Master Thesis

University of Natural Resources and Life Sciences, Vienna

Institute of Surveying, Remote Sensing and Land Information

November 2015

SOIL MAPPING USING REMOTE SENSING TECHNIQUES

Author: Manuel Scherrer, BSc
Supervisor: Univ.Prof. Dr.rer.nat. Clement Atzberger
Co-Supervisor: Dr. Francesco Vuolo

CONTENTS

1. Introduction 1

1.1. Soils and Soil Mapping 1

1.2. Remote Sensing 3

1.3. Objective 5

2. Materials and Methods 6

2.1. Site Description 7

2.2. Data Acquisition & Preparation 12

2.3. Reflectance based indicators 14

2.4. Unsupervised classification 19

2.5. Post-processing and Analysis 22

2.5.1. Internal Cluster Analysis 23

2.5.2. External Cluster Analysis 23

3. Results and Analysis 27

3.1. Satellite data pre-processing 27

3.2. Input features for classification 33

3.3. Maps and Cluster Evaluation 39

3.3.1. "Band medians - dataset" 40

3.3.2. "Fitted polynomial function - dataset" 43

3.3.3. "Soil line - dataset" 46

3.3.4. "Fitted polynomial function - dataset" - automatic / manual merge 51

4. Discussion and Conclusions 62

4.1. Outlook 64

4.2. Lessons learned 65

Bibliography 66

ABSTRACT

Traditional soil surveys are conducted via in-situ measurements and provide among other uses information for soil maps. Satellite images and remote sensing data can help to enhance these maps by addressing different soil attributes and providing better spatial resolution. Their temporal and spatial availability enables users to assess any remote region, without prior or deeper knowledge about it. This thesis aims to evaluate the feasibility of soil mapping with free earth observation satellite data using open source software. It is conducted in the Marchfeld region in Lower Austria which is characterized by high soil variability. The author generates three bare-soil-reflectance based indicators (median over time of spectral reflectance, coefficients of a polynomial function fitted on the spectral reflectance of the pixel, intercept and slope of the soil line) on the basis of 16 Landsat-7 images. They serve as input features for an unsupervised classification, using the "self-organizing map" (SOM) and the "kMeans" algorithm. The resulting maps are compared analytically and visually with an existing soil map of the area of interest. It is shown that the maps produced are more dependent on the input feature used, than on the algorithm applied, with one reflectance based indicator - the coefficients of a polynomial function fitted on the spectral reflectance of the pixel - outperforming the other indicators. Coarse patterns and shapes can be delineated from the results and a better spatial resolution compared to the existing soil map is provided. For an improvement of the results, additional soil specific information - like in-situ measurements - can be used as an input feature or for a supervised classification.

Keywords: digital soil mapping, remote sensing, unsupervised classification, Marchfeld, Landsat, R

1. INTRODUCTION

Earth's surface is under constant change in land use and land cover. Due to demographic pressure and climate change these environmental changes are likely to continue and have impact on the earth's pedosphere (cf. Mulder et al., 2011). Its importance for food security and for being under stress - mainly caused by human activities - has led to risen awareness from society in the last years - manifesting for example in the "International Year of Soils 2015" (cf. “The International Union of Soil Sciences - IUSS,” n.d.). Soil sealing and agricultural practices are the main reasons for the loss of fertile soil. Environmental changes but also anthropological impacts have great influence on soils as they not only determine its physical and chemical composition and condition but also its availability for different uses. The first soil maps were published in the middle of the 19th century (cf. Hartemink et al., 2013) and since then soil maps provide information used for policy-making, land resource management and for monitoring the environmental impact of development.

The following chapters will give a short historic review on soil science in general and introduce theoretical background to soils, soil mapping and remote sensing.

1.1. Soils and Soil Mapping

The Russian Vasily Dokuchaev was one of the first prominent soil scientists in the 19th century and he is considered as the father of pedology. His main effort was classifying different soil categories and visualizing them in a map. He established pedology and soil science as an independent but interdisciplinary discipline within science, having strong relations to chemistry, physics, geology and microbiology for example. The first journal related to soil science was named "Pochvovedenie" (Russian for "soil science") and appeared in 1899 (cf. Hartemink et al., 2001, p. 218).

In 1941, Hans Jenny - a Swiss scientist - introduced a new conceptual equation and paradigm which is still followed today:

$$S = f\,(cl, o, r, p, t, \dots)$$

It states: soil (S) is a function of climate (cl), organisms (o), relief or topography (r), parent material (p), time (t), and unspecified factors (...) including human activities (cf. Jenny, 1941). This concept was the first that treated the formation of soil as an aggregate of many interrelated physical, chemical and biological processes. The conceptual shift in those days could also be observed in the new focus on individual detailed soil attributes and grain-to-grain relationships instead of gross attributes of the whole soil. Nowadays the three main topics (according to the number of published papers) in soil science are firstly: "soil genesis, classification and mapping"; secondly: "soil chemistry" and thirdly: "soil physics"(cf. Hartemink et al., 2001, p. 235). It shows that the main emphasis is still in soil classification and mapping.

Soil forms part of the basic and crucial elements of life. It is a finite resource, meaning that its loss and degradation cannot be recovered in a human lifespan. As one of the key life

support systems it is responsible for the performance of major ecosystems services and functions such as:

- climate regulation;
- flood regulation;
- nutrient cycling;
- water purification and soil contaminant reduction;
- carbon sequestration;
- habitat for organisms;
- source of raw materials;
- archive of geological and archeological heritage;
- biomass production in agriculture and forestry and
- foundation for human infrastructure and activities (cf. Stolbovoy et al., 2008; FAO, 2015).

Soils are very much under pressure from different uses such as agriculture, forestry and through urbanization. As well the intensification of these uses is contributing to the loss of fertile soil. The FAO (Food and Agriculture Organization of the United Nations) evaluated that "33 percent of land is moderately to highly degraded due to the erosion, compaction, acidification and chemical pollution of soils" and further that the "projected growth in global population (to exceed 9 billion by 2050) are estimated to result in a 50 percent increase in demand for food, feed and fibre by 2050" (FAO, 2015, p. 1). FAO also indicates that there is very little space left for agricultural expansion and emphasize that "soils need to be recognized and valued for their productive capacities as well as their contribution to food security and the maintenance of key ecosystem services". As shown, the responsible institution does not give a very promising insight. Regarding these facts, it is surprising that until now soils were not considered as a highly valuable resource and were often overlooked (cf. Miehlich, 2009).

