1973 nm

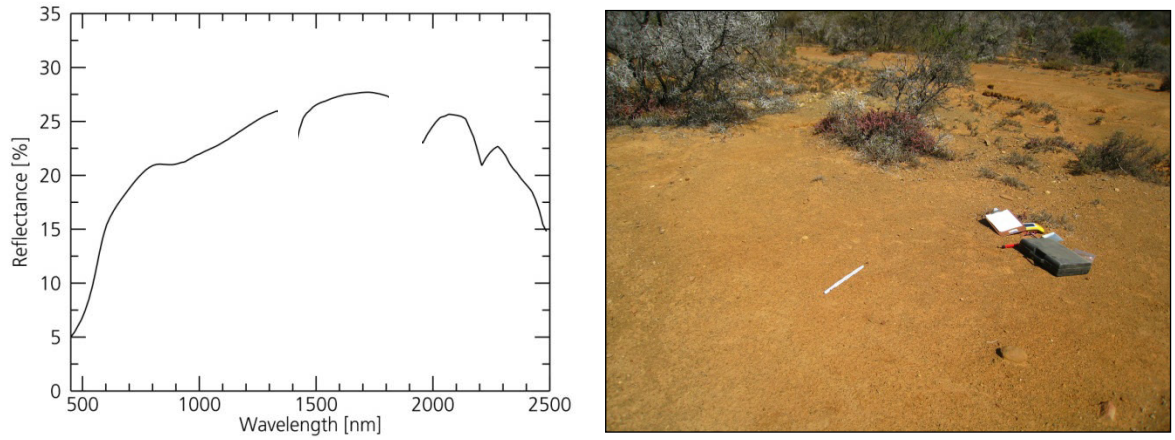


Figure B-4: Field spectrum and image of the ground reference site used for inflight calibration.

Spectral Polishing

Remaining inaccuracies in the data are reduced within spectral polishing. Table B-8 lists the wavelength ranges where local spectral polishing is performed within pre-processing. The given center wavelengths are of the bands which reflectance values were changed during this processing. The bands are first linearly interpolated and then smoothed with an interpolation in the given window. The kernel of the filter exceeds the bands to be smoothed depending on its kernel size.

Leveling of flight lines

The leveling of the flight lines corrects for brightness and radiometric differences between the individual HyMap flight lines. The applied approach uses the overlaps between the flight lines to derive a bandwise correction factor for each line. It differs from other leveling approaches in not using one master line but instead all flight lines are used equally. A further extended version of the leveling approach is given in Rogge et al. (2012). For the leveling of the South African hyperspectral imagery. no additional scattering correction was applied, since BRDF effects should be small due to almost N-S oriented flight lines. The crucial steps of the approach are summarized here.

The overlap regions for all pairs of flight lines are determined first. For every pair of adjacent flight lines, the mean spectrum of both overlap regions ($M_{A\ olap}$, $M_{B\ olap}$) and the average of the two ($M_{AB\ olap}$) is calculated. A multiplicative bandwise correction factor for each region is derived correcting the mean spectrum of each

Table B-8: Wavelength ranges where spectral polishing is applied.

Center wavelength of first and last band	Size of smoothing filter kernel	Comments
747 – 777 nm	3	Oxygen absorption around 763 nm
823 – 953 nm	5	Inaccuracies between detectors 1 and 2
1032 – 1093 nm	3	Oxygen absorption around 1063 nm
1252 – 1280 nm	3	Oxygen absorption around 1266 nm
1460 – 1794 nm	3	Short-wave infrared ranges between major water absorptions

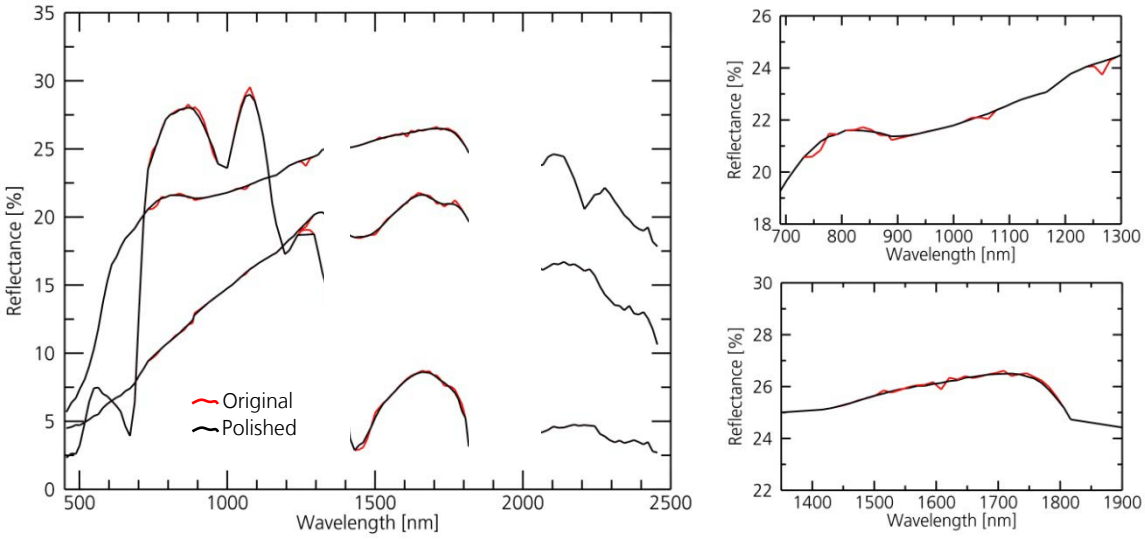


Figure B-5: Correction of local inaccuracies by spectral polishing for examples of image spectra dominated by bare soil, green and dry vegetation. The close-ups show the influence of the operation on the soil dominated spectrum.

overlap region to the joint average. This is done for all pairs of lines. The correction parameters are further combined and for each line a correction factor is derived correcting the line to the joint average. Figure B-6 shows an example for the leveling procedure on three flight lines and the derivation of correction factors for each line.

To level the hyperspectral imagery of the South African study area, the flight lines of the northern part (1 to 3) and the ones of the southern part (4 to 6) are leveled individually. To visualize the effect of the leveling operation on each line, two subsets of the original and leveled imagery were selected, one within the northern part and the other within the southern part (Figure B-7 and Figure B-8). Since in the leveling the lines were processed as a whole, the trend of the changes shown in the presented subsets here are representative for the entire lines they are part of. Figure B-7 shows the mean reflectance spectra of each flight line for the selected subsets before and after the leveling. To show the pixelwise influence of the

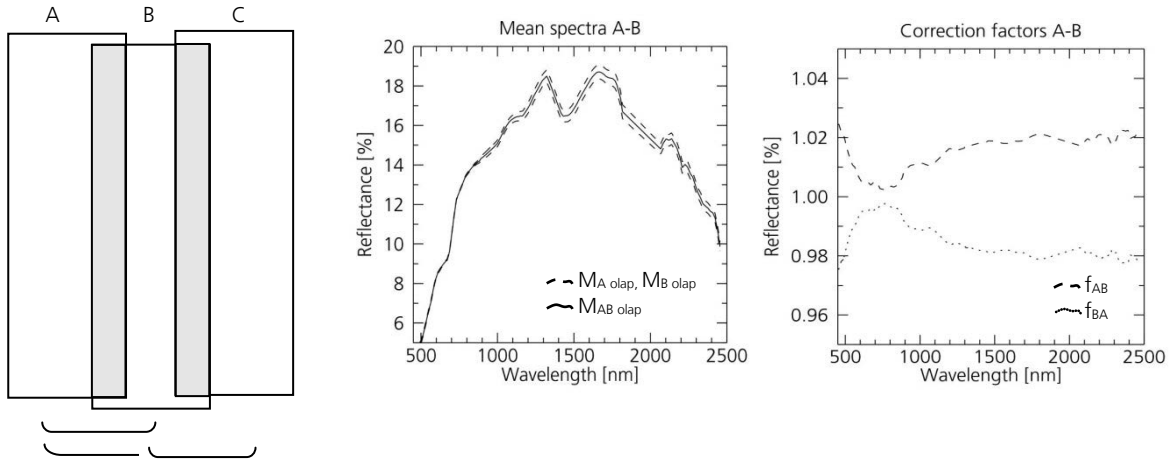


Figure B-6: Leveling procedure showing example flight line layout, mean spectra of overlap regions and derived correction factors for lines A and B (redrawn with hyperspectral imagery of study area based on Rogge et al. 2012).

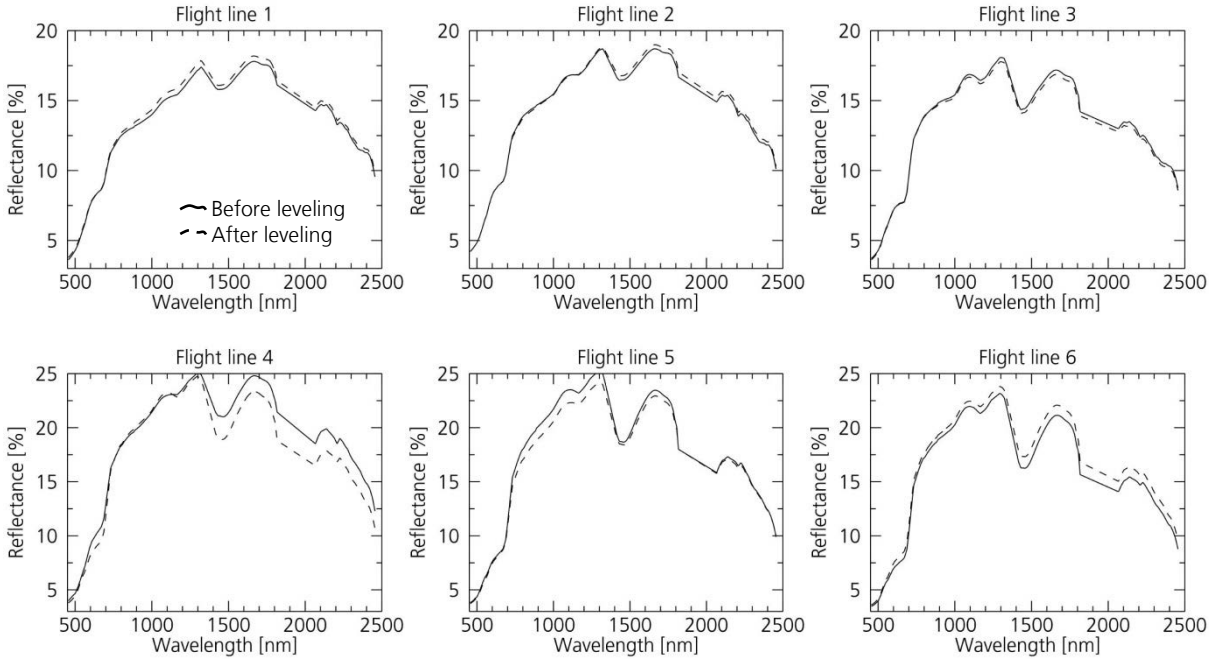


Figure B-7: Change in mean spectra for each three flight lines due to the leveling process calculated for two selected subsets (Line 1 to 3 in the northern part, 4 to 6 in the southern part).

leveling operation, the mean reflectance of each pixel as average of all bands before and after the leveling operation was compared in Figure B-8. In the northern part the deviation is positive for flight lines 1 and 2, thus their reflectance spectra were brightened in the leveling operation (compare also Figure B-7). The deviation images show that the mean reflectance of the pixels in line 1 rose an average at about 0.25 %, and the pixels of line 2 rose about 0.12 %. In contrast, the mean reflectance of the pixels in line 3 was reduced by the leveling operation and thus shows a negative deviation. In an average the mean reflectance of each pixel of flight line 3 was reduced about 0.20 %. In the southern part the differences between the flight lines were larger. The pixels of line 6 were brightened by the leveling operation (in an average around 0.70 % in reflectance for the given subset), while the average reflectance of the pixels of lines 4 and 5 were reduced for around 1.00 % and 0.55 %, respectively.

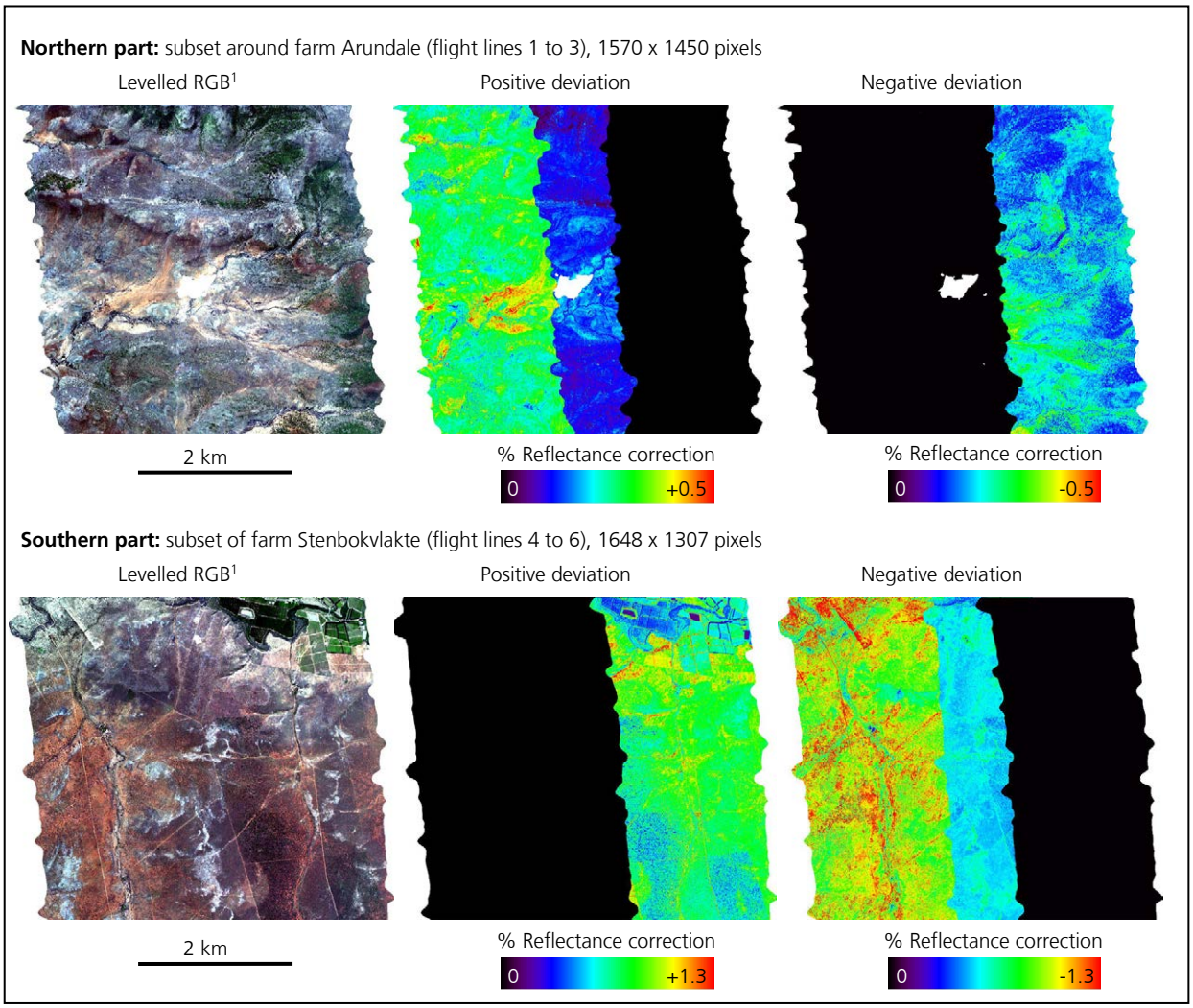


Figure B-8: Changes in per pixel mean reflectance within the leveling of the flight lines for two subsets in the northern and the southern transect. The given deviation of the individual flight lines of the subsets are representative for the entire lines they are part of (mean reflectance of each pixel of leveled data – mean reflectance of pixel of non-leveled data).
[1]True color image with R: 640 nm, G: 549 nm, B: 456 nm.

Mosaicking and masking of artificial objects

The individual flight lines of every transect part are mosaicked based on their georeference. This is done with line 2, respectively 4 for the southern part, in the background of the outside lines 1 and 3, respectively 4 and 6, since the data quality of particularly line 2 appeared slightly reduced compared to the other lines.

As a last step within pre-processing, artificial objects are masked in order to limit the consideration within this study to natural surfaces. Concrete streets and gravel roads, farm areas, barns, water ponds, etc. are masked based on their visibility in the mosaicked imagery. This mask was updated after spectral unmixing, when remaining small objects, that were not obvious in RGB or CIR imagery during the first masking, showed up because of a higher unmixing RMSE, which were mostly very small water ponds. In the southern transect part also the production panels and pasture land of the Kirkwood Correctional Service located in the north of this part is excluded since it differentiates significantly from natural land surface coverage.

IAS filtering

The leveled and mosaicked HyMap imagery is spatially filtered with an iterative adaptive spatial filtering approach (IAS, Rogge and Rivard 2010). IAS was designed to minimize random noise and local intra-class spectral variability while retaining edges. The filter iteratively smoothes pixel spectra according to their spectral similarity to the adjacent pixels. Thus, in a homogenous region, the filter acts like a mean filter while it is a spatially adaptive filter in heterogeneous regions. The process is similar to diffusion filters (e.g. Wang et al. 2010, Velasco-Forero and Manian 2009). Spectral similarity is determined using standard hyperspectral similarity measures (RMSE and spectral angle). The application of the RMSE focuses on differences in the amplitude, while using spectral angle, the filter corrects for random noise but preserves spectral features. Spectral angle is used here since the filter operation is applied in order to reduce non-significant changes in brightness and thus the natural and measurement-dependent variability within spectral classes. This ensures the preservation of spectral features that are the base for this study. The IAS filter applied here is a slightly modified version form the one described in Rogge and Rivard (2010).

IAS first produces a gradient map for each pixel based on the spectral angle of the center pixel compared to the pixels in a 3 x 3 window around it. The spectral angle of the center pixel is set to 1 and the adjacent pixels graded between 0 and 1. The gradient map is used to derive a weighted kernel for the given window. If the difference in spectral angle is above a given threshold value (SA_{min}), the pixel gets a weight of 0. With the weighted kernel window the given pixel is averaged with the other pixels in the window according to their weights.

$$\rho_{xyIAS} = \frac{\sum \rho_{xy} \cdot w_{xy}}{w_{total}} \quad [-] \qquad (B\text{-}1)$$

with the original and smoothed reflectance of the center pixel ρ_{xy} and $\rho_{xy\,IAS}$, the weight of the center pixel w_{xy} and the summarized weights of all pixels in the 3 x 3 window. An iterative approach allows for a progressive smoothing with more influence of pixels farther away. The degree of smoothing decreases exponentially since in each iteration the threshold value SA_{min}, determining pixels that contribute to the smoothing, is reduced by its half. An appropriate default value of the minimum spectral angle SA_{min} that can be determined by a routine deriving the mean spectral angle of all image pixels compared to their adjacent pixels. Table B-9 gives the mean spectral angles that were calculated for each subset of 1000 lines of the HyMap imagery of both transect parts. The threshold should match the degree of brightness and spectral

Table B-9: Mean spectral angles of the HyMap imagery determined to define an appropriate threshold in spectral angle used for IAS processing.

Northern transect		Southern transect	
Lines	**Mean spectral angle**	**Lines**	**Mean spectral angle**
1-1 000	0.070276	1-1 000	0.107061
1 001-2 000	0.054180	1 001-2 000	0.067138
2 001-3 000	0.061832	2 001-3 000	0.068778
3 001-4 000	0.104648	3 001-4 000	0.066245
4 001-5 000	0.096578	4 001-5 000	0.153832
5 001-6 000	0.091176	5 001-6 000	0.138923
6 001-7 000	0.141115	6 001-7 400	0.111688
7 001-8 000	0.090262		
8 001-9 000	0.118584		
9 001-10 000	0.066139		
10 001-11 000	0.050417		
11 001-12 000	0.098956		
12 001-13 000	0.102067		
13 001-14 000	0.078437		
14 001-15 000	0.061679		
15 001-16 000	0.074447		
16 001-17 400	0.073085		
Mean of northern part	0.0843457	Mean of southern part	0.1019522
Mean of entire data	0.0894809		
Applied spectral angle	0.04474046	Half of calculated average value	

variance within the imagery. To process the HyMap imagery, the half of this value is used to keep the influence of the filter operation low. Figure 3-12 of section 3.2.2 gives an example for the effect of the IAS filter operation on the HyMap imagery.

B.3 Simulated EnMAP data

Table B-10 gives details on the pre-processing of the simulated EnMAP satellite data. The spectral ranges applied in this pre-processing are adapted to the spectral ranges covered by the HyMap sensor. Figure B-9 shows examples for the spectral polishing of local inaccuracies remaining in the spectra after atmospheric correction. The polishing consists of a band interpolation and subsequent smoothing.

Table B-10: Details for pre-processing of simulated EnMAP data.

Pre-processing step	Bands	Comments
Removal of overlapping regions	904-994 nm	Band overlap between first and second detector Bands of second detector used
Band interpolation	771	Oxygen absorption around 763 nm
	872-975	Removal of artifacts remaining from atmospheric correction, 12 bands
	1109-1155	Removal of artifacts remaining from atmospheric correction, 5 bands
	1347-1406	6 bands influenced by water vapor absorption around 1400 nm
	1824-2026	22 bands influenced by water vapor absorption around 1900 nm
Spectral smoothing	440-1287 1488-1783 2062-2424	Spectral smoothing over entire spectrum with 3 x 3 (ranges of prominent water bands are excluded to prevent their influence at the borders)
Band selection	455-1335 1418-1814 2035-2446	Removal of first 6 bands because not covered by HyMap data, Removal of bands within the major water bands 198 bands selected for further analysis.

