Using Machine Learning to derive Soil Information from Soil Reflectance Composites (SCMaP)

Wissenschaftliche Arbeit zur Erlangung des akademischen Grades

Bachelor of Science

im Studiengang

Geographie

am Institut für Geographie der Universität zu Köln

Autor:

Simon Donike

Matrikelnummer 5976642

22. November 2019

Erstgutachter:	Zweitgutachterin:
Prof. Dr. Georg Bareth	Dr. Uta Heiden
Universität zu Köln	Deutsches Zentrum für Luft- und Raumfahrt e. V.

Table of Contents

i. Acknowledgement

Firstly, I would like to thank Prof. Dr. Georg Bareth. He allowed me to explore the many possibilities within the German Aerospace Center (DLR), enabling me to find my own scientific interest.

Most importantly, I express my gratitude to the German Aerospace Center in Oberpfaffenhofen, Germany. I am thankful for the trust extended to me and my capabilities, especially by Dr. Uta Heiden and Prof. Dr. Claudia Künzer for giving me the opportunity to spend the time of this thesis at the DLR.
Being surrounded by a group of extremely knowledgeable and motivated experts who never ignored any questions that arose deeply furthered my understanding of the scientific framework and the methods used in this thesis. I express my gratitude to Dr. Marianne Jilge, Dr. Stefanie Holzwarth, Dr. Nicole Pinnel, Chaonan Xi, Martin Habermayer, Martin Bachmann and especially Simone Zepp, who was unfortunate enough to share an office with me, having to bear the brunt of my (sometimes trivial) questions.
Also, the Bavarian State Offices for Agriculture and Environment made this thesis possible by kindly providing non-public soil samples from their databases.

The Institute for Geography at the University of Cologne, Germany as well as the University of Turku, Finland enabled me to develop the knowledge and skills necessary to conduct this project. I am very thankful for the opportunities provided by these institutions, which allowed me to explore the field of Geography and find my personal interests and strengths within.

Lastly, my deepest gratitude and appreciation goes to my parents, who gave both me and my sister the proverbial “roots to grow and wings to fly”, as well as supported me in every step of my life.

ii. List of Figures

iii. List of Tables

iv. List of Equations

v. List of Abbreviations

ATCOR – Atmospheric and Topographic Correction

BGR – Bundesanstalt für Geowissenschaften und Rohstoffe, Federal Office for Geosciences and Raw Materials

BKKA – Bodenkundliche Kartieranleitung, German Soil Systematic

CLC – CORINE Land Cover

CORINE – Coordination of Information on the Environment

C_{org} – Organic Carbon

DLR – Deutsches Zentrum für Luft- und Raumfahrt, German Aerospace Center

ETM+ - Enhanced Thematic Mapper Plus (Landsat Sensor)

LDA - Linear Discriminant Analysis

LfL – Bayrisches Landesamt für Landwirtschaft, Bavarian State Office for Agriculture

LFU – Bayrisches Landesamt für Umwelt, Bavarian State Office for the Environment

NDVI – Normalized Difference Vegetation Index

PC1/PC2/PC3 – Principal Component 1/2/3

PCA – Principal Component Analysis

PV – Photosynthetically active vegetation index

SCMaP – Soil Composite Mapping Processor

SRC –Soil Reflectance Composite

SSL – Soil Sampling Location

RDF – Random Decision Forest

TM – Thematic Mapper (Landsat Sensor)

1. Introduction

1.1 Significance of Soil Earth Observation

The soils on earth are an existential and fundamental resource for humanity (DOMINATI et al. 2010: 1858–1868), as well as for the planet's ecosystem (YOUNG et al. 2004: 113–132). The 2005 Millennium Ecosystem Assessment identified four major services granted by soils (Millennium Ecosystem Assessment: 2005):

- providing direct and indirect food, as well as water, wood, fiber and fuel,
- regulating services regarding water, climate, floods and erosion,
- cultural services in the domains of recreation, spirituality and aesthetics.
- supporting services as to the nutrition cycle, providing a habitat and supporting biodiversity.

Thus, soils play an important regulating and limiting role for the atmosphere, lithosphere, hydrosphere and biosphere (SZABOLCS: 1994: 33–39). Because of its many variations, soils are "one of the most complex biomaterials on the planet" (ADHIKARI et al. 2016: 101–111), also playing a role in climate change (OMUTO et al. 2013: 81). Since soils are important for many economic and ecological factors, information about them needs to be embedded into political decision making and supported by accurate and cost-effective scientific data (DAILY et al. 1997: 113–132). In the past, information about the soils and their dynamics have been mostly gained through in-situ sampling, which is rather expensive and time-consuming (HANKS et al. 1962: 530; RUBIN et al. 1963: 247–521). Since the opening of the Landsat Archives (WOODCOCK et al. 2008: 1011), large scale images over long time periods are available and used for land monitoring (HANSEN et al. 2012: 66–74). A great effort has been made using multispectral satellite and aircraft imagery to gain soil information (MULDER et al. 2011: 1–19), in order to expand existing soil databases (BEN-DOR et al. 2008: 321–392). In the age of precision farming, where soil information is essential, satellite-based information can fill the gap between small scale measurements (CANDIAGO et al. 2015: 4026–4047) and very coarse information as the Harmonized World Soil Database (NACHTERGAELE et al. 2009: 33-37). Not only providing this information on a large spatial scale, satellite-based information

can also provide continuous monitoring on a much higher temporal scale by providing images and their analysis in short intervals.

1.2 Visibility of Soil Parameters in Earth Observation

The research conducted thus far mostly focuses on aircraft multi- and hyperspectral imagery, since these data acquisition methods provide a higher spectral resolution and fewer disturbances (OCHSNER et al. 2013: 1888–1919; STEVENS et al. 2008: 395–404). In the past, extensive studies have been conducted regarding the detection and quantification of the organic carbon content of the soil, as well as for soil texture and soil type detection (BARNES et al. 2000: 731–741; JARMER et al. 2003: 115–123). The visibility of soil parameters in multispectral and hyperspectral imagery was investigated by Bayer et al in 2016 for the C_{org} content (BAYER et al. 2016: 3997–4010), Lakshmi et al in 2015 for soil texture (LAKSHMI et al. 2015: 1452–1460) and Nanni et al in 2012 for a soil type discrimination (NANNI et al. 2012: 103–112) and found to be generally possible.

This thesis also attempts to identify distinguishing reflectance characteristics but based on the Soil Composite Mapping Processor (SCMaP). Since these distinctions have been successfully identified previously, this thesis tries to recreate these results based on the SCMaP output.

1.3 Soil Composite Mapping Processor (SCMaP)

1.3.1 SCMaP

The basis for this thesis is the product derived from the Soil Composite Mapping Processor (SCMaP) as developed by Derek Rogge et al. (2018) at the German Aerospace Center (DLR). This processor utilizes Landsat imagery from 1984 until 2014 to create an exposed soil map on a regional scale, using the temporal dimension of the data to overcome soil exposure inconsistencies and atmospheric influences by looking at vegetation cycles (ROGGE et al. 2018: 1–17).

Landsat-data from missions 4 (Sensor: TM), 5 (Sensor: TM) and 7 (Sensor: ETM+) was used due to their free availability and most importantly their long-term continuous data acquisition (WOODCOCK et al. 2008: 1011). This processor is then

applied to create the aforementioned soil exposure map, which is used as the data basis for this thesis.

SCMaP was initially developed for Germany based on 36 path/row combinations with a scene size of 170 km north-south and 183 km east-west (paths 192-197, rows 22-27). The TM sensor onboard of Landsat 4 and 5 collects spectral information on six bands ranging from 0.45 to 2.35 µm with a spectral resolution of 30 m. On board of the Landsat 7 satellite, the ETM+ sensor was placed, which has a similar setup as the TM with the addition of a panchromatic band at 15m spatial resolution. After the scan line corrector (SLC) failure of Landsat 7 in May 2003, 22% of any given scene is missing. Different areas are affected by the failure for every flyover, thus image stacking can be used to provide continuous coverage.

The data type of the downloaded images is the compressed GeoTIFF format in an Universal Transverse Mercator Projection with the WGS 84 datum (ROGGE et al. 2018: 1–17). The dataset is made of scenes classed as "Tier 1 precision and terrain corrected" (L1T) and hence is deemed appropriate and accurate enough for time-series analysis by the USGS.

In order to achieve a high quality of detection of exposed soils, artifacts and visual obstructions need to be removed. According to Zhang et al. (ZHANG et al. 2005: 357–371), the global annual mean cloud coverage in Landsat satellite image datasets is approximately 66%. Identifying clouds, as well as their shadows, proved to be crucial, since their darkening and brightening effects complicate many remote sensing applications. The "Fmask" algorithm of Zhu and Woodcock (2012) is used, which has a detection rate of 96.4% (ZHU et al. 2005: 357–371).

Secondly, the ATCOR software introduced in 2008 by R. Richter et al. is used to minimize the effects of atmospheric disturbances such as water vapor and aerosols. Starting from the "digital number" value, the individual pixel metadata consisting of coordinates, illumination angle and the date of acquisition are taken into account to derive the bottom of atmosphere reflectances via the top of atmosphere radiances (RICHTER et al. 2006: 2077–2085).

Due to snow- and cloud cover during the winter, images taken within the first and last 30 days of the year are eliminated. After the application of the processor for Germany as done by the DLR, 9,331 Landsat scenes remain.

After the preprocessing, the scenes are transformed into 1° by 1° tiles and broken up into six time frames of five years. The only exception is the first timeframe, which contains the data of 6 years (1. 1984-89, 2. 1990-94, 3. 1995-99, 4. 2000-04, 5. 2005-09, 6. 2010-14). Additionally, all data from 1984 until 2014 are combined into a seventh timeframe by the DLR, which is the final SCMaP product showing the exposed soils used in this thesis (ROGGE et al. 2018: 1–17). This product will be refered to as SRC.

Fundamentally, SCMaP tries to identify exposed soil pixels (without vegetation) in the images. Other than in very rare conditions in the Bavarian Alps above the tree line or along the shore of oceans and rivers, naturally exposed soil occurs very rarely. A key characteristic of exposed soil is a very low vegetation index, such as the Normalized difference vegetation index (NDVI), which has been used since the early days of Earth Observation (ROUSE et al. 1974: 309–317). This vegetation index uses the high reflectance rate of healthy green plants in the near infrared range from 0.7 to 1.1 µm to create a photosynthetically active vegetation index (PV) (KRIEGLER et al. 1969).

Rogge et al. decided to use a modified version of the NDVI for the highest/lowest PV selection. The blue channel was introduced to the Index, in order to reduce the higher reflectance effect of thin haze in pixels, which were not correctly revised during the preprocessing stage.

$$PV = \frac{NIR - RED}{NIR + RED} + \frac{NIR - BLUE}{NIR + BLUE}$$

Equation 1: PV determination, modified NDVI for SCMaP

A low value of the aforementioned vegetation index value by itself is not sufficient to detect exposed soils, since water and other anthropogenically influenced regions as well as non-photosynthetically active vegetation have a low PV value as well. Since exposed soils most commonly occur on agriculturally used land (in Germany), their distinguishing characteristic is a high seasonal variability of the PV. Due to growing and harvesting/plowing seasons, agriculturally active areas experience at least one, if not several distinct changes in PV every given year (ROGGE et al. 2018: 1–17). Because of the high temporal resolution of the Landsat data, these changes can

be detected and quantified in order to create a composite image. This composite image encompasses both the highest and lowest PV values for each pixel.

The next step is overlaying the composite with the Coordination of Information on the Environment (CORINE) Land Cover data set (CLC). After eliminating those pixels, which underwent a change in the CLC classification, 5000 pixels were randomly selected and then plotted to show their maximum and minimum PV value, as well as their CLC classification group (**Figure 1**) (ROGGE et al. 2018: 1–17).

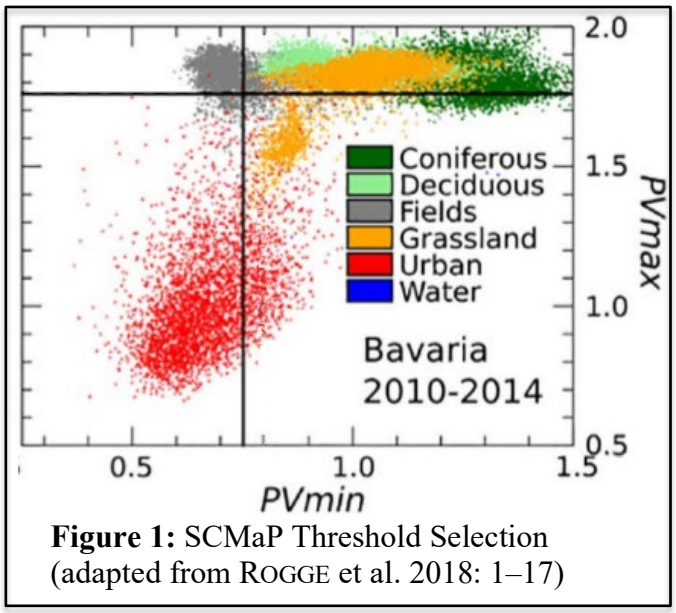

Figure 1: SCMaP Threshold Selection (adapted from ROGGE et al. 2018: 1–17)

Pixels classed as agricultural land can be clearly distinguished from other pixels by their significant difference between maximum and minimum PV, a clear cluster of the "fields" classification is visible. Different classifications such as coniferous and deciduous trees have a similar maximum PV value, but the lower minimum PV value increases the variability, enabling the distinction between the classes.

Finally, thresholds are manually chosen to "box in" this cluster of soil pixels after careful examination of the five test tiles. An automatization of the threshold selection process is currently under development (ROGGE et al. 2018: 1–17).

1.3.2 SCMaP Product –Soil Reflectance Composite

The main result of SCMaP is the SRC, which shows the mean reflectance value of those pixels, which were identified to have undergone significant changes in PV and are thus classed as exposed soil. Those composites are created for each timeframe and show the average reflectance value for each band of the stack of images at any given soil pixel. The error rate for identifying exposed soils of the SRC over the whole 30 year period is stated as 2.24% (Rogge et al. 2018: 1–17).

1.4 Research Objectives

The aim of this thesis is to extract a maximum of soil related information from the Soil Composite Mapping Processor product, using the technique of machine learning. More specifically, the SRC of the years 1984-2014 is used.
SCMaP, as previously described, is a method for identifying bare soil pixels from averaged reflectance data from these points; the reflectance composite thus contains spectral information about the soils. By examining the SRC pixel colours (cf. "2.2 Reflectance Soil Composite"), areas of conformity can be clearly distinguished from other areas, even though all pixels represent bare soil. Based on this observation, the idea to try to extract information about soil parameters was conceived and this hypothesis formed. Variances in reflectance could possibly be traced down to local differences in soil parameters.
Based on in-situ soil samples and their reflectance value of the according SRC pixel, predictions will be made for all other SRC pixels regarding their soil parameters.
Since the pixels have to be assigned to predefined classes, the machine learning method of random decision forests was chosen in order to make these predictions.

The objectives can be more precisely articulated with the following questions:
-Do the reflectance values of SCMaP pixels correlate with the soil parameters of carbon organic content, soil type and soil texture?
-Can a random decision forest perform an accurate classification of the entirety of the SCMaP product?

Figure 2 shows the workflow of this thesis. After acquiring and preprocessing the input data, a database is built, containing the spectral information of the sampling locations as well as their soil parameters. This database is then analysed for statistical correlations between the reflectance and the soil parameters, checking to what degree the classes are spectrally distinguishable. Using the input database, a random decision forest machine learning classification method is built, which predicts the soil parameters for all pixels of the SCMaP composite. The predictions are consequently transformed into three soil parameters maps. The previously obtained statistical results are used to check the accuracy of the predictions.

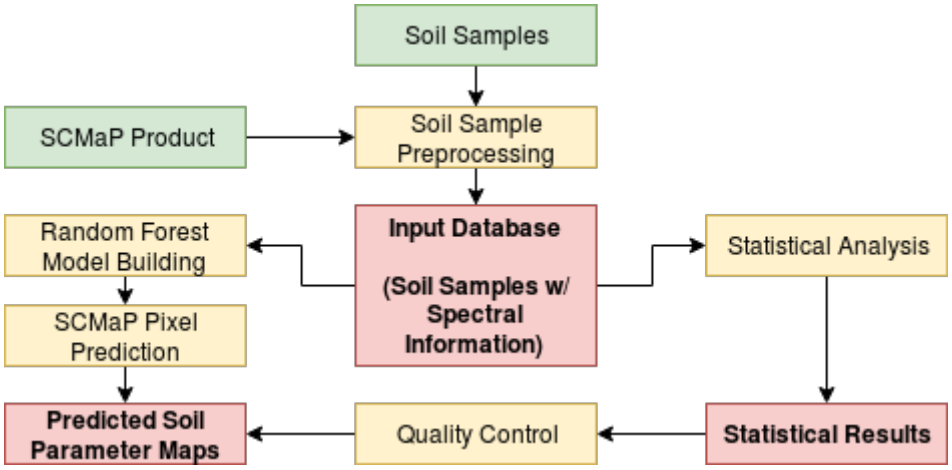

Figure 2: Workflow
Green: inputs, *yellow:* processing steps, *red:* results

2. Data and Study Area

2.1 Study Area

Based on the 1° by 1° tiles of the SCMaP output, twelve of these tiles were chosen as the study area (**Figure 3**).