In contrast to "climate", "air" and "water", "soil" has not been part of society's focus in the past and it is just in recent years - due to current efforts - gaining importance. According to Miehlich (2009) this little awareness might be caused by sociological aspects:

- **Soils are "invisible":** in urban areas soils are mostly sealed and elsewhere they are covered by vegetation. The only exception are crops, where bare soil is visible after harvesting.
- **Soils are "uniform / equal":** their qualitative attributes and monetary value cannot be retrieved via visual or sensual examination but has to be analyzed scientifically. This makes it hard to estimate its value and importance as they seem like a constant or a continuum, making it difficult to separate different soil classes visually.
- **Soils develop very slowly:** soils change very little in a human life span - they seem static and constant to us.
- **Soils are not an "eye-catcher":** other environmental issues can be shown very well using emotional aspects. "Dirty" soils cannot be illustrated as a cute, little thing which moves as you wave. As a result soil issues are very seldom in our media today.
- **The relation between human activities and soils is very complex:** our lifestyle and demand for products has indirect relation to the impact we have on our soils. We often do not know about the provenience of our products and the role they

play for soil composition and soil conditions. This impact is indirect and not very obvious, leading to little awareness about the role of our activities.

The European Commission for instance recognized this gap of soil consciousness and initiated various programs. For example, in 2006 it provided a "Thematic Strategy for Soil Protection" (TSSP) which includes an important focal point for soil research and awareness raising. Applying different methods and conducting different activities it tries to reach stakeholder groups and policy-makers. "Unfortunately, the TSSP has not yet been followed up with a legally binding Framework Directive mainly because of political barriers" (Bouma et al., 2012, p. 552). Nevertheless there is pinned big hope in soil atlases and soil maps presenting soil information to (partly) tackle the problem of unawareness. Soil monitoring and soil mapping are regarded as key factors to provide information on soils, as they are able to illustrate soil conditions and their spatial and temporal variations. They play a big role for soil protection and conservation, since the choice of the applied management techniques and preservation methods relies on the provided information. Therefore attention is being paid on its relevance to major environmental problems and to local socio-economic conditions, rather than focusing on soil information as such - as in traditional soil survey reports (cf. Bouma et al., 2012; European Commission - Joint Research Center, 2014).

Soil mapping nowadays is working quite different than in its beginnings (cf. Bockheim et al., 2005). It still shows the diversity of soil types and its properties, but there are more attributes that can be assessed, so the maps are richer in content and have higher spatial resolution. This was made possible by better monitoring instruments such as earth observation satellite data and through advances in computer technology and processing of large data sets. In-situ measurements play a big role in successful soil mapping. Wulf et al. (2015) show in their work that they provide high accuracy and feasibility to obtain soil specific information and therefore outperform satellites in some uses. Their main disadvantages lie in the cost-effectiveness and their spatial availability. With digital development, information about soils is stored in rasters. They can be visualized in GIS (geographic information systems) like the desktop-based open source software "QGIS" (cf. QGIS, 2015) or proprietary products like "ArcGIS" (cf. ArcGIS, 2015) and "ERDAS IMAGINE" (cf. ERDAS IMAGINE, 2015). This software enables to combine data from different sources efficiently and perform spatial analysis (geostatistics, modelling,...) on the data set. Nowadays most of the mapping process is done remotely, instead of being at the area of interest. This is made possible through data either provided by in-situ measurements or by remote sensing techniques that collect data from above - from air or space.

1.2. Remote Sensing

Remote sensing is the technique of data acquisition without making physical contact with the object. It can be divided into passive and active remote sensing. Passive remote sensing uses solar radiation that is reflected or emitted by the land surface. Active remote sensing on the other hand is "actively" emitting energy with wavelengths ranging between 0.8 cm and 100 cm to detect the objects backscattered radiation (cf. Schowengerdt, 2007). Furthermore it can be distinguished between airborne and spaceborne remote sensing, differing in the platform which carries the instruments. As with many new technologies, the military sector was the driving force. First, remote sensing was mainly used to fulfill

national security missions, like using aerial photography during World War I. Later the first civil applications, such as mapping land cover and photogrammetry were introduced. In the 1950s multispectral sensors have been developed enabling further non-military applications. With the threat of cold war remote sensing was brought to new levels: The first satellites were launched and in 1960 the first photograph of the earth's surface from space was taken by a satellite of the US-Corona program. The most prominent and longest running program for civil uses of remote sensing is called Landsat. It was initiated in the U.S. in 1966 under the name "Earth Resources Technology Satellites Program" and renamed in 1975 to "Landsat" (cf. Estes, 2005). Since then eight satellites have been launched - latest being Landsat 8 in February 2013 - providing the most complete and continuous time series of images of the earth's surface for free. During the years there has been improvement in temporal, spatial and spectral resolution with the result that Landsat 8 has a spatial resolution of 30 meters in 8 spectral bands - four bands covering the visible and four bands the infrared spectrum. Landsat 8 also features a panchromatic band and two thermal infrared bands and the revisit time is 16 days. The availability of free and open source data is going to improve in the coming months thanks to the Copernicus program, an initiative of the European Commission in cooperation with ESA (European Space Agency) to provide a set of satellites "Sentinels" for monitoring the status of the land surface. With the launch of "Sentinel 1A" (radar domain) in 2014, "Sentinel 2A" (optical domain) in 2015, an upcoming launch of "Sentinel 3A" in 2016 and further launches from 2017 on of "Sentinel 1B", "Sentinel 2B" and "Sentinel 3B" it ensures good data quality for the future (cf. ESA, 2015a, 2015b). The Copernicus program shall provide data for several environmental and civil purposes. For example with "Sentinel 2A" and its high spatial and temporal resolution, a revisit time of up to 2-3 days at mid-latitudes and its distribution policy (data use for free), there is big hope on broad use and lots of applications of this dataset. The monitoring of changes to vegetation within the growing season is one example (cf. ESA, 2015b, 2015c).