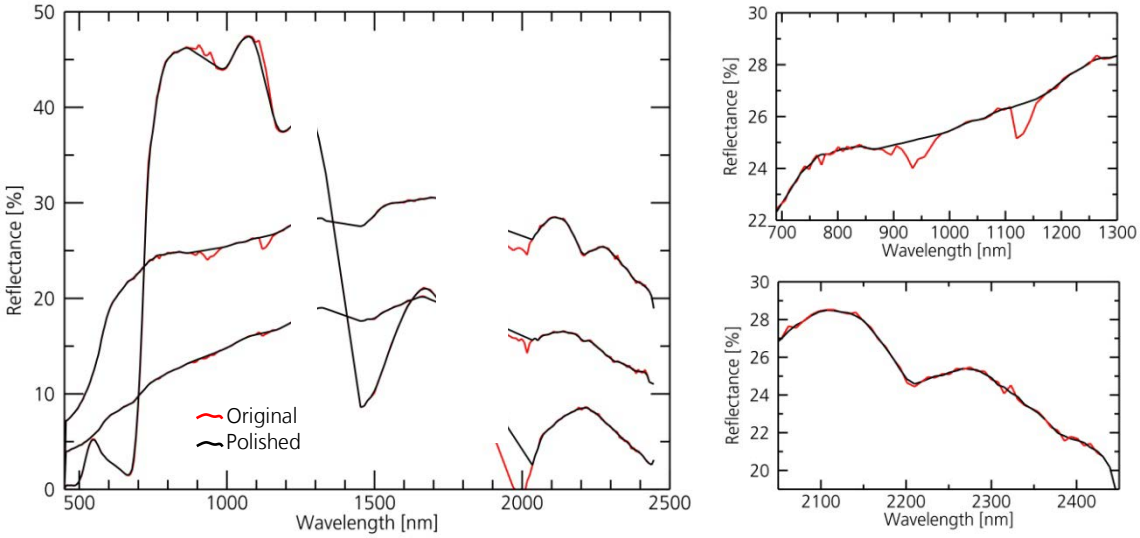


Figure B-9: Correction of local inaccuracies remaining from atmospheric correction by spectral polishing for examples of EnMAP image spectra dominated by bare soil, green and dry vegetation. The close-ups show the influence of the operation on the soil dominated spectrum.

C Quality indicators for multivariate analysis

In the following, common quality indicators of multivariate analysis are presented which are used for comparison of regression models. They may be calculated for calibration (index "Cal") and cross-validation (index "Val"). Among these, model performance is rated in sections 4.3 and 6.1.1 based on R^2, RPD and RMSE only, though the program code also returns SEC, Bias and PRESS. Further, the AIC is used as indicator within cross-validation of PLS models (section 4.3).

Coefficient of determination (R^2)

R^2 gives the correlation of the predicted and the reference value for the dependent variables (y) of a regression model. It is an indicator for the variance, the model explains compared to the total variance of the reference y values. Its values range between 0 and 1, with 1 indicating residuals equal zero and a perfect correlation.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_{mi} - y_{ci}\right)^2}{\sum_{i=1}^{n} \left(y_{mi} - y_{mmean}\right)^2} \qquad \text{(Kessler 2007)} \qquad [-] \qquad \text{(C-2)}$$

with n as the number of samples, y_m are the reference values of the i dependent variables y and $y_{m\,mean}$ as their average, and y_c are the dependent variables as predicted by the regression model.

Root mean square error (RMSE)

The individual differences between the measured (y_{mi}) and the predicted variables (y_{ci}), called residuals, are aggregated in the RMSE. It thus gives the typical vertical distance of the data to the regression line. The RMSE is a good measure of the accuracy of the predicted variables and is given in the unit of the dependent variables y, which in this study is %.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{mi} - y_{ci}\right)^2}{n}} \qquad \text{(Kessler 2007)} \qquad [\%] \qquad \text{(C-3)}$$

The relation between RMSE and R^2 can be expressed using the standard deviation of the reference samples. This shows that the RMSE ranges between 0 (when R^2=1) and the standard deviation of the reference values of the y variables (when R^2=0).

$$RMSE = \sqrt{1 - R^2} \cdot \sigma_{ym} \qquad [\%] \qquad \text{(C-4)}$$

Ratio of performance to deviation (RPD)

The RPD is calculated as the ratio of the standard deviation of the reference method to the RMS.

$$RPD = \frac{\sigma_{ym}}{RMS} \qquad \text{(Chang et al. 2001)} \qquad [-] \qquad \text{(C-5)}$$

with the standard deviation $\sigma_{y\,m}$ of the reference values of the dependent variables y_m, and the root mean square error RMS.

Standard Error of Calibration (SEC)

The standard error is comparable to the RMSE. While it is SEC in calibration procedures, in prediction it is often referred to as SEP (standard error of prediction). The SEC/SEP is the standard deviation of the residues.

$$SEC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{mi} - y_{ci}\right)^2}{n-1}} \qquad \text{(Kessler 2007)} \qquad [\%] \qquad \text{(C-6)}$$

Systematic Error (Bias)

The Bias is the mean of the residues, which is close to zero in a good calibration.

$$Bias = \sum_{i=1}^{n}\frac{\left(y_{mi} - y_{ci}\right)}{n} \qquad \text{(Kessler 2007)} \qquad [\%] \qquad \text{(C-7)}$$

Predicted Residual Sum of Squares (PRESS)

PRESS gives the sum of the residuals' squares between the measured and predicted values.

$$PRESS = \sum_{i=1}^{n}\left(y_{mi} - y_{ci}\right)^2 \qquad \text{(Kessler 2007)} \qquad [\%] \qquad \text{(C-8)}$$

Akaike Information Criterion (AIC)

The AIC is used in PLS analysis as a second indicator besides the RMSE in order to select the optimal number of PLS factors to use for modeling. The smallest AIC gives the best model.

$$AIC = n \cdot \ln RMSE + 2p \qquad \text{(Viscarra Rossel and Behrens 2010)} \qquad [-] \qquad \text{(C-9)}$$

where n is the number of samples and p the number of PLS factors used in the prediction.

D Technical realization of the workflow

This section gives details of the implementation of the workflow as it was described in section 5. This includes the realization of the calibration, prediction and validation procedures as own developments (D.1 to D.4) in the IDL language. An overview and examples of the most important files generated during the fully automated workflow are given. D.5 gives details of settings within the applied endmember and spectral unmixing approaches. Finally some processing times for all individual steps are given (D.6).

D.1 Calibration workflow

Input data for calibration is a spectral dataset in ENVI Spectral Library format. Subsequently, the user decides which soil parameter will be modeled and whether or not he would like to run it as batch job without any further manual interaction. If the user decides for continuous batch job processing, all additional output files are created (predominantly ASCII files comprising details of crucial processing steps), feature variables from parameterization are standardized before MLR and identified outliers and insignificant variables are excluded for final regression analysis. If the user declines, he can prevent each file output during the calibration procedure, can run calibration without a standardization of the variables, and has the option to prevent the removal of all or of a part of the identified outlier samples. Results of all MLR, cross-validation and subsequent identification of outlier samples and insignificant variables are given in the IDL console as information for the user on which the user can base the decision for an intervention in the workflow if necessary.

Calibration starts with parameterization of spectral features. Names of spectral variables comprise the modeled soil parameter, the type of spectral feature (AF, CF, or HF) together with an index number, and a wavelength specification for a simple identification of the particular feature (e.g. corg_AF2_2350, iron_CF1_550/590, etc.). Hull features are named without the modeled parameter as they represent the background reflection which is inherent to several soil constituents (e.g.HF_VIS_albedo). Calibration can also be run with inclusion of features in the bands of atmospheric water vapor to include the two water features around 1400 and 1900 nm, characteristic for clay minerals. This is a possibility relevant only for a calibration based on laboratory spectra.

After the procedure registers the spectral features corresponding to the selected soil parameter to be modeled, the input data are checked for parameters which are necessary for the following analysis – (1) the spectral extension of the input data has to cover the spectral ranges of included features and (2) the number of samples has to be higher than the number of spectral variables, as a different set-up would result in an underdetermined equation system, not solvable in regression analysis. The calibration procedure is only applied if both criteria are fulfilled. The bands of all wavelengths given as specifications of spectral features (CR range, λ_{lit}) are determined as the closest bands to the given position. Continuum removal for AF and the calculation of the continuum is conducted using IDL's own routine. In the same way the line fits used in parameterization of CF and HF (routine *poly_fit*) are produced and often performed calculations of mean or standard deviation are executed. An increment of 0.0001 is used to determine the invariant part of a spectral variable in initial quality checks. For regression analysis IDL's function for a multiple linear regression fit is applied. Outliers and insignificant variables are removed by default in the calibration

workflow but removal can manually be prohibited completely or for defined samples only, for instance to exclude samples known to be questionable for some reason. An overview file gives all results of regression analysis and steps in between. Measured vs. calculated parameter contents are given as table and graph. The calibrated functional relationship is given as one file stating coefficients of the regression equation. It also includes details of the ingoing dataset such as number of bands, wavelength of first and last bands, number of included variables, the unit of chemical reference values, range of chemical reference values, a flag if the spectral feature variables are standardized and the IDs of the spectral variables included in final MLR. Table D-11 gives an overview of the files that are created during the calibration and prediction procedures.

Table D-11: Overview of files saved during calibration and prediction procedures. To every file name a time string is attached as simple identification criterion. All files are output in one created folder.

		File name with ID	Type	Output file
Calibration		_Cal_1a_PARoverview	.txt	Overview of parameterization with input files and parameters of spectral data, number, types and details of used features, and overview of initial quality control tests including investigated parameters for all spectral variables
		_Cal_1b_PARfeaturevariables	.txt	Table of calculated feature variables for each input spectrum
		_Cal_2a_PARcrspectra	.txt	Table of continuum removed spectra of each interval around an absorption feature
		_Cal_3a_MLRcombchemspectra	.txt	Table of chemical reference samples combined with correlated spectra
		_Cal_3b_MLRstandardizedvariables	.txt	Table of standardized spectral variables and details of the performed standardization
		_Cal_4a_MLRoverviewallMLR	.txt	Overview of all regression analyses with given accuracy measures, details of identified and excluded outlier samples and insignificant spectral variables
		_Cal_4b_INPUT_MLRregressionEQ	.txt	Final functional regression relationship with parameters of regression equation and additional information like spectral characteristics, calibration limits, number and IDs of spectral variables used in MLR, and accuracy measures of calibration - This file is input for the prediction procedure
		_Cal_4c_MLRplotMLRallsamples	.jpg	Plot measured vs. calculated of first MLR including all samples
		_Cal_4d_MLRplotfinalMLR	.jpg	Plot measured vs. calculated of final MLR established without insignificant variables and potentially outlier samples removed
		_Cal_4e_MLRdetailsLOO	.txt	Table of detailed statistics of leave-one-out analysis for every sample
		_Cal_4f_MLRcrossval	.txt	Table of detailed statistics of cross validation (leave-10-out) analysis for every sample set
		_Cal_5_MLRmeasVScalc	.txt	Table of measured vs. calculated parameter contents
Test set validation	Point dataset	_Val_1_OverviewValidation	.txt	Overview file of validation procedure with input datasets, processing steps and file outputs
		_Val_1b_PARfeaturevariables	.txt	Table of feature variables calculated for each sample
		_Val_2_CalcChemValues	.txt	Table of calculated vs. measured soil parameters for all samples
		_Val_3b_MLRstandardizedvariables	.txt	Table of standardized spectral variables and details of standardization
Prediction	Point	_Pred_1_OverviewPrediction	.txt	Overview file of prediction procedure with ingoing datasets, processing steps and file outputs
		_Pred_2_CalcChemValues	.txt	Table of calculated soil parameters for all samples
	Image[1]	_Pred_1_OverviewPrediction	.txt	Overview file of prediction procedure with ingoing datasets, processing steps, file outputs and statistics of image
		_Pred_2_CalcChemValues	.bsq	5 band image file of calculated soil parameters: Band 1: Originally calculated parameter contents Band 2: Calculated contents within calibration limits Band 3: Smoothed by spatial filter with filter size 3 Band 4: Smoothed by spatial filter with filter size 5 Band 5: Pixel mask, flagging pixel outside calibration limits and pixels interpolated in 3 x 3 filter operation

[1]In the prediction routine it was desisted from an output of the table of feature variables calculated in parameterization (output 1b) and the table of standardized spectral variables (output 3b), because especially for image data they make up to a hundred megabytes and are not used with this dimension of data.

In the following, examples of the most important files are given. Figure D-10 shows the ASCII text file giving the details of the regression analyses within the calibration procedure (file 4a in Table D-11). It is supposed to give an overview of the crucial steps within model calibration and in addition some further explanations for the user. The selected example shows the model calibration for soil organic carbon based on field spectra in HyMap's spectral resolution.

```
------------------------------------------------------------
Detailed results of multiple linear regression analysis
Based on feature variables calculated for parameter
               CORG
------------------------------------------------------------

* File paths:
   Input file Spectral Library:   Fsp07u09_bs_v3_8HyMapNeucalBbr3.slb
   Input file feature variables:  _Calv1_1b_PARfeaturevariables_12Jan2012_1416.txt
   Input file chemical values:    Tab_chemAnalysis_allsamples_4to164_MASTERdonotmove.txt
   Output folder:                 /Regressions/Fsp07u09_bs_v3_8HyMapNeucalBbr3
                                  _CORG_Cal+Val_12Jan2012_1416
   Procedure:                     /users/baye_an/programms/idl/Programmdateien/
                                  Soilparameters/v1_org/calibration.pro

* Chemical unit of input analysis: All values in %

* Number of samples and variables:
   From input files, after Parametrization respectively:
       Total number of samples/spectra (for PAR):           125
       Total number of calculated spectral variables (PAR):  16
   For final regression analysis (after removal of outliers and unsignificant variables):
       Final number of samples/spectra:                     125
       Final number of input spectral variables:             13

* Standardisation:
   Feature variables were standardised before regression analysis:
   stand. Var = mean substracted from original variable and afterwards divided by stdev
   Mean =0, Stdev=1 for all variables

* Calculation of influence of variable on regression (results LOO):
   abs(regress_coeff)/(total(abs(regress_coeff))))*100
   (only posible if variables are standardised)

* Test if number of samples is sufficient for calculation of MLR
       (over-/under-determined equation system)
       Final number of samples/spectra:                     125
       Final number of input spectral variables:             13
       - > no_variables < no_samples AND no_variables < no_samples/2
           Overdetermined equation system: MLR regression possible
           Variable limits (e.g. regarding determination of not significant variables,
           outliers) should be fine!
           Accuracy of variable limits increases with number of samples

* File Output:
   Details of final regression equation wrote to file
   Fsp07u09_bs_v3_8HyMapNeucalBbr3_CORG_Calv1_4b_INPUT_MLRregressionEQ_12Jan2012_1416.txt
   File contains IDs and names of variables, regression coefficients and mean and
   stdev of each variable for standardisation in VAL analogous to CAL routine

------------------------------------------------------------

RMSE: error in unit of input reference data -> %
RPD: dimensionless

------------------------------------------------------------
```

```
------------------------------------------------------------
1. Result of multiple linear regression (without LOO)
------------------------------------------------------------
   125 samples, 15 spectral variables

   Multiple linear correlation coefficient,            R^2:     0.8870
   Root mean squared error,                           RMSE:     0.4174
   Ratio of performance to deviation,                  RPD:     2.1739
   Systematic error,                                  Bias:    -0.0000
   Predicted residual error sum of squares,          PRESS:    21.7826
   Standard error of calibration,                      SEC:     0.4191

Measures per feature variable:     Influence    Reg_coeff     Lin_cor        Sigma
   depth_max_corg_AF1_1730            9.89         0.48         0.62         0.26
   wvl_depth_max_corg_AF1_1730       -1.34        -0.07         0.01         0.08
   width_corg_AF1_1730                0.02         0.00         0.42         0.09
   area_corg_AF1_1730                -6.81        -0.33         0.55         0.25
   asym_fac_corg_AF1_1730             0.94         0.05         0.05         0.06
  depth_max_corg_AF2_2330            -1.73        -0.08        -0.29         0.15
   wvl_depth_max_corg_AF2_2330       -5.25        -0.26         0.24         0.10
   depth_lit_corg_AF2_2330           -2.78        -0.14        -0.36         0.14
   width_corg_AF2_2330               -0.49        -0.02         0.07         0.06
   area_corg_AF2_2330                 0.36         0.02        -0.24         0.14
   asym_fac_corg_AF2_2330             7.82         0.38         0.47         0.08
   slope_CO1_vis                    -16.98        -0.83        -0.15         0.09
   height_CO1_vis                   -11.14        -0.54        -0.27         0.10
   slope_CO2_swir1                   -6.88        -0.34        -0.01         0.07
   height_CO2_swir1                  27.57         1.35         0.10         0.15

   Total of influences (for verification, should be 100)  :   100.00 %
   Influence = Influence of variable on regression, [%]
   Reg_coeff = Regression coefficients (coeff)
   Lin_corr = Linear correlation coefficient (R^2)
   Sigma = Standard Error, 1-sigma uncertainty (sigma)

------------------------------------------------------------
2. Leave-one-out cross validation
------------------------------------------------------------
Every sample subsequently removed from sample set (full cross validation),
calibration model established with remaining samples -> statistics and influence of particular
variables (significance of variables)
then soil information predicted for removed samples -> Prediction ability, influence of singular
samples  (removal of outliers)

   125 samples, 15 spectral variables

   Mean multivariate correlation coefficient,          R^2:     0.8407
   Mean Root mean squared error,                      RMSE:     0.4906
   Ratio of performance to deviation,                  RPD:     1.8495
   Systematic error,                                  Bias:    -0.0042
   Predicted residual error sum of squares,          PRESS:    30.0921
   Mean standard error of calibration,                 SEC:     0.4926

Measures per feature variable: Mean_Inf   Sig.   Mean_RC  Stdev_RC  Mean_lincorr Mean_sigma
   depth_max_corg_AF1_1730       9.88     yes      0.48      0.03      0.61         0.27
   wvl_depth_max_corg_AF1_1730   1.34     yes      0.07      0.01      0.01         0.08
   width_corg_AF1_1730           0.10     -        0.01      0.01      0.42         0.09
   area_corg_AF1_1730            6.81     yes      0.33      0.03      0.55         0.25
   asym_fac_corg_AF1_1730        0.94     yes      0.05      0.00      0.05         0.06
   depth_max_corg_AF2_2330       1.72     yes      0.08      0.02      0.29         0.15
   wvl_depth_max_corg_AF2_2330   5.25     yes      0.26      0.01      0.24         0.10
   depth_lit_corg_AF2_2330       2.77     yes      0.14      0.02      0.36         0.14
   width_corg_AF2_2330           0.49     yes      0.02      0.01      0.07         0.06
   area_corg_AF2_2330            0.38     -        0.02      0.01      0.24         0.15
   asym_fac_corg_AF2_2330        7.81     yes      0.38      0.01      0.47         0.08
   slope_CO1_vis                16.96     yes      0.83      0.01      0.15         0.09
   height_CO1_vis               11.13     yes      0.54      0.01      0.27         0.10
   slope_CO2_swir1               6.87     yes      0.34      0.01      0.01         0.07
   height_CO2_swir1             27.54     yes      1.35      0.02      0.10         0.15

   Total of influences (for verification, should be 100)  :   100.00 %
   Mean_Inf = Mean Influence of variable on regression, [%]
   Sig. =  Variable significant for regression
   Mean_RC = Regression coefficients (mean(abs(coeff))
   Stdev_RC = Variation of coefficients (stdev(coeff))
   Mean_Lincorr = Mean linear correlation coefficient (mean(R^2))
   Mean_Sigma = Mean standard Error, 1-sigma uncertainty (mean(sigma))
```

```
------------------------------------------------------------
3. Removal of not significant variables based on results of LOO analysis
------------------------------------------------------------

   Significance of variable for regression (s. Kessler2007, p.97):
   Variable is significant for regression if abs(mean_reg_coeff)-2*stdev(reg_coeff_LOO) > 0
   (where reg_coeff_LOO includes only samples included in final regression)

   No. of variables removed as insignificant:   2

   Details of removed variables (deviation above 2*stdev):
                          Mean(reg_coeff)     Stdev(reg_coeff)     Deviation >2*stdev
   width_corg_AF1_1730        0.0051            0.0085               -0.0120
   area_corg_AF2_2330         0.0188            0.0120               -0.0052

------------------------------------------------------------
4. Identification of outlier samples based on results of LOO analysis
------------------------------------------------------------

   All 0 detected outliers removed

   Citerion to detect sample as outlier: deviation of R^2 >/= 2*stdev
   (R^2 of regression models of leave-one-out cross validation)
   with:    mean R^2: 0.887012
            stdev:    0.00386970 = 0.436262 %
   Detected: 3
   Removed:  0

   Outliers detected but not removed and deviation of mean R^2 [%], sorted by deviation:
   180_Cbs_mean:  dev=4.00078 %
   022_Csv_mean:  dev=2.20635 %
   011_Cbs_mean:  dev=1.10482 %

------------------------------------------------------------
5. Result of final multiple linear regression (outliers determined via LOO are removed)
------------------------------------------------------------

   125 samples, 13 spectral variables

   Accuracy measurements (and derivation to first regression including all samples):
   Multiple linear correlation coefficient,            R^2:    0.8869 (-   0.0000)
   Root mean squared error,                           RMSE:    0.4175 (+   0.0000)
   Ratio of performance to deviation,                  RPD:    2.1737 (-   0.0001)
   Systematic error,                                  Bias:   -0.0000 (-   0.0000)
   Predicted residual error sum of squares,          PRESS:   21.7855 (+   0.0029)
   Standard error of calbration,                       SEC:    0.4192 (+   0.0000)

   Measures per feature variable:
                                      Influence   Reg_coeff    Lin_corr      Sigma
   depth_max_corg_AF1_1730               9.99        0.48        0.62        0.26
   wvl_depth_max_corg_AF1_1730          -1.38       -0.07        0.01        0.07
  area_corg_AF1_1730                    -6.84       -0.33        0.55        0.25
   asym_fac_corg_AF1_1730                0.98        0.05        0.05        0.06
   depth_max_corg_AF2_2330              -1.50       -0.07       -0.29        0.12
   wvl_depth_max_corg_AF2_2330          -5.25       -0.25        0.24        0.10
   depth_lit_corg_AF2_2330              -2.72       -0.13       -0.36        0.14
   width_corg_AF2_2330                  -0.40       -0.02        0.07        0.05
   asym_fac_corg_AF2_2330                7.84        0.38        0.47        0.08
   slope_CO1_vis                       -17.11       -0.83       -0.15        0.09
   height_CO1_vis                      -11.25       -0.55       -0.27        0.09
   slope_CO2_swir1                      -6.93       -0.34       -0.01        0.07
   height_CO2_swir1                     27.80        1.35        0.10        0.15

   Total of influences (for verification, should be 100)  :   100.00 %
   Influence = Influence of variable on regression, [%]
   Reg_coeff = Regression coefficients (coeff)
   Lin_corr = Linear correlation coefficient (R^2)
   Sigma = Standard Error, 1-sigma uncertainty (sigma)
```

```
------------------------------------------------------------
6. Evaluation with cross-validation (leave-10-out)
------------------------------------------------------------

   125 samples, 13 spectral variables
   NOTE: Results shown here are calculated as average of 100 single L10o runs!
   Part of sample set removed, calibration model established with remaining samples,
   then soil information predicted for removed samples -> prediction accuracy

   No. of repetitions of L10o-CV:                  100

   10 samples left out for cross validation
   No. of test sets:                               13
   No. of samples taken out in first 12 test sets:  10
   No. of samples taken out for last test set:       5

   Accuracy measurements (and comparison with FINAL regression):
   (calculated from all samples predicted in separate L10o test sets)
   Multivariate correlation coefficient,            R^2(CV):    0.7599 (-   0.1271)
   Root mean squared error,                        RMSE(CV):    0.4740 (+   0.0565)
   Ratio of performance to deviation,               RPD(CV):    1.5719 (-   0.6018)
   Systematic error,                               Bias(CV):    0.0044 (+   0.0044)
   Predicted residual error sum of squares,       press(CV):   30.0031 (+   8.2175)
   Mean standard error of calibration,              SEC(CV):    0.4759 (+   0.0567)

   Mean_Influence = Mean Influence of variable on regression, [%]
   Mean_Regcoeff = Regression coefficients (mean(abs(coeff))
   Stdev_Regcoeff = Variation of coefficients (stdev(coeff))
   Mean_Lincorr = Mean linear correlation coefficient (mean(R^2))
   Mean_Sigma = Mean standard Error, 1-sigma uncertainty (mean(sigma))
```