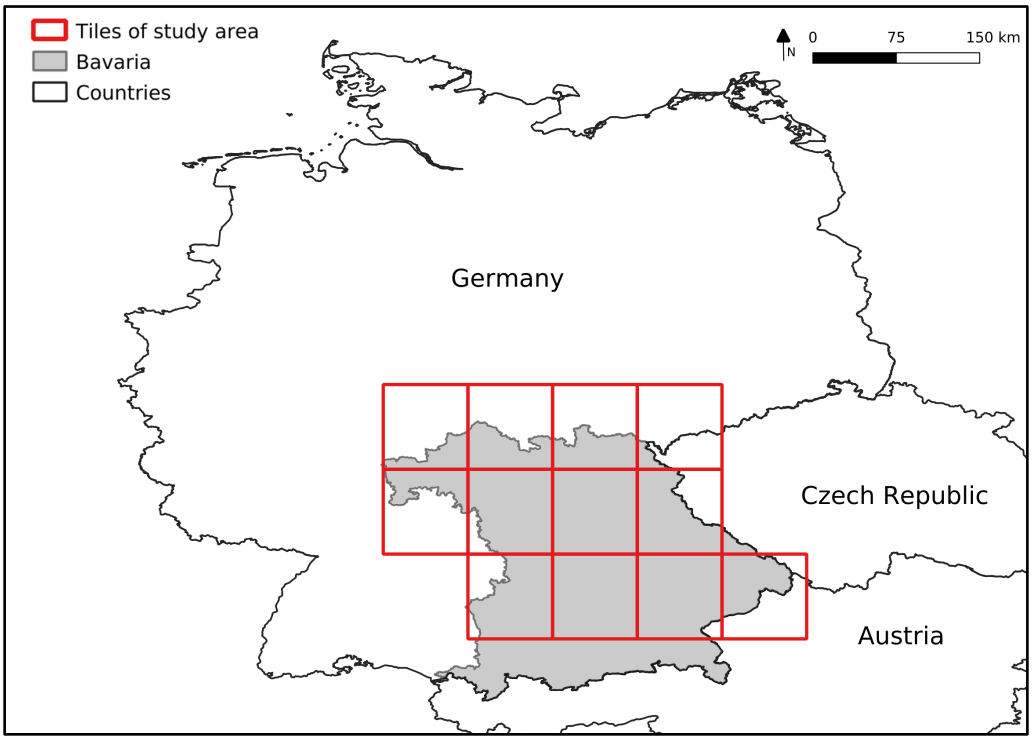

Figure 3: Study area in relation to Germany and the state of Bavaria

These tiles were selected because they mostly cover Bavaria and exclude the southernmost region of the state where the Alps are located, preventing alpine

samples with different soil genesis and sediment configurations from distorting the dataset. It is focused on Bavaria because of the availability of geocoded soil samples (cf. "2.3 Soil Sample Point Information"). Some areas of the tiles also cover territories of the German federal states of Baden-Württemberg in the west, Hesse, Thuringia, and Saxony in the north as well as parts of the Czech Republic and Austria in the east.

As a result of the central European location of the study area, the climate is classed as "temperate oceanic" according to Köppen (Cfb). Mean annual temperatures range from 9°C to 4°C with a precipitation of 550 mm up to 2500 mm (CHEN D. et al. 2015: 69–79).
Geomorphologically the area consists of a sedimentary filled basin formed by glacial advances during the last ice age at the foot of the Alps in the south surrounding Munich. In the north of Munich are the mountain ranges of the Franconian Jura in the northwest and the Bavarian and Bohemian Forests in the northeast. The northernmost area encompasses the South German Scarp lands in the northwest as well as parts of the Thuringian-Franconian Highlands. These are relatively low mountain ranges with peaks just shy of 1000m, formed by the variscan orogeny.
Mostly, soils in the study area are classed as Cambisols and Luvisols (this includes their German classification of *brown earths*), covering about 45% of the area. With 15% coverage Regosols are the next most frequent, followed by water stagnation soils like Stagnosols at 12%, according to the BKKA (AD-HOC AG Boden: 2005: 173-175) , and their equivalent soil classes of the World Reference Base for Soil Resources system (IUSS Working Group WRB 2006).
The CLC classification shows that 31% of the study area is cropland, 37% is covered by forests and 18% is grassland.

2.2 Soil Reflectance Composite

As a spectral database, this thesis uses the Soil Reflectance Composite, as processed by the DLR after the SCMaP workflow, combined for the years 1984 until 2014. The long timespan is chosen for this thesis because of the higher number of individual pixels classed as fields and cropland over the larger timespan. This results in a larger repository for comparing reflectance values to the SSLs. In the 5-year time frames on the other hand, the total amount of pixels is smaller and thus the likelihood of a soil

sample intersecting a soil pixel is decreased. Additionally, most of the SSLs have been taken at dates distributed from the 1980s to the 2000s, choosing a 5-year period would limit the available SSLs.

Most importantly, averaging out the reflectance values over a greater amount of time helps reduce variability due to medium-term weather influences such as droughts, exceptionally long summers or extended wet periods.

The spectral basis for this thesis consequently consists of about 63 million 30 m by 30 m pixels from Landsat missions 4, 5, and 7 which have been identified as exposed soil, containing spectral bands (**Table 1**) averaged out over the timeframe 1984-2014. Band 6 is a thermal band with a range from 10.40 - 12.50µ and was therefore removed. An excerpt of the final SRC shows the pixels identified as exposed soils as an RGB image, all other pixels are removed (white background) (**Figure 4**).

Table 1: Landsat Bands used in SCMaP

Band Number	Band Name	Range
Band 1	Blue	0.45 - 0.52 µm
Band 2	Green	0.52 - 0.60 µm
Band 3	Red	0.63 - 0.69 µm
Band 4	Near-Infrared	0.77 - 0.90 µm
Band 5	Short-Wave Infrared 1	1.55 - 1.75 µm
Band 7	Short-Wave Infrared 2	2.09 - 2.35 µm

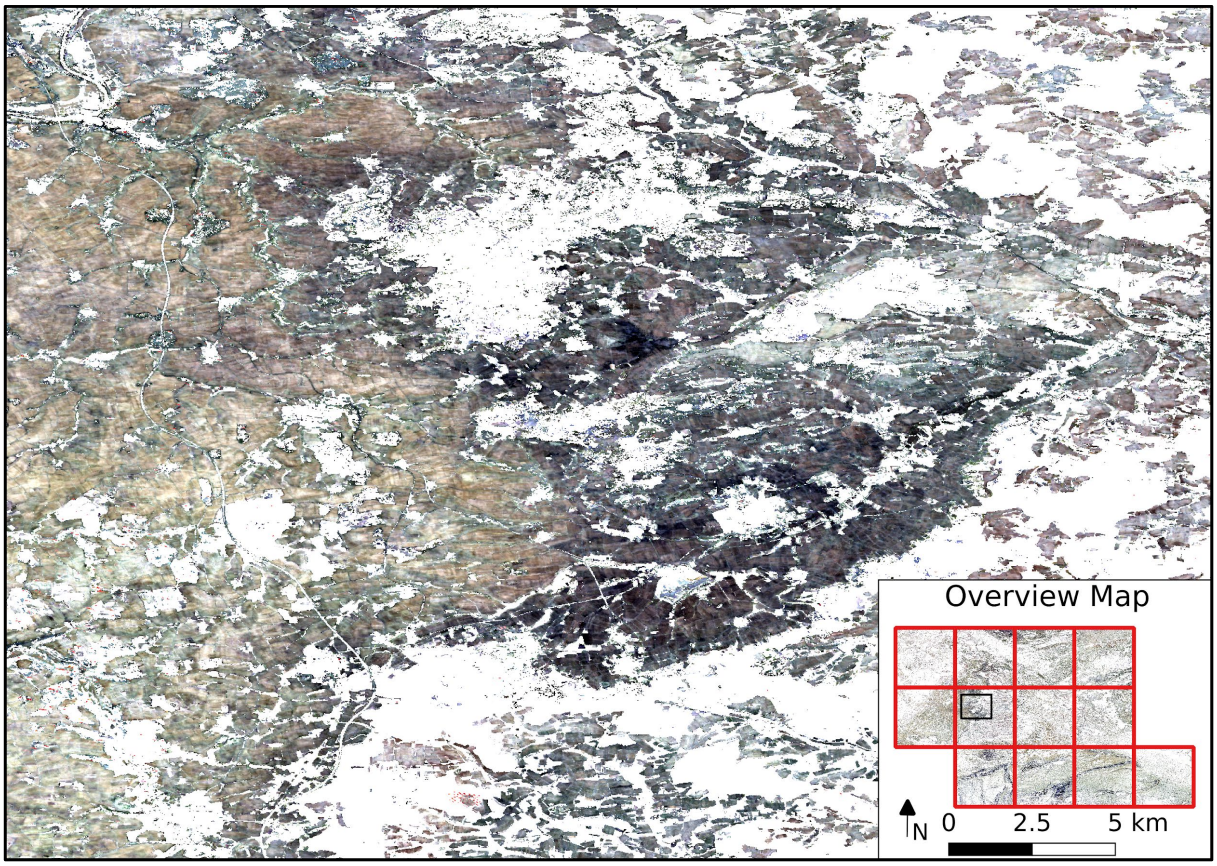

Figure 4: Excerpt of area of SRC for Bands 7 (represented as R), 5 (represented as G) and 3 (represented as B). False-color image. Areas of conformity in color are visible.

2.3 Soil Sample Point Information

2.3.1 Point Samples

In order to create an input database with sample locations and their known soil parameters, several data sources are compiled together. Two state agencies provided their non-public data points to this thesis. Additionally, a database from the European Statistical Office is used. The different sources, their methods of acquiring the data and its representation is explained below.

LFU

The Bavarian State Office for the Environment (Bayerisches Landesamt für Umwelt - LFU) maintains a network of sampling locations in Bavaria. Sites are selected based on an 8 km by 8 km grid which is drawn over the state, exact sampling locations are then picked within a 500-meter radius around the grid nodes to find a suitable position. Special attention is given to choosing sites as homogeneously as possible regarding altitude, relief, soil type, parent rock and vegetation (WIESMEIER et al. 2014: 208–220). Only sample locations which are situated within an agricultural field were made available by the LFU, 2086 in total.

The soil type attribute of these points is given in abbreviations according to the German soil systematic BKKA (AD-HOC AG Boden: 2005: 173-175), for example "RR-BB" for rendzina/brown earth, as well as their highest-tier classification of the soil group, in the aforementioned example "B" for brown earth.

Top soil texture is given in four classes according to the BKKA (AD-HOC AG Boden: 2005: 132-133),: sand, with a diameter of 0.063 mm to 2 mm; silt, with a diameter of 0.002 mm to 0.063 mm; clay, with a diameter smaller than 0.002 mm and loam, which is an even mixture of the three diameter classes.

Appended are finer distinguishing prefixes, such as "Sl" to indicate the presence of other diameter classes, with the dominant class written in a capital letter.

The unit in which the carbon organic content is listed is mg/g.

LfL

The Bavarian State Office for Agriculture (Bayerische Landesanstalt für Landwirtschaft - LfL) is part of a network of permanent soil sampling sites, aggregated and maintained by the German Federal Environmental Agency. All points

located within the state of Bavaria and on land classified as agriculturally used were provided to this study by the LfL (in total 118 sampling locations). More than 400 parameters are recorded for each location, including all of the soil parameters relevant in this study, as well as for example land use, in-depth soil analysis, and radioactive measurements (SCHILLI et al. 2011: 16) .The network exists since the 1980s, each location is revisited once every 1 to 10 years, depending on the individual point (KAUFMANN-BOLL et al. 2011: 4).

Since several measurements at different points in time are present for any given point, they need to be modified. Samples taken outside of the timeframe were removed, the remaining C_{org} values were averaged and then transformed into their according classes (cf. "2.3.2 Preprocessing"). None of the points showed a change in soil type or soil texture, thus the classification could be easily transferred. After the transformation, one single value was left for each soil parameter used in this thesis.

LUCAS

The European Statistical Office (Eurostat) conducts the "Land Use/Cover Area frame Survey" (LUCAS) since 2001 and is available for all EU member states since 2012. In this survey, sample points are placed over the European Union in a grid with a width of 2 km in between nodes. Each of the 1.1 million points is sampled and classed individually. Many different measurements are taken, including the humus content of the top soil, but not the soil class or the soil texture (TÓTH et al. 2013: 3-6). The humus content can be easily converted into the organic content of the topsoil, so that the LUCAS information can be used to expand the C_{org} training dataset, but not the soil type and soil texture datasets. Unlike the other two data sources, the LUCAS dataset is not limited to Bavaria, meaning that there are also points located within areas which fall outside of the state into other German federal states as well as into the Czech Republic and Austria. A total of 505 LUCAS points are available for the twelve test tiles.

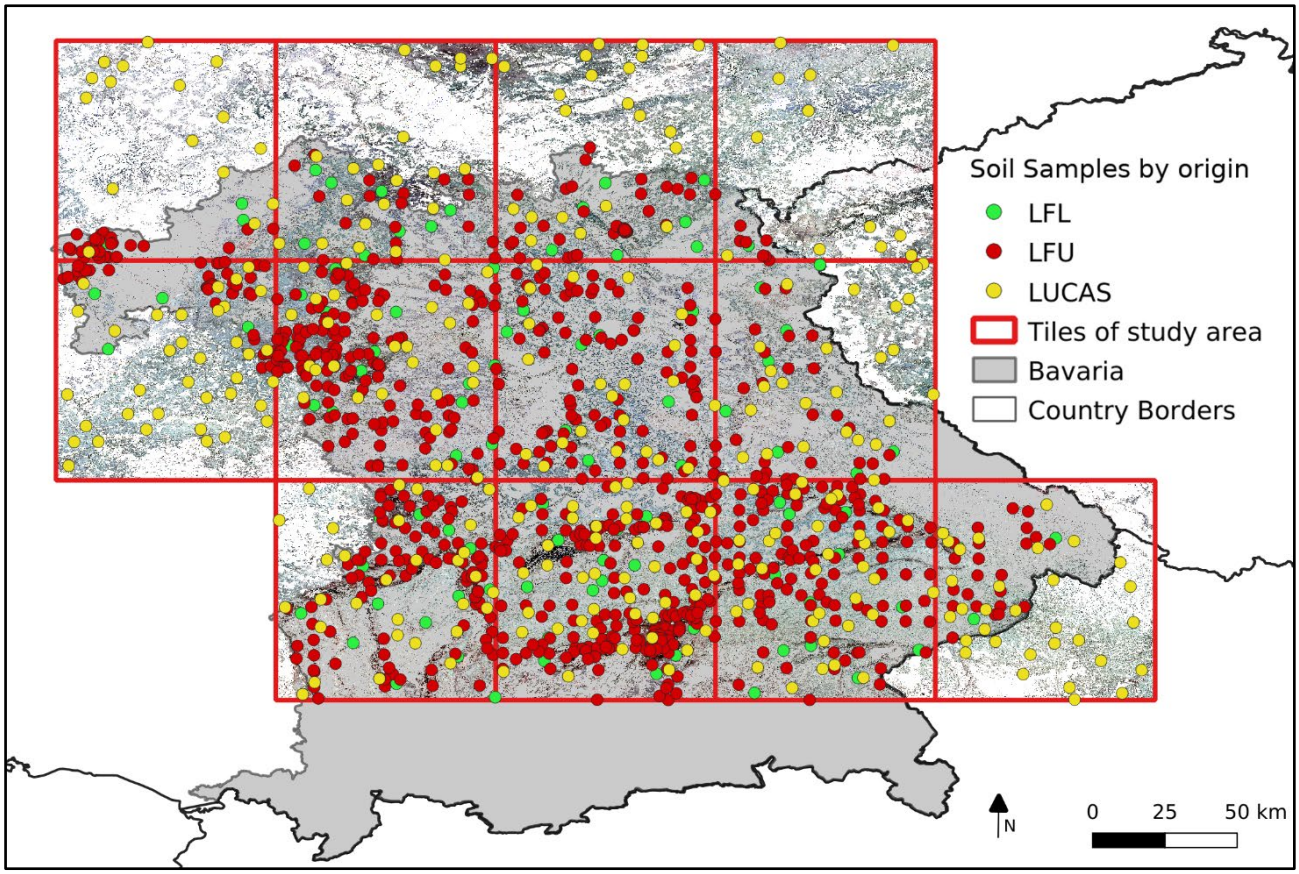

Figure 5: Distribution of Soil Sampling Locations, color-coded by origin.

2.3.2 Preprocessing

The data from the different sources need to be unified into the same format. Merging the data from all sources is only possible if the classifications for the soil parameters follow the same structure. As mentioned in the data descriptions the classification methods differ from each other.

Information regarding soil type is converted to the highest tier classification as given in the German soil systematic BKKA (AD-HOC AG Boden: 2005: 173-175). The classes are presented in **Table 2**.