The technological developments in remote sensing can enable advances in soil mapping. In-situ measurements can be combined with high spatial resolution remote sensing data to enrich models and soil maps - different approaches such as regression trees or generalized linear models are presented by Mulder et al. (2011). Small spatial patterns and soil variability are easy to detect and not only indicated well, but can also be extended to much larger regions than without the use of remote sensing data. It also gives the opportunity to do time series and analyze the development of the soils. Therefore the use of remote sensing for (digital) soil mapping can be quite evident: it has the advantage of being both time and cost efficient and retrieving better spatial variability compared to conventional soil sampling and mapping. Mulder et al. (2011) have evaluated the use of remote sensing in soil mapping and did a literature review. The study classifies the feasibility of soil mapping with remote sensing as "medium" with the main issues in the "spectral resolution being too coarse" and in finding pure bare soil pixels - not having a vegetation cover over 20 percent. However, using remotely sensed imagery as additional input data (secondary information), accuracy and efficiency show a significant increase. Especially the spatial segmentation can be improved (cf. Mulder et al., 2011, p. 12; Wulf et al., 2015, p. 22). With future development remote sensing will gain more importance and play a bigger role in the field of soil mapping.

1.3. Objective

Conventional soil mapping and especially the in-situ sampling of soil is considered to be very time-consuming and expensive (cf. Wulf et al., 2015). For this reason and as a result of this work the following question should be answered:

> ***Is it feasible to map soil types and soil patterns using Landsat time series data and an unsupervised classification algorithm approach?***

The approach is tested for a region of interest located in Lower Austria (Marchfeld). The results are evaluated using an existing soil map of the region of interest. Recognizing and emphasizing the importance of cost effectiveness, the thesis also tries to make a point with showing that it is possible to use only free open source software for processing, computation and illustration - the experience is reported in 4.2 Lessons learned.

2. MATERIALS AND METHODS

The scope of the thesis is to evaluate the feasibility of automatic soil mapping using Landsat data. It is conducted in the Marchfeld region, which is known for its intensive agricultural use and is in need of spatial high resolution soil maps providing a good base for decision-making. The advantage of an automatic approach is that theoretically any (remote) area can be assessed no matter how much ground sampled data and knowledge about the area is available, as it looks for underlying structures and patterns in the dataset and maps them accordingly. The Marchfeld site was chosen to enable evaluation between the results of the thesis and the existing soil map.

Various Landsat images from different dates over a period of four years are chosen to minimize the variability of each image. In remote sensing temporal variability can be caused by changes in atmospheric conditions, variations in soil moisture and has great influence on the recorded pixel reflectance. In this approach those effects are minimized by including 16 Landsat images, in order to show permanent soil patterns. They all capture the same area (test site) but have different acquisition dates and though differ in their recorded reflectance. This enables to delineate time-consistent patterns and minimize temporal variability.

On the basis of the 16 Landsat images several masks (layers that mask and hide areas that are not of interest) are applied to generate a dataset only containing bare soil pixels. Those masks exclude Slovakian territory, urban areas, forests and crops from further processing with only agricultural bare soil being left. The following indicators are derived using the reflectance of bare soil pixels:

1. The median over time (all 16 images) of each Landsat Band (Band 1:5, 7) was calculated - respectively the median of the temporal variability of each Landsat Band.
2. The coefficients of a polynomial function fitted to the reflectance of each pixel offer additional information. The coefficients represent information lying in between the recorded reflectance values of the bare soil and can enable a better differentiation between soil types.
3. The slope and intercept of the soil line concept which was introduced by (Richardson and Wiegand, 1977) are calculated. The concept is considered as a robust method which is influenced little by varying effects like soil roughness, soil moisture or higher reflectance values due to solar irradiation and angle and is therefore considered as a reliable parameter.

These indicators derived from the spectral/temporal reflectance are used for automatic mapping. Two different mapping algorithms are applied leading to different results. They are evaluated and validated with the existing soil map.

For most of the processing work the open source software "R" was used. It is designed for statistical computing, data mining and data analysis. As there are lots of powerful packages for different implementations and uses available, nearly everything can be computed and processed without leaving the R environment. Working in R enables the use of the latest methods and algorithms but requires coding skills and knowledge about the structure of the R language.

For visual examination, data evaluation and plotting of maps the open source software "QGIS" was used. It is a geographic information system (GIS) comparable to ArcGIS providing a way better graphical user interface than R. Therefore it was mainly used for graphical design of the maps. Both R and QGIS are open source alternatives to conventional programs - like Matlab and ArcGIS.

The next chapters introduce the test site and explain the methodology in detail - a methodological workflow is shown in Figure 14.

2.1. Site Description

The Marchfeld region forms part of the Vienna Basin and it is one on the country's main agricultural regions. Its whole area - shown in Figure 1 - is about 900 km^2 and it stretches from Vienna in the West to Slovakia in the East and borders the Danube in the South. It is a flat area with minor variations in elevation, ranging from about 143 to 178 meter above sea level.

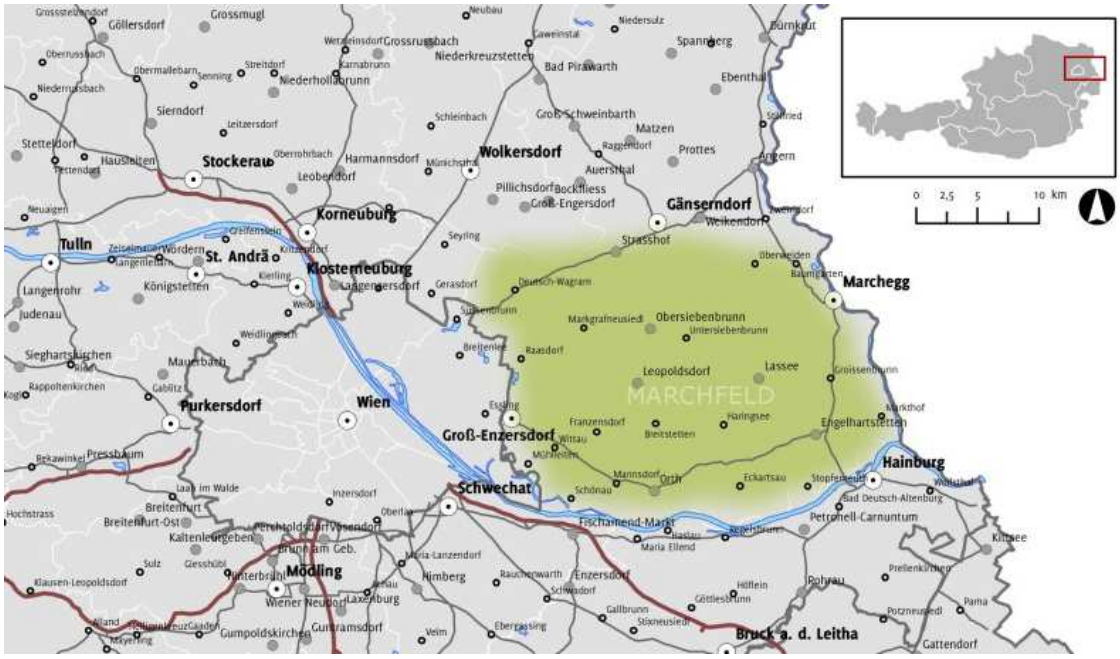


Figure 1: Location of the Marchfeld agricultural region (BOKU, n.d.).