Figure D-10: Example for the ASCII text file giving a summary of each crucial step performed during model calibration (file 4a in Table D-11).

Figure D-11 gives an example for the ASCII text file giving the established final regression relationship (file 4b in Table D-11). It includes details of the input spectral library, details of the processing (if standardization was applied, unit of chemical reference used, etc.) and finally the spectral variables and their regression coefficients ($c_{0...n}$). Comments are given for generic files lacking detailed explanations due to a focus on machine readability. This file is input to the prediction procedure and used to predict soil constituents with the calibrated regression model.

0	Details of spectral dataset	no_bands	116.0	0.0	0.0
0		first_band	456.0	0.0	0.0
0		last_band	2455.0	0.0	0.0
0		no_var_sfter_par	16.0	0.0	0.0
0		cal_limit_min	0.21 (calibration limits)	0.0	0.0
0		cal_limit_max	5.85	0.0	0.0
0		const	c_0 1.138480	0.0	0.0
0		Standardisation	1 (Standardization was performed)	0.0	0.0
0	accuracies of final MLR and CV	R2(finalMLR)	0.89	0.0	0.0
0		RPD(finalMLR)	2.17	0.0	0.0
0		R2(CV)	0.76	0.0	0.0
0		RPD(CV)	1.57	0.0	0.0
0		%	0	0.0	0.0

variable ID	variable name	Regression coefficients c_i	$mean_i$	sd_i
0	depth_max_corg_AF1_1730	0.484167	0.002297	0.003184
1	wvl_depth_max_corg_AF1_1730	-0.066841	1741.432007	18.843729
4	area_corg_AF1_1730	-0.331618	0.131271	0.227785
5	asym_fac_corg_AF1_1730	0.047476	1.294800	0.658579
6	depth_max_corg_AF2_2330	-0.072927	0.035194	0.013840
7	wvl_depth_max_corg_AF2_2330	-0.254655	2337.760010	13.269160
8	depth_lit_corg_AF2_2330	-0.132049	0.028977	0.014542
9	width_corg_AF2_2330	-0.019564	111.279999	16.578064
11	asym_fac_corg_AF2_2330	0.380082	0.948566	0.674316
12	slope_CO1_vis	-0.829382	6.394707	1.484815
13	height_CO1_vis	-0.545125	1931.990356	414.834900
14	slope_CO2_swir1	-0.335696	0.597391	0.202797

Figure D-11: ASCII text file giving the regression equation of the established MLR relationship.

D.2 Prediction procedures

The procedure to predict soil properties from spectra using a calibrated regression relationship was developed in the first instance for application to hyperspectral imagery. But since the method in general was developed to be also applicable to other datasets, also the possibility to predict soil constituents from spectral libraries is included in the IDL procedure. In the following, the workflow for spectral libraries for unknown and known corresponding chemistry (test set validation) is described. The applied workflow is comparable for both spectral libraries and hyperspectral imagery. In any case the input spectra must be of same spectral resolution and band position as the data used for calibration (see section 5.3). Usually, spectral characteristics of the calibration dataset (e.g. a spectral library measured in the field) are customized to fit the target dataset of prediction (e.g. hyperspectral imagery), as in most of the cases high resolution spectroradiometer data is used for calibration which can easily be resampled to any other spectral resolution.

Prediction of soil constituents for spectral libraries (including test set validation)

The prediction of soil constituents for a spectral library where no corresponding chemistry is available is directly comparable to the workflow for hyperspectral imagery as presented in section 5.3. This leads to a calculated content of the considered soil constituent for each sample, i.e. spectrum. They are stored as table in ASCII format. In this procedure, the calculated parameter contents are provided as they were calculated, even if they exceed the ranges of the calibration dataset. If soil constituents are predicted for a spectral library where corresponding chemistry is available, the predicted contents are further compared to measured soil constituents and the correlation accuracy R^2_{Val} is determined. Both are given in the output table. This procedure may also be used for test set validation. This allows a user to split a set of spectra where corresponding chemistry is available into a training set on which a model is calibrated (commonly including around two thirds of the sample population) and an independent test set of the remaining samples where the calibrated model is transferred to and the predictability is assessed. In this study validation of the established regression model is performed by internal cross-validation and not with an independent set of spectra to benefit from the usage of all samples within calibration.

Realization of the prediction procedures in IDL

First, a dataset is selected as ENVI Spectral Library file or ENVI bsq image file. In case of a point spectral dataset the user is asked to decide if chemical reference data are available or not (prediction or validation). Image data are always treated as dataset for a prediction of an unknown dataset. The comparison of spatially continuously predicted soil constituents to local measurements is performed in another step. The file of the regression relationship established during calibration is selected by the user and automatically tested for consistency of data parameters like spectral resolution, band position, etc. The specific soil constituent for which the selected regression relationship was calibrated is read from the file and used as the command variable for which soil constituents' concentrations are calculated. Image spectra are arranged to 2D format without considering background pixels where no soil signature was approximated and then processed like spectral libraries. Sample/pixel spectra are parameterized, standardized using the same parameters as in calibration (mean values and standard deviations), the spectral variables found significant during calibration are extracted and the calibrated regression relationship is transferred to each sample's/pixel's variables to calculate a soil constituent's content for each of them. A prediction overview file in ASCII text format is saved including file names of input data and details of the procedure. Prediction

of soil parameters for image data results in a five band file, the contents of which are specified in section 5.3. Statistics of all bands (calculated without pixels where no soil information could be calculated) are in addition included in the overview file. They show the number and range of pixel values exceeding calibration limits and the influence of the filter operations on statistics of the result. The contents calculated for spectral library samples are given as table.

In case of prediction for a spectral library, in the same way as in calibration, tables of the calculated spectral feature variables and of the standardized variables are saved, to allow of a control of the workflow and the calculation of additional statistics. In the prediction routine for image data it was desisted from an output of the table of feature variables calculated in parameterization and after standardization (output 1b and 3b in Table D-11), because especially for image data they make up to a hundred megabytes and are not used with this dimension of data.

Batch job processing is also possible in the prediction procedure. In general it is only needed for the prediction of soil constituents from image data. Similar to within the calibration procedure it leads to the automatically storage of all overview files (see Table D-11) and applies default filter kernel sizes. Manual intervention can prevent the storage of any additional text files and allows applying user-defined kernel sizes.

D.3 Confidence assessment

Table D-12 gives an overview of the data layer and thresholds applied to calculate the confidence layer as quality indicator.

Table D-12: Detailed calculation of the confidence layer with all applied thresholds.

Image band	Step	Data layer	Layer thresholds	Assigned value	Range	Comment
Rel 1	Spectral Unmixing	RMSE	0 - 3 3 - 6 6 - 8 8 – 12 > 15	10 9 8 6 3	3 - 10	Reliability is decreased where unmixing accuracy is close to the given threshold
Rel 2	Approximation of soil residues	Unmixing cover fraction of bare soil (BS)	> 0.65 0.64 - 0.55 0.54-0.45 < 0.44	10 8 6 4	4 – 10	Reliability is decreased where bare soil fraction is low and close to the given threshold
Rel 3	Approximation of soil residues	Unmixing cover fraction of vegetation (PV + NPV)	< 0.25 0.26-0.35 0.36-0.45 > 0.45	10 8 6 4	4 - 10	Reliability is decreased where fractional cover of vegetation is close to the given thresholds. Here, the summed abundance of PV and NPV is considered
Rel 4	Quantification of soil parameters	C_{org} (band 3 + 5)	Band 3 </> mean+/-2*stdev Band 3 >/< mean+/-2*stdev Band 5 = 2 Pixel with both criteria	10 7 4 1	1 – 10	Reliability is decreased where pixel values close to the calibration borders are calculated (band 3) and/or the pixel mask (band 5) indicates the interpolation of the pixel from neighborhood (by moving window filter)
Rel 5	Quantification of soil parameters	Fe (band 3 + 5)	Band 3 </> mean+/-2*stdev Band 3 >/< mean+/-2*stdev Band 5 = 2 Pixel with both criteria	10 7 4 1	1 – 10	Reliability is decreased where pixel values close to the calibration borders are calculated (band 3) and/or the pixel mask (band 5) indicates the interpolation of the pixel from neighborhood (by moving window filter)
Rel 6	Data acquisition	Local incidence angle	99° - 81° 109 - 100° or 80 - 71° 119 - 110° or 70 - 61° 120 - 124° or 60 - 56° 129-125° or 80-71° >130° or <50°	10 9 8 6 4 1	1 – 10	Reliability is decreased where the angle of incidence is significantly differing from nadir observations, Angle of incidence for acquisition of data calculated from .sca and .gps files and elevation model
	Overall confidence index	Confidence layer 1 - 6	Mean of all single bands[1]		1 – 10	[1]Additional constraint: can be at most 2 higher than the lowest singular band index
	Final confidence classification	Overall confidence	7-10 5-6 1-4	3 2 1	1-3	Overall confidence classified to 3 classes: 3 - green: Good 2 – yellow: Medium 1 – red: Reduced

D.4 Comparison of spatially predicted soil constituents to ground reference data

The procedure to compare the map of predicted soil constituents established based on hyperspectral imagery to local ground reference measurements is based on three input datasets: the 5 band image file of the predicted values of one soil constituent, the classification file of the confidence index (3 class final confidence) and a table in ASCII format giving coordinates and chemical reference values of each sampled site. For each ground reference site and its surrounding the following information is determined for the analyzed soil constituent:

- Measured content of the soil sample taken from this location,
- Contents predicted for the pixel including the sample's location (band 3 of output map),
- Deviation of these two values in units of the chemical analysis (%),
- Pixel values and confidence flags in a 3 x 3 surrounding around the center pixel,
- Interpolated content as average of the pixels within the 3 x 3 surrounding of the center pixel, with a double weight given to the center pixel.

Based on this information the included sample sites are assigned to three classes:

- "Valid sample points": Constituent content could be predicted for the center pixel and confidence of all pixels in the 3 x 3 surrounding is flagged as "good". Two pixels are allowed to show a medium confidence.
- "Influenced sample points": Constituent content could be predicted for the center pixel, but confidence of the center pixel is not flagged as "good", or
  in the 3 x 3 neighborhood three or more pixels show a "medium" or "reduced" confidence, or
  iIn the 3 x 3 neighborhood two or more pixels show no values.
- "No value calculated": No value was calculated for the center pixel

Figure D-12 gives an example for the output file for the comparison of ground reference samples with predicted soil constituents based on hyperspectral imagery.

In addition to this detailed comparison, a table in ASCII format is delivered giving the in-situ measured content and the predicted content for the valid and also the influenced sample points. This table is used to generate scatter plots of measured vs. calculated contents (see e.g. Figure F-21) and to derive the correlation of them.

```
-----------------------------------------------------
      Comparison of measured and calculated values
-----------------------------------------------------

Input files:
Image file (calc chem contents):           … .bsq
Classification file of reliability index:  … _ReliabilityClass.bsq
Table of measured chemical contents:       … .txt

Interpolated values include only valid pixels

-----------------------------------------------------
Number of sample points where measurements exist included in subset:    12
Number of sample points where a value was calculated (NE 0):            6
Number of sample points where with valid reliability (Rel_class EQ 3):  4
-----------------------------------------------------
Correlation coefficient determined of all 9 valid pairs of calculated and measured values (calc
NE 0)       R² = 0.34
Correlation coefficient determined of 6 pairs of calculated and measured values
                  R² = 0.32
-----------------------------------------------------
* Details of sample points, where measurements exist and no influences could be detected
                     4 samples
-> Value is claculated and no reduced reliability at point and in neighbourhood
meas = chemical value measured from field sample at this location
calc = chemical value calculated by delineation from hyperspectral data
int  = value interpolated from all valid points (NE 0) in direct surrounding_calc
       (given in matrix)
dev  = deviation calc-meas in % ((calc - meas)/meas)

* 150
      meas:     0.26
      calc:     0.73
      int :     0.79 (9)
      dev (abs): +   0.47 %
      Surrounding:
                0.719388    0.877678     1.00770
                0.626613    0.728353     0.90369
                0.738834    0.687607     0.90333
      Reiability flags of surrounding:
                        3           3           3
                        3           3           3
                        3           3           3

…

-----------------------------------------------------
* Details of outlier sample points, where influence of any ingoing data could be detected
                    2 samples
Outliers: where - reliability index of pixel not equal "good" (flag 3) OR
                - in 3x3 neighbourhood at least 1
 pixel(s) show no values (0) or reduced reliability (flag 1)
                - in 3x3 neighbourhood more than 3
 pixels  show medium reliability (flag 2) OR
-> then influence of any ingoing data (i.e. influence by neighbourhood to "bad" point) is
expected

* 64
      meas:     0.22
      calc:     1.13
      int :     1.09 (7)
      Surrounding:
                        -           -     1.56500
                  1.03363     1.12516     1.14372
                 0.950373    0.894371    0.893651
      Reiability flags of surrounding:
                        0           0           2
                        3           3           3
                        3           3           3
…
```

Figure D-12: ASCII text file giving an example for the comparison of local ground reference data with the constituent contents which were predicted based on hyperspectral imagery.

D.5 Spectral unmixing

Endmember extraction and spectral unmixing is performed using routines of existing approaches developed by Rogge et al. (2006 and 2007). Table D-13 lists some settings applied for processing.

Table D-13: Settings for the endmember extraction and spectral unmixing. For further details see Rogge et al. (2006 and 2007) and references therein.

	Input parameter	Value	Comment
SSEE	Window size	20 x 20 pixels	Size of the spatial subsets, in which SSEE determines the local EMs
	Svd threshold	0.99	Form each subset, the eigenvectors that account for 99 % of the total spectral variance are retained
	Spectral angle	1	Differences between EMs are determined based on spectral angle
ISMA	Maximum no. of EMs	5	Maximum number of endmember to use for each pixel
	RMSE threshold	0.10	Minimum difference of two successive iterations which determines the critical point for the optimal endmember set for the pixel

D.6 Processing times

In the following some processing times for the application of the individual steps of the workflow are given. They refer to the processing of the subsets of 1000 lines, the entire dataset was split in (each between 1 600 and 1 750 samples, see section 5.3). They can be used as rough indicator as they depend on RAM and available computing memory. IAS spatial-spectral filtering needed processing times between 12 and 26 minutes for each subset. The ISMA unmixing routine takes between 2 and 3 hours for the iterative unmixing of 1000 lines of imagery. Approximating the soil residuum based on original HyMap imagery and unmixing cover fractions takes about 4 to 6 minutes for each subset. Calibration of the functional regression relationship based on spectral libraries of 125 and 163 samples takes about 2 seconds. The processing duration of the final delineation of soil constituents from the image data of approximated soil signatures depends on the number of pixels for which soil signatures could be approximated within each subset of 1 000 lines. For the most subsets, less than 10 minutes were needed, while one subset where exceptionally many pixels are included showing a valid soil signature took about 31 minutes for being processed.

E Details on processing results

This appendix provides details on the generated results, including statistics that are referred to in the main text.

E.1 Calibrated regression models

The following tables show parameters of the regression models established with the proposed feature-based regression technique for soil organic carbon, iron oxides and clay content based on the bare soil field spectra in HyMap's spectral resolution. The five most important variables in the regression relationship are highlighted.

Table E-14: Details for regression relationship for soil organic carbon established by MLR based on spectral feature variables.

	Influence on regression [%]	Regression coefficient	Details for standardization	
			Mean	Standard deviation
Constant factor		1.14		
depth_max_corg_AF1_1730	9.99	0.48	0.002	0.003
wvl_depth_max_corg_AF1_1730	-1.38	-0.07	1741.43	18.84
area_corg_AF1_1730	-6.84	-0.33	0.13	0.23
asym_fac_corg_AF1_1730	0.98	0.048	1.29	0.66
depth_max_corg_AF2_2330	-1.50	-0.07	0.04	0.01
wvl_depth_max_corg_AF2_2330	-5.25	-0.26	2337.76	13.27
depth_lit_corg_AF2_2330	-2.72	-0.13	0.03	0.02
width_corg_AF2_2330	-0.40	-0.02	111.28	16.58
asym_fac_corg_AF2_2330	7.84	0.38	0.95	0.67
slope_HF1_vis	-17.11	-0.83	6.40	1.49
height_HF1_vis	-11.25	-0.55	1931.99	414.84
slope_HF2_swir1	-6.93	-0.34	0.60	0.20
height_HF2_swir1	27.80	1.35	3865.84	711.60
Sum of influences	100.00			

Table E-15: Details for regression relationship for iron oxides established by MLR based on spectral feature variables.

	Influence on regression [%]	Regression coefficient	Details for standardization	
			Mean	Standard deviation
Constant factor		2.80		
depth_max_iron_AF1_550	-10.59	-2.19	0.14	0.06
wvl_depth_max_iron_AF1_550	-1.91	-0.39	506.27	10.91
depth_lit_iron_AF1_550	-7.98	-1.65	0.09	0.06
width_iron_AF1_550	-1.39	-0.29	136.43	15.68
area_iron_AF1_550	16.34	3.38	11.36	5.83
asym_fac_iron_AF1_550	1.63	0.34	0.76	0.28
depth_max_iron_AF2_700	10.11	2.09	0.004	0.01
wvl_depth_max_iron_AF2_700	0.19	0.04	687.18	13.16
width_iron_AF2_700	-1.39	-0.29	57.16	42.16
area_iron_AF2_700	-9.28	-1.92	0.22	0.73
depth_max_iron_AF3_900	10.92	2.26	0.02	0.01
wvl_depth_max_iron_AF3_900	-0.92	-0.19	954.90	88.80
depth_lit_iron_AF3_900	2.60	0.54	0.01	0.01
width_iron_AF3_900	0.48	0.10	332.30	142.44
area_iron_AF3_900	-12.04	-2.49	3.62	4.18
asym_fac_iron_AF3_900	-0.77	-0.16	1.12	0.80
slope_iron_CV1_550/590	-3.23	-0.67	11.29	3.40
slope_CO1_vis	5.18	1.07	6.40	1.49
height_CO1_vis	-3.07	-0.63	1931.99	414.84
Sum of influences	100.00			

Table E-16: Details for regression relationship for clay established by MLR based on spectral feature variables.