Table 2: Soil Type Classes

Symbol	German Name (BKKA)	Translated Name
A	Auenboden	Floodplain Soil
B	Braunerde	Brown Earth
D	Pelosol	Pelosol
G	Gley	Gley
L	Lessivés	Lessivés
M	Marsch	Marshland Soil
P	Pelosol	Pelosol
R	Rendzina	Rendzina
S	Stauwasserböden	Water Stagnation Soils
T	Tschernosem	Chernozem
Y	Terr.-Anthr. Boden	Anthrosol

Soil texture classifications are also converted to their highest-tier classes by keeping the majuscule letter of the abbreviation given in the BKKA (AD-HOC AG Boden: 2005: 173-175), see **Table 3**.

Table 3: Soil Texture Classes

Symbol	Name	Grain Size
S	Sand	0.063 mm to 2 mm
U	Silt	0.002 mm to 0.063 mm
T	Clay	< 0.002 mm
L	Loam	Even mixture of the three

C_{org} values were given in different units by the different data sources. The LFU used mg/g, while LfL used percentage. The conversion to percent is easily done, but the LUCAS database gives C_{org} values only indirectly by specifying the humus content of the top soil. According to the BKKA, humus is made up of 58% organic carbon (AD-HOC AG Boden: 2005: 107). Therefore, multiplying the humus content by the factor 0.58 returns the C_{org} content in its original unit, which was also percent. All of the C_{org}-information is thus converted to the same unit and format (percentage of soil mass).

As seen in **Table 4**, the values are classed according to the carbon content classification of the BKKA (AD-HOC AG Boden: 2005: 109).

Table 4: Soil Organic Carbon Content Classes

Symbol	Organic Carbon Content Range
k1	<0.5 %
k2	0.5 - 2.0 %
k3	2.0 - 5.0 %
k4	5.0 - 15.0 %
k5	15.0 - 30.0 %
K	>30.0 %

The tables from the data providers can be merged because of the data conformity after the aforementioned restructuring and conversion. LUCAS points are given the soil type and soil texture value of "*nA*", simplifying their detection and exclusion from soil type and texture classifications further down the line.

In order to create a joined spatial database with correct location data, the point information is reprojected into the same map projection. Since both Bavarian agencies use the Gauß-Krüger coordinate system, they are warped into WGS1984 EPSG:4326 to fit the LUCAS data as well as the WGS1984 formatted *.tiff* SCMaP mosaics which are used in this thesis. Also, unique IDs are given to each point location.

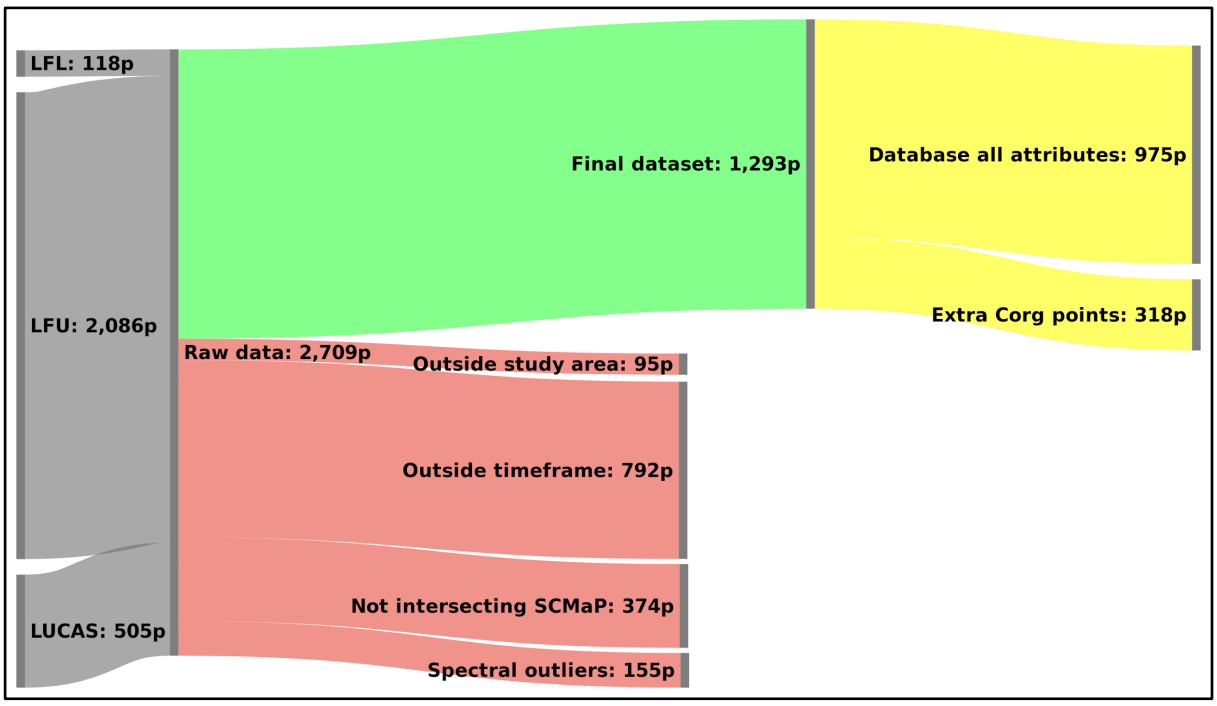

Figure 6: Preprocessing of Soil Sample Locations.
Sankey Diagram, width of bars is proportional to amount of SSLs.
Grey: data sources, *red:* eliminated SSLs, *green:* final dataset, *yellow:* composition of final dataset.

Following the conversions and attribute-related changes, the 2,709 data points need to be adapted to the spectral data and the study area.

As shown in the Sankey Diagram (**Figure 6**), 95 points are eliminated because they are not within the twelve tiles. 792 sample points are excluded because they were not taken within the 1984 - 2014 timeframe. Even though it is not likely that soil type, C_{org} content or soil texture will change during a few decades, it is still preferred that the sample point was actually once depicted in the spectral data.

Afterwards, an intersect is performed to retain only those locations which are situated on a pixel of the SRC. Another 374 points are discarded during this step.

Within the SRC pixels, severe spectral outliers are expected. At a pixel size of 30 m, the pixel might consist of 50% cropland and 50% road or pavement. To eliminate distortion by these outliers, a matrix of 3 by 3 pixels around the SSLs is created. SSL pixels which display a deviation of more than twice the standard deviation compared to the surrounding pixels in the matrix are removed (**Figure 7**). By doing so, another 155 pixels are excluded.

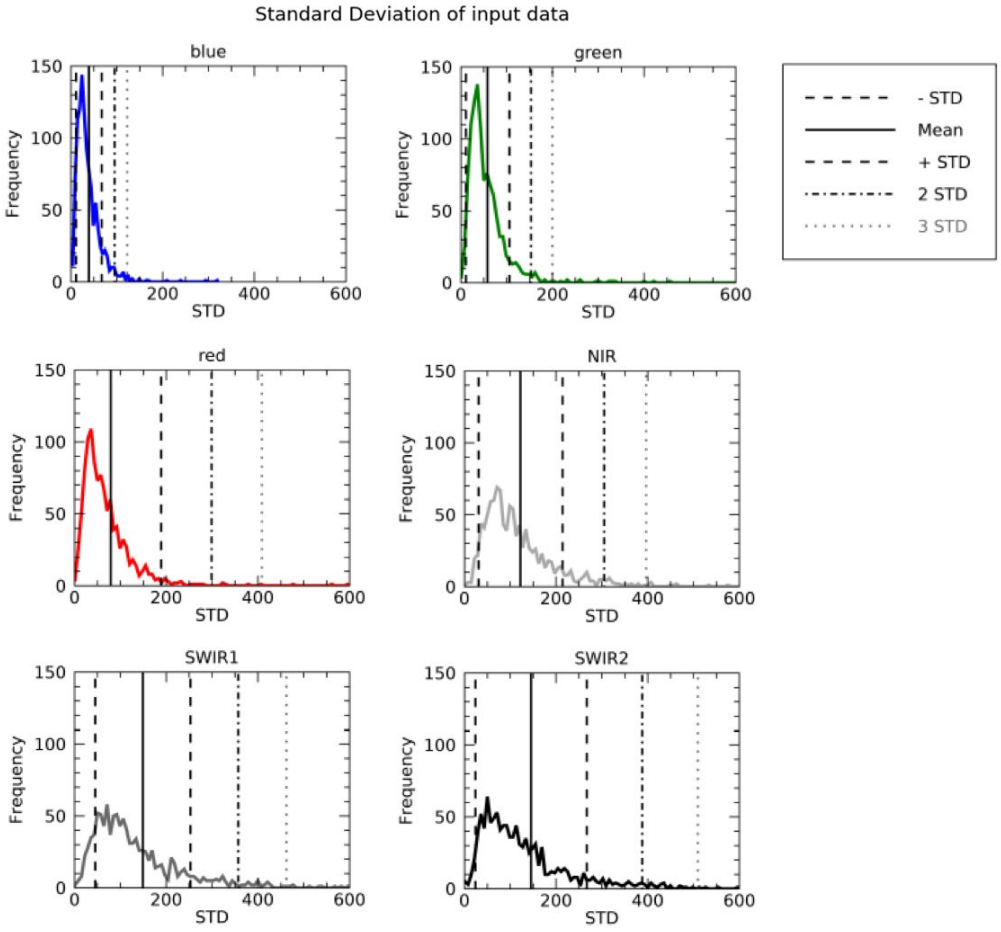

Figure 7: Histogram of pixels in 3x3 matrix around original locations, also showing standard deviation multiples.

After all ground-sampled data is merged, they need to be combined with the spectral data. Reflectance data from the SRC is now gathered by sampling the pixel reflectance value of every location from the list. Values for all six bands are added to each respective point as attributes in the list.
The final input database is hence an indexed list containing x and y coordinates, their classifications regarding soil type, C_{org} content and soil texture aggregated and uniformized; as well as the six local reflectance values extracted from the SRC for 1293 entries. A column showing the C_{org} numerical value is also included.
975 of these points have attributes regarding soil type and/or texture, due to the LUCAS database almost all 1293 locations have an attribute regarding their C_{org} content. It is important to note that due to in-field sampling errors, some SSLs do not have a valid data entry. The attribute itself is not removed because it has entries for the other parameters. The statistical methods and Random Decision Forest methods used automatically ignore these fields ("no entry" percentage: C_{org} 0,15%, soil type 2,36%, soil texture 1,33%)

3. Methodology

3.1 Analysis of distributions within dataset

Initially, the input dataset is analysed with a focus on the soil parameters, gaining knowledge on the distribution of the classes is essential. Regarding the three different soil parameters, it is important to know how the classes are distributed within the dataset. An overwhelming majority of one or several classes could drain out other classes or limit the validity of their spectral boundaries if there are only a few points available in the dataset for a given class. Therefore, the class distributions are visualised as bar graphs.

Then, the aggregated training dataset with their spectral information is statistically analysed in order to visually determine correlations between different classes and their reflectance values. All points are plotted as individual graphs. Plotting all points together results in a visualisation of the variance within the dataset. Combining the

points and adding graphs showing the mean reflectance as well as the standard deviation can show the variance inside the classes, as well as expose possible outliers.

Since the C_{org} soil parameter also carries a numerical value, a scatterplot can further visualize the distribution of those values. Consequently, a scatterplot is created for each band, plotting the C_{org} value against the reflectance value of the bands. Colouring the points according to their class shows the distribution of reflectance values of the classes in direct comparison to the other classes of this individual band, enabling a visual judgement as to how far the classes are varying.

Then, for each soil parameter, the mean spectra of all classes are plotted together, allowing for a visual comparison of the graphs, as well as enabling the visual identification of the distinguishability of the classes.

In order to back up the previous visual observations with numbers, a table showing the differences between the classes is created. For each soil parameter, the classes are plotted against each other showing the differences in reflectance between them for their manually chosen most distinctive single band. The differences between the classes are shown as percentages, enabling the cross-referencing of all classes for this individual band. This allows for the identification of classes with a higher percentage in spectral difference in singular band and thus have a higher chance of correct prediction.

3.2 Correlation Coefficients

Two statistical correlation methods are implemented to find possible correlations in the dataset where numerical values are available (C_{org}), the Pearson (Pearson R) and Spearman correlation coefficients (Spearman R). Since the other soil parameters are classed and without numerical values, correlation cannot be examined in those cases. For each of the six bands, a correlation between C_{org} value and reflectance is examined using the "SciPy" open source scientific computing library for Python 3.7 (MILLMAN et al. 2011: 9–12).

The *scipy.stats.pearsonr* utility calculates the correlation according to the method derived by Karl Pearson (PEARSON 1896: 253–318). As a measure of linear strength

of association between two observations (bivariable), the coefficient ranges from -1 and 1, which represent the highest correlation values for negative and positive correlation, to 0, which indicates no correlation (CHEN et al. 1981: 135; CROUX et al. 2010: 497–515).

$$R = \frac{covariance(x, y)}{(\sigma x)(\sigma y)}$$

Equation 2: Pearson correlation coefficient

As this simplified formula (**Equation 2**) shows, the Pearson correlation coefficient (R) is the covariance of each set of the bivariable (*x* and *y*), divided by their multiplied standard deviations (σx, σy) (PARK: 2018: 213–265). Pearson R is applied to the training data in order to evaluate a possible direct linear correlation.

Spearman R works similar to Pearson's method, but the variable sets are ranked beforehand. Each variable is given it's rank along the axis as its new value, starting at the lowest value with rank one. This replacement of absolute values with their rank from lowest to highest removes the variability and is thus useful in detecting monotinic trends within the dataset.
With the values changed to their corresponding ranks, the Pearson R method is executed (WEAVER et al.: 2018: 435–448). The resulting Spearman correlation coefficient takes the same form as the Pearson coefficient, meaning +1 and -1 stand for highest positive and negative correlation, while 0 indicates no correlation (CROUX et al. 2010: 497–515). Due to its ranking system, the Spearman R is able to more accurately detect monotonic relationships, thus relationships where while variable *x* declines in value, variable *y* never increases and vice versa (BORKOWF 2002: 271–286). Removing the absolute values makes this coefficient less vulnerable regarding outliers and as long as the *x* and *y* values develop in the direction of the same algebraic sign (regardless of their absolute value), a correlation is detected. (ROWE 2015: 311–335). Spearman R is chosen to detect a possible monotonic trend relationship between reflectance and C_{org} content, in case there is a correlation between rising reflectance values and rising C_{org} content or vice versa.

3.3 Component Analysis

3.3.1 Principal Component Analysis (PCA)

While dealing with multispectral data such as the training dataset, it is challenging to visualize the variance inherent in the spectral bands. Each band stands for one dimension in which the data might vary. In order to reduce the dimensionality, a Principal Component Analysris is performed (PEARSON 1901: 559-572). PCA has established itself as a statistical tool for multivariate analysis and is widely used in research dealing with large datasets, including spectroscopy (JAMES: 2013: 127-173). The aim to break down the attributes and identify the most significant data dimensions is achieved by linearly transforming the data to uncover the principal scatter direction of each dimension.

The best transformation is found by visualising the attributes as a scatterplot. Then, the centre of all data points from all classes is found and the origin of the graph is moved to this location (DE SILVA 2017: 15–17). The next step is drawing a line, which will later become the new axis, intersecting the origin and most accurately describing the axis of variance of the points, followed by the reprojection of the points onto that line. The squared sum of all distances from each original point towards the new axis is called “eigenvalue”, the vector describing the tilt of the original axis to the newly created axis is called “eigenvector” (PEARSON 1901: 559-572). The line with the best eigenvalue score is chosen. The line’s eigenvector shows the influence of the original attributes in regard to the importance of the examined feature (JEFFERS 1964: 225–236). The line itself represents the Principal Component 1 and is the first axis of the new (tilted) coordinate system. To find the second axis, all possible lines perpendicular to PC1 are assessed for their eigenvalues as done in the step before. The line which fits the variance best is chosen as axis for PC2. For each additional dimension in the original data, the line among the perpendicular plane (of the previous PCs) with the best eigenvalue is found and added to the principal components list (JOLLIFFE: 2002: 2-4). Once all PCs are determined, the graph is rotated so that PC1 is at the x axis. Since the sample points are still projected onto each PC axis, they are now reprojected to the position the points along the PC axes intersect (HOTELLING 1933: 417–441 and 498-520). Because the eigenvectors show the shift of the axis towards the PCs and thus towards an axis of more importance, they indicate the

proportion of importance for each former attribute for the new PC (WOLD et al. 1987: 37–52). The proportions of each attribute in the new PC is called the "loading score", comparing the loading scores for all PCs allows a numerical ranking of importance for all PCs (PEARSON 1901: 559-572).
Because of the transformation into a covariance matrix, the data points have been transformed from absolute to relative units (JEFFERS 1964: 225–236). The results are changed graph axes, with the dataset having new coordinates within this graph. Finally, a three-dimensional scatterplot is created from the three previously identified most important attributes. Ideally, the different classes of the dataset should group together as separable clusters, with their variance spread out along one of the axes (KESHAVA 2003: 55–78).