Climate, agriculture and soil conditions: The region is influenced by a semi-arid climate, having cold winters with periods of frosts and limited snow cover. Summer is characterized by being very hot and dry. Marchfeld is one of the driest regions in Austria having an average annual precipitation of 550 mm. The average annual temperature is about 10 °C, and approximately 1.900 hours of sunshine are provided throughout the year.

Despite of its dry climate and low precipitation, it is a very prominent region for agriculture and it is considered as the country's bread basket due to one of its main crops - winter wheat. Prognosis stress that with climate change winter wheat will gain importance in the region, because CO_2 compensation overcomes the crops high sensibility to drought. It will be more favored in the future than maize for example - leading to

further increment and popularity of winter wheat among the farmers (cf. Thaler et al., n.d.). The area has a long tradition for intensive agriculture, as the topology enables an easy mechanization. It is affected by wind erosion and, due to high water demand, by sinking groundwater body. The first problem was treated with wind breakers and the latter with the construction of the Marchfeldkanal.

The Marchfeldkanal is a channel that gives path to water from the Danube to the Marchfeld region, maintaining and improving the water supply of the area. It was constructed between 1986 and 2004 and is about 100 kilometers long (cf. "Marchfeldkanal," n.d.). Figure 2 gives an overview of the area and shows the Marchfeld channel in the center.

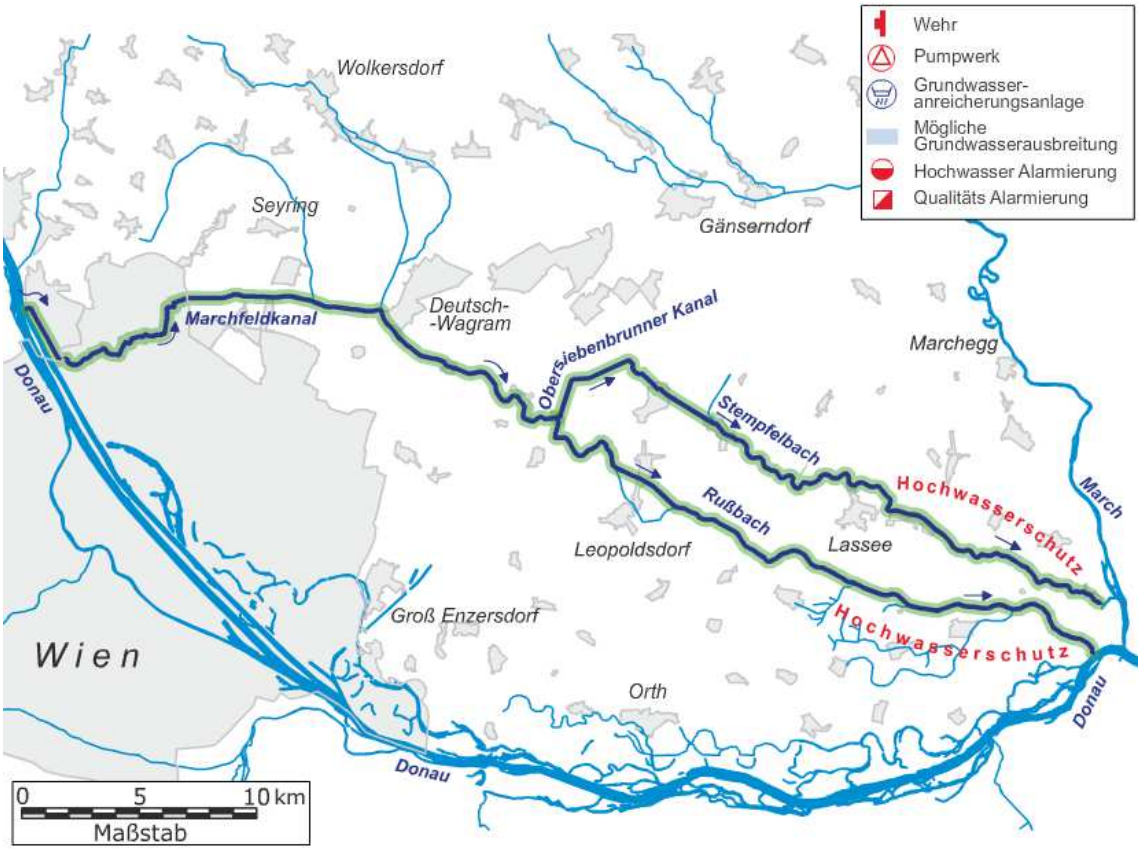


Figure 2: Overview of the area with the Marchfeld channel ("Marchfeldkanal," n.d.).

The channel maintains and improves the water quality and benefits the region as a recreational area. It is further used as a flood control and reestablishes the ground water body of the Marchfeld, which is suffering water loss due to intensive irrigation. Sprinkler systems and travelling gun systems are the most common irrigation systems (cf. Sommer et al., 2009). Soil conditions make irrigation systems indispensible. Due to the areas importance for agriculture, there is great need for reliable information on soil conditions and small-scale spatial variability. On the basis of this information different decisions concerning crops, fertilizer, amount of irrigation or management techniques can be taken in a more efficient way. They have great influence on the ecological development and sustainability of the region and on the profit of the farmers. Hence there is great need for better soil mapping in the Marchfeld region.

Around 90% of the soils of the region were formed by the Danube and this has an impact on the pedological conditions (cf. Sommer et al., 2009). Soils are mostly alluvial soils, Chernozems Fluvisols and colluvial soils with high humus content and very different loam and loess content. Generally a humus-rich A horizon and a sandy C horizon can be found. The effect of the soil being formed and influenced by the Danube can be seen in Figure 3. The characteristic meandering patterns of streams in their potamal are clearly visible in the satellite image.