	Influence on regression [%]	Regression coefficient	Details for standardization	
			Mean	Standard deviation
Constant factor		5.73		
depth_max_clay_AF1_2200	8.50	0.62	0.07	0.02
wvl_depth_max_clay_AF1_2200	1.52	0.11	2194.61	7.89
depth_lit_clay_AF1_2200	-6.28	-0.46	0.06	0.02
area_clay_AF1_2200	17.82	1.31	4.27	1.16
asym_fac_clay_AF1_2200	-3.76	-0.28	0.85	0.24
wvl_depth_max_clay_AF2_2340	-5.50	-0.40	2338.02	13.38
depth_lit_clay_AF2_2340	-10.09	-0.74	0.03	0.02
width_clay_AF2_2340	13.56	0.99	108.40	15.42
area_clay_AF2_2340	6.21	0.45	1.80	0.78
asym_fac_clay_AF2_2340	-3.37	-0.25	0.93	0.60
slope_CO1_vis	-1.26	-0.09	6.40	1.49
slope_CO2_swir1	-12.18	-0.89	0.60	0.20
height_CO2_swir1	9.94	0.73	3865.84	711.60
Sum of influences	100.00			

E.2 Endmember extraction

The followings figures show the endmember sets of the scale 1, 2 and 3 subsets as they were manually selected from the semi-automatically derived EMs using the SSEE approach.

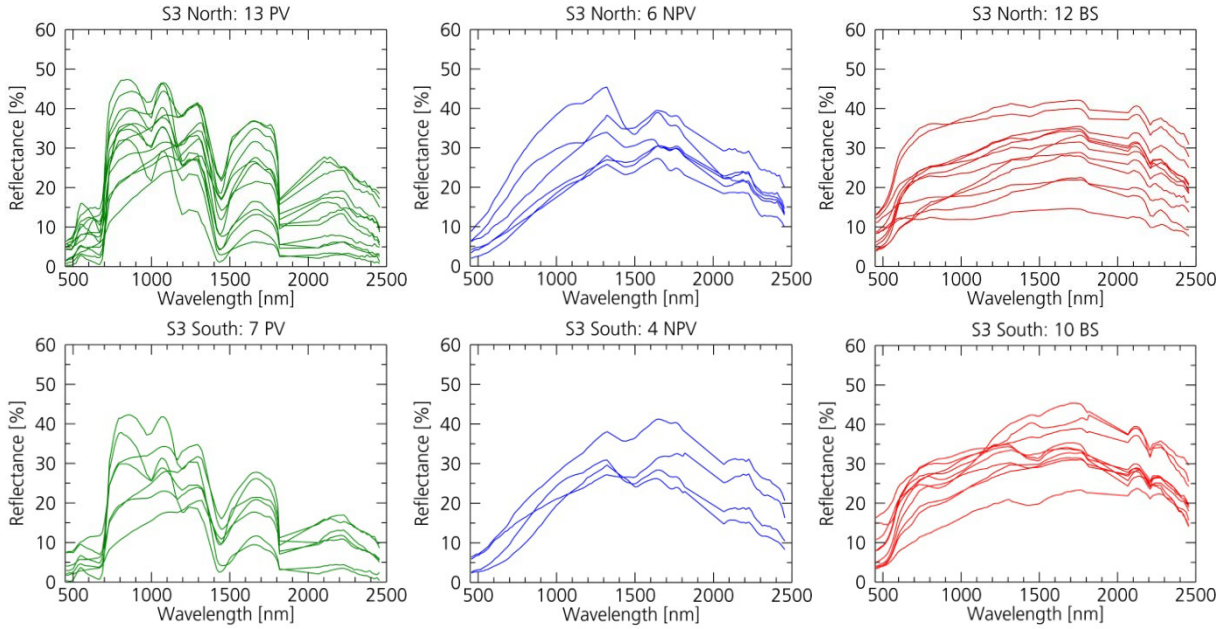
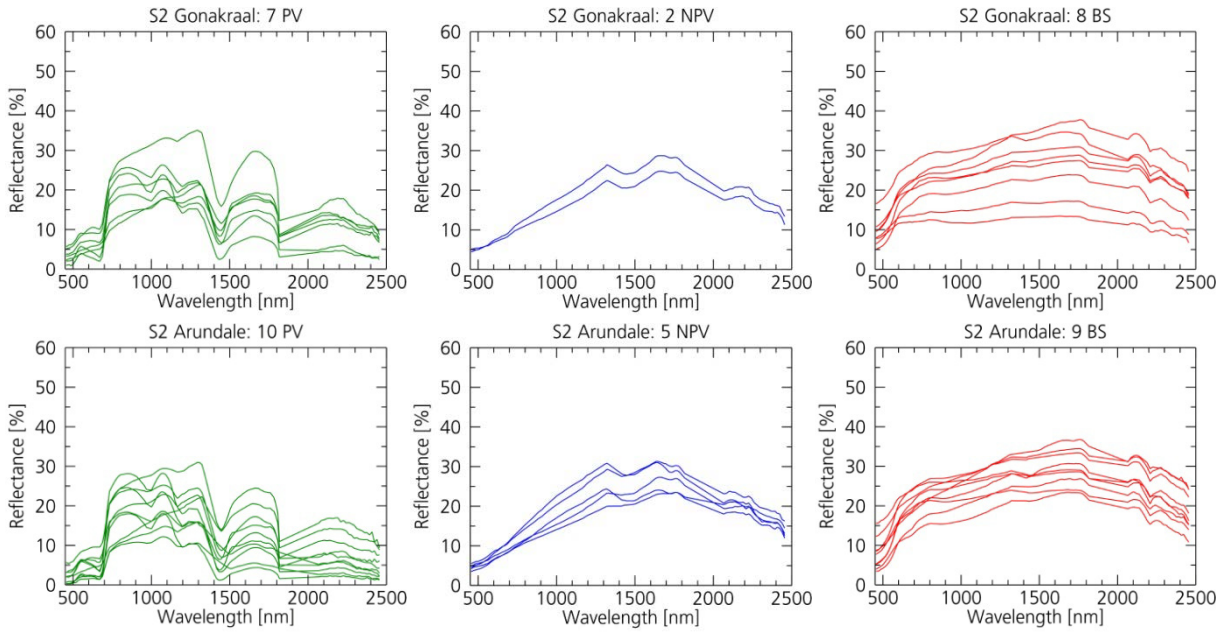


Figure E-13: Endmember sets for scale 3 subsets.



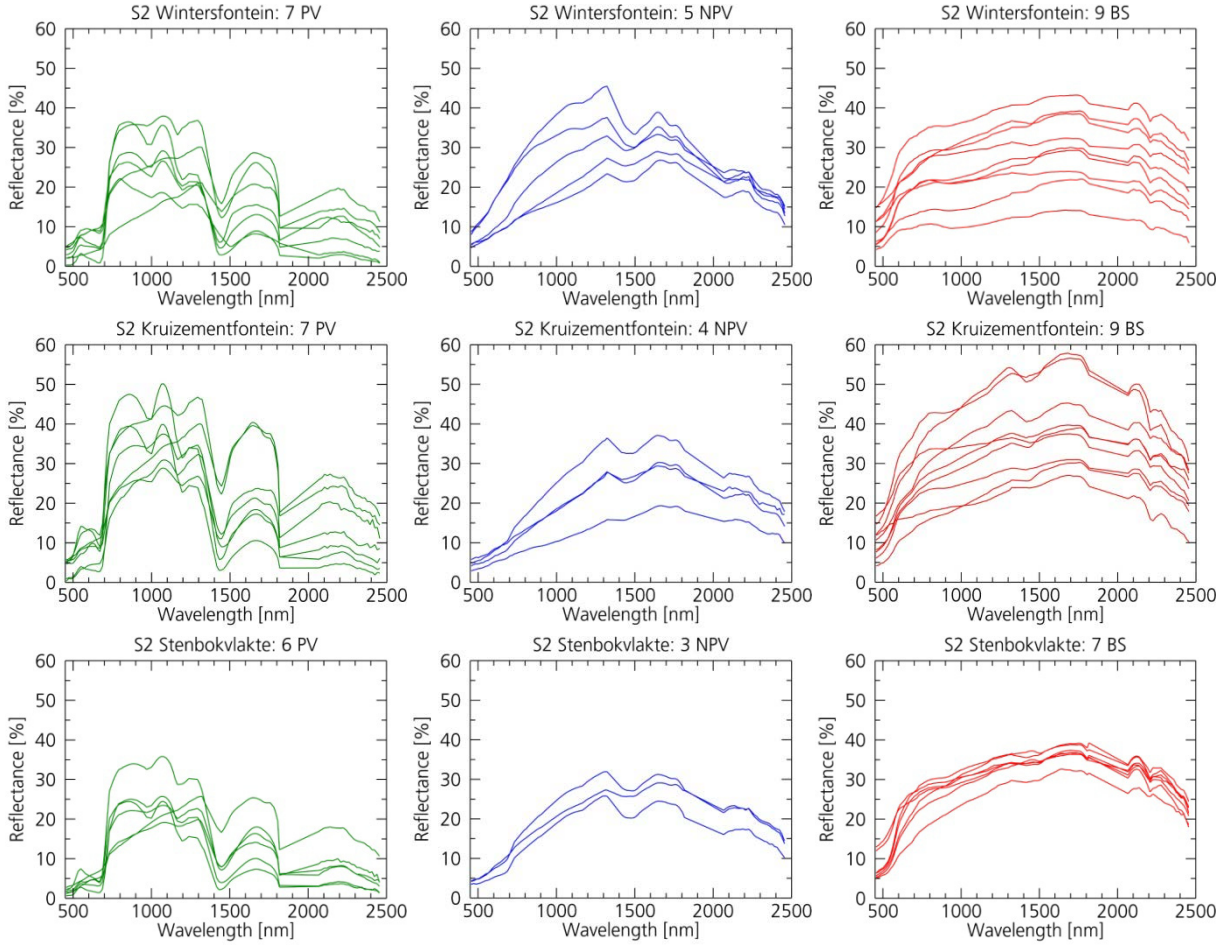
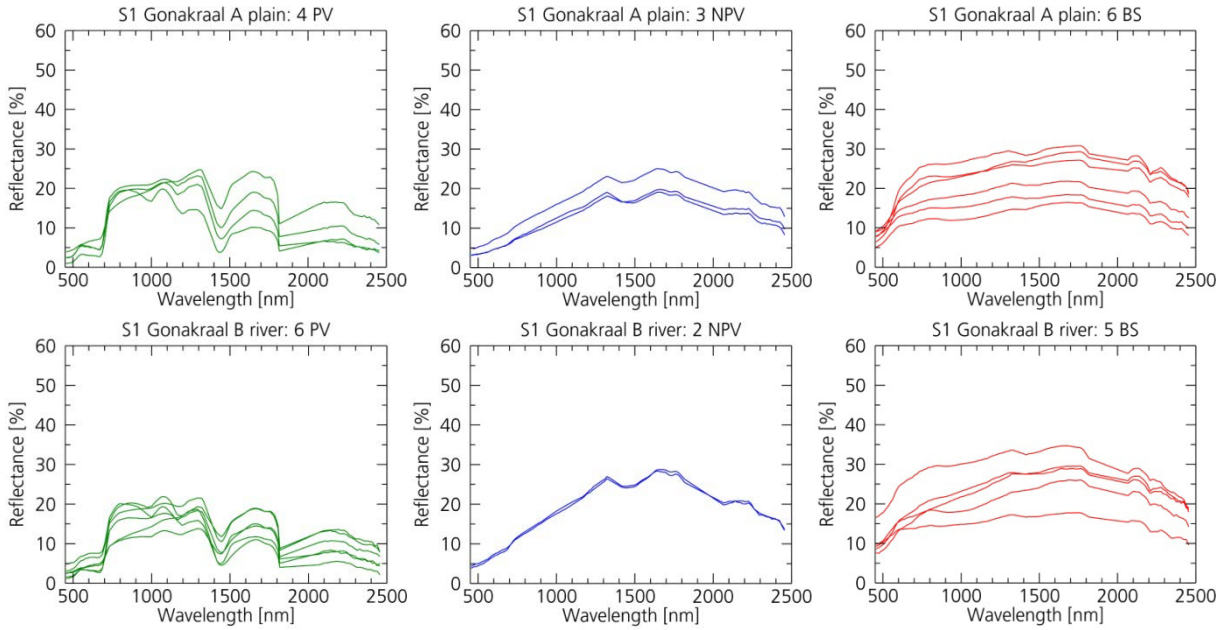


Figure E-14: Endmember sets for scale 2 subsets.



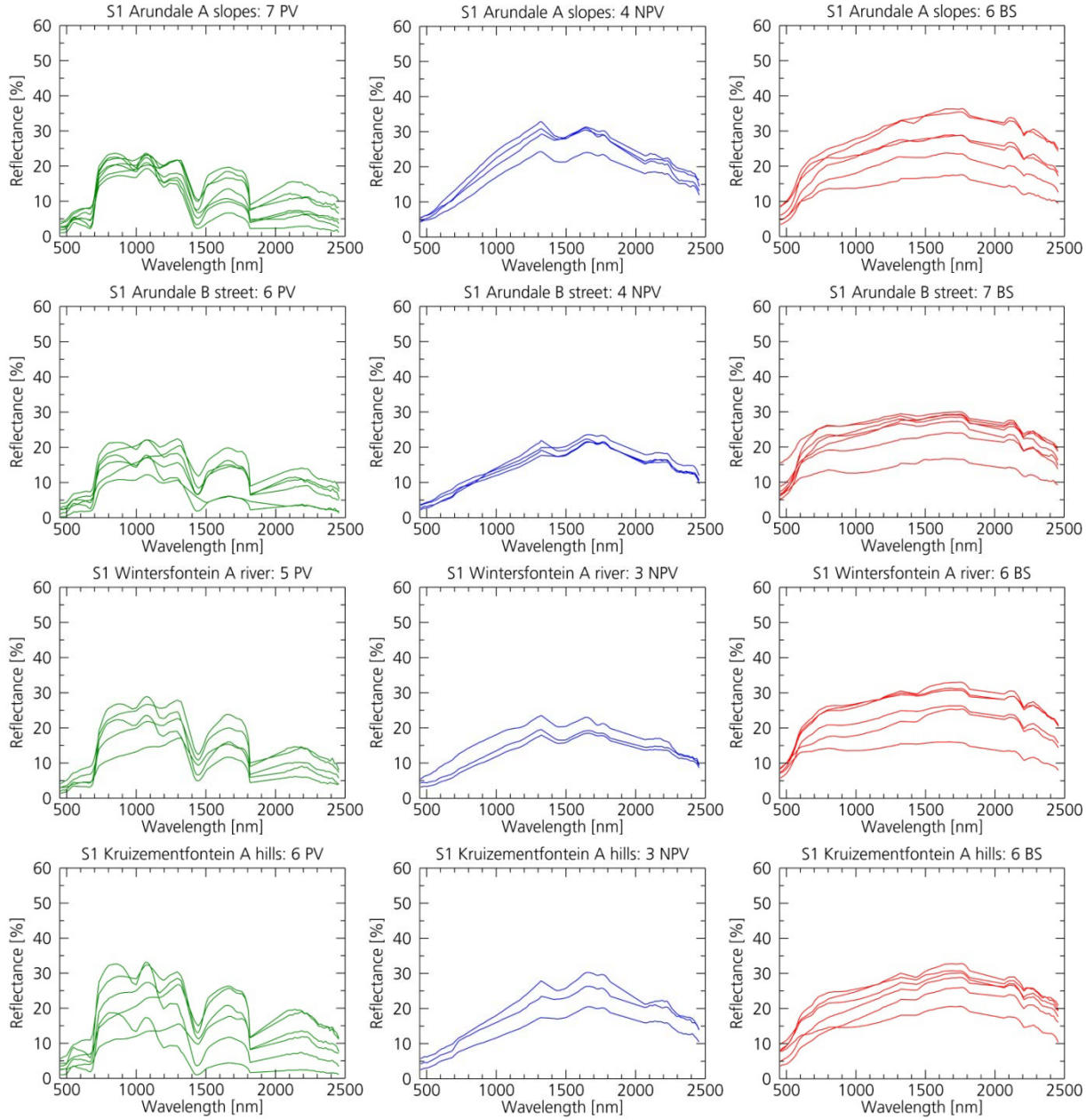


Figure E-15: Endmember sets for scale 1 subsets.

E.3 Spectral mixture analysis

Table E-17: Statistics of ISMA spectral unmixing results for the three scales and their subsets. Distinctive features mentioned in section 6.2.1 are highlighted.

	Subsets	Ground cover fractions PV	NPV	BS	Summed cover fractions	RMSE
Scale 1	Gonakraal -A plain	Mean: 0.10 Stdev: 0.10 Min: 0.00 Max: 1.54	Mean: 0.34 Stdev: 0.15 Min: 0.00 Max: 1.62	Mean: 0.59 Stdev: 0.21 Min: 0.00 Max: 2.18	Mean: 0.96 Stdev: 0.14 Min: 0.34 Max: 2.51	Mean: 1.30 Stdev: 0.59 Min: 0.00 Max: 9.06
	Gonakraal -B river	Mean: 0.11 Stdev: 0.11 Min: 0.00 Max: 1.28	Mean: 0.20 Stdev: 0.11 Min: 0.00 Max: 1.52	Mean: 0.57 Stdev: 0.20 Min: 0.00 Max: 1.75	Mean: 0.77 Stdev: 0.14 Min: 0.16 Max: 2.61	Mean: 1.53 Stdev: 0.69 Min: 0.00 Max: 8.29
	Arundale -A slopes	Mean: 0.23 Stdev: 0.27 Min: 0.00 Max: 1.44	Mean: 0.21 Stdev: 0.11 Min: 0.00 Max: 1.08	Mean: 0.57 Stdev: 0.19 Min: 0.00 Max: 2.09	Mean: 0.83 Stdev: 0.11 Min: 0.10 Max: 2.62	Mean: 1.74 Stdev: 0.66 Min: 0.00 Max: 22.95
	Arundale - B street	Mean: 0.16 Stdev: 0.16 Min: 0.00 Max: 1.93	Mean: 0.28 Stdev: 0.15 Min: 0.00 Max: 1.46	Mean: 0.50 Stdev: 0.20 Min: 0.00 Max: 1.35	Mean: 0.83 Stdev: 0.16 Min: 0.11 Max: 2.15	Mean: 1.08 Stdev: 0.44 Min: 0.00 Max: 6.67
	Wintersfontein - A river	Mean: 0.08 Stdev: 0.14 Min: 0.00 Max: 1.35	Mean: 0.28 Stdev: 0.15 Min: 0.00 Max: 1.44	Mean: 0.57 Stdev: 0.21 Min: 0.00 Max: 2.58	Mean: 0.86 Stdev: 0.15 Min: 0.27 Max: 3.27	Mean: 1.23 Stdev: 0.61 Min: 0.00 Max: 10.21
	Kruizementfontein - A hills	Mean: 0.23 Stdev: 0.21 Min: 0.00 Max: 1.42	Mean: 0.41 Stdev: 0.16 Min: 0.00 Max: 1.57	Mean: 0.40 Stdev: 0.19 Min: 0.00 Max: 2.75	Mean: 0.83 Stdev: 0.18 Min: 0.07 Max: 3.20	Mean: 2.34 Stdev: 1.15 Min: 0.00 Max: 12.93
Scale 2	Gonakraal	Mean: 0.13 Stdev: 0.16 Min: 0.00 Max: 1.29	Mean: 0.29 Stdev: 0.11 Min: 0.00 Max: 1.51	Mean: 0.44 Stdev: 0.19 Min: 0.00 Max: 2.16	Mean: 0.79 Stdev: 0.12 Min: 0.16 Max: 2.62	Mean: 1.63 Stdev: 0.84 Min: 0.00 Max: 10.33
	Arundale	Mean: 0.30 Stdev: 0.22 Min: 0.00 Max: 2.01	Mean: 0.35 Stdev: 0.17 Min: 0.00 Max: 1.11	Mean: 0.36 Stdev: 0.18 Min: 0.00 Max: 2.10	Mean: 0.82 Stdev: 0.15 Min: 0.11 Max: 2.79	Mean: 1.35 Stdev: 0.64 Min: 0.00 Max: 17.59
	Wintersfontein	Mean: 0.14 Stdev: 0.17 Min: 0.00 Max: 2.19	Mean: 0.18 Stdev: 0.12 Min: 0.00 Max: 1.89	Mean: 0.57 Stdev: 0.21 Min: 0.00 Max: 2.80	Mean: 0.74 Stdev: 0.16 Min: 0.08 Max: 2.82	Mean: 1.49 Stdev: 0.82 Min: 0.00 Max: 23.55
	Kruizementfontein	Mean: 0.19 Stdev: 0.20 Min: 0.00 Max: 1.38	Mean: 0.37 Stdev: 0.18 Min: 0.00 Max: 4.11	Mean: 0.22 Stdev: 0.14 Min: 0.00 Max: 2.20	Mean: 0.58 Stdev: 0.19 Min: 0.00 Max: 4.27	Mean: 1.71 Stdev: 0.78 Min: 0.00 Max: 36.90
	Stenbokvlakte	Mean: 0.33 Stdev: 0.28 Min: 0.00 Max: 1.51	Mean: 0.39 Stdev: 0.14 Min: 0.00 Max: 1.32	Mean: 0.30 Stdev: 0.15 Min: 0.00 Max: 2.16	Mean: 0.85 Stdev: 0.09 Min: 0.13 Max: 2.91	Mean: 1.94 Stdev: 1.15 Min: 0.00 Max: 33.65
Scale 3	Northern section	Mean: 0.15 Stdev: 0.16 Min: 0.00 Max: 1.46	Mean: 0.25 Stdev: 0.12 Min: 0.00 Max: 2.12	Mean: 0.43 Stdev: 0.21 Min: 0.00 Max: 4.39	Mean: 0.65 Stdev: 0.17 Min: 0.00 Max: 4.96	Mean: 2.07 Stdev: 1.29 Min: 0.00 Max: 190.14
	Southern section	Mean: 0.42 Stdev: 0.29 Min: 0.00 Max: 1.51	Mean: 0.33 Stdev: 0.15 Min: 0.00 Max: 1.24	Mean: 0.27 Stdev: 0.18 Min: 0.00 Max: 3.08	Mean: 0.79 Stdev: 0.13 Min: 0.07 Max: 3.08	Mean: 2.22 Stdev: 1.18 Min: 0.00 Max: 39.98

E.4 Delineation of soil constituents

All bands of the maps of delineated iron oxides and clay contents are shown in Figure E-16 and Figure E-17. They show the influence of the post-processing including the selective filter operation.