The PCA performs a covariance analysis to find the most important attributes in a dataset in order to lower dimensionality, before transforming their axis in a way that most accurately reflects their linear direction of variance.
In this thesis, the *sklearn.decomposition.PCA* utility from the Scikit-learn machine learning library for Python 3.7 is used and the result plotted to a 3D scatterplot for the three most important attributes (PEDREGOSA et al. 2011: 2825–2830).

3.3.2 Linear Discriminant Analysis (LDA)

Another statistical method for dimensional reduction used in earth sciences (TAHMASEBI et al. 2010: 564–576) is the LDA, which is closely related to the PCA.
It builds upon Fisher's linear discriminant (FISHER 1936: 179–188) and also searches for linear combinations within the attributes, but contrary to PCA, takes not just the attribute values, but the actual classes themselves into account (BÜYÜKÖZTÜRK et al. 2008: 73–92).The search for linear dependency is thus not focussed on the values, but on the distribution of the classes themselves by projecting the classes onto a line, therefore maximising separability among the categories (WANG X. et al. 2003: 2429–2439). The resulting principal components are then plotted to axes according to their rank of importance, like in the PCA.
The axes lines are formed by considering two separate criteria simultaneously. First, when plotting the variables, the mean within each variable dataset is found, and then the line is drawn in a way that maximises the distance between the means of the

classes (FISHER 1936: 179–188). Since more than two dimensions are considered, the distance is not measured between the means of the two classes, but between each of the class means and the central point of all classes combined (BÜYÜKÖZTÜRK et al. 2008: 73–92)(BÜYÜKÖZTÜRK et al. 2008: 73–92).

The second factor consequently taken into account is the minimization of variation ("scatter") of the classes within the dataset along the axis (FISHER 1936: 179–188). The optimal ratio is then calculated by the following formula (**Equation 3**), where d_1, d_2 and d_3 stand for the mean distances to the central point of the classes and s_1, s_2 and s_3 for the distance of scatter along the newly created axis. The distances and scatter values are squared so that negative values do not cancel out positive ones.

$$\frac{d_1^2 + d_2^2 + d_3^2}{s_1^2 + s_2^2 + s_3^2}$$

Equation 3: Variation ratio for axis location (LDA)

Ideally, the numerator would be very large, representing a big distance between the means; and the denominator very small indicating a low scatter. All possible axes are assessed, then ultimately an axis is drawn at the line with the best ratio as returned by the equation aforementioned. The result is a dataset with coordinates along a relative scale, whose axes were altered from their original position in order to better differentiate the classes (FISHER 1936: 179–188). Drawing a three-dimensional scatterplot from this dataset is expected to result in a clustering of the classes.

In this thesis, the *sklearn.discriminant_analysis.LinearDiscriminantAnalysis* utility from the Scikit-learn machine learning library for Python 3.7 is used (PEDREGOSA et al. 2011: 2825–2830).

3.4 Random Decision Forest Classifier

3.4.1 Random Decision Forest Methodology

In supervised learning procedures, the algorithm is given a database of inputs, which are also called predictors, and their according output label. The predictors contain attributes, which are values that are corresponding with the output label (HASTIE et al.: 2009: 9-41). This "input database" is the basis which the learning process uses to create a decision model which rates the importance of predictor values based on the

known output labels. After creation of the model, classifications for predictors without a given output label can be made using the previously created decision model. These methods can be used for both classifications and regressions, but only classifications will be discussed and needed in this thesis.

Singular Decision Trees

Singular Decision trees follow the *recursive binary splitting* approach, which is a top down method splitting the tree into two following nodes at every node (**Figure 8**). Starting from the root node, a decision between two options is made regarding a singular attribute value, which in turn leads to further nodes. These nodes are called internal nodes, and just like the root node, divide the tree into two mutually exclusive regions. While from the root node, all possible nodes and node regions are accessible via the paths taken within the model, terminal nodes represent the end and final labelling decision of the tree (SONG et al. 2015: 130–135). The terminal nodes are thus the lowest level of the tree, at which the algorithm is sufficiently confident to classify the predictor with the according label. Terminal nodes exist for all possible labels, so that given the appropriate predictor values, the predictor can be assigned every possible label present in the dataset (HASTIE et al.: 2009: 101-137).

The definition of the most binary partition (which splits the tree into two paths at this node) is determined by evaluating all possible splits for all possible values of all available predictor attributes, then choosing the splits with the lowest degree of impurity, meaning the most accurate and thus most stringent decision path between root node and terminal node (ISHWARAN 2007: 519–537). Node impurity refers to the mean decrease in accuracy, thus misclassification, as measured by the Gini impurity after validating the final labels assigned. The final model is thus a decision tree created by the already known predictor values and their labels, which can then be extrapolated onto different predictors and their values, where the output labels are unknown (CUTLER et al. 2012: 157–175).

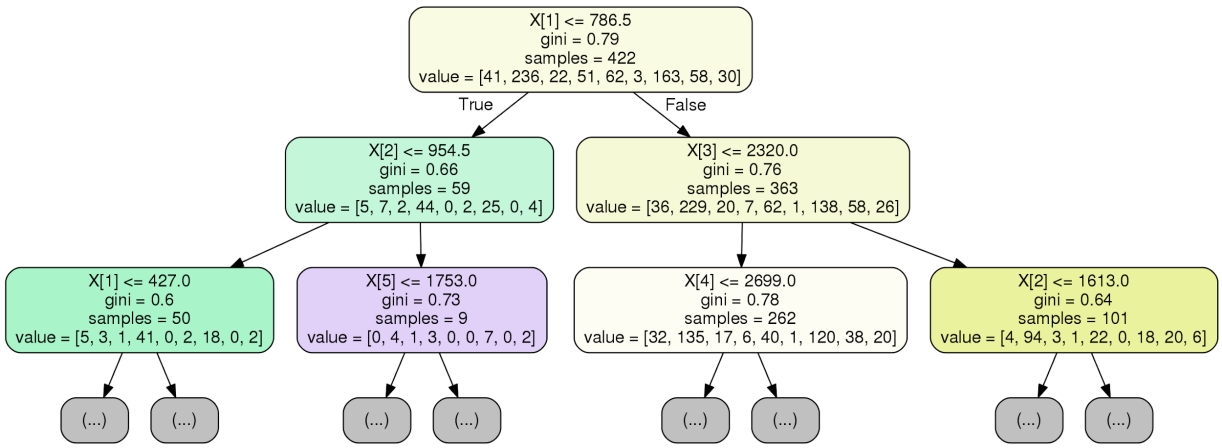

Figure 8: First 3 layers of singular decision tree.
The Random Decision Forest used in this thesis has 500 of these trees with a depth of up to 13 layers.

Decision trees come with the inherent disadvantages of possible "overfitting" as well as a high sensitivity to changes in the input data (BURNHAM & ANDERSON: 2002: 1-9).

The term overfitting describes instances where the model is very closely dependent on the input data, thus distinguishing between more statistical parameters than the "real-life" model. This means the decision making is based on statistical noise within the data, as if this noise is representative of the classification (BURNHAM & ANDERSON: 2002: 1-9).

Due to this close correspondence to the data ("overfitting"), a single decision tree is very sensitive to changes in the training data. This leads to inaccurate split parameters at the nodes, which are based too closely on the training data, and therefore a tendency to mislabel predictors (CUTLER et al. 2012: 157–175).

Random Decision Forest

The RDF is a supervised learning algorithm first created by Tin Kam Ho in 1996 (HO 1995: 278–282) and later extended by Leo Breiman (BREIMAN 2001: 5–32), which is based on singular decision trees. Since its inception, this method has found widespread use and acceptance in many research fields including spectral surface classification due to its easy implementation and reliable results (GISLASON et al. 2006: 294–300; RICHARDS 2013: 343–380).

Random Decision Forests enables to reduce overfitting by combining multiple decision trees and Bootstrap aggregating or "bagging".

"Bagging", as introduced by Breiman, extracts a smaller training set from the original training data at random, usually 75%. For each of the trees in the random decision forest, a different "bagged" dataset is extracted and a decision tree model built upon this input data created. Repeating the decision tree method many times with varying inputs from the "bagged" original training data results in slightly different trees with a different node layout and decision values. This randomization of the input dataset aims to reduce date dependability of the model by using many trees with consequently different layouts (BREIMAN 2001: 5–32).

The unused part of the training dataset can then be used to verify the accuracy of classification of each individual tree (BREIMAN 2001: 5–32). Subsequently, the decisions made by each individual tree for a single predictor are aggregated and averaged. Having multiple singular decision trees with their own structure for each predictor as well as their classification result is then used to choose the final labelling decision, based on the number of instances the predictor was assigned a certain label (JAMES: 2013: 127-173).

"Bagging" results in lower training data bias and overfitting and consequently in a higher accuracy model (BREIMAN 2001: 11).

3.4.2 Random Decision Forest implementation and parameters

In this thesis, the *Scikit-learn* implementation of a Random Decision Forest for Python 3.7 is used. *Scikit-learn* is an open source Python machine learning library, featuring many classification, regression and clustering algorithms and is publicly available since 2011 (PEDREGOSA et al. 2011: 2825–2830).

Before starting to build the trees, 30% of the training data is removed and stored for the verification of the forest, using the *sklearn.model_selection.train_test_split* utility ("Bagging"). Usually, only 25% are removed, but due to the high frequency of several classifications such as "k2" in the C_{org} dataset and "B" in the soil dataset, increasing the verification sample size increases the likelihood of inclusion for different classes, making the verification process more accurate. A smaller sample size would result in a large presence of one singular class, displacing more infrequent classes and

consequently lead to an artificially inflated accuracy percentage. 30% “bagging” proved to be a reasonable compromise between reducing tree building sample size and maintaining a high accuracy of verification (Scikitlearn User Guide 2019: 2331). Tree building is then handled by the *sklearn.ensemble.RandomForestClassifier* utility, which features numerous possible parameters to influence model building. Out of all possible parameters, the following were considered while building the model for this thesis:

- *n_estimators=500*
- *bootstrap=True*
- *max_depth=None*
- *max_features = auto*.

All other parameters remain in their standard value, as described in the documentation for the *RandomForestClassifier*.

The *n_estimators* parameter sets the number of trees built. The default amount of 100 was increased to 500, expecting an increase in model accuracy.
Limiting the number of internal tree nodes and thus the depth of the tree can be done using the *max_depth* command but is set to be unlimited in this model. Not limiting the maximum tree depth might result in overfitting but increases the distinguishability of smaller variances in the training data, as present in this thesis. Different settings for this parameter were considered and experimented with, ultimately it was set to *unlimited* because of its higher prediction accuracy.
As mentioned in the previous paragraph, ”bootstrapping” is a valuable tool in increasing accuracy, thus it is enabled by setting *bootstrap=True*.
Using an integer as the parameter *max_features* results in no feature subset selection within the trees, which according to the documentation of *scikitlearn* is best used in regression tasks. Setting the parameter to *auto* is advised in classification tasks such as this implementation, meaning the size of the randomized subset of features to be considered for splitting an internal node is set to the square root of the remaining training dataset (Scikitlearn User Guide 2019: 2331).

The aforementioned parameters were set to many different combinations, with the current setup achieving the highest level of prediction accuracy.

Further restraining tree size using commands such as *min_samples_split* and *min_samples_leaf* was considered but decided against due to the risk of lower distinguishability regarding features with only small variances, but as a trade-off slightly increases the risk of overfitting.
Other more in-depth settings like rating node splits by entropy instead of the Gini Index using the *criterion* parameter or influencing the node split creation method using the *splitter* parameter were not changed and left in their default states (Scikitlearn User Guide 2019: 2331).

3.5 Validation

Verifying the results obtained by the Random Decision Forest is a necessary step in ensuring the quality of the output. Machine learning methods are quite capable of producing results and even high prediction accuracy scores within its dataset, but the results might not correlate with the reality. It is not necessarily the case that the classifier bases its decision on those dataset variations, which are also responsible for the "real-life" classifications. This "overfitting" is a common example of a situation where the classifier evaluates itself with high accuracy scores, but the classification process has lost its link to natural circumstances (BURNHAM & ANDERSON: 2002: 1-9)

The most important metric in judging the method used in this thesis is the internal RDF evaluation. As described before, 30% of the training data is set aside before the classification, the classifier is then tested on the verification data and an accuracy percentage produced. Since every individual tree uses a different 30% of the training data and the outputted accuracy is a mean of all accuracies from each individual tree, it is still an important indicator as to how accurate the process is.
The German Federal Institute for Geosciences and Resources (BGR) produces maps regarding all of the soil parameters ("HUMUS1000", "BUEK1000", "BOART1000"). Unfortunately, these maps are heavily interpolated. Before considering these maps as basis for validation, it is checked if all the SSLs used in this thesis are identically classified in the according soil maps. Unfortunately, for all classes, only 55% of data points received the same classification by both the BGR and the data providers. Evaluating the prediction using these maps returns very low

accuracy percentages (C_{org} 19.46%, type 48.34% and texture 20.49%). With such high discrepancy in classification for validation- and training, the maps will not be used as a method of validation.

Also, the Topsoil Organic Carbon Content for Europe Map by the European Soil Data Centre was considered, but not used due to its low resolution of 1,000 m by 1,000 m per pixel and therefore high interpolation. Additionally, this map is built upon the LUCAS dataset, meaning the validation would use its own data to validate itself. The sample location density is also increased over the LUCAS database by adding the LfU and LfL datapoints to the training dataset. The LUCAS map is thus coarser in density than the training dataset. The idea of using map products as a method of verification is discarded due to the previously mentioned factors.

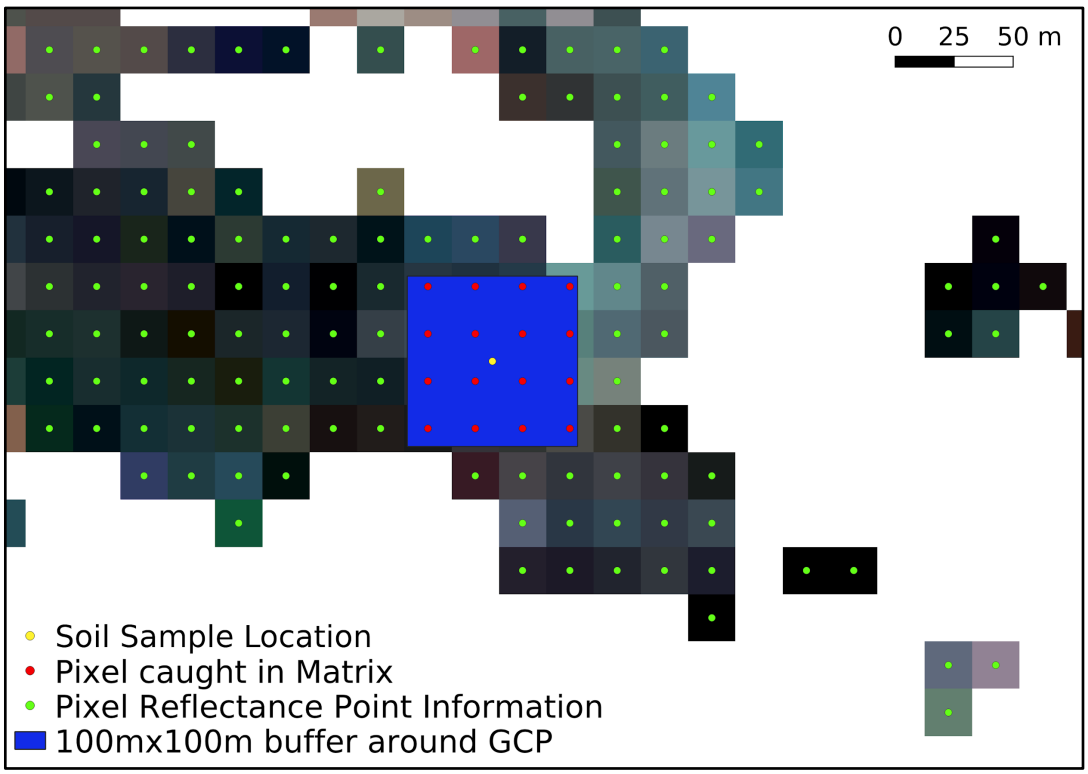

Figure 9: Examplary location of „quality control" matrix.
SCMaP pixels are in the background, their information values saved in the green point features. Original Soil Sample Location in yellow, surrounded by the 100x100m matrix in blue with the pixels "caught" in this matrix in red.

Basing on the SSLs, a different method for quality control is derived.
Assuming that neighbouring pixels to the SSLs only differ marginally from the SSL's soil parameter classifications, a 100 m by 100 m rectangle around each soil sample location is spanned. Assigning the "validation parameter" of the SSL to all other SRC pixels caught in the matrix leads to 15,385 pixels which can be used for quality control purposes (Figure 9). The central pixel of each matrix, which has been used to build the model, is excluded from verification. Since this method is based on the unsure assumption of unchanged attributes around the sample location it cannot be seen as a "true" verification method, more as a quality control measure. Also, because the data upon which the RDF is built is used to validate itself, the resulting accuracy score is of limited validity.