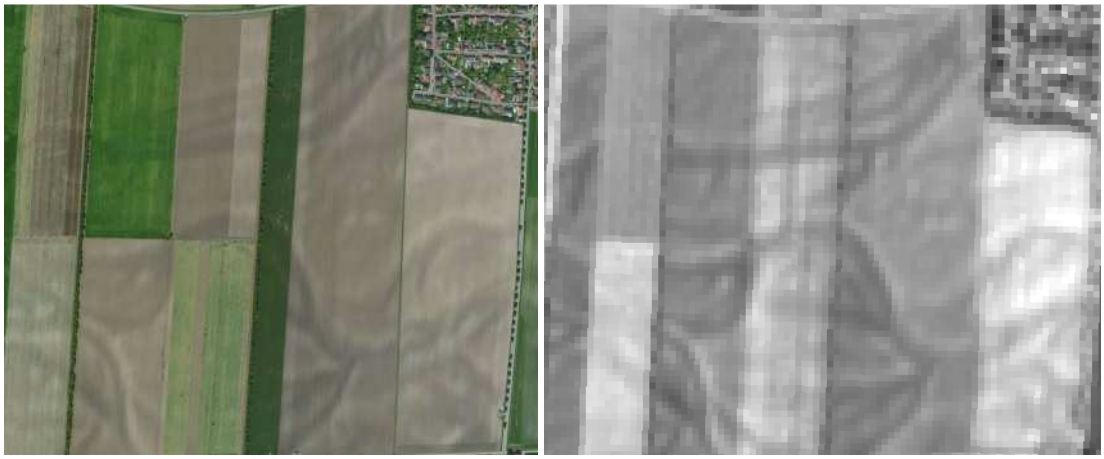


Figure 3: The meandering shapes of the Danube are clearly visible at the soil surface and have big impact on soil condition and composition. Bing Aerial Maps was used for the image on the left side. The right side is a panchromatic Landsat image of the 10th of October 2013. Both show the same spatial extent.

Against a general opinion that considers the area as very productive, Rötzer (2004) just partly agrees on that, arguing that precipitation is very little and makes irrigation obligatory. Although Rötzer (2004) further claims that good soils in Marchfeld just get a value of 60 out of 100 from the "Finanzbodenschätzung" (estimation of the monetary value of the soil), a comparison with the corresponding map (see: gis.lebensministerium.at/eBod; layer: "Wertigkeit Ackerland") shows that there are indeed soils with very low values but nevertheless the major part of the region is classified between "medium" and "very good" in terms of the "Finanzbodenschätzung". A high spatial diversity even on a small scale is noticeable.

To confirm the statement from Rötzer (2004) the following map (Figure 4) illustrates the field capacity. It is an essential parameter as it indicates very well water availability for the plants. Large parts of the Marchfeld just are classified as "medium" or "low". Classification was done according to the standards of AG Boden (1994).

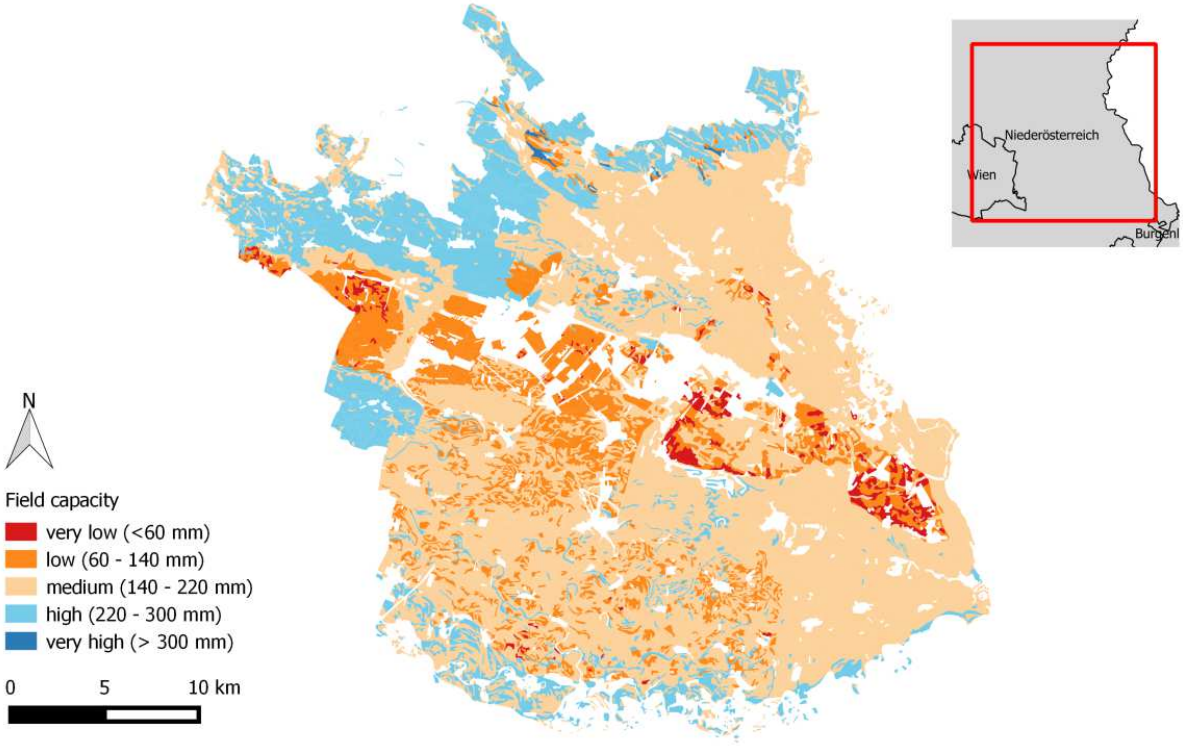


Figure 4: Field capacity in the Marchfeld region. Classification is done according to AG Boden (1994) standards.

Local conditions are very variable due to soil composition which ranges from sandy and gravel containing soils, having the lowest field capacity with around 70 mm, to colluvial Chernozem with over 300 mm field capacity. The five different main soil categories used for agriculture, according to Kromp-Kolb et al. (2007), are:

- **Class 1:** area: 14 km^2 (1,9 % of agricultural area); very low field capacity; soil depth: 30 cm; clayey sand; soil type: Parachernozem; unimportant for agricultural use.
- **Class 2:** area: 112 km^2 (14,7 % of agricultural area); low field capacity; soil depth: 30 - 60 cm; sandy clay; soil types: Parachernozems and Chernozems above sand and gravel. inferior agricultural land.
- **Class 3:** area: 466 km^2 (61,4 % of agricultural area); medium field capacity; soil depth: 80 - 120 cm; sandy clay; soil types: Chernozems and Fluvisols; medium to high quality agricultural land.
- **Class 4:** area: 166 km^2 (21,9 % of agricultural area); high field capacity; soil depth: 80 - 120 cm; clayey silt; soil types: Chernozems and Fluvisols; medium to high quality agricultural land.
- **Class 5:** area: 1,3 km^2 (0,2 % of agricultural area); very high field capacity; soil depth: 150 cm; clayey sand (from 70cm sandy clay); soil type: cullovial Chernozem; medium to high quality agricultural land (due to big soil depth high field capacity - needs crops with long roots to take advantage of) (cf. Kromp-Kolb et al., 2007).