Maps of spectral variables

The spatial variation of the spectral variables can be used to visually assess the significance of each variable for the regression model and to identify local artifacts. Figure E-18 and Figure E-19 show example maps of the 13 derived spectral variables for soil organic carbon prediction on the S2 subset Gonakraal. For each pixel, (1) the spectral variables are determined from the spectra of the hyperspectral imagery of the approximated soil residues, (2) standardized in the same way as the field spectra used for calibrating the regression relationship and (3) multiplied with the regression coefficient of the variable. Thus, for each pixel the sum of the spectral variables shown in Figure E-18 and Figure E-19 gives the predicted soil organic carbon content. Besides giving the spatial distribution of the variables, this way of presentation allows the direct assessment of the influence of each variable on the finally predicted contents. In Figure E-18 and Figure E-19 the variables are ordered according to their influence on the regression relationship. Each variable is given in % C_{org}.

Statistics of predicted soil constituents

The following tables show statistics of the predicted contents for the entire dataset (Table E-18) and the individual subsets of the three scales (Table E-19, Table E-20, Table E-21). The tables show the little changes of the statistics introduced by the filter operation and the interpolation of missing pixels.

Table E-18: Overview of histogram statistics for all predicted soil constituents for the entire dataset.

Parameter	All values (band 1)	Unfiltered (band 2)	Filter kernel 3 x 3 (band 3)	Filter kernel 5 x 5 (band 4)
Soil organic carbon	Mean: 1.20 Stdev: 0.62 Min: -1.13 Max: 6.45	Mean: 1.22 Stdev: 0.60 Min: 0.21 Max: 5.84	Mean: 1.22 Stdev: 0.53 Min: 0.21 Max: 5.84	Mean: 1.23 Stdev: 0.49 Min: 0.21 Max: 5.84
Iron oxides	Mean: 4.17 Stdev: 1.85 Min: -9.33 Max: 29.97	Mean: 4.10 Stdev: 1.58 Min: 0.90 Max: 10.62	Mean: 4.12 Stdev: 1.42 Min: 0.90 Max: 10.62	Mean: 4.14 Stdev: 1.32 Min: 0.90 Max: 10.62
Clay content	Mean: 6.92 Stdev: 2.75 Min: -7.96 Max: 24.04	Mean: 7.12 Stdev: 2.45 Min: 0.00 Max: 22.33	Mean: 7.11 Stdev: 2.23 Min: 0.00 Max: 22.33	Mean: 7.06 Stdev: 2.10 Min: 0.00 Max: 22.30

Table E-19: Overview of histogram statistics for the prediction of topsoil organic carbon for scales 2 and 3.

	Subsets	All values (band 1)	Unfiltered (band 2)	Filter kernel 3 x 3 (band 3)	Filter kernel 5 x 5 (band 4)
Scale 2	Gonakraal	Mean: 1.04 Stdev: 0.47 Min: -0.57 Max: 4.92	Mean: 1.05 Stdev: 0.46 Min: 0.21 Max: 4.92	Mean: 1.05 Stdev: 0.41 Min: 0.21 Max: 4.92	Mean: 1.07 Stdev: 0.38 Min: 0.21 Max: 4.62
	Arundale	Mean: 1.34 Stdev: 0.65 Min: -1.23 Max: 33.24	Mean: 1.35 Stdev: 0.62 Min: 0.21 Max: 5.82	Mean: 1.36 Stdev: 0.57 Min: 0.21 Max: 5.66	Mean: 1.37 Stdev: 0.54 Min: 0.21 Max: 5.55
	Wintersfontein	Mean: 1.74 Stdev: 0.86 Min: -0.77 Max: 34.58	Mean: 1.74 Stdev: 0.85 Min: 0.21 Max: 5.85	Mean: 1.76 Stdev: 0.76 Min: 0.21 Max: 5.85	Mean: 1.80 Stdev: 0.71 Min: 0.21 Max: 5.76
	Kruizementfontein	Mean: 1.41 Stdev: 0.79 Min: -1.19 Max: 6.92	Mean: 1.48 Stdev: 0.74 Min: 0.21 Max: 5.67	Mean: 1.48 Stdev: 0.69 Min: 0.21 Max: 5.48	Mean: 1.47 Stdev: 0.65 Min: 0.21 Max: 5.48
	Stenbokvlakte	Mean: 1.89 Stdev: 0.71 Min: -0.60 Max: 6.57	Mean: 1.92 Stdev: 0.68 Min: 0.21 Max: 5.40	Mean: 1.91 Stdev: 0.60 Min: 0.21 Max: 5.40	Mean: 1.91 Stdev: 0.54 Min: 0.21 Max: 5.40
Scale 3	Northern section	Mean: 1.12 Stdev: 0.56 Min: -1.13 Max: 6.44	Mean: 1.14 Stdev: 0.54 Min: 0.21 Max: 5.84	Mean: 1.14 Stdev: 0.47 Min: 0.21 Max: 5.84	Mean: 1.15 Stdev: 0.42 Min: 0.21 Max: 5.84
	Southern section	Mean: 1.94 Stdev: 0.66 Min: -0.59 Max: 6.45	Mean: 1.94 Stdev: 0.65 Min: 0.21 Max: 5.59	Mean: 1.94 Stdev: 0.57 Min: 0.21 Max: 5.59	Mean: 1.94 Stdev: 0.52 Min: 0.21 Max: 5.59

Table E-20: Overview of histogram statistics for the prediction of iron oxides for scales 2 and 3.

	Subsets	All values (band 1)	Unfiltered (band 2)	Filter kernel 3 x 3 (band 3)	Filter kernel 5 x 5 (band 4)
Scale 2	Gonakraal	Mean: 3.94 Stdev: 1.09 Min: -7.21 Max: 14.16	Mean: 3.95 Stdev: 1.07 Min: 0.90 Max: 10.58	Mean: 3.95 Stdev: 0.93 Min: 0.90 Max: 10.47	Mean: 3.96 Stdev: 0.85 Min: 0.90 Max: 10.47
	Arundale	Mean: 4.01 Stdev: 1.68 Min: -83.20 Max: 20.22	Mean: 4.01 Stdev: 1.59 Min: 0.90 Max: 10.62	Mean: 4.02 Stdev: 1.49 Min: 0.90 Max: 10.62	Mean: 4.04 Stdev: 1.41 Min: 0.90 Max: 10.62
	Wintersfontein	Mean: 3.46 Stdev: 1.62 Min: -42.26 Max: 33.02	Mean: 3.43 Stdev: 1.26 Min: 0.90 Max: 10.62	Mean: 3.44 Stdev: 1.08 Min: 0.90 Max: 10.62	Mean: 3.46 Stdev: 1.00 Min: 0.90 Max: 10.62
	Kruizementfontein	Mean: 4.07 Stdev: 1.60 Min: -12.70 Max: 20.86	Mean: 4.07 Stdev: 1.46 Min: 0.90 Max: 10.62	Mean: 4.07 Stdev: 1.32 Min: 0.90 Max: 10.62	Mean: 4.06 Stdev: 1.21 Min: 0.90 Max: 10.62
	Stenbokvlakte	Mean: 3.29 Stdev: 1.73 Min: -3.21 Max: 15.18	Mean: 3.55 Stdev: 1.48 Min: 0.90 Max: 10.59	Mean: 3.54 Stdev: 1.38 Min: 0.90 Max: 10.59	Mean: 3.50 Stdev: 1.29 Min: 0.90 Max: 10.59
Scale 3	Northern section	Mean: 4.06 Stdev: 1.73 Min: -9.34 Max: 29.97	Mean: 4.02 Stdev: 1.52 Min: 0.90 Max: 10.62	Mean: 4.04 Stdev: 1.36 Min: 0.90 Max: 10.62	Mean: 4.06 Stdev: 1.26 Min: 0.90 Max: 10.62
	Southern section	Mean: 5.20 Stdev: 2.50 Min: -4.63 Max: 27.63	Mean: 4.93 Stdev: 1.90 Min: 0.90 Max: 10.62	Mean: 4.93 Stdev: 1.74 Min: 0.90 Max: 10.62	Mean: 4.92 Stdev: 1.61 Min: 0.90 Max: 10.62

Table E-21: Overview of histogram statistics for the prediction of clay content for scales 2 and 3.

	Subsets	All values (band 1)	Unfiltered (band 2)	Filter kernel 3 x 3 (band 3)	Filter kernel 5 x 5 (band 4)
Scale 2	Gonakraal	Mean: 8.79 Stdev: 1.86 Min: -1.38 Max: 18.57	Mean: 8.79 Stdev: 1.86 Min: 0.01 Max: 18.57	Mean: 8.78 Stdev: 1.64 Min: 0.01 Max: 16.74	Mean: 8.73 Stdev: 1.51 Min: 0.02 Max: 16.74
	Arundale	Mean: 7.22 Stdev: 2.37 Min: -7.53 Max: 19.71	Mean: 7.25 Stdev: 2.32 Min: 0.00 Max: 19.71	Mean: 7.22 Stdev: 2.08 Min: 0.00 Max: 17.42	Mean: 7.16 Stdev: 1.93 Min: 0.00 Max: 17.42
	Wintersfontein	Mean: 6.38 Stdev: 2.46 Min: -6.32 Max: 19.81	Mean: 6.46 Stdev: 2.35 Min: 0.00 Max: 19.81	Mean: 6.41 Stdev: 2.05 Min: 0.00 Max: 18.99	Mean: 6.34 Stdev: 1.91 Min: 0.00 Max: 16.74
	Kruizementfontein	Mean: 5.65 Stdev: 2.11 Min: -4.03 Max: 16.87	Mean: 5.66 Stdev: 2.09 Min: 0.00 Max: 16.87	Mean: 5.65 Stdev: 1.94 Min: 0.00 Max: 16.87	Mean: 5.65 Stdev: 1.83 Min: 0.00 Max: 16.87
	Stenbokvlakte	Mean: 1.91 Stdev: 2.33 Min: -5.75 Max: 16.57	Mean: 2.78 Stdev: 1.71 Min: 0.00 Max: 16.57	Mean: 2.78 Stdev: 1.53 Min: 0.00 Max: 16.57	Mean: 2.76 Stdev: 1.38 Min: 0.00 Max: 16.57
Scale 3	Northern section	Mean: 7.43 Stdev: 2.21 Min: -7.96 Max: 20.08	Mean: 7.46 Stdev: 2.16 Min: 0.00 Max: 20.08	Mean: 7.45 Stdev: 1.90 Min: 0.00 Max: 20.08	Mean: 7.41 Stdev: 1.75 Min: 0.00 Max: 18.54
	Southern section	Mean: 2.10 Stdev: 2.71 Min: -7.50 Max: 24.04	Mean: 3.11 Stdev: 2.13 Min: 0.00 Max: 22.33	Mean: 3.10 Stdev: 1.88 Min: 0.00 Max: 22.33	Mean: 3.10 Stdev: 1.73 Min: 0.00 Max: 22.33

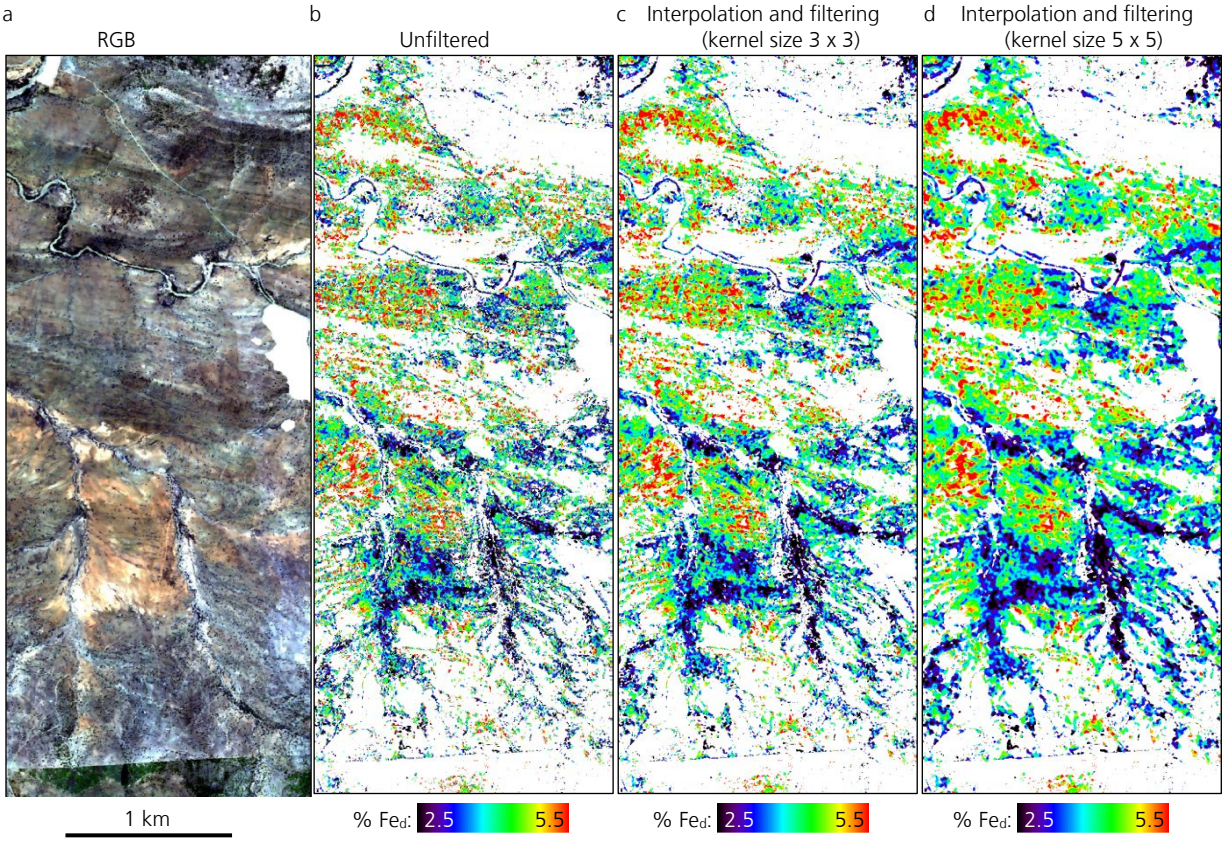


Figure E-16: Unfiltered and filtered maps of predicted topsoil iron oxides content for the Gonakraal area (scale 2 data). True color image with R: 640 nm, G: 549 nm, B: 456 nm.

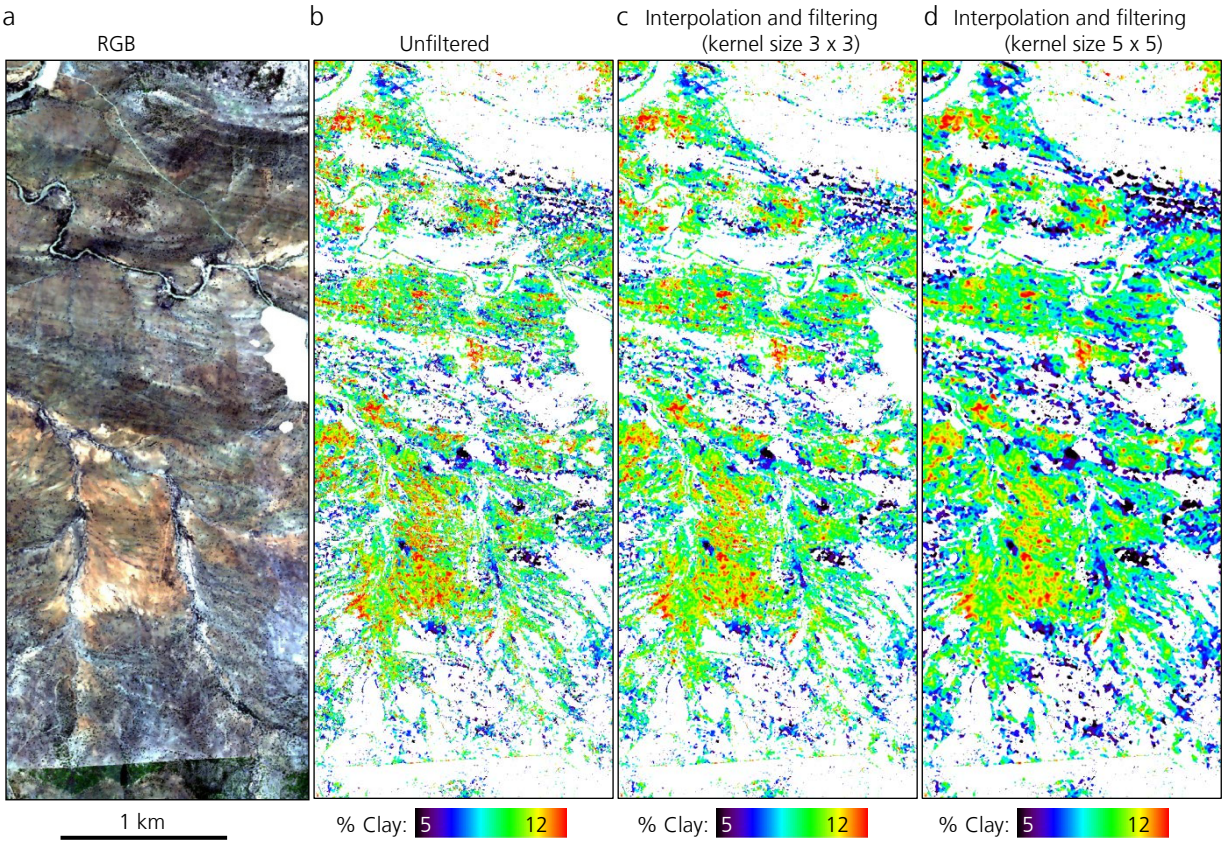


Figure E-17: Unfiltered and filtered maps of predicted topsoil clay content for the Gonakraal area (scale 2 data). True color image with R: 640 nm, G: 549 nm, B: 456 nm.

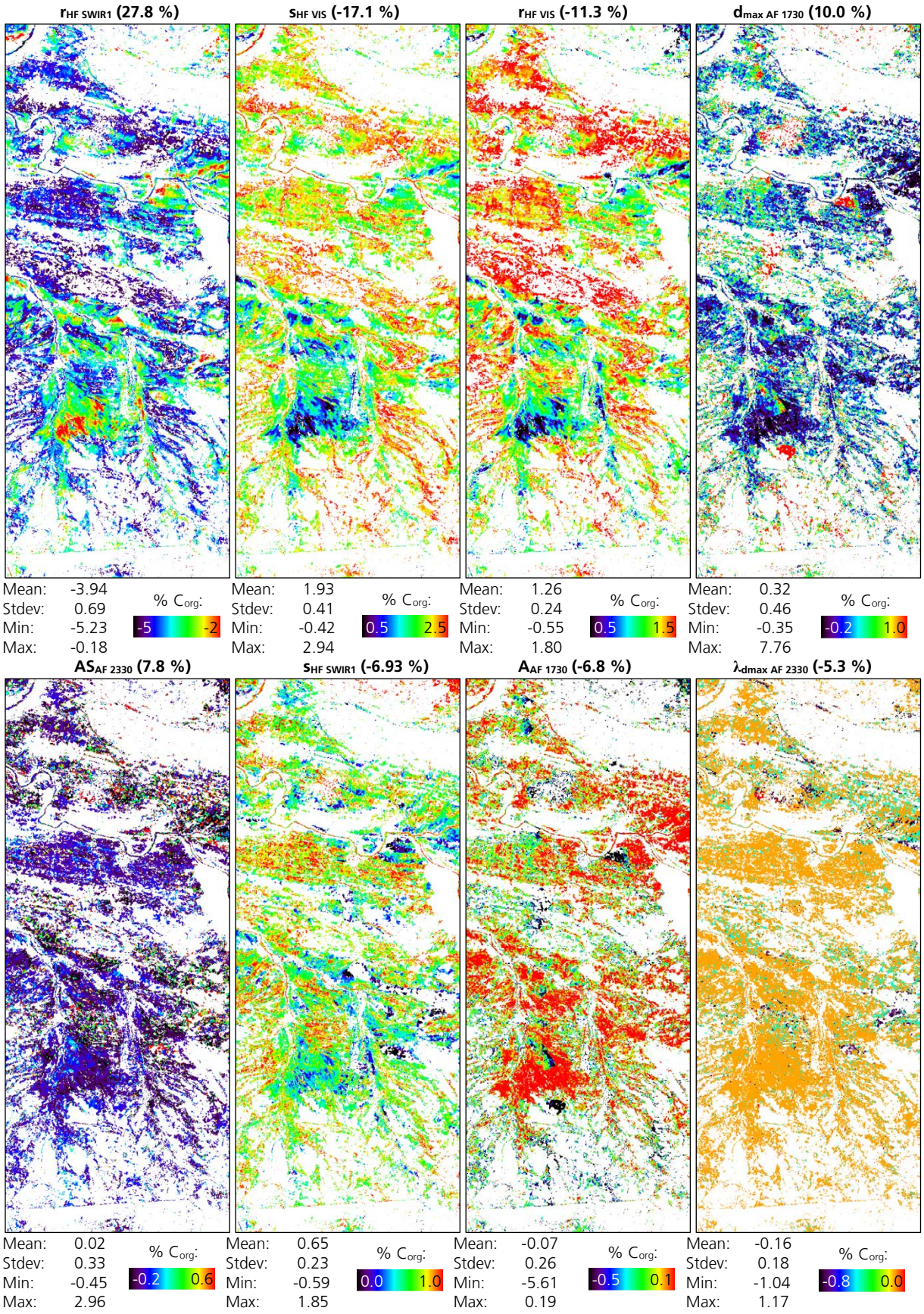


Figure E-18: Maps of spectral variables after standardization and multiplication with the regression coefficient of the S2 Gonakraal subset, Part I.
Symbols: A_{AF}: area, $d_{max\ AF}$: maximum depth, $\lambda_{dmax\ AF}$: wavelength of $d_{max\ AF}$, AS_{AF}: asymmetry factor of absorption features (AF). r_{HF}: mean reflectance in interval, s_{HF}: slope in interval of hull features (HF).