A scientifically valid method to verify the prediction accuracy cannot be implemented. The aforementioned methods possess limited explanatory power by themselves but taken together do give an overall impression of the validity. These methods are to be considered as "quality control", not as true validation methods.

4. Results

4.1 Statistical Results

Before presenting the predictions made by the RDF as well as their measurements of validity, the statistical results are shown.
Firstly, the input database is analysed regarding its internal composition and the distribution of the different soil parameter classes. This gives an insight into how well the different classes are represented in the input dataset. Secondly, the distribution of reflectance values within those classes is visualized, showing how or if the different classes have spectral characteristics in common. After that, the correlation coefficients and the results of the PCA and LDA are presented.

4.1.1 Representation of Soil Parameter Class Distributions

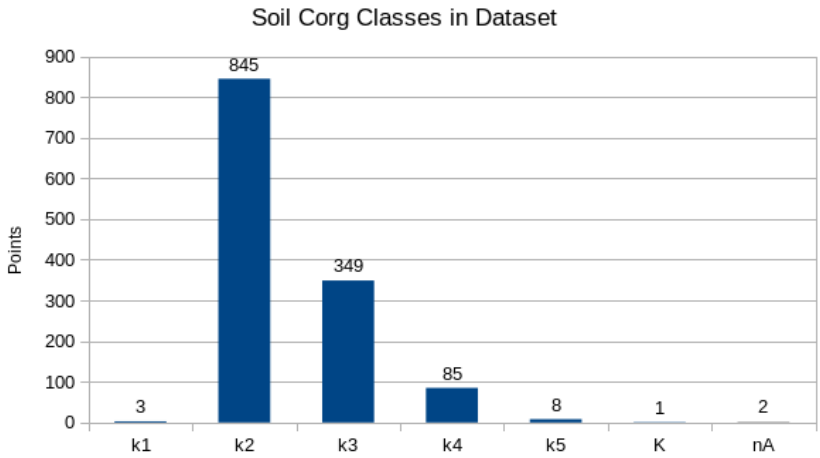

Figure 10: Distribution of C_{org} classes in soil samples

Examining the C_{org} class distribution (**Figure 10**), a very clear dominance of the *k2* class is visible since 65.35% of all points fall into that category. The second largest class, *k3*, makes up 26.99% of the dataset. The other four classes share the remaining 7.66%. The total number of available points for this class is 1291.

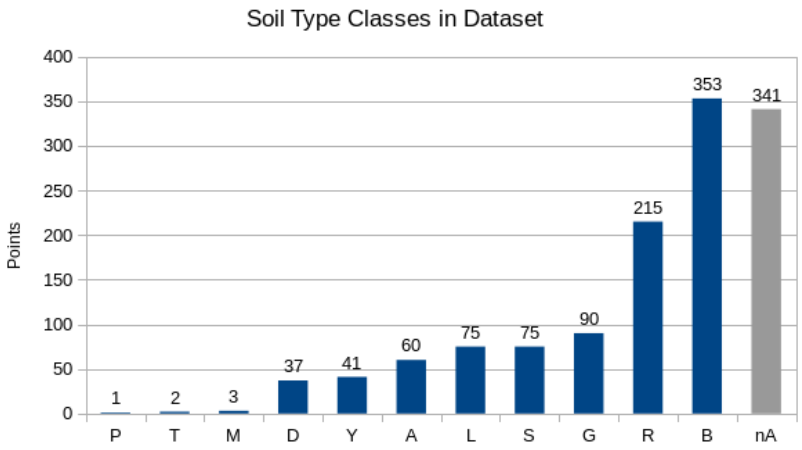

Figure 11: Distribution of Soil Type classes in soil samples

Examining the categorization of soil types (**Figure 11**), the total number of available points is 952. The dominance of a single class is not as prevalent as in the Soil C_{org} category. The largest class is brown earths (*B*) with 37.08%, followed by rendzinas

(*R*) with 26.37%. All other classes are represented by a percentage of under 10% of all samples, with classes *M*, *T* and *P* consisting of only a few points.

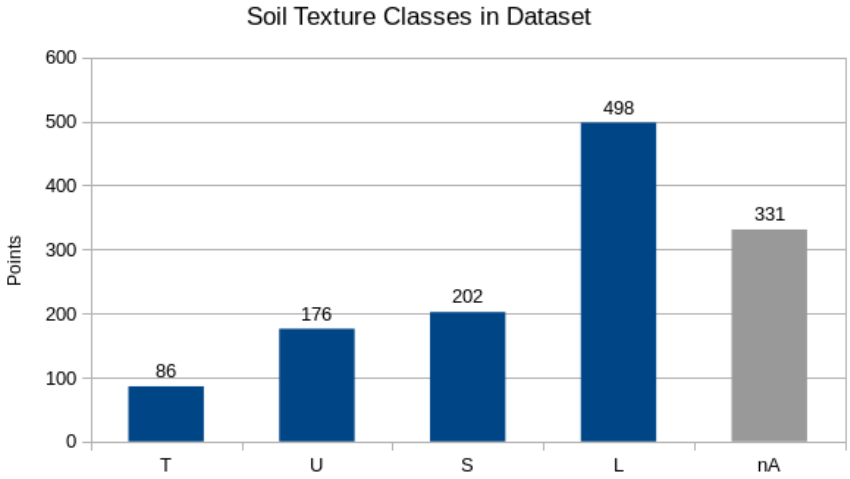

Figure 12: Distribution of Soil Texture classes in soil samples

Analysing the soil texture classes (**Figure 12**), 52.31% are classed as *L* (Loam), 22.22% as *S* (Sand), 18.49% as *U* (Silt) and 9.03% as *T* (Clay). A dominance of class *L* is visible, the total number of available SSLs is 962.

4.1.2 Spectral Distribution of Parameters by Classes

Carbon Organic Content

After examining the class distribution within the training dataset, the spectral information for each class within the soil parameters is analysed. Firstly, the soil C_{org} class is visualized.

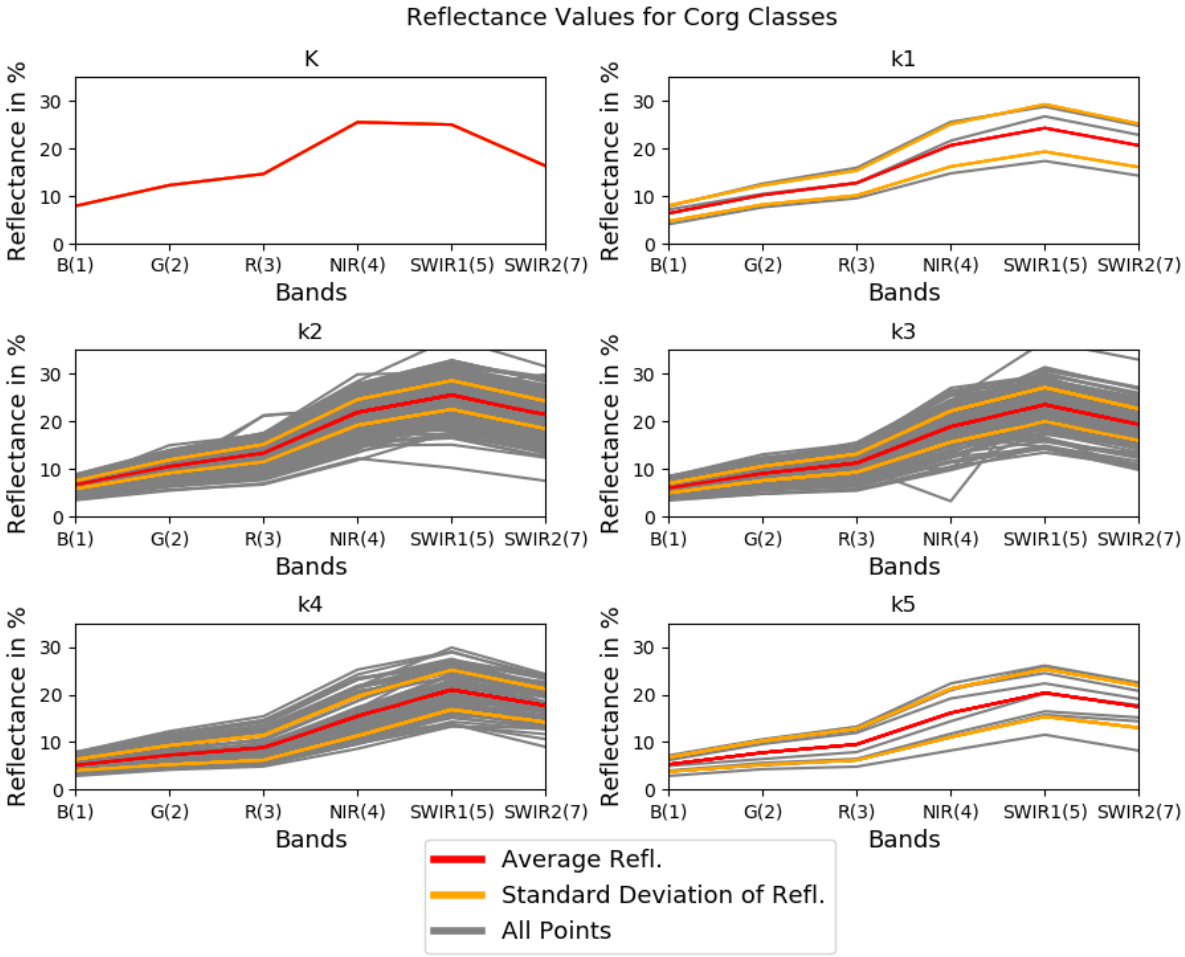

Figure 13: Reflectance Values for C_{org} Classes

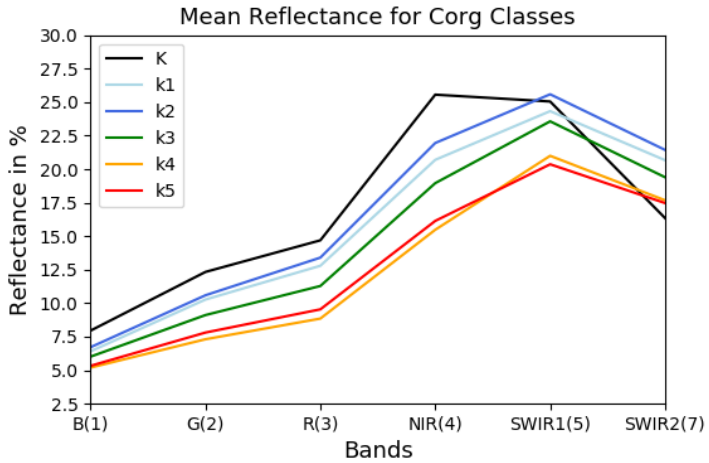

Figure 14: Mean Reflectances by C_{org} Class

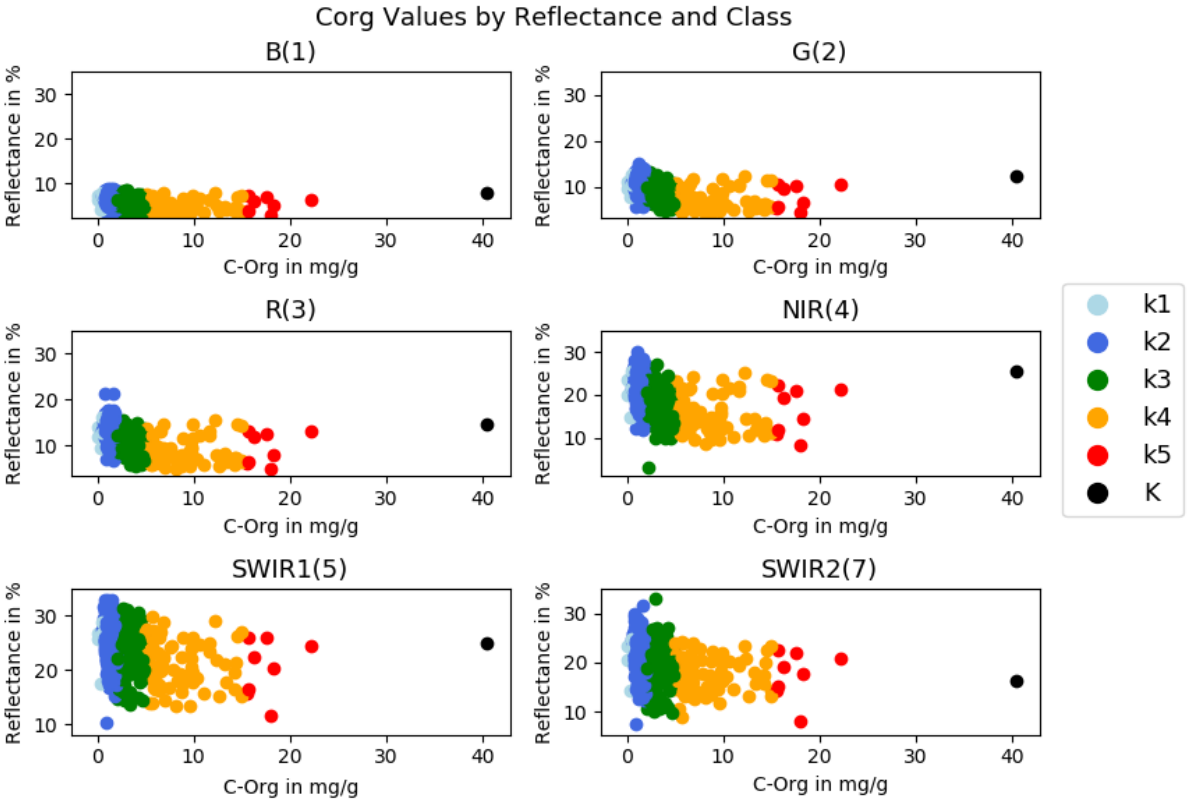

Figure 15: Scatterplot for C_{org} numerical values per Band

Table 5: Difference Matrix of all C_{org} classes for Band 4 (NIR)

Band 4 – NIR

	k1	k2	k3	k4	k5	K
k1	0	1.26	-1.81	-3.27	-0.64	0.36
k2	-1.26	0	-3.06	-4.53	-1.9	-0.9
k3	1.81	3.06	0	-1.46	1.17	2.16
k4	3.27	4.53	1.46	0	2.63	3.63
k5	0.64	1.9	-1.17	-2.63	0	1
K	-0.36	0.9	-2.16	-3.63	-1	0

The reflectance plot (**Figure 13**) shows, for each C_{org} class, the reflectance value for each point. Also included are the average and standard deviations up- and downwards. The spectral lines for each class do not have significant features exhibited in their spectral lines which allow an easy distinction between the classes.

Of importance is the high reflectance variability and scatter between the classes, as well as the high standard deviations. All lines from the input points seem to blend together without showing any considerable difference. Even though some small differences, like the lower maximum of bands 4 and 5 are visible, the high variability makes an easy distinction impossible.

Creating a plot of all mean spectra shows more interpretable results (**Figure 14**). The spectral lines are "layered", it is visible that with the exemption of class K, the lines barely intersect and are ordered from higher value class at the bottom and lower value class at the top.

Since this soil parameter also has numerical values, a scatterplot of the reflectance against the organic carbon content can be created (**Figure 15**). The resulting plot does not show differences between the classes, but instead that soil samples are spread out along the y-axis. Ideally, similar values and thus same-class pixels would cluster together at different y-values than other classes. In general, this graph shows the high intraclass variability and low interclass distinguishability.

By calculating the mean spectral difference of each class to all other classes, the visual observations can be supported (**Table 5**). A larger percentage shows a bigger difference in the mean spectrum and thus a higher distinguishability. The highest difference is between classes *k2* and *k4* with 4.53%, the next highest difference between *k2* and *k3* is 3.06%. The difference in reflectance average of the two most dominant classes *k2* and *k3* in the NIR band implies a distinguishability.

Soil Type

Plotting the reflectance values for the soil types classes shows no significant characteristics, but a high variance and thus also a high standard deviation (**Figure 16**).

Examining the mean spectra of the classes (**Figure 17**) leads to a similar conclusion. Some classes, such as *G* and *L* differ quite a lot from each other in their mean spectra while almost all other classes blend reasonably closely together, especially keeping in mind that these lines represent the averages and that the standard deviations are fairly high. Even though some lines intersect each other, it seems like in general a class mean spectrum is either higher or lower than the other classes. Without such a high variance within the dataset as observed in (**Figure 16**), the mean class spectra graph might lead to the conclusion that a correct classification is possible.