Soil profiles of the five classes are illustrated in Figure 5. From left to right: Class 1 - Class 5.

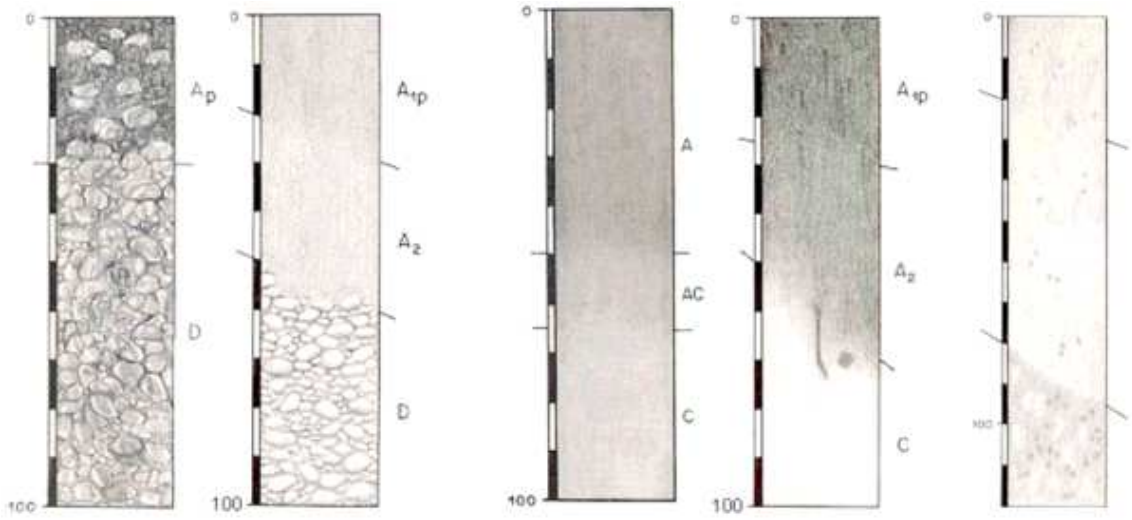


Figure 5: Soil profiles from Class 1 (left) to Class 5 (right) (Kromp-Kolb et al., 2007).

This work is limited to the eastern boundaries of the Marchfeld region around the city of Gänserndorf. The test site is indicated in the red box in Figure 6. In the east - next to the river March, it limits with Slovakia. The following soil map shows five different soil types for the test site and is used as a reference map in the final validation. Slovakian territory is depicted by the non-colored area in the east, whereas the gaps in the soil map are forest or urban areas.

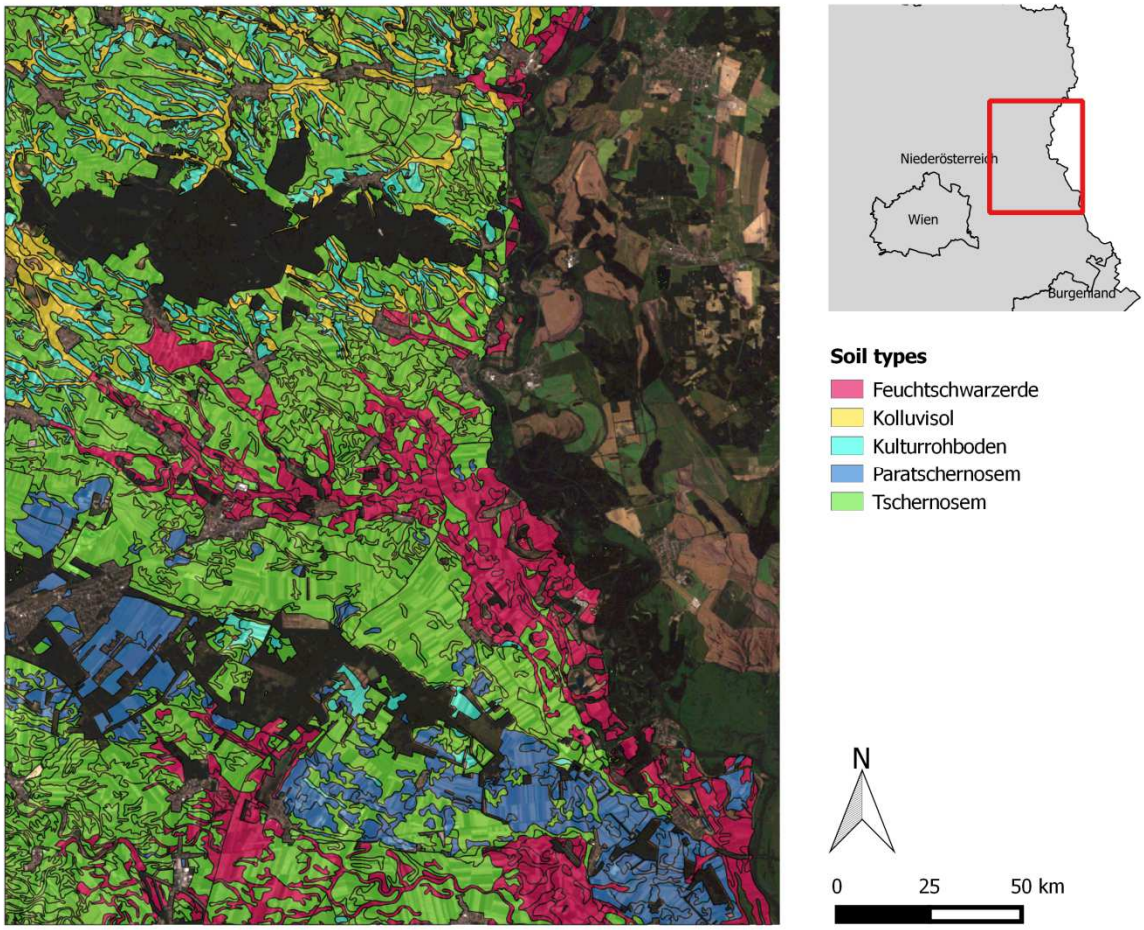


Figure 6: Soil map of the test site. Five different soil types are distinguished.