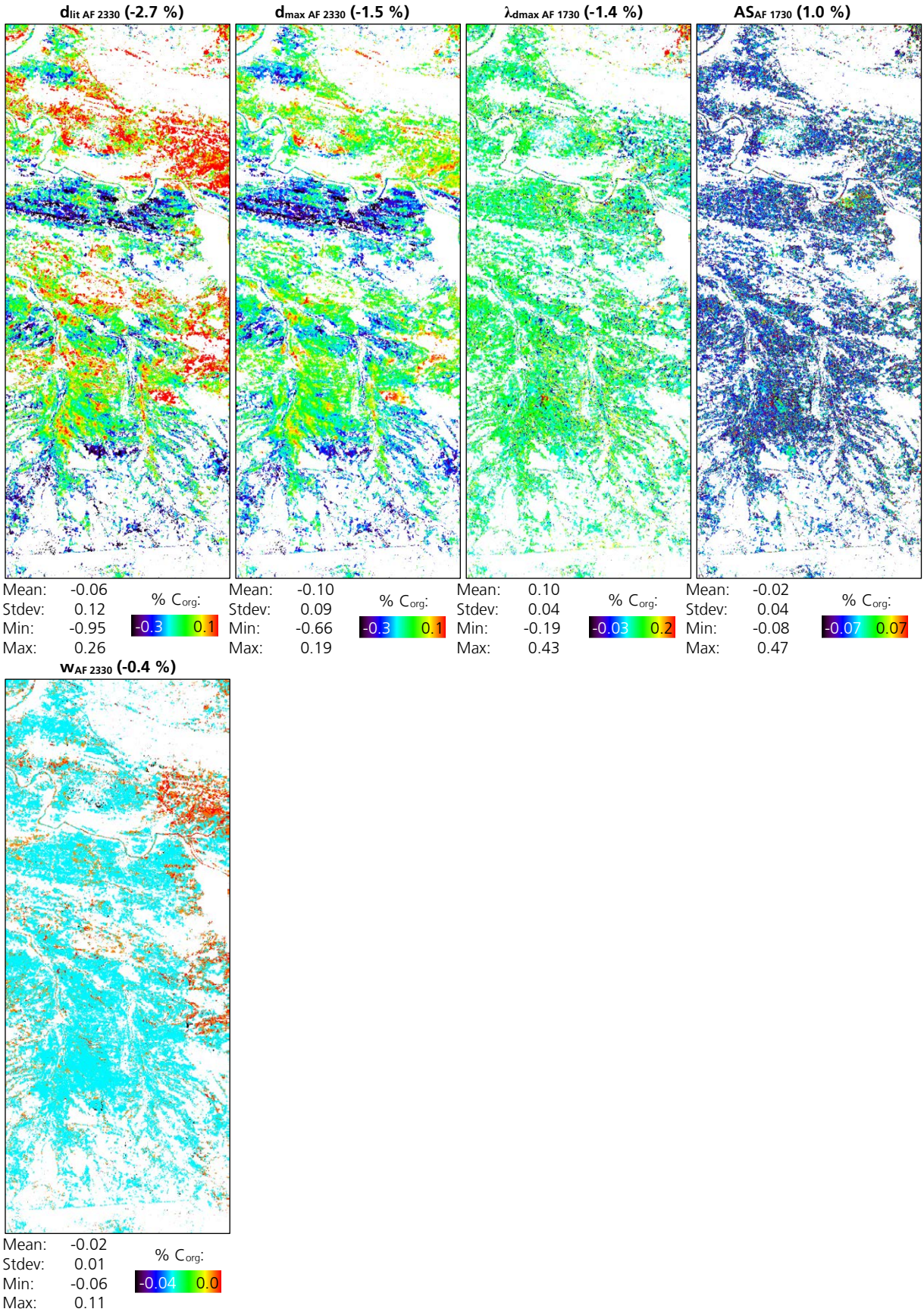


Figure E-19: Maps of spectral variables after standardization and multiplication with the regression coefficient of the S2 Gonakraal subset, Part II.
Symbols: $d_{max\ AF}$: maximum depth, $\lambda_{dmax\ AF}$: wavelength of $d_{max\ AF}$, $d_{lit\ AF}$: depth at wavelength position given in literature, w_{AF}: feature width, AS_{AF}: asymmetry factor of absorption features (AF).

E.5 Classification of prediction confidence

Table E-22 shows the assessed confidence in % of pixels for the subsets within each scale. Within the entire study area of 320 km² 78 % of the pixels show a good and only 2 % a reduced accuracy.

Table E-22: Statistics of confidence assessment for the subsets of each scale and the entire dataset.

		Confidence assessment [% of pixels]		
	Subsets	**Good**	**Medium**	**Reduced**
Scale 1	**Gonakraal -A plain**	98	1	1
	Gonakraal -B river	97	2	1
	Arundale -A slopes	96	3	1
	Arundale - B street	98	2	0
	Wintersfontein - A river	98	2	0
	Kruizementfontein - A hills	96	3	1
Scale 2	**Gonakraal**	76	22	2
	Arundale	94	5	1
	Wintersfontein	97	2	1
	Kruizementfontein	98	1	1
	Stenbokvlakte	82	17	1
Scale 3	**Northern section**	80	18	2
	Southern section	63	28	9
	Entire dataset	78	20	2

Figure E-20 shows an example of the individual layers within the confidence assessment contributing to final classification layer flagging good, medium and reduced confidence of predictions.

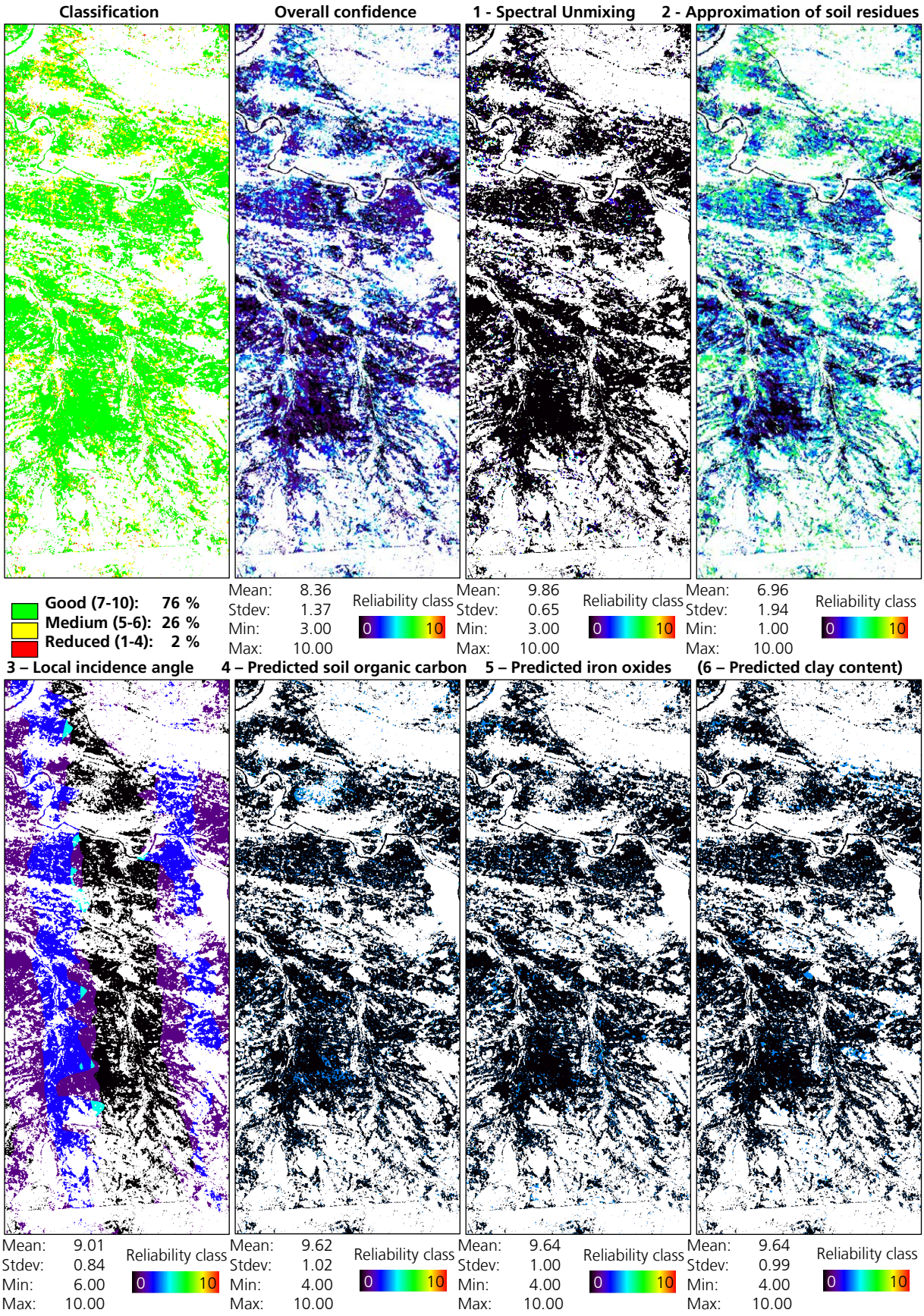


Figure E-20: Example of individual and classified confidence layers for the Gonakraal subset (scale 2).

F Details on performance evaluation

F.1 Validation with ground reference measurements

Figure F-21 shows scatter plots of all sample sites, soil constituents were predicted for based on the HyMap imagery and their correlation with the ground reference measurements. Thus, besides the valid sample sites (mostly good confidence assessment of pixel and surrounding, see section 5.3.2) here also the sites classified as influenced are included. They follow the same trends as when only including the valid sample sites (figure 6-22), though point clouds are denser and correlation coefficients a bit lower. Table F-23 lists the number of reference samples within each subset and their classification in the confidence assessment.

To test for an existing influence of vegetation on the accuracy of soil constituents' predictions, the deviations of measured and predicted contents of all three constituents are plot against the determined vegetation cover (Figure F-22). Therefore the fractional cover of green and dry and the sum of both are used. The existence of such a correlation could not be determined for any of the three constituents.

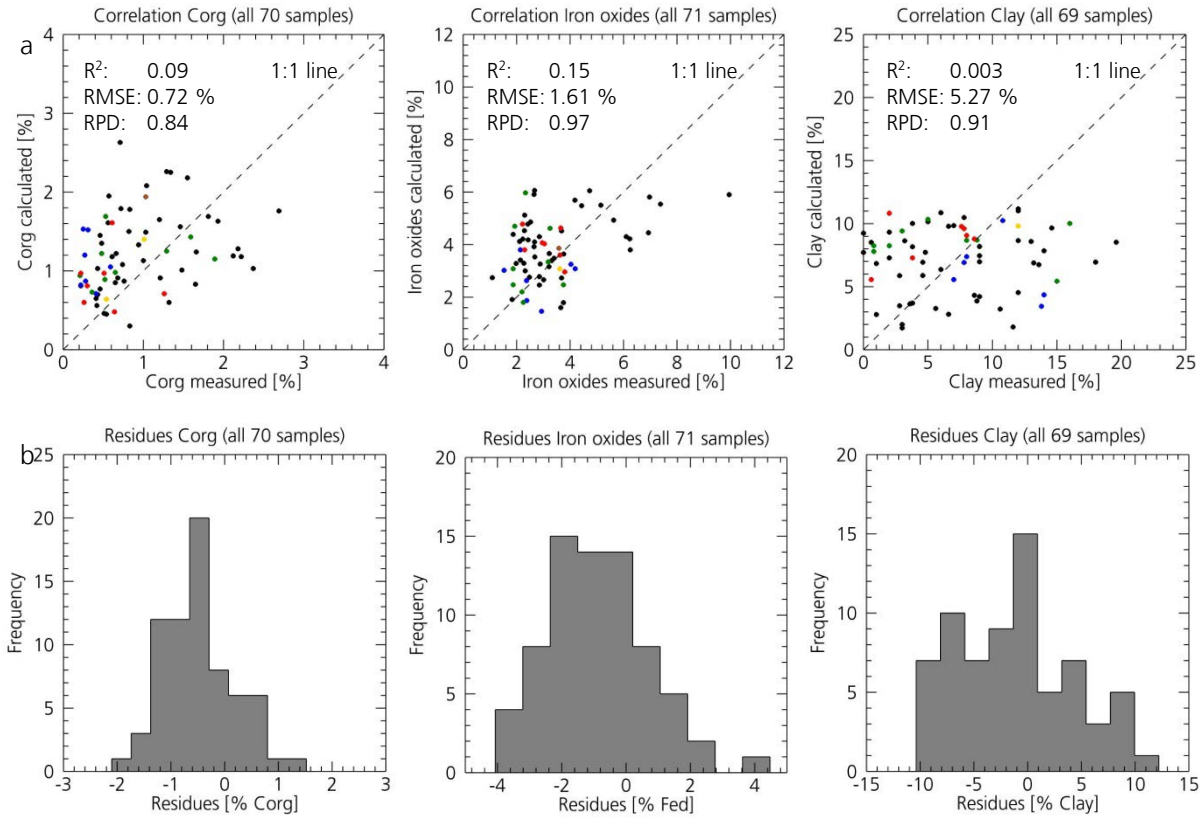


Figure F-21: Scatter plots for in-situ measured versus from spectral analysis predicted soil constituents (a) and plots of modeling residues (b).
Color coding in a: • Validation sites outside the scale 2 subsets, sites within scale 2 • Gonakraal, • Arundale, • Wintersfontein, • Kruizementfontein and • Stenbokvlakte area.

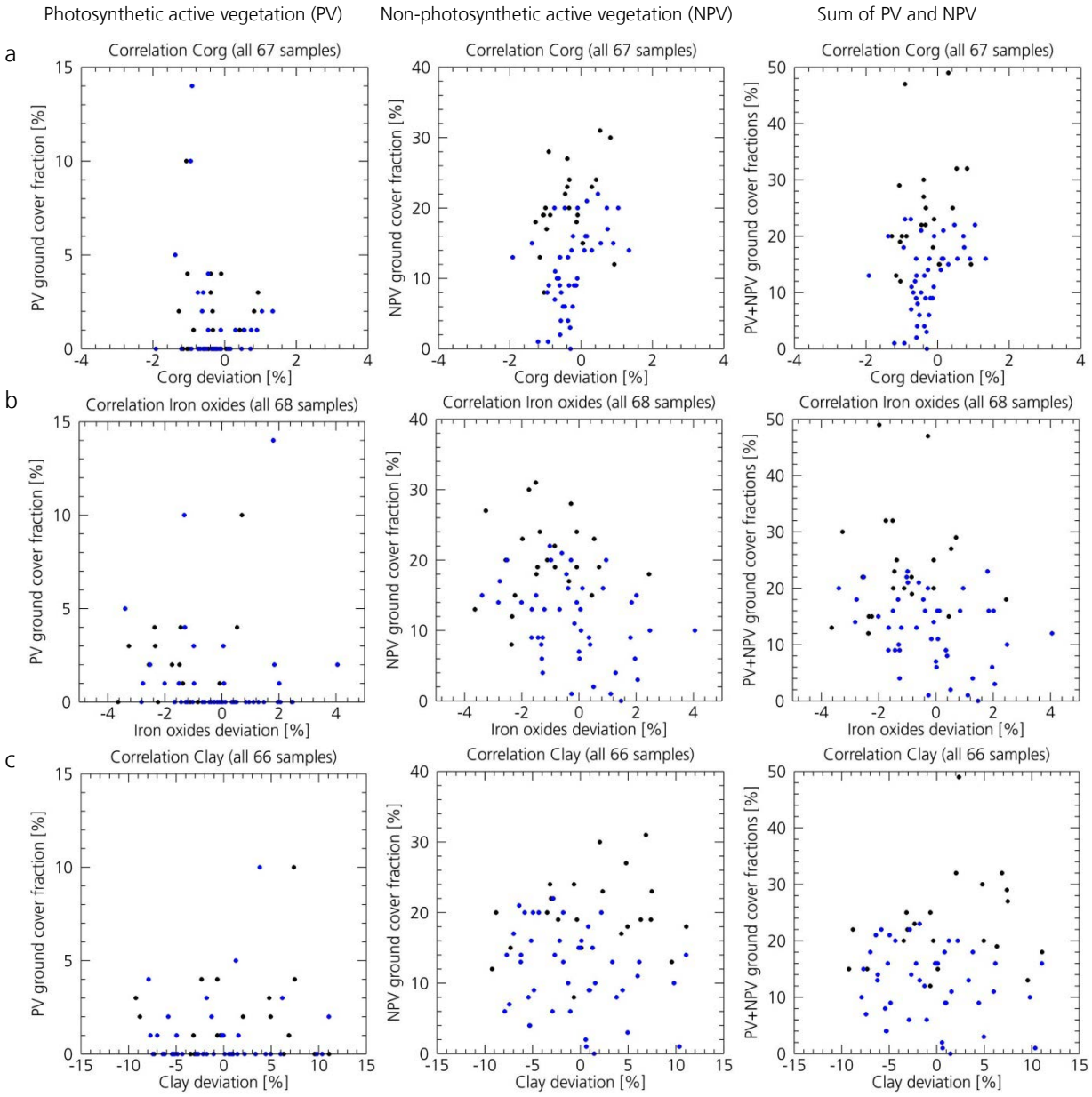


Figure F-22: Scatter plots showing the correlation between the deviation of measured and calculated soil organic carbon (a), iron oxides (b) and clay contents and the ground cover fractions of both vegetation types determined for the sample locations in spectral mixture analysis.
Color coding: • valid sample sites, • other sites for which soil constituents were predicted.

Table F-23: Number of validation samples in subsets.

		No. samples in subset	No. of samples soil constituents were predicted for (valid + influenced samples)	No. of samples of good quality (valid samples)
Scale 1	**Gonakraal - A plain**	4	3/2[1]	2
	Gonakraal - B river	3	2	1
	Arundale - A slopes	2	2	2
	Arundale - B street	0	0	0
	Wintersfontein - A river	3	3	3
	Kruizementfontein - A hills	1	1	0
Scale 2	**Gonakraal**	12	6	4
	Arundale	13	9	6
	Wintersfontein	8	8	7
	Kruizementfontein	9	1	1
	Stenbokvlakte	7	1	0
Scale 3	**North**	118	57/56[1]	38
	South	45	12/14[1]	9/10[1]
	Entire dataset	163	69-71[1]	47/48[1]

[1]The number of samples differs for each soil constituent due to differences in pixels, that were removed because of unreliable values or being interpolated, either at the determined site location or its surrounding.

F.2 Investigation of scaling issues

This appendix gives details on the investigation of the dependency of soil constituents' predictions on size and characteristics of the imagery. The results are summarized in section 6.4.2 in the main text.

Table F-24: Statistics of spectral unmixing in the different spatial scales for the extension of the scale 2 (data from processing of scale 2 and 3) and scale 1 subsets (data from processing of scale 1, 2 and 3). The values present an average of all subsets with the extension of scale 1 and 2.

	Subsets	Ground cover fractions PV	NPV	BS	Summed cover fractions	RMSE
Medium subsets	Statistics of all **scale 2** subsets	Mean: 0.23 Stdev: 0.23 Min: 0.00 Max: 2.19	Mean: 0.28 Stdev: 0.17 Min: 0.00 Max: 4.15	Mean: 0.43 Stdev: 0.23 Min: 0.00 Max: 2.80	Mean: 0.75 Stdev: 0.18 Min: 0.00 Max: 4.27	Mean: 1.56 Stdev: 0.84 Min: 0.00 Max: 36.91
	Statistics of **scale 3** data with the extension of the scale 2 subsets	Mean: 0.19 Stdev: 0.20 Min: 0.00 Max: 1.46	Mean: 0.23 Stdev: 0.12 Min: 0.00 Max: 2.01	Mean: 0.45 Stdev: 0.23 Min: 0.00 Max: 4.39	Mean: 0.70 Stdev: 0.18 Min: 0.03 Max: 4.96	Mean: 1.94 Stdev: 1.17 Min: 0.00 Max: 29.90
Small subsets	Statistics of all **scale 1** subsets	Mean: 0.16 Stdev: 0.19 Min: 0.00 Max: 1.93	Mean: 0.28 Stdev: 0.16 Min: 0.00 Max: 1.62	Mean: 0.54 Stdev: 0.21 Min: 0.00 Max: 2.75	Mean: 0.85 Stdev: 0.16 Min: 0.07 Max: 3.27	Mean: 1.56 Stdev: 0.82 Min: 0.00 Max: 22.92
	Statistics of **scale 2** data with the extension of the scale 1 subsets	Mean: 0.16 Stdev: 0.17 Min: 0.00 Max: 2.02	Mean: 0.30 Stdev: 0.17 Min: 0.00 Max: 3.73	Mean: 0.45 Stdev: 0.21 Min: 0.00 Max: 2.14	Mean: 0.78 Stdev: 0.16 Min: 0.02 Max: 3.89	Mean: 1.43 Stdev: 0.70 Min: 0.00 Max: 17.59
	Statistics of **scale 3** data with the extension of the scale 1 subsets	Mean: 0.12 Stdev: 0.15 Min: 0.00 Max: 1.03	Mean: 0.22 Stdev: 0.12 Min: 0.00 Max: 1.88	Mean: 0.50 Stdev: 0.22 Min: 0.00 Max: 3.43	Mean: 0.71 Stdev: 0.17 Min: 0.03 Max: 3.88	Mean: 1.93 Stdev: 1.21 Min: 0.00 Max: 24.19

Differences within the processing of the three scales

Reasonable endmember sets could be derived for all data scales, because the extent and location of the subsets of the lower scales were selected to include all land cover types. Only for dry and in some cases also for green vegetation only few EM were found by the SSEE algorithm in the scale 1 subsets. When comparing the results of different scales for the same subsets (e.g. medium-sized subsets processed in scale 2 and 3), the performance of spectral unmixing showed little variance (Table F-24). Nonetheless the sums of the cover fractions are higher for the lower scale subsets and unmixing RMSE are lower for the lower scales (see also Figure F-23). It is noted that RMSE are lowest in the processing of the scale 2 imagery, indicating that this scale is best suited for unmixing, because the extension of the imagery is large enough to allow the identification of EMs for all land cover classes but in addition is small enough to allows the EMs to well represent the local mixtures. Figure F-23 illustrates the differences between unmixing results given in Table F-24.