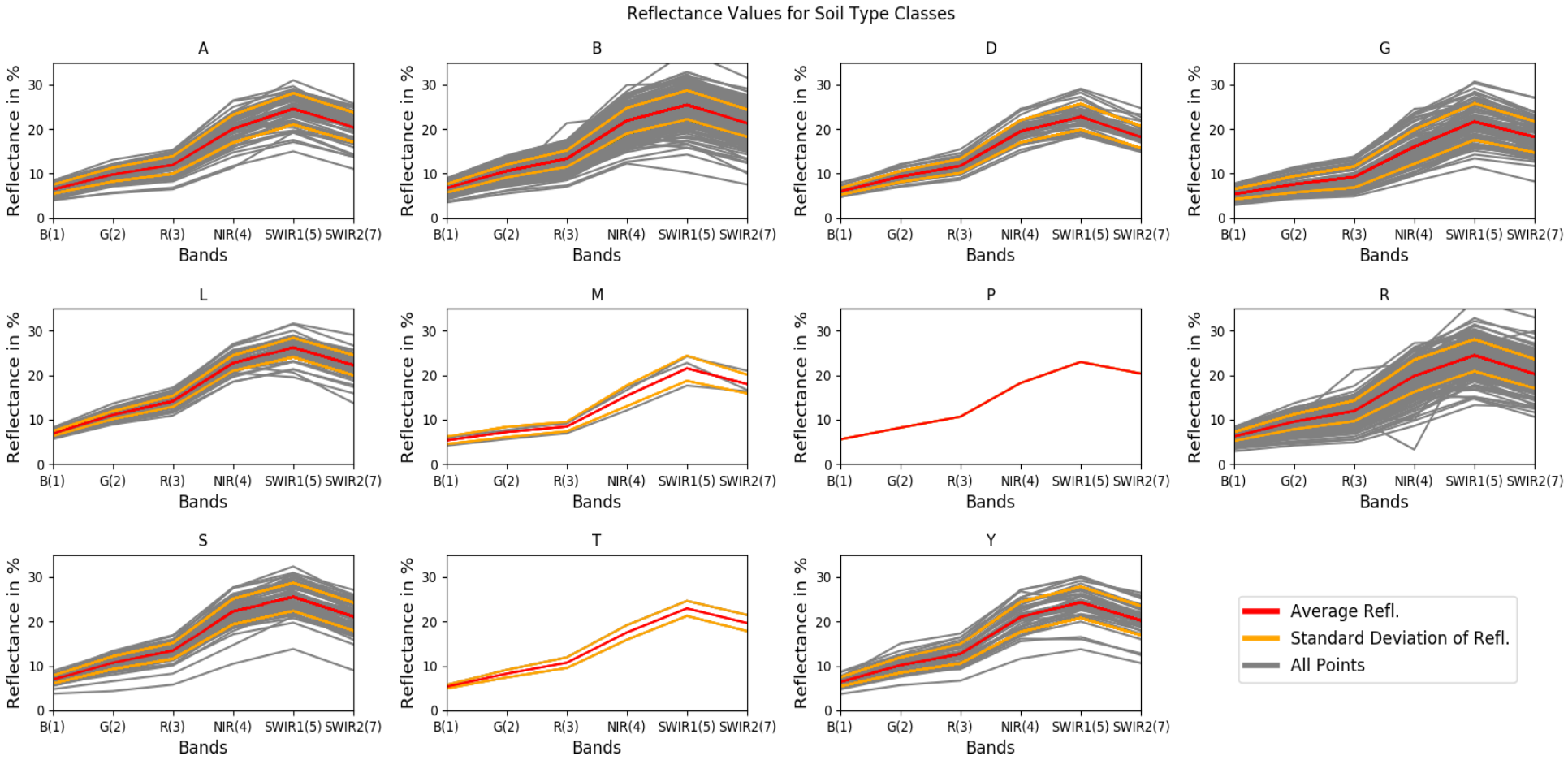

Figure 16: Reflectance Values for Soil Type Classes

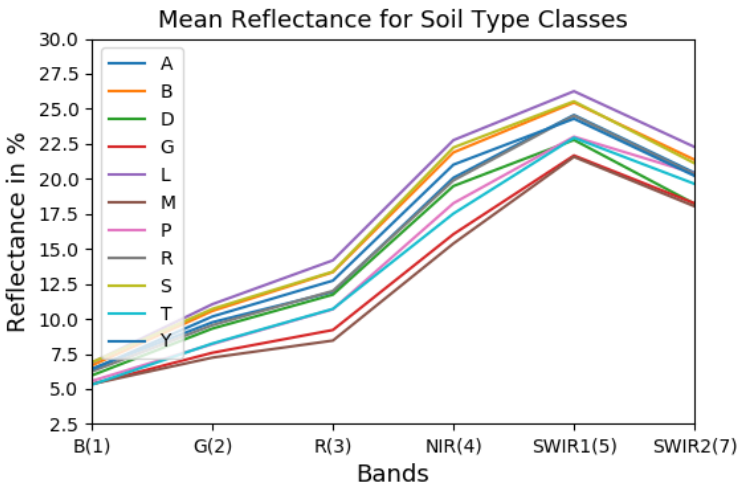

Figure 17: Mean Reflectancesc by Soil Type Classes

Table 6: Difference Matrix of all Soil Type classes for Band 4 (NIR)

Band 4 – NIR

	A	B	D	G	L	M	P	R	S	T	Y
A	0	1.81	-0.59	-4.03	2.69	-4.69	-1.82	-0.19	2.16	-2.56	0.94
B	-1.81	0	-2.4	-5.84	0.88	-6.5	-3.63	-2	0.35	-4.37	-0.87
D	0.59	2.4	0	-3.44	3.28	-4.1	-1.23	0.4	2.75	-1.97	1.53
G	4.03	5.84	3.44	0	6.71	-0.67	2.21	3.84	6.19	1.47	4.96
L	-2.69	-0.88	-3.28	-6.71	0	-7.38	-4.5	-2.88	-0.52	-5.24	-1.75
M	4.69	6.5	4.1	0.67	7.38	0	2.88	4.5	6.85	2.14	5.63
P	1.82	3.63	1.23	-2.21	4.5	-2.88	0	1.63	3.98	-0.74	2.75
R	0.19	2	-0.4	-3.84	2.88	-4.5	-1.63	0	2.35	-2.37	1.13
S	-2.16	-0.35	-2.75	-6.19	0.52	-6.85	-3.98	-2.35	0	-4.72	-1.23
T	2.56	4.37	1.97	-1.47	5.24	-2.14	0.74	2.37	4.72	0	3.49
Y	-0.94	0.87	-1.53	-4.96	1.75	-5.63	-2.75	-1.13	1.23	-3.49	0

This is supported partially by calculating the mean differences between all classes at the NIR band represented in (**Table 6**). Between some classes, like the combination of classes *L* and *G*, the mean percentual difference is reasonably high, with 7.38 percent. Other combinations are within the 3 to 4% range, with *S* <-> *G* on the higher end reaching 6.19%. Classes *M* and *L* show high percentages but since their sample size is very small, the validity is extremely limited.

Soil Texture

Lastly, the spectral distribution of the parameter soil texture is investigated.

The plotted reflectance values show similar characteristics the other soil parameters: low interclass variance but high intraclass variance (**Figure 20**(Appendix)). The illustrations for this soil parameter can be found in the appendix.

Reviewing the mean spectra on the other hand (**Figure 21**(Appendix)) reveals a slight correlation, similar to the correlation within the C_{org} parameter. The soil texture classes follow a gradient where clay (*T*) represents the smallest grain size, silt (*U*) the medium and sand (*S*) the coarsest material. Not taking into account the mixture of the three (*L*), the coarsest material has the highest reflectance while the finest one has the lowest, with the "middle" class right in between the other two lines. This indicates a correlative relationship between finer material and higher absorption, which cannot be backed by correlation coefficients since no numerical data is available for this class.

Investigating the mean difference table (**Figure 11**(Appendix)) shows lower values than the other classes, with a maximum of 1.864% for classes T and S. This can be explained by a higher variance in in the classes. Investigating all the spectral distribution results for the soil texture parameter, the distinguishability seems to be the lower than the other soil parameters.

4.1.3 Correlation Coefficients

Regarding the C_{org} content of the points, the Pearson correlation coefficients, calculated by the previously described method, for each SRC band are shown in **Table 7**.

Table 7: C_{org} and Reflectance Pearson correlation coefficients per Band

Band	Pearson R
Band 1 (B)	+0.00871
Band 2 (G)	-0.00197
Band 3 (R)	-0.01491
Band 4 (NIR)	-0.00765
Band 5 (SWIR1)	-0.00403
Band 7 (SWIR2)	-0.00846

Clearly, all coefficients of all bands are very close to zero, indicating no linear correlation between the pixel's reflectance value and the C_{org} content measured at these locations. The results show that from a linear perspective, the classes are not able to be distinguished by their reflectance value by looking at singular bands only. The statistical observations support the spectral analysis as done before.
What this method does not take into account are possible patterns across bands, thus a correct classification is harder but not impossible.

Calculating the correlation coefficients with Spearman's method gives the results represented in **Table 8**.

Table 8: C_{org} and Reflectance Spearman correlation coefficients per Band

Band	Spearman R
Band 1 (B)	-0.45939
Band 2 (G)	-0.53166
Band 3 (R)	-0.57061
Band 4 (NIR)	-0.53722
Band 5 (SWIR1)	-0.40953
Band 7 (SWIR2)	-0.43661

Spearman's correlation coefficient, which also takes the classes into account, reach a higher score of correlation. Bands 1, 5 and 7 are between -0.40 and -0.50 and thus at the upper end of the "low negative correlation" interpretation. Bands 2, 3 and 4 are even greater than -0.50 and therefore classified to have a "moderate negative correlation" (HINKLE et al.: 2002: 17-26).
From these numbers, it can be concluded that there is a small trend for higher C_{org} value pixels to have lower reflectance values, especially in bands 3, 4 and 5. This supports and gives mathematical background to the observation of the plot (**Figure**

14), where the mean spectral lines of higher C_{org} classes seem to be partially "stacked" in negative order.
This relationship is expected to be recognized by the Random Decision Tree Classifier.

4.1.4 Principal Component Analysis and Linear Discriminant Analysis

Executing the Principal Component Analysis and creating three-dimensional graph out of the principal components for each soil parameter shows no clear clusters (**Figure 18** - PCA). Ideally, the different colours, which represent the classes, should group together and form distinguishable clusters.
The C_{org} graph is quite a bit more spatially stretched than the others, but when changing the viewing angle, it becomes clear that there is also no clear cluster to be observed. A few outliers are still present. The dominance of one colour is only due to the drawing order of the points since the classes are plotted one by one, changing the viewing angle consequently also changes the drawing order.

The LDA seems to show a slightly better result, at least the C_{org} soil parameter (**Figure 18** - LDA). Even though all points are within the same cluster, one hemisphere of the cluster "sphere" is dominated by the *k2* class, but the other hemisphere is still quite mixed. Expectantly the distinguishability for this soil parameter might be slightly higher.
The other soil parameters on the other hand are indistinguishably scattered within a singular cluster, suggesting a low chance of successful classification.

Both PCA and LDA were also conducted with the following Python scaling methods:
-*sklearn.preprocessing.**MinMaxScaler***
-*sklearn.preprocessing.**normalize***
-*sklearn.preprocessing.**scale***
Datasets which span over a wide range of values can be transformed by e.g. scale magnification and/or reduction in order to reduce their order of magnitude. Since the scaling methods did not show an improvement over the non-scaled data, the original data is depicted, and the scaled datasets are discarded (**Figure 18**).

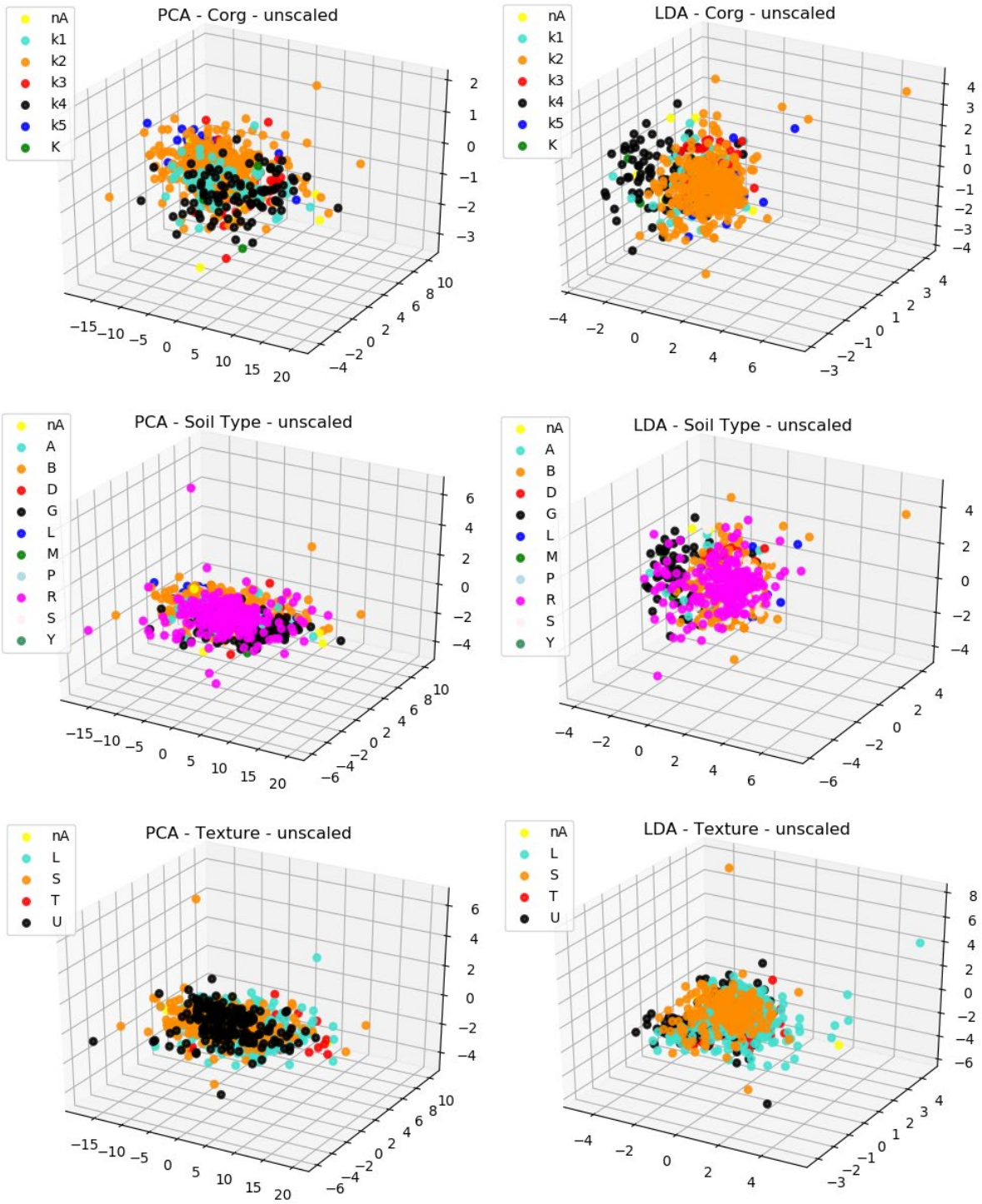

Figure 18: 3D PCA and LDA graphs for each soil parameter.
Values are dimensionless. PC1 on *x* axis, PC2 on *y* axis, PC3 on *z* axis.

4.2 Validation and Quality Control

A scientifically accurate validation method cannot be implemented. The following accuracy metrics have only limited scientific significance.

Using the internal RDF quality control as described (cf. "3.5 Validation"), the classification accuracy can be approximated (**Table 9**)

Table 9: Internal RDF Prediction Verification

Soil Parameter	Internal RDF Prediction Accuracy
C_{org} Content	74.55%
Soil Type	41.09%
Soil Texture	18.22%

As previously discovered in the statistical results, the C_{org} content shows the highest accuracy, while the other soil parameters have lower scores.
The classification results for soil type and texture are only slightly better (by about 8% for soil type, 2% for soil texture) than results achieved by guessing and can be thus interpreted as failed.
The C_{org} classifier on the other hand does show some correlation. The internal verification by itself cannot be taken as absolute truth because overfitting might have occurred. Also, since a considerable majority of points belong to the *k2* class, assigning each and every point the *k2* classification might already result in a high accuracy percentage.

After drawing a buffer around the sample locations and assigning each pixel within this buffer the same *truth value* as the sample location itself, a total of 15,385 pixels can be used to assess the quality of prediction. The accuracy percentages achieved are shown in **Table 10**.

Table 10: Matrix "quality control" accuracy

Soil Parameter	Matrix "quality control" accuracy
Soil C_{org} Content	59.59%
Soil Type	47.17%
Soil Texture	40.49%

Soil parameter maps, which are published by German authorities, where not used for validation because of the aforementioned reasons (cf. “3.5 Validation”).

4.3 RDF Classification Results – Map Products

A map which shows the predicted classes for the whole study area is produced by fitting the Random Decision Forest Classifier model with the reflectance value of all ≈63 million SRC pixels, for each soil parameter. Three maps in a *tiff* format with the pixel size of 30 m x 30 m are created, one map for each soil parameter.

Especially the C_{org} map shows regional clustering of classes, which supports a correct classification since the organic carbon content of the soil varies, but only on a scale of several hundred meters to a few kilometres. Comparing the classification result (**Figure 19**) with the SRC, it becomes clear that the pixel colour correlates with the assigned C_{org} class.

Additionally, darker areas in the SRC correlate with higher-class C_{org} content areas in the predicted map. This reaffirms the observations made previously that a higher C_{org} content correlates with a lower reflection value, thus darker SRC pixels.