Zonal statistics of the soil map are shown in Table 1:

Table 1: Zonal statistics for the soil map.

Soil type	Number of Pixels	Area [in km^2]	% of the Area
Feuchtschwarzerde	78484	70.6	18.0
Kolluvisol	22265	20.0	5.1
Kulturrohböden	29994	27.0	6.9
Paratschernosem	51600	46.4	11.8
Tschernosem	253690	228.3	58.2

2.2. Data Acquisition & Preparation

As input data Landsat 7 images were used. The main advantage lies in their free access and their relatively high temporal resolution in comparison with other (mostly non-free) satellite datasets. Landsat 7 has an 16 days revisit time, some areas - including the area of interest - which are covered by two satellite paths are captured more frequently leading to a higher temporal resolution. The Landsat images consist of eight spectral bands, each one covering different wavelengths. Figure 7 shows the different bands of Landsat 7 and which

part of the electromagnetic spectrum they cover. Band 1, 2 and 3 cover the visible range, whereas Band 4 covers the Near Infrared and Band 5 and 7 cover the short wave Infrared spectrum. Band 8 is a panchromatic channel with better spatial resolution - 15 instead of 30 meters seen in the other bands.

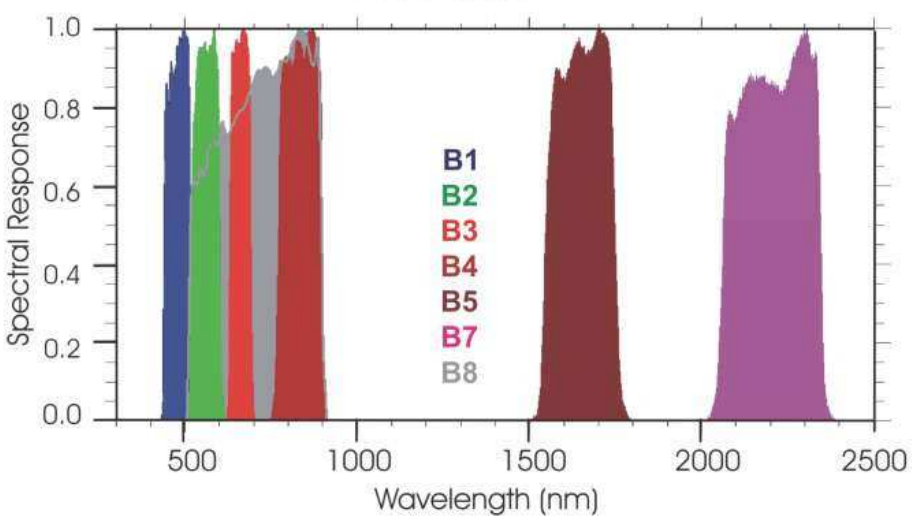


Figure 7: Spectral response of Landsat 7. B1-B8 representing Landsat Band 1 - Band 8("Spectral Response of Landsat 7," n.d.).

To mask bare-soil pixels only we used the following approach: One of the areas main crops is winter wheat, which is harvested in the summer months and seeded again in autumn - beginning at the end of September (cf. bmlfuw - Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, 2015). We can expect to capture bare soil pixels between End of July and September as the crops are cut and removed from the fields and bare soil is visible. Dematte et al. (2009) stressed the importance of the time period for bare soil delineation. Due to the knowledge of the composition of the crops - a successful delineation can be expected in this study. Data acquisition of Landsat images was limited to this period - exactly from the 29th of July until the 6th of September - from 2010 to 2013 and led to 16 images in total. On these 16 "Landsat 7" images three masks

1. Agricultural-Area
2. Cloud Mask
3. Vegetation Mask

were applied partly following the concept of bare soil detection and discrimination by (Dematte et al., 2009). The first mask excludes all non agricultural land, the second mask excludes clouds and shadows and the third mask excludes the presence of green vegetation. After the application of the three masks the only observations left in the dataset are bare soil pixels. In the following part the three masks are described more in detail:

1. Agricultural-Area: A shapefile representing the agricultural area of the region was used to exclude urban spaces, forests, rivers,... - short: the non-agricultural-area. As the test site also limits with Slovakian territory this part of the Landsat images was also excluded. After the application of the mask only agricultural area is left over.

2. Cloud Mask: this mask was used additionally to exclude all pixels covered by clouds and cloud shadows. It is provided by the USGS - U.S. Geological Survey (provider of the Landsat data). Being applied on each of the 16 Landsat images, it ensures that just agricultural area not representing clouds and shadows are left over in the dataset.

3. Vegetation Mask: The Normalized Differenced Vegetation Index (NDVI) was used to mask green vegetation. This index helps to detect live green plant canopies in multispectral remote sensing data. It uses a relationship between the red (RED) and the near-infrared channel (NIR). The calculation of the NDVI index is done according to the following formula:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

Healthy, green vegetation absorbs a great amount of radiation in the red spectrum - its pigment chlorophyll uses it for photosynthesis. Vital plants on the other hand strongly reflect near-infrared light. This characteristic difference in reflectance between the red and the near-infrared channel helps to differentiate vital green plants from other land cover types such as water bodies, urban areas or soils (cf. Fox et al., 2004, p. 2). Figure 9 shows the characteristic curve of green vegetation and its difference to other land cover types like soil.

The NDVI index can take up values between -1 and 1. Negative NDVI values indicate water bodies, snow and ice, slightly positive ones soils and urban spaces and high NDVI values indicate vegetation - the higher the value the more vital, dense and "greener" the vegetation (cf. USGS, 2015). Therefore the index can be used to map bare soil pixels using a simple threshold. Various values of NDVI can be found in literature for bare soils. The range spans between a NDVI value of 0.05 and 0.3 (cf. Holben, 1986; “Institut Cartogràfic i Geològic de Catalunya,” n.d., “ArcGIS - Using the NDVI process,” n.d., “UW-Madison Satellite Meteorology,” n.d.). Most authors agree on the necessity and importance of an individual and site-specific assessment. Through visual examination and evaluation of our dataset, the range of the NDVI value of bare soil pixels was defined between 0.19 and 0.28. So every pixel not meeting the NDVI-threshold must not be considered "bare soil" and thus has to be excluded from further processing. After the application of this last mask the only area on the 16 Landsat images left is bare soil.