The thresholds selected within the approximation of soil residues are more depending on the characteristics of the individual subsets (e.g. high amounts of dry vegetation within the Kruizementfontein area) than on the spatial scale (see table 6-5). Within scale 1 and 2 the thresholds could be very well adapted to separate reliable from biased soil signatures. In contrast for the scale 3 data including both transect parts, the setting of the thresholds was often a compromise between removing too many or too few pixels. Due to the large variance in characteristics in these datasets the thresholds were not sufficient to separate reliable from biased approximated soil signatures.

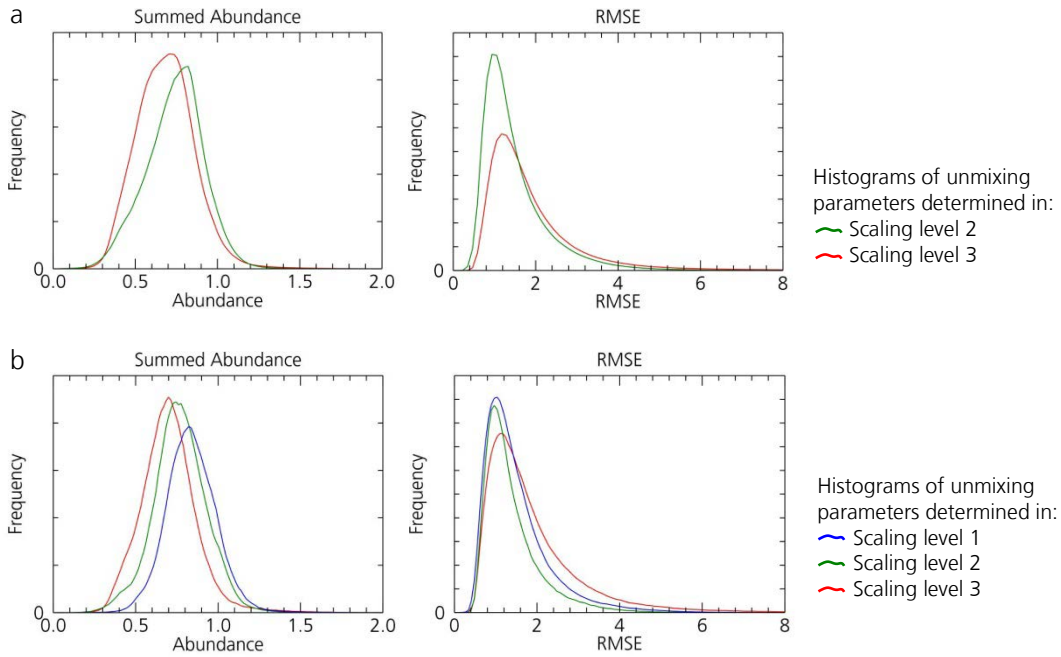


Figure F-23: Histograms of unmixing parameters sum of abundance and unmixing RMSE for medium (a) and small sized subsets (b) determined in the processing of the different scales.

Validation accuracies over the three spatial scales

Figure F-24 and Figure F-25 show scatter plots for ground truth samples in scale 2 subsets predicted in the scales 2 and 3 and for the samples within scale 1 subsets that were predicted in all three scales. It seems that predictions are better for the lower scales (1 and 2) but due to the low number of samples especially in the small scale 1 subsets, correlation coefficients are not significant.

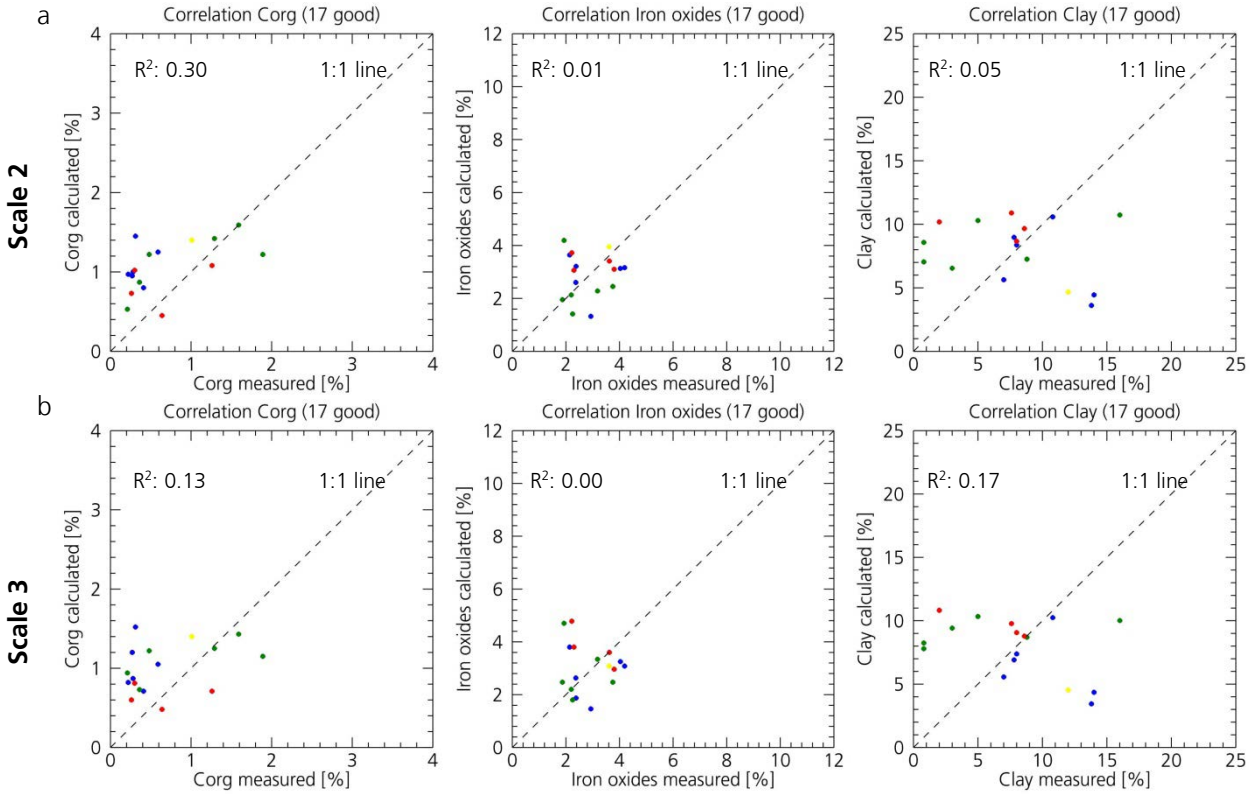


Figure F-24: Scatter plots for in-situ measured vs. from spectral analysis predicted soil constituents for samples that were predicted in scale 2 (a) and scale 3 processing (b) and in both scales labeled with a good confidence.
Color coding: Validation sites within scale 2 • Gonakraal, • Arundale, • Wintersfontein and • Kruizementfontein area.

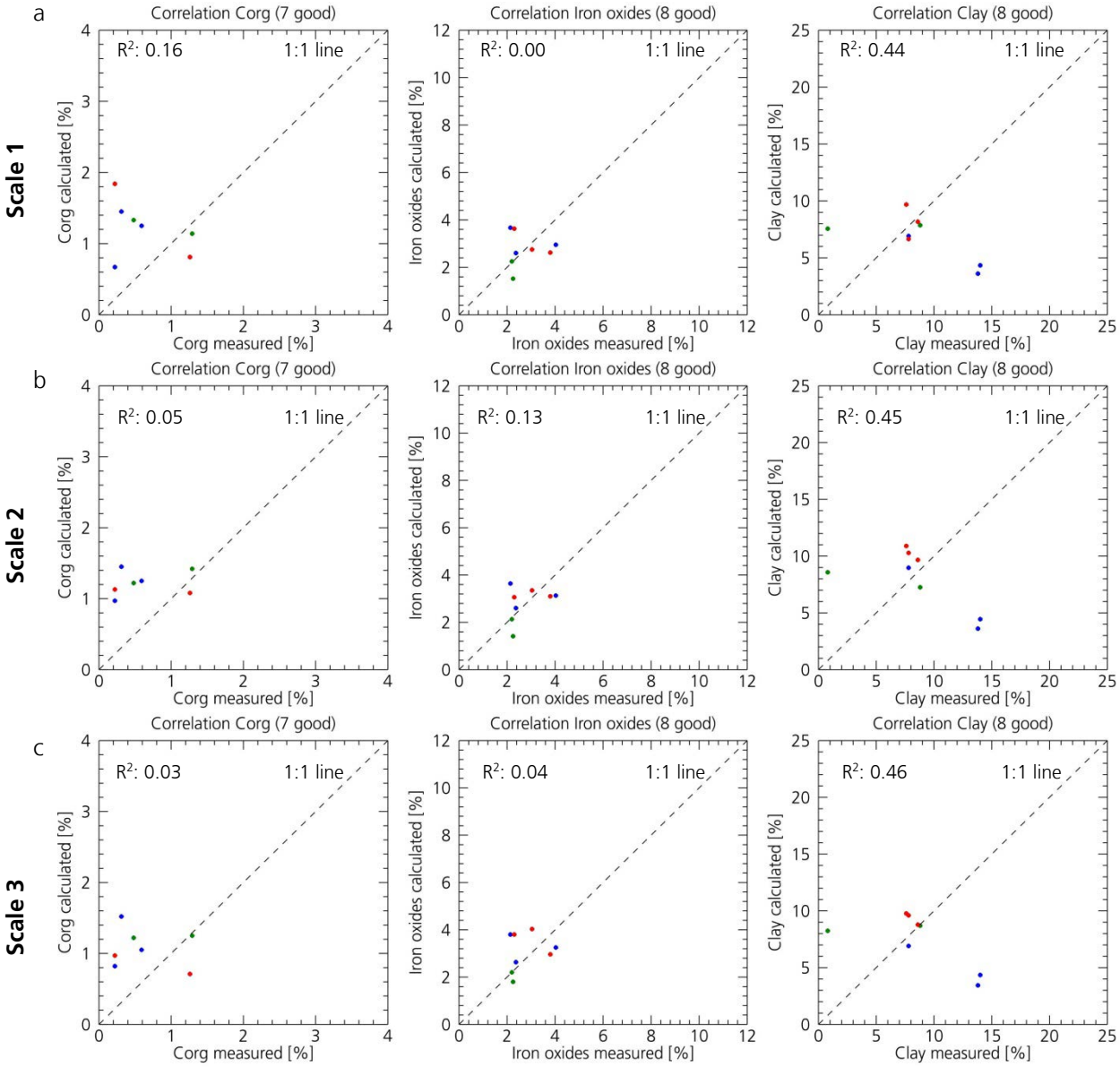


Figure F-25: Scatter plots for in-situ measured vs. from spectral analysis predicted soil constituents for samples that were predicted in scale 1 (a), scale 2 (b) and scale 3 processing (c) and in all scales labeled with a good confidence.
Color coding: Validation sites within scale 2 • Gonakraal, • Arundale, • Wintersfontein and • Kruizementfontein area.

Confidence assessment for the three spatial scales

Table F-25 shows statistics of the confidence assessment of the processing of each scale. To compare the performance for multiple smaller subsets (scale 1 and 2) when processed in the different scales, from the larger scale data, e.g. of the northern transect which was processed in scale 3, the smaller subsets of scale 1 and 2 are clipped.

Table F-25: Statistics of confidence assessment for the subsets of each spatial scale and the entire dataset.

		Confidence assessment [% of pixels]								
		Processing scale 1			Processing scale 2			Processing scale 3		
	Subset extension	Good	Medium	Reduced	Good	Medium	Reduced	Good	Medium	Reduced
Small subsets	Gonakraal -A plain	99	1	0.1	82	16	2.2	85	14	0.7
	Gonakraal -B river	97	3	0.3	75	23	2.1	85	14	0.7
	Arundale -A slopes	96	4	0.2	96	4	0.4	81	17	1.5
	Arundale - B street	98	2	0.2	97	3	0.1	84	15	0.6
	Wintersfontein - A river	99	1	0.1	98	2	0.1	90	10	0.3
	Kruizementfontein - A hills	97	3	0.1	99	1	0.2	80	19	1.3
Medium subsets	Gonakraal				76	22	2.6	83	16	1.1
	Arundale				94	5	0.5	75	23	2.0
	Wintersfontein				98	2	0.1	87	13	0.6
	Kruizementfontein				98	1	0.4	75	23	1.8
	Stenbokvlakte				82	18	0.4	55	33	12.3
Entire data	Northern section							80	19	1.5
	Southern section							63	29	8.5
Averages	All small subsets	97	3	0.2	91	8	0.8	85	14	0.8
	All medium subsets				95	5	0.4	79	19	2.0
	Entire dataset							78	20	2.2

G Prediction of soil constituents for simulated EnMAP data

The following tables and figures give details for the processing of the simulated EnMAP data towards a prediction of soil constituents. Results and statistics are compared to the HyMap data of the corresponding subsets (scale 2 data).

Derivation of image endmembers

Table G-26 gives an overview of the EM sets derived from the simulated EnMAP data. The finally selected EMs are a combination of EnMAP extracted EMs and re-sampled HyMap derived EMs (end of Table G-26 and Figure G-26).

Table G-26: Overview of SSEE derived EM sets for the simulated EnMAP data.

	Subsets	Size in pixels	EMs extracted by SSEE	EMs selected for S2 Unmixing			
				PV	NPV	BS	all
Medium HyMap subsets	**Gonakraal**	546 x 1325	105	7	2	8	17
	Wintersfontein	1662 x 2000	264	7	5	9	21

	Subsets	Size in pixels	EMs extracted by SSEE	EnMAP derived suitable EMs			
				PV	NPV	BS	
EnMAP subsets	**Gonakraal**	60 x 149	29	1	0	4	5
	Wintersfontein	170 x 219	45	3	2	6	11

	Subsets			EMs used for EnMAP unmixing			
				PV[1]	NPV[1]	BS[1]	all[1]
EnMAP subsets	**Gonakraal**			0+7	0+2	2+8	19
	Wintersfontein			2+5	1+5	2+6	21

[1]No. of EMs derived from simulated EnMAP data + No. of resampled HyMap EMs.

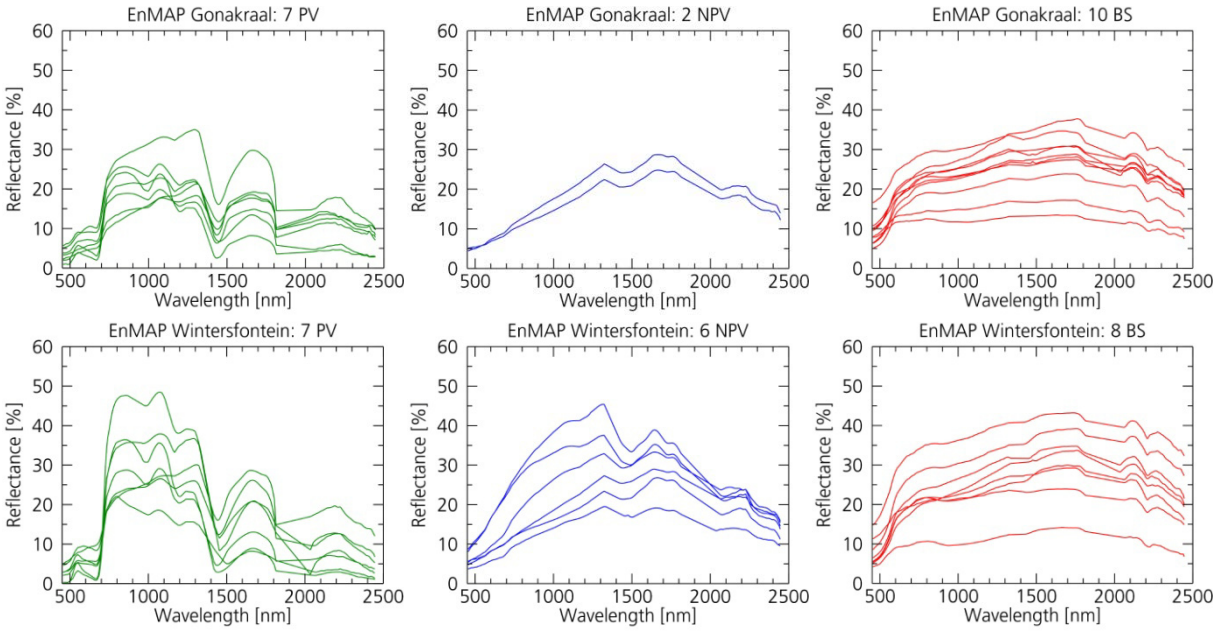


Figure G-26: Endmember sets for simulated EnMAP subsets.

Spectral unmixing and approximation of soil residues

Figure G-27 shows a comparison of HyMap and EnMAP derived ground cover fractions for the two subsets EnMAP data were simulated for. The statistics of the spectral unmixing results are given in Table G-27. Table G-28 shows the thresholds applied in the approximation of soil residues and statistics of the results.

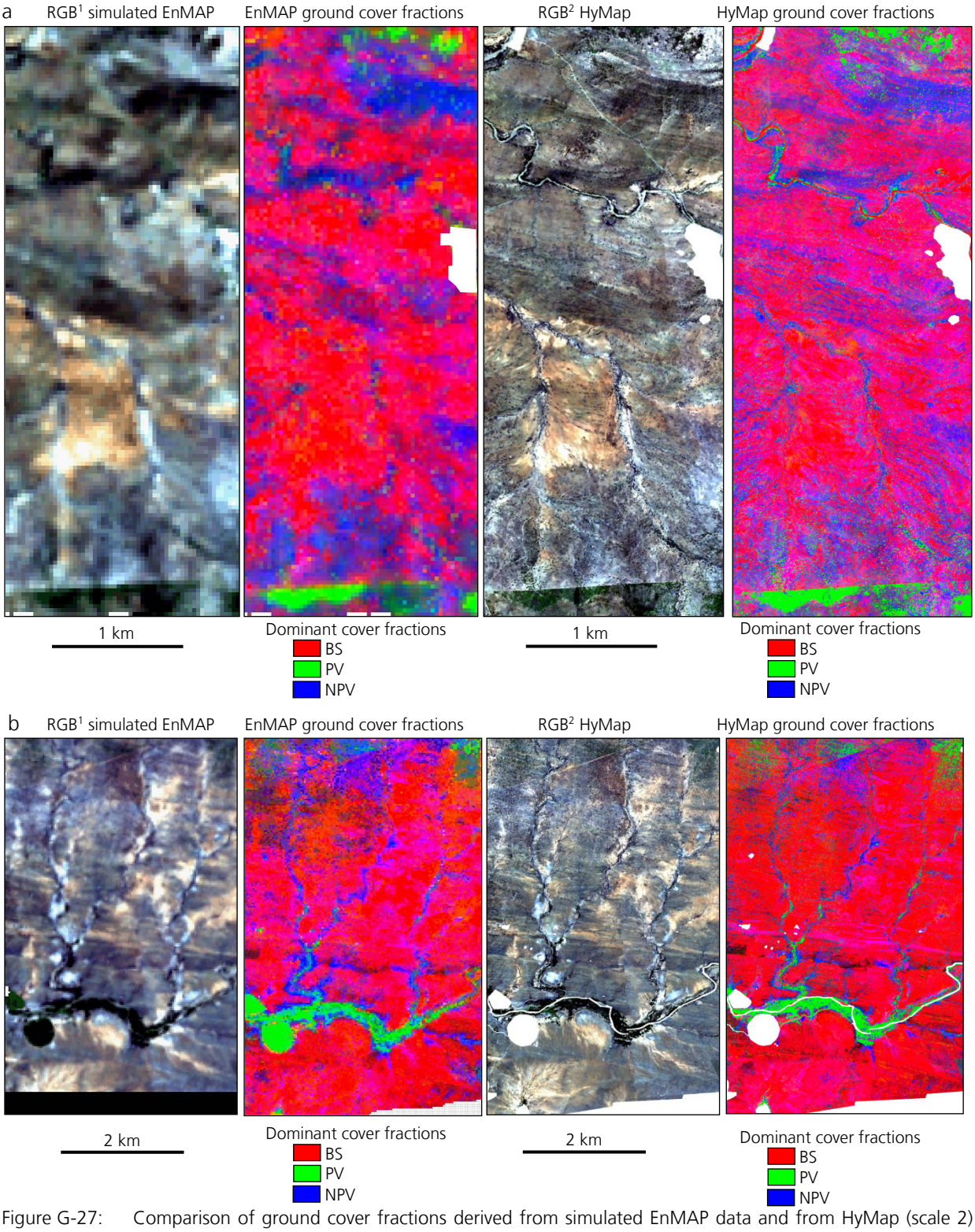


Figure G-27: Comparison of ground cover fractions derived from simulated EnMAP data and from HyMap (scale 2) for the Gonakraal (a) and Wintersfontein areas (b).
True color composites with R: 640 nm, G: 549 nm, B: 456 nm.