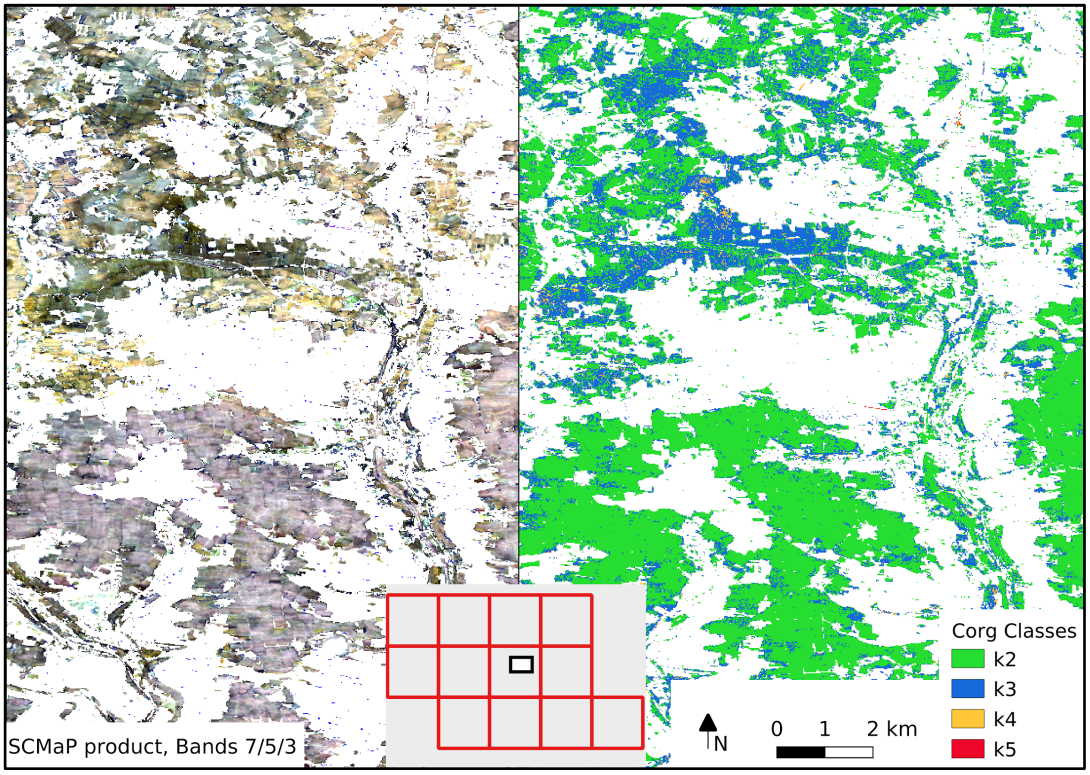

Figure 19: Side-by-side comparison of SRC and predicted C_{org} content map. SRC (SCMaP product): false color image with band 7 as R, 5 as G and 3 as B.

The soil type map shows an overwhelming presence of brown earth, scattered with different classes in between, comparable to white noise. This suggests that the regionally different soils were mostly incorrectly classified, but some very general spatial differences in the prevalence of some classes can be observed (c.f. **Figure 22(A):** Predicted Soil Type Map, excerpt).

The soil texture map presents a more or less random scattering of the classes along the SRC pixels, indicating that the classification has failed. (c.f. **Figure 23(A):** Predicted Soil Texture Map, excerpt)

5. Discussion

The following section discusses the validity of the results obtained with the methods used, as well as possible improvements.

Starting with the training dataset, the nominal number of SSLs is sufficient for RDF model building. Unfortunately, due to the dominance of single classes in all three soil parameters, other classes are severely underrepresented. This causes accuracy problems due to an absence of enough training data for certain classes, possibly resulting in mislabelling. It would be possible to increase the number of available points for all classes significantly by including the 792 points, which were taken outside of the timeframe, making the assumption that the locational parameters did not change. Also, scaling the input dataset in a way that results in similarly sized classes might result in better class separation.

During the next step, the statistical analysis, several methods to attempt to identify correlations between soil parameters and reflectance are used. Firstly, visualising the classes and their reflectance gives a first impression of the difficulty in distinguishing the classes. The first conclusions are already visible, and then consequently supported by the statistical methods.

The Pearson R values for the C_{org} parameter are very close to zero, indicating no linear correlation between the SRC pixel and the C_{org} value of the SSL. Following the correlation coefficient interpretation guide by Hinkle, all bands show “negligible correlation” (HINKLE et al.: 2002: 17-26).. The Spearman R results, which also take the classes into account, reach higher values. Bands 1, 5 and 7 are interpreted as “low

negative correlation", while bands 2,3 and 4 are observed to have "moderate negative correlation" (HINKLE et al.: 2002: 17-26). These results support the observation that the higher the C_{org} concentration, the lower the reflectance value. The moderate correlation suggests that a correct classification by the RDF is not impossible.
For the numerical values inherent in the C_{org} parameter, the correlation coefficients support a small but noticeable negative linear relationship already visible in the mean spectra, while the PCA and LDA methods were not able to differentiate the classes in any meaningful way. Given the current dataset, the best attempt to separate the classes is made and a good understanding of the dataset is achieved, but a concise statistical separation remains unsuccessful. Further attempts using more sophisticated methods of multivariate statistical analysis could possibly identify stronger correlations, such as the Partial Least Square method (NOCITA et al. 2014: 337–347). Possible improvements in expressing the differences in spectral means of the data could also be achieved by using methods such as the Spectral Angle Mapper (PARK et al. 2007: 323–333) or the Spectral Correlation Angle (DENNISON et al. 2004: 359–367)

The C_{org} classes, as well as the other soil parameter classes used in this thesis are designed for the needs of soil experts. Keeping spectral distinguishability in mind, it is possible to redefine the classes, for example join soil types with similar topsoils together or classing C_{org} content on a nominal scale.
C_{org} classes for example vary greatly in their range; analysing the change in reflectance of different C_{org} contents and adjusting the class sizes accordingly (for example into 2% steps) might produce an improved result by providing more uniform class ranges. Another possibility would be to perform a Random Decision Forest Regression instead of a classifier, using the values themselves instead of classes.
The topsoil layers of the soil type classes used in this thesis were not analysed for their spectral reflection, it is thus likely that trying to separate classes with similar reflectance features lowers the accuracy of prediction. Combining these similar classes might produce higher accuracy predictions, but simultaneously lower the informative value of the final map product. Nani et al, for example, achieved an accuracy increase of 30% by grouping classes together (NANNI et al. 2012: 103–112).
It is also problematic that the input database only takes reflectance values into account as distinguishing factors. Other components, such as the bedrock and genesis of the soil, wetness, altitude and the intensity of potential agricultural use are ignored as

factors in this classification. Using a wetness index based on a digital elevation model to exclude permanently moist soils has the potential to lower variability within the classes, leading to a more accurate prediction. Quantifying or classifying these local circumstances and consequently creating SSL groups of situational conformity will most likely greatly improve accuracy but is very time consuming and demands an extensive knowledge in soil studies and thus exceeds the limit of this thesis.

After the statistical analysis of the classes, a RDF is built in this thesis. Using the aforementioned improvement attempts such as removing spectral outliers and tweaking the RDF settings, the internal RDF accuracy was improved by 11% (C_{org}), 4% (soil type) and 1% (soil texture) from its original state to its final state. Still, the risk of overfitting is accepted by not limiting the decision tree size in order to improve distinguishability. This might not be the perfect settings for the RDF model, but through experimentation these settings proved to be optimal for the given training dataset.

The validation of the results poses a great challenge. Similar studies, especially the one conducted by (LAKSHMI et al. 2015: 1452–1460) and (NANNI et al. 2012: 103–112), validated their results by entirely mapping and sampling their study area and comparing the results obtained to the local sampling data. This requires extensive and expensive field work and is also limited to a much smaller study area. Due to the nature of this thesis and the size of the study area, this method is not viable.

The soil parameter maps are not used due to their limitations; the method of using a matrix around the SSLs bases on the uncertain assumption that the soil parameters do not change within the SSLs proximity. The internal RDF validation, usually quite an important metric, only has limited validity since the majority validation dataset consist of mostly one class. Also, this method cannot detect the effects of "overfitting". Visually examining the resulting prediction maps does give the impression that the classification has worked to a reasonable degree (at least for the C_{org} soil parameter), but this method does not produce a quantifiable accuracy score. Consequently, several accuracy metrics, individually only of limited significance, taken together can give a vague impression of overall accuracy. The internal RDF validation, the method of creating a matrix "quality control" method and manually

examining the resulting maps and looking for areas of conformity taken together can give only an accuracy estimation.
Even though promising results for the C_{org} content prediction are visible, these observations cannot be sufficiently backed by statistical information because an accurate and scientific validation method is unavailable.

6. Conclusion

This thesis analysed the spectral properties of a Landsat soil composite by using the Spearman and Pearson correlation coefficients as well as a PCA and LDA. This was done in order to investigate if the reflectance values correlate with the soil parameters of organic carbon content, soil type and soil texture. Based on these spectral properties, a prediction on all ca. 63 million SRC pixels was executed using an RDF.

Regarding the possible correlation between SRC reflectance values and soil parameters, a considerable correlation for the C_{org} parameter can be observed and consequently statistically proven, showing that the organic carbon content does correlate with the SRC's spectral information. Examining the other two soil parameters, no correlation could be established using the aforementioned methods.
As to the prediction validation, results of uncertain accuracy were achieved. A concise verification method, showing clear statistical results of the prediction accuracy could not be implemented. Nevertheless, a considerable prediction accuracy for the C_{org} soil parameter can be visually observed in the resulting maps and is backed by the internal RDF validation as well as the implemented "quality control" measures. The other two soil parameter predictions (soil type and soil texture) can be considered failed.
The RDF was able to show its potential and produce an output map of reasonable quality for the C_{org} soil parameter, but not for the other parameters. The ability of the RDF to build a model upon the very small correlation within the dataset shows its efficiency in satellite-based soil classification applications. Given an improved quality of the training data, the RDF is a viable method of classifying SRC pixels.

Ultimately it can be stated that a large amount of information was extracted from the input dataset in its current state, but improvements are most definitely possible. Statistical applications have been extensively used to gain information about the dataset and connect the spectral information to the soil properties, a continuous effort to improve the dataset and its statistical analysis are promising.
The resulting maps possess a limited validity since they are built upon an imperfect dataset and the aforementioned assumptions, with the C_{org} soil parameter prediction promising the highest accuracy and validity. Expanding the statistical research as well as using soil expertise to investigate the details of soil reflectance under different circumstances, leading to an improved quality of the input data as well as the prediction, will most likely result in enhanced classification and prediction accuracies.

A growing population and its need for nutrition, the ongoing climate change as well as the changing ecosystems pose great challenges to scientists and politicians around the world. Observing and analysing the soils of this planet, which play an important role in these challenges, provides decision-makers with valuable information. Continuing the effort of extracting soil information from multispectral images and composites promises to improve the availability of large-scale soil information.

Bibliography

Adhikari, K. & Hartemink, A. E. 2016. Linking soils to ecosystem services — A global review. Geoderma. Vol 262, 101–111.

AD-HOC AG Boden 2005. Bodenkundliche Kartieranleitung. Bundesanstalt für Geowissenschaften und Rohstoffe. Stuttgart: 107; 109; ;132-133;173-175.

Barnes, E. & Baker, M. 2000. Multispectral data for mapping soil texture: Possibilities and limitations. Applied Engineering in Agriculture. 731–741.

Bayer, A., Bachmann, M., Rogge, D., Müller, A., Kaufmann, H. 2016. Combining Field and Imaging Spectroscopy to Map Soil Organic Carbon in a Semiarid Environment. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. Volume 9 Issue 9, 3997–4010.

Ben-Dor, E. & ET. AL. 2008. Imaging spectrometry for soil applications. Advances in Agronomy. Volume 97, 321–392.

Borkowf, C. B. 2002. Computing the nonnull asymptoticvariance and the asymptotic relative efficiency of Spearman's rank correlation. Computational Statistics & Data Analysis. Volume 39 Issue 3, 271–286.

Breiman, L. 2001. Random Forests. Machine Learning. Vol 45, 5–32.

Burnham, K. P. & Anderson, D. R. 2002. Model Selection and multimodel Inference: A practical information-theoretic Approach. Springer. New York: 1-9.

Büyüköztürk, Ş. & Çokluk-Bökeoğlu, Ö. 2008. Discriminant function analysis: Concept and application. Eurasian Journal of Educational Research. Volume 33, 73–92.

Candiago, S., Remondino, F., Giglio, M., Dubbini, M., Gattelli, M. 2015. Evaluating Multispectral Images and Vegetation Indices for Precision Farming Applications from UAV Images. Remote Sensing of Environment. Volume 7, 4026–4047.

Chen, D. & Chen, H.W. 2015. Using the Köppen classification to quantify climate variation and change: an example for 1901-2010. Environmental Development 6: 69-79

Chen, P. Y. & Popovich, P. M. 1981. Quantitative Applications in the Social Sciences. Social Science Information Studies. Volume 1 Issue 2, 135.

Croux, C. & Dehon, C. 2010. Influence functions of the Spearman and Kendall correlation measures. Statistical Methods and Applications. Vol 19, 497–515.

Cutler, A., Cutler, D. R., Stevens, J. R. 2012. Random Forests. Ensemble Machine Learning. 157–175.

Daily, G. C., Matson, P. A., Vitousek, P. M. 1997. Ecosystem services supplied by soil. Nature's Services: Societal Dependence on Natural Ecosystems. 113–132.

De Silva, C. C. 2017. Principal component analysis (PCA) as a statistical tool for identifying key indicators of nuclear power plant cable insulation degradation. Dissertation - Iowa State University. 15–17.

Dennison, P. E., Halligan, K. Q., Roberts, D. A. 2004. A comparison of error metrics and constraints for multiple endmember spectral mixture analysis and spectral angle mapper. Remote Sensing of Environment,. Volume 93 Issue 3, 359–367.

Dominati, E., Patterson, M., Mackay, A. 2010. A framework for classifying and quantifying the natural capital and ecosystem services of soils. Ecological Economics. Vol 69, 1858–1868.

Fisher, R. A. 1936. The use of multiple measures in taxonomic problems. Annals of Eugenics. Volume 7, 179–188.

Gislason, P. O., Benediktsson, J. A., Sveinsson, J. R. 2006. Random Forests for land cover classification. Pattern Recognition Letters. 4, 294–300.

Hanks, R. & Bowers, S. 1962. Numerical Solution of the Moisture Flow Equation for Infiltration into Layered Soils. 1. Soil Science Society of America Journal. Volume 26, 530.

Hansen, M. C. & Loveland, T. R. 2012. A review of large area monitoring of land cover change using Landsat data. Remote Sensing of Environment. 122, 66–74.

Hastie, Trevor, Tibshirani, Robert,Friedman, J. H. 2009. The elements of statistical learning. Springer. New York: 9-41; 101-137.

Hinkle, D. E., Wiersma W.,Jurs S.G. 2002. Applied Statistics for the Behavioral Sciences: 17-26.

Ho, T. K. 1995. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, QC: 278–282.

Hotelling, H. 1933. Analysis of a complex of statistical attributes into principal components. Journal of Educational Psychology. Volume 24, 417–441 and 498-520.

Ishwaran, H. 2007. Variable Importance in binary Regression Trees and Forests. Electronic Journal of Statistics. Volume 1, 519–537.

IUSS Working Group WRB. World reference base for soil resources. A framework for international classification, correlation and communication. Food and Agriculture Organization of the United Nations. 2006. Rome.

James, G. 2013. An introduction to Statistical Learning with Applications in R. Springer. New York, NY: 127-173.

Jarmer, T., Uedelhoven, T., Hill, J. 2003. Möglichkeiten zur Ableitung boden-bezogener Größen aus multi-und hyperspektralen Fernerkundungsdaten. Photogrammetrie – Fernerkundung – Geoinformation. Volume 2, 115–123.

Jeffers, J.N.R. 1964. Two Case Studies in the Application of Principal Component Analysis. Journal of the Royal Statistical Society (Applied Statistics). Volume 16 Issue 3, 225–236.

Jolliffe, I. T. 2002. Principal component analysis. Springer. New York: 2-4.

Kaufmann-Boll, C. & Rinklebe, J. 2011. Auswertzung der Veränderung des Bodenzustands für Boden-Dauerbeobachtungsflächen (BDF) und Validierung räumlicher Trends unter Einbeziehung anderer Messnetze. Teil A: Methodencode und Umgang mit Verfahrensweiswechseln. UMWELTFORSCHUNGSPLAN DES BUNDESMINISTERIUMS FÜR UMWELT, NATURSCHUTZ UND REAKTORSICHERHEIT. 4.

Keshava, N. 2003. A survey of spectral unmixing algorithms. Vol 14, 55–78.

Kriegler, F. J., Malila, W. A., Nalepka, R. F.,Richardson, W. 1969. Preprocessing transformations and their effects on multispectral recognition. Proceedings of the Sixth International Symposium on Remote Sensing of Environment.

Lakshmi, V., James, J., Soundariya, S., Vishalini, T., Pandian, K. 2015. A Comparison of Soil Texture Distribution and Soil Moisture Mapping of Chennai Coast using Landsat ETM+ and IKONOS Data. Aquatic Procedia. Vol 4, 1452–1460.