Table G-27: Statistics of ISMA spectral unmixing results for the simulated EnMAP data and for comparison of the HyMap data (scale 2).

	Subsets	Ground cover fractions PV	Ground cover fractions NPV	Ground cover fractions BS	Summed cover fractions	RMSE
Medium HyMap subsets	Gonakraal	Mean: 0.13 Stdev: 0.16 Min: 0.00 Max: 1.29	Mean: 0.29 Stdev: 0.11 Min: 0.00 Max: 1.51	Mean: 0.44 Stdev: 0.19 Min: 0.00 Max: 2.16	Mean: 0.79 Stdev: 0.12 Min: 0.16 Max: 2.62	Mean: 1.63 Stdev: 0.84 Min: 0.00 Max: 10.33
	Wintersfontein	Mean: 0.14 Stdev: 0.17 Min: 0.00 Max: 2.19	Mean: 0.18 Stdev: 0.12 Min: 0.00 Max: 1.89	Mean: 0.57 Stdev: 0.21 Min: 0.00 Max: 2.80	Mean: 0.74 Stdev: 0.16 Min: 0.08 Max: 2.82	Mean: 1.49 Stdev: 0.82 Min: 0.00 Max: 23.55
EnMAP subsets	Gonakraal	Mean: 0.11 Stdev: 0.09 Min: 0.00 Max: 0.82	Mean: 0.26 Stdev: 0.10 Min: 0.00 Max: 0.96	Mean: 0.43 Stdev: 0.16 Min: 0.06 Max: 1.31	Mean: 0.76 Stdev: 0.10 Min: 0.49 Max: 1.58	Mean: 1.24 Stdev: 0.47 Min: 0.00 Max: 4.75
	Wintersfontein	Mean: 0.12 Stdev: 0.14 Min: 0.00 Max: 1.38	Mean: 0.24 Stdev: 0.14 Min: 0.00 Max: 1.07	Mean: 0.49 Stdev: 0.19 Min: 0.02 Max: 2.20	Mean: 0.76 Stdev: 0.14 Min: 0.23 Max: 3.00	Mean: 1.11 Stdev: 0.55 Min: 0.00 Max: 16.49

Table G-28: Overview of thresholds and results of the approximation of soil signatures for the two simulated EnMAP datasets compared to the corresponding scale 2 subsets. Background pixels are excluded for the calculation of the percentages.

		Thresholds						% of pixels					
	Subsets	Maximal RMSE	Minimal fraction BS	Sum of cover fractions	Maximal fraction PV	Maximal fraction NPV	Minimal mean reflectance [%]	Valid Pixels	Maximal RMSE	Minimal fraction BS	Sum of cover fractions	Maximal fraction PV and NPV	Minimal mean reflectance
Medium HyMap subsets	Gonakraal	5.81[1]	0.35	0.4-1.4	0.20	0.30	10	40.8	0.2	37.1	0.3	19.3	2.3
	Wintersfontein	5.59[1]	0.35	0.4-1.4	0.30	0.35	9	72.7	0.4	20.5	0.1	2.7	3.8
EnMAP subsets	Gonakraal	3.12[1]	0.35	0.4-1.4	0.20	0.25	10	37.9	0.6	30.7	0.1	28.6	2.2
	Wintersfontein	3.60[1]	0.35	0.4-1.4	0.25	0.30	9	57.8	0.6	24.0	0.0	13.9	3.8

*Factor 5 of RMSE

Re-calibration of regression models

Due to EnMAP's sensor characteristics differing from HyMap's, a re-calibration of the regression models was necessary. In Table G-29 the regression models developed based on the bare soil field spectra in HyMap's and EnMAP's spectral resolution are compared. Differences are small and in the same range as the differences between the models in full ASD spectral resolution and lower HyMap resolution.

Delineation of soil constituents

The following tables give statistics of predicted topsoil constituents based on HyMap and simulated EnMAP data. Figure G-28 shows soil clay contents predicted based on the EnMAP data, which were not included in the main text.

Table G-29: Calibration and validation accuracies for the feature based regression approach established based on bare soil field spectra in the spectral resolution matching the pre-processed simulated EnMAP data.

		No. spectral variables	Calibration			Validation		
			R^2_{Cal}	$RMSE_{Cal}$	RPD_{Cal}	R^2_{Val}	$RMSE_{Val}$	RPD_{Val}
HyMap (115 bands)	**C_{org}**	13 of 16	0.79	0.42	2.17	0.66	0.46	1.77
	Iron oxides	19 of 21	0.61	0.77	1.61	0.30	0.93	1.19
	Clay	13 of 16	0.14	4.47	1.08	0.09	4.92	0.93
EnMAP (198 bands)	**C_{org}**	14 of 16	0.77	0.44	2.09	0.62	0.50	1.56
	Iron oxides	16 of 21	0.59	0.79	1.57	0.29	0.87	1.18
	Clay	14 of 16	0.17	4.39	1.10	0.09	4.90	0.92

Table G-30: Overview of histogram statistics for the prediction of topsoil organic carbon for the simulated EnMAP data in comparison with the corresponding HyMap data (scale 2).

	Subsets	All values (band 1)	Unfiltered (band 2)	Filter kernel 3 x 3 (band 3)	Filter kernel 5 x 5 (band 4)
Gonakraal	**EnMAP**	Mean: 1.52 Stdev: 0.61 Min: -0.02 Max: 7.18	Mean: 1.52 Stdev: 0.60 Min: 0.22 Max: 5.14	Mean: 1.53 Stdev: 0.46 Min: 0.43 Max: 4.58	Mean: 1.54 Stdev: 0.38 Min: 0.50 Max: 4.29
	Medium HyMap subsets	Mean: 1.04 Stdev: 0.47 Min: -0.57 Max: 4.92	Mean: 1.05 Stdev: 0.46 Min: 0.21 Max: 4.92	Mean: 1.05 Stdev: 0.41 Min: 0.21 Max: 4.92	Mean: 1.07 Stdev: 0.38 Min: 0.21 Max: 4.62
Wintersfontein	**EnMAP**	Mean: 1.49 Stdev: 0.65 Min: -0.002 Max: 7.15	Mean: 1.49 Stdev: 0.64 Min: 0.22 Max: 5.73	Mean: 1.50 Stdev: 0.48 Min: 0.23 Max: 5.73	Mean: 1.51 Stdev: 0.41 Min: 0.23 Max: 5.73
	Medium HyMap subsets	Mean: 1.74 Stdev: 0.86 Min: -0.77 Max: 34.58	Mean: 1.74 Stdev: 0.85 Min: 0.21 Max: 5.85	Mean: 1.76 Stdev: 0.76 Min: 0.21 Max: 5.85	Mean: 1.80 Stdev: 0.71 Min: 0.21 Max: 5.76

Table G-31: Overview of histogram statistics for the prediction of iron oxides for the simulated EnMAP data in comparison with the corresponding HyMap data (scale 2).

	Subsets	All values (band 1)	Unfiltered (band 2)	Filter kernel 3 x 3 (band 3)	Filter kernel 5 x 5 (band 4)
Gonakraal	**EnMAP**	Mean: 2.69 Stdev: 1.16 Min: -2.10 Max: 9.24	Mean: 2.81 Stdev: 1.05 Min: 0.90 Max: 9.24	Mean: 2.81 Stdev: 0.82 Min: 0.92 Max: 8.19	Mean: 2.81 Stdev: 0.71 Min: 0.92 Max: 8.19
	Medium HyMap subsets	Mean: 3.94 Stdev: 1.09 Min: -7.21 Max: 14.16	Mean: 3.95 Stdev: 1.07 Min: 0.90 Max: 10.58	Mean: 3.95 Stdev: 0.93 Min: 0.90 Max: 10.47	Mean: 3.96 Stdev: 0.85 Min: 0.90 Max: 10.47
Wintersfontein	**EnMAP**	Mean: 2.48 Stdev: 1.25 Min: -1.78 Max: 15.56	Mean: 2.65 Stdev: 1.12 Min: 0.90 Max: 10.60	Mean: 2.63 Stdev: 0.89 Min: 0.91 Max: 10.38	Mean: 2.64 Stdev: 0.81 Min: 0.90 Max: 10.38
	Medium HyMap subsets	Mean: 3.46 Stdev: 1.62 Min: -42.26 Max: 33.02	Mean: 3.43 Stdev: 1.26 Min: 0.90 Max: 10.62	Mean: 3.44 Stdev: 1.08 Min: 0.90 Max: 10.62	Mean: 3.46 Stdev: 1.00 Min: 0.90 Max: 10.62

Table G-32: Overview of histogram statistics for the prediction of clay content for the simulated EnMAP data in comparison with the corresponding HyMap data (scale 2).

	Subsets	All values (band 1)	Unfiltered (band 2)	Filter kernel 3 x 3 (band 3)	Filter kernel 5 x 5 (band 4)
Gonakraal	**EnMAP**	Mean: 7.61 Stdev: 2.01 Min: -0.77 Max: 14.60	Mean: 7.62 Stdev: 1.99 Min: 0.47 Max: 14.60	Mean: 7.60 Stdev: 1.31 Min: 0.47 Max: 14.55	Mean: 7.58 Stdev: 1.00 Min: 1.47 Max: 13.49
	Medium HyMap subsets	Mean: 8.79 Stdev: 1.86 Min: -1.38 Max: 18.57	Mean: 8.79 Stdev: 1.86 Min: 0.01 Max: 18.57	Mean: 8.78 Stdev: 1.64 Min: 0.01 Max: 16.74	Mean: 8.73 Stdev: 1.51 Min: 0.02 Max: 16.74
Wintersfontein	**EnMAP**	Mean: 5.98 Stdev: 2.50 Min: -7.00 Max: 15.59	Mean: 6.10 Stdev: 2.33 Min: 0.00 Max: 15.59	Mean: 6.09 Stdev: 1.52 Min: 0.02 Max: 15.59	Mean: 6.11 Stdev: 1.23 Min: 0.03 Max: 13.23
	Medium HyMap subsets	Mean: 6.38 Stdev: 2.46 Min: -6.32 Max: 19.81	Mean: 6.46 Stdev: 2.35 Min: 0.00 Max: 19.81	Mean: 6.41 Stdev: 2.05 Min: 0.00 Max: 18.99	Mean: 6.34 Stdev: 1.91 Min: 0.00 Max: 16.74

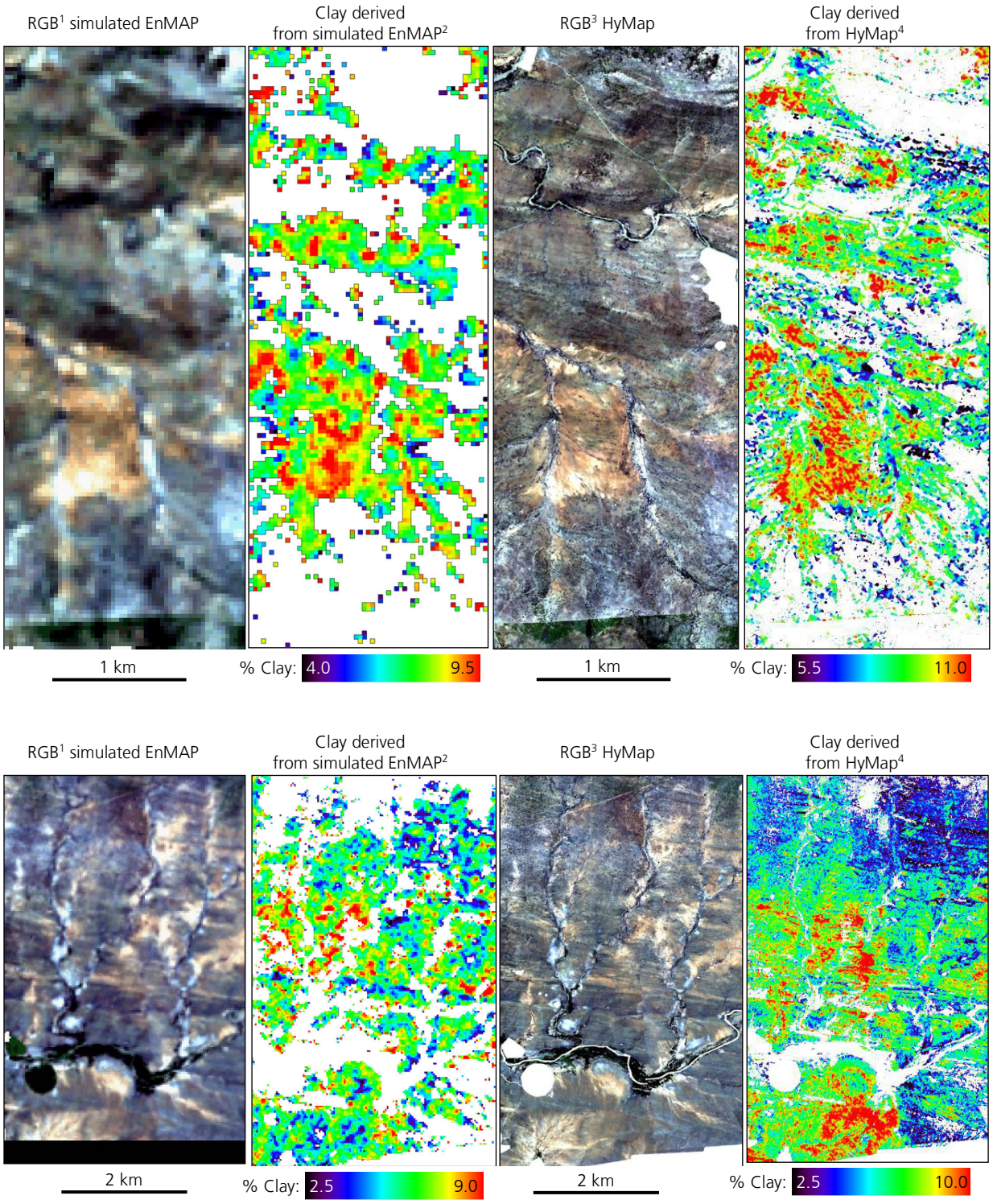


Figure G-28: Maps of topsoil clay contents derived from simulated EnMAP data and HyMap data (scale 2) in comparison for the Gonakraal area (a), and the main part of the Wintersfontein area (b).
[1]True color EnMAP images with R: 640 nm, G: 549 nm, B: 460 nm, [2]Predicted contents after interpolation and filtering with kernel size 3 x 3, [3]True color HyMap images with R: 640 nm, G: 549 nm, B: 456 nm, [4]Predicted contents after interpolation and filtering with kernel size 5 x 5.

Erklärung

Die vorliegende Arbeit ist bisher an keiner anderen Hochschule eingereicht worden und ist selbständig und nur mit den angegebenen Hilfsmitteln erstellt worden. Wörtliche oder inhaltliche Zitate sind kenntlich gemacht.

München, den 24. Oktober 2012

Acknowledgements

This work has been funded within the PhD Program of the Helmholtz EOS Network and conducted at the German Remote Sensing Data Center (DFD) within the German Aerospace Center (DLR). Several people deserve recognition for their support for the completion of this work.

- Prof. Dr. Herrmann Kaufmann (University of Potsdam and Deutsches GeoForschungsZentrum Potsdam, GFZ) for taking up the task of the first reviewer. He always encouraged me in my work and its value, although it is all about only obtaining two letters in front of a name.

- Prof. Dr. Stefan Dech (University of Würzburg and DLR) for being my second supervisor and his continuous efforts to keep up and enlarge the surpassing research environment within DFD, providing great opportunities for PhD projects like this.

- Andreas Müller and Dr. Martin Bachmann (DLR) who introduced me to the field of Imaging Spectroscopy, gave me the opportunity to carry out this research and finish it in the given timeframe. Despite your busy schedule as head of the department/team you always tried to make some time available for me, my ideas and problems and for the discussions of concepts, methods and results. Your guidance, scientific advice and continuous support enabled me to complete this thesis.

- A warm thank you goes to the colleagues at Imaging Spectroscopy work group at DLR for all their advice and support in topics that are not less crucial for the success of a study but can only be touched in a thesis like this. I'd like to express my thanks especially to:
  Dr. Christian Fischer, the continuously busy "Projektameise", for his preparedness to solve every arising problem, taking the time to discuss manuscripts in evening hours and his support since my diploma thesis,
  Dr. Derek Rogge for his valuable contribution in form of many pieces of program code, experience and advice regarding pre-processing and analysis of hyperspectral imagery,
  Dr. Rolf Richter and Stefanie Holzwarth who took me with my data to the depth of atmospheric correction and kindly also back out of it,
  Derek, Martin, Christian B., Christian F., Sabine, Mike and Frank for the time-consuming cross-reading that lead to a significant improvement of the manuscript and my English skills,
  The PhD and post-doc colleagues at DFD-LAX Annekatrin Metz, Christina Eisfelder, Dr. Xingjuan Wang, Christoph Ehrler, Grégoire Kerr, Dr. Markus Tum, Markus Niklaus, Dr. Uta Heiden, Martin Habermeyer and Dr. Nicole Pinnel for the valuable discussions in lunch and coffee breaks.
  I further want to thank the helping hands that supported the field data acquisition, the one and only "Messknecht" Basti Rößler and Sebastian Weide, Elisabeth Kompter and Falco Stier.

- Prof. Dr. Hartmut Asche (University of Potsdam) for offering to be the third reviewer, being open for any new and interesting.

- The colleagues at GFZ Potsdam, first of all Dr. Sabine Chabrillat, who provided me best advice on several occasions during the study.

- Several people from the PRESENCE network and universities within South Africa, who work with amazing enthusiasm on environmental and social problems. In particular:
  Silvia Weel and Dieter van den Broeck (PRESENCE network and learning village, Patensie) who put enormous efforts in the coordination of student work on rehabilitation of the Albany Thicket Biome and helped me with building contacts to the right persons, provided base geodata and a lot of organization before the field phase,
  Mike Powell (Ecological Restoration Capital and Rhodes University, Grahamstown). Thicket can be happy having such an incredibly dedicated person like you. I am sorry for your car in case its life ended shortly after our off-road tour. Thank you for the valuable feedback, literature and your unlimited support, although I'm only a geologist having no time for plants when in the field.
  Dr. Freddie Ellis from Stellenbosch University for lending us technical equipment for soil profile measurements in 2011.

- Frank Westad (CAMO Software, Norway), the "old Viking" that is perfectly adapted to Bavarian habits, for his advice in terms of statistical analysis techniques and cross-reading.

- Pieter Kruger, a farmer in the Baviaanskloof Valley, South Africa. You encouraged me in believing in the value of studies like this, although far from operational status are worth for everyday's life in a region like the Thicket. The time will come when more farmers become aware of what it's worth to care about.

- Finally, I would like to express my thanks to my parents and Chriss, simply for being there, their encouragements and for their long-term support in the background forming my way not only to this PhD degree.

Curriculum Vitae

This page contains personal data and therefore is not part of the online publication.

Scientific publications

BAYER, A., BACHMANN, M., MÜLLER, A. & KAUFMANN, H.: A comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African ecosystem. *Applied and Environmental Soil Science*, 2012, Vol. 2012, Article ID 971252, 20 p.

BAYER, A., BACHMANN, M. & MÜLLER, A.: Delineation of Soil Parameters to assess Ecosystem Degradation using Spectral Mixture Analysis. In: Proceedings of 34th International Symposium on Remote Sensing of Environment, 10-15 April 2011, Sydney, Australia.

FISCHER, C., EHRLER, C. & BAYER, A.: Modelling Surface Energy Release to Validate ASTER night time imagery in Latest Developments in Coal Fire Research. In: Proceedings of ICCFR2 Second International Conference on Coal Fire Research, 19-21 May 2010, Berlin, Germany.

BAYER, A., BACHMANN, M. & HABERMEYER, M.: Quantitative derivation of soil parameters from hyperspectral remote sensing data. Quantitative applications on Soil Spectroscopy Workshop, 15-16 April, 2010, Potsdam, Germany.

BAYER, A., MÜLLER, A., BACHMANN, A., RÖßLER, S. & WEIDE, S.: Quantitative derivation of key soil parameters on the basis of hyperspectral remote sensing data. In: Proceedings of Hyperspectral Workshop 2010. ESTEC. Hyperspectral Workshop 2010, 17-19 March 2010, Frascati, Italy.

BAYER, A., HELDENS, W. & FISCHER, C.: Surface Energy Balance Analysis to improve Land Surface Temperature Information retrieved by TIR satellite data. Thermal Remote Sensing meeting within the 29th EARSeL Imaging Europe Conference, 15-18 June 2009, Chania, Crete.

BAYER, A.: Modellierung der Temperaturentwicklung oberflächennaher Boden- und Gesteinsschichten arider Gebiete unter Berücksichtigung der solaren Einstrahlung und der Topographie. Diploma thesis, University of Stuttgart, Germany, 2008.