Millennium Ecosystem Assessment 2005. The Millennium Ecosystem Assessment: Ecosystems and human well-being. The Island Press. Washington D.C.

Millman, K. Jarrod & Aivazis, Michael 2011. Python for Scientists and Engineers. Computing in Science & Engineering. 2, 9–12.

Mulder, V. L., Bruin, S. de, Schaepman, M. E., Mayr, T. R. 2011. The use of remote sensing in soil and terrain mapping—A review. Geoderma. Volume 162, 1–19.

Nachtergaele, F., van Velthuizen, H., Verelst, L., Batjes, N., Dijkshoorn, K. et al. Harmonized World Soil Database – Version 1.1. Food and Agriculture Organization of the United Nations. 2009: 33-37.

Nanni, M. R., Demattê, J.A.M., Chicati, M. L., Fiorio, P. R., Cézar, E. et al. 2012. Soil surface spectral data from Landsat imagery for soil class discrimination. Acta Scientiarum - Agronomy. 103–112.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B. et al. 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biology and Biochemistry. Volume 68, 337–347.

Ochsner, T. E., Cosh, M. H., Cuenca, R. H., Dorigo, W. A., Draper, C. S. et al. 2013. State of the Art in Large-Scale Soil Moisture Monitoring. Soil Science Society of America Journal. Volume 77 Issue 6, 1888–1919.

Omuto, C., Nachtergaele, F.,Rojas, R. State of the Art Report on Global and Regional Soil Information: Where Are We? Where to Go? Food and Agriculture Organization of the United Nations. 2013. Rome: 81.

Park, B., Windham, W. R., Lawrence, K. C., Smith, D. P. 2007. Contaminant Classification of Poultry Hyperspectral Imagery using a Spectral Angle Mapper Algorithm. Biosystems Engineering. Volume 96 Issue 3, 323–333.

Park, Kun Il 2018. Fundamentals of Probability and Stochastic Processes with Applications to Communications - Applications. Springer International Publishing: 213–265.

Pearson, K. 1896. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences. 253–318.

Pearson, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. Volume 2 Issue 11, 559-572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, A., Thirion, B. et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12, 2825–2830.

Richards, J. A. 2013. Feature Reduction. Remote Sensing Digital Image Analysis. 343–380.

Richter, R., Schläpfer, D., Müller, A. 2006. An automatic atmospheric correction algorithm for visible/NIR imagery. International Journal of Remote Sensing. Vol 27, 2077–2085.

Rogge, D., Bauer, A., Zeidlera, J., Mueller, A., Esch, T. et al. 2018. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984-2014). Remote Sensing of Environment. 205, 1–17.

Rouse, J. W., Haas, R. H., Schell, J. A., Deering, D. W. 1974. Monitoring vegetation systems in the Great Plains with ERTS. International Journal of Remote Sensing. Vol 27, 309–317.

Rowe, P. 2015. Ordinal and non-normally distributed data. Essential Statistics for Pharmaceutical Sciences. 311–335.

Rubin, J. & Steinhardt, R. 1963. Soil water relations during rainfall infiltration. Soil Science Society of America Journal. Volume 27, 247–521.

Schilli, C., Rinklebe, J., Lischeid, G., Kaufmann-Boll, C.,Lazar, S. Auswertung der Veränderungen des Bodenzustands für Boden-Dauerbeobachtungsflächen (BDF). Teil B: Datenauswertung und Weiterentwicklung des Monitorings. Umweltbundesamt. 2011: 16.

Scikitlearn User Guide, 2019. , release 0.21.3. 2019: 2331.

Song, Y. Y. & Lu, Y. 2015. Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry. Volume 27 Number 2, 130–135.

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B. et al. 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma. Volume 144 Issues 1-2, 395–404.

Szabolcs, I. 1994. Soil Resilience and Sustainable Land Use. CAB International. Wallingford, UK: 33–39.

Tahmasebi, E., Hezarkhani, A., Mortazavi, M. 2010. Application of Discriminant Analysis for Alteration Separation; Sungun Copper Deposit, East Azerbaijan, Iran. Australian Journal of Basic and Applied Sciences. Vol 6 Issue 4, 564–576.

Tóth, G., Jones, A.,Montanarella, L. LUCAS Topsoil survey methodology, data and results. European Commision. 2013: 3-6.

WANG X. & Paliwal, K. K. 2003. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recognition. Volume 36, 2429–2439.

Weaver, K. F., Morales, V. C., Dunn, S. L., Godde, K.,Weaver, P. F. 2018. An introduction to Statistical Analysis in Research: With Applications in the Biological and Life Sciences. John Wiley & Sons: 435–448.

Wiesmeier, M., Schada, P., Lützowa, M. von, Poeplau, C., Spörlein, P. et al. 2014. Quantification of functional soil organic carbon pools for major soil units and land uses in southeast Germany (Bavaria). Agriculture, Ecosystems and Environment. Vol 185, 208–220.

Wold, S., Esbensen, K., GELADI P. 1987. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems. Volume 2, 37–52.

Woodcock, C. E., ALLEN, A. A., ANDERSON, M., Belward, A. S., Bindschadler, R., Cohen, W. B. 2008. Free access to Landsat Imagery. Science. Volume 320, 1011.

Young, I. & Crawford, J. 2004. Interactions and Self-Organisation in the Soil-Microbe Complex. Science. Volume 304, 113–132.

Zhang, Q., Xiao, X., Braswell, B., Linder, E., Baret, F. et al. 2005. Estimating light absorption by chlorophyll, leaf and canopy in a deciduous broadleaf forest using MODIS data and a radiative transfer model. Remote Sensing of Environment. 99, 357–371.

Zhu, Z. & Woodcock, C. E. 2005. Continuous change detection and classification of land cover using all available Landsat data. Remote Sensing of Environment. 99, 357–371.

Appendix

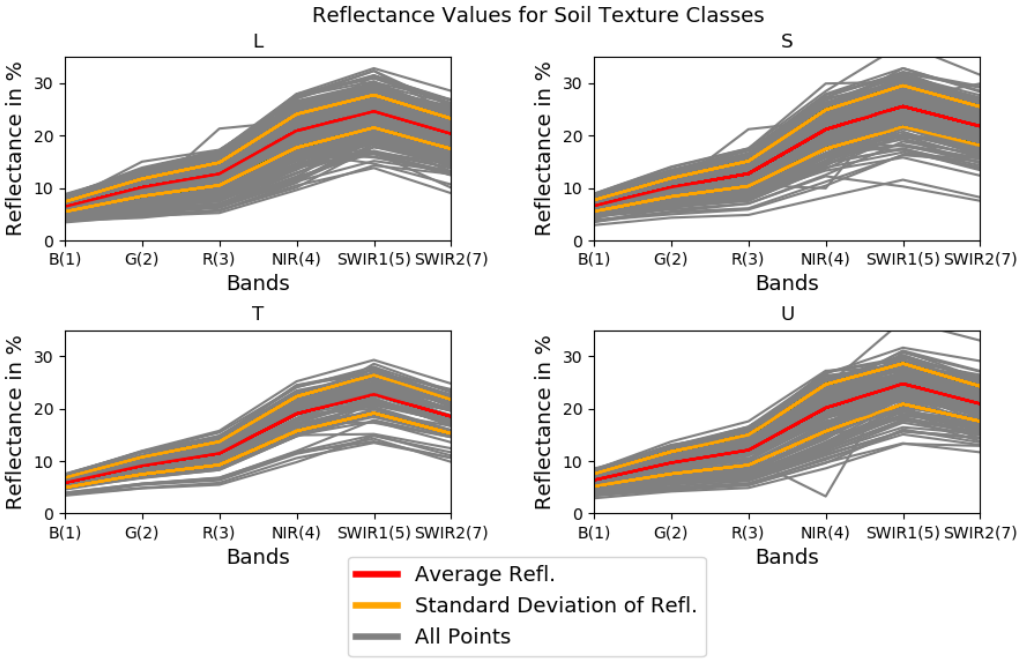

Figure 20(A): Reflectance Values for Soil Texture Classes

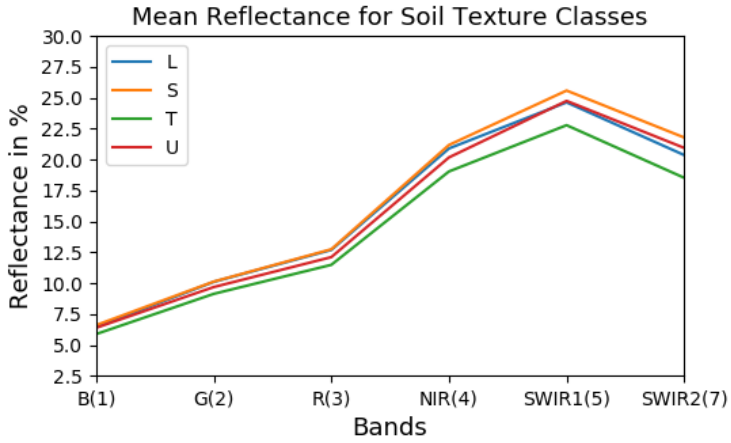

Figure 21(A): Mean Reflectance for Soil Texture Classes

Table 11(A): Difference Matrix of all Soil Texture Classes for Band 5 (SWIR1)

Band 5 – SWIR1

	S	U	T	L
S	0	-0.84	-2.8	-0.95
U	0.84	0	-1.96	-0.11
T	2.8	1.96	0	1.85
L	0.95	0.11	-1.85	0

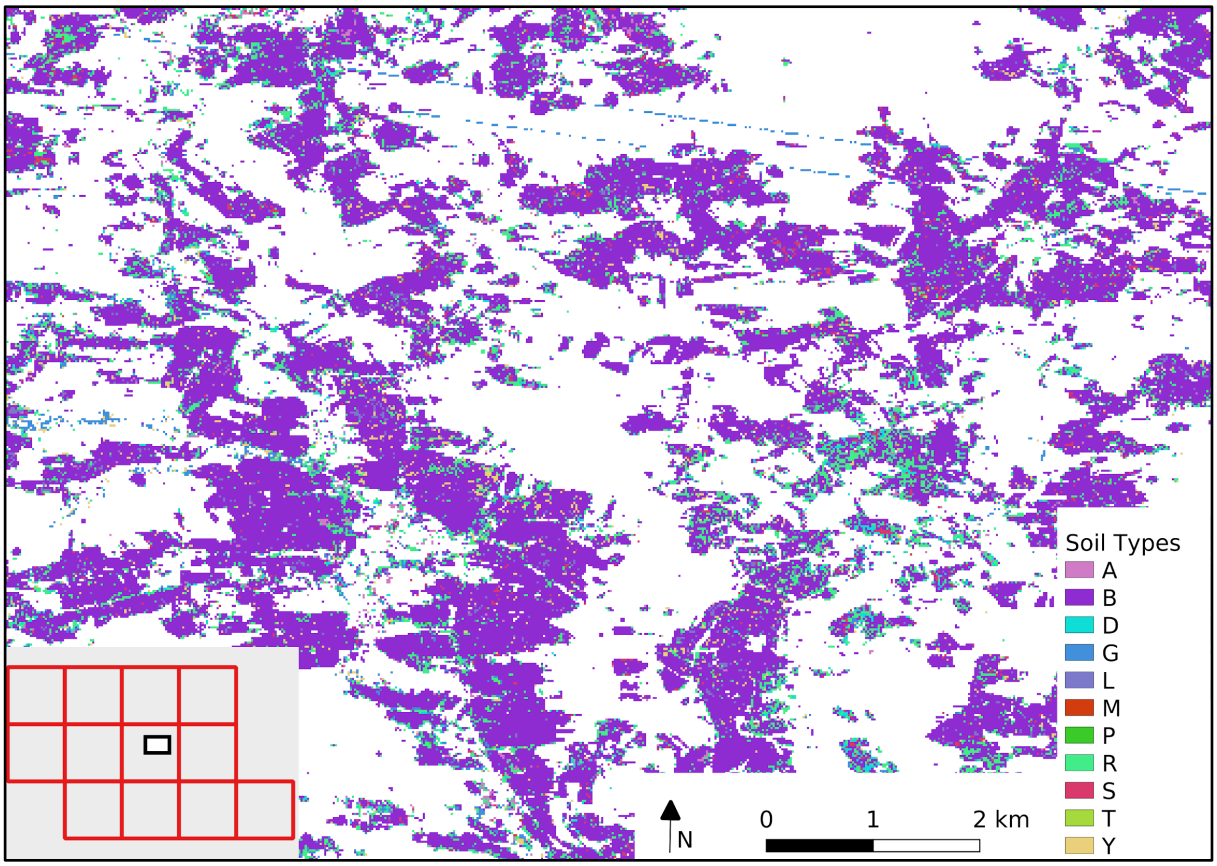

Figure 22(A): Predicted Soil Type Map, excerpt

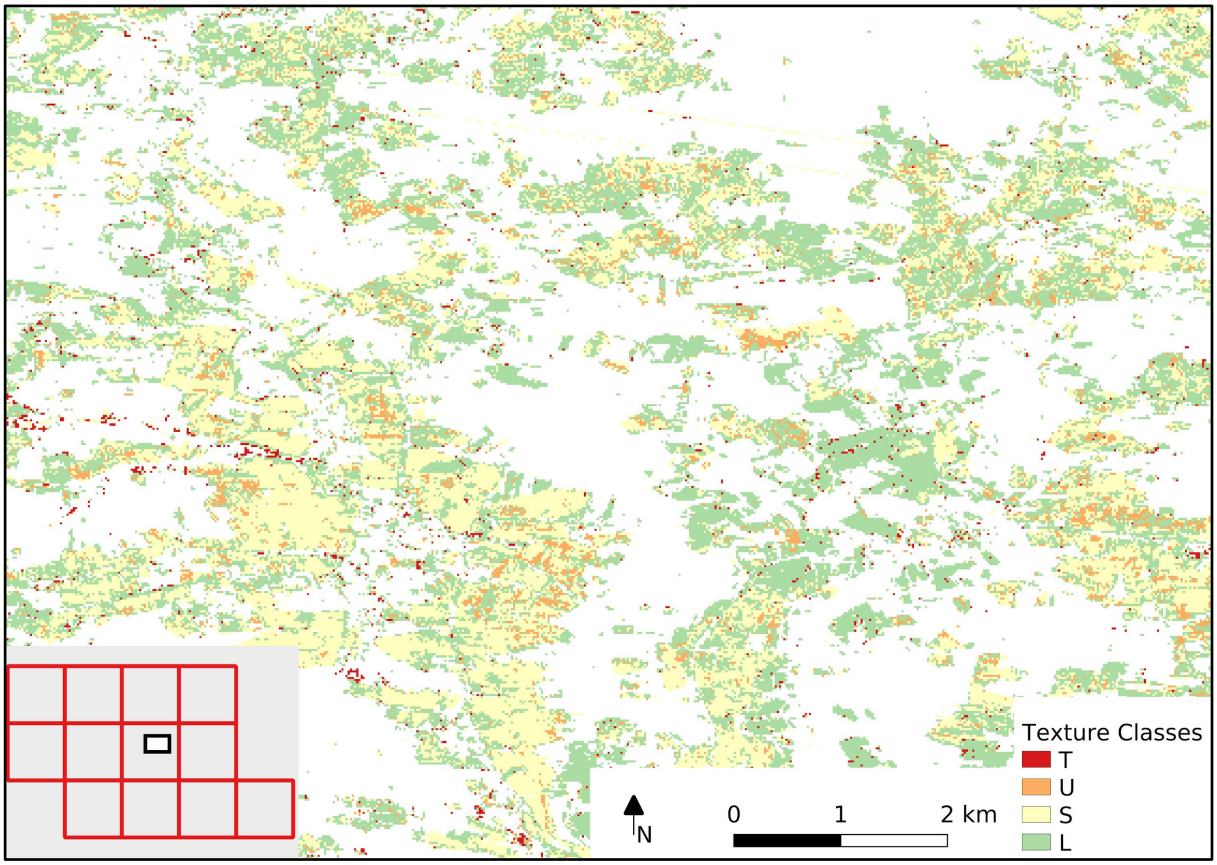

Figure 23(A): Predicted Soil Texture Map, excerpt

Eidesstaatliche Erklärung

UNIVERSITÄT ZU KÖLN
Albertus-Magnus-Platz
50923 Köln

Eidesstattliche Erklärung

Hiermit erkläre ich,

Simon Donike,
geboren am 05.04.1995,
Matrikelnummer 5976642

an Eides statt, dass die vorliegende, an diese Erklärung angefügte Arbeit selbständig und ohne jede unerlaubte Hilfe angefertigt wurde, dass sie noch keiner anderen Stelle zur Prüfung vorgelegen hat und dass sie weder ganz noch im Auszug veröffentlicht worden ist. Die Stellen der Arbeit – einschließlich Tabellen, Karten, Abbildungen etc. – die anderen Werken und Quellen (auch Internetquellen) dem Wortlaut oder dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall als Entlehnung mit exakter Quellenangabe kenntlich gemacht.

München, den 22.11.2019

_____

Simon Donike