2.3. Reflectance based indicators

Reflectance based indicators were generated starting from the multi-spectral and multi-temporal reflectance to provide more reliable and soil specific information. Simple reflectance values may not illustrate enough the desired soil patterns as they are influenced by varying effects like atmospheric conditions, solar illumination angle and soil moisture. Band transformations instead can provide additional information that is more related to the goal of mapping soil classes. Different concepts are applied - as described in the following - trying to reduce undesired information and noise. The generated indicators are:

Median of spectral reflectance: the median over time (all 16 images) of each Landsat Band (Band 1:5, 7) is calculated. Output are six layers (Band 1:5, 7) representing the median of all 16 images. Compared with the mean value, the median has the advantage of

being more robust and less sensitive to outliers. For one pixel (point 849) Figure 8 shows six box plots, each representing one Landsat Band. They show the variability of reflectance recorded over all 16 Landsat images. The median values - marked for point 849 by the thick horizontal lines inside the boxes - represent the reflectance based indicator.

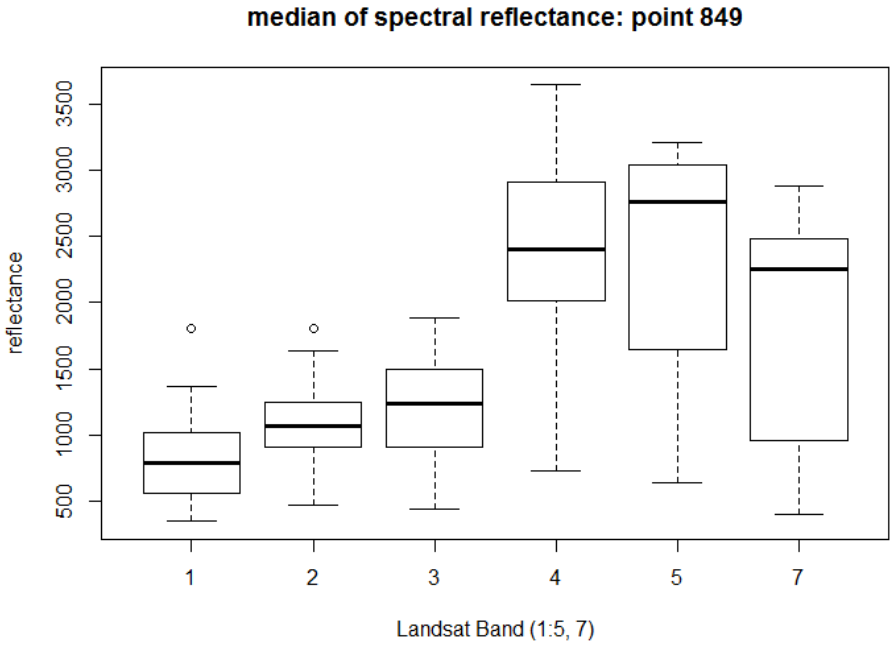


Figure 8: The box plots indicate the variability of the spectral reflectance over all 16 Landsat images for each Landsat Band of point 849. The thick line inside the box marks the median for the Landsat Band.

Polynomial function fitted to reflectance data: The reason behind the calculation of this indicator lies in the idea of taking account of the information lying in between the bands. Spectral information of a pixel is just detected at the wavelength of each band (see Figure 7) and misses out the rest. The fitting of a polynomial function and the extraction of its coefficients, helps to describe the spectral signature of a pixel. The spectral signature is unique for each pixel and contains additional and more information than all the bands. Examples for typical spectral signatures can be seen in Figure 9. Different land cover types have different spectral signatures and thus can be discriminated. The six gray bars indicate the position of the Landsat bands.

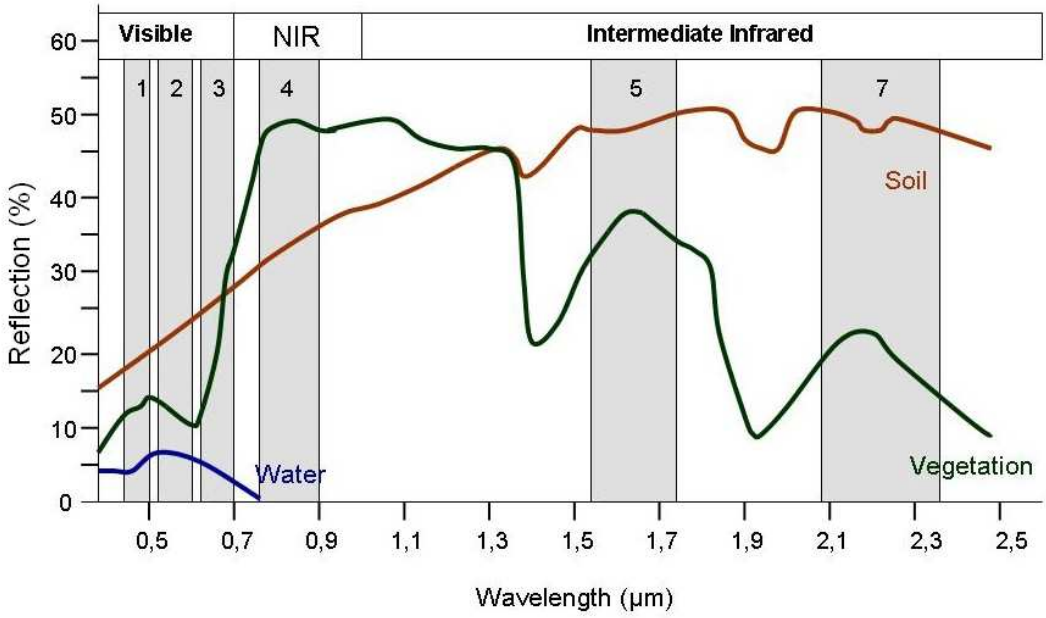


Figure 9: Typical spectral signature of three land cover types (Water, Vegetation and Soil). The gray bars indicate the position of the Landsat bands (SEOS Project, n.d.).

In Figure 10 the red curve represents a 3rd order polynomial function fitted on a pixels (point 28 and 849) detected spectral values. The dots mark the observed reflectance from Landsat Band 1 (left) to Band 7 (right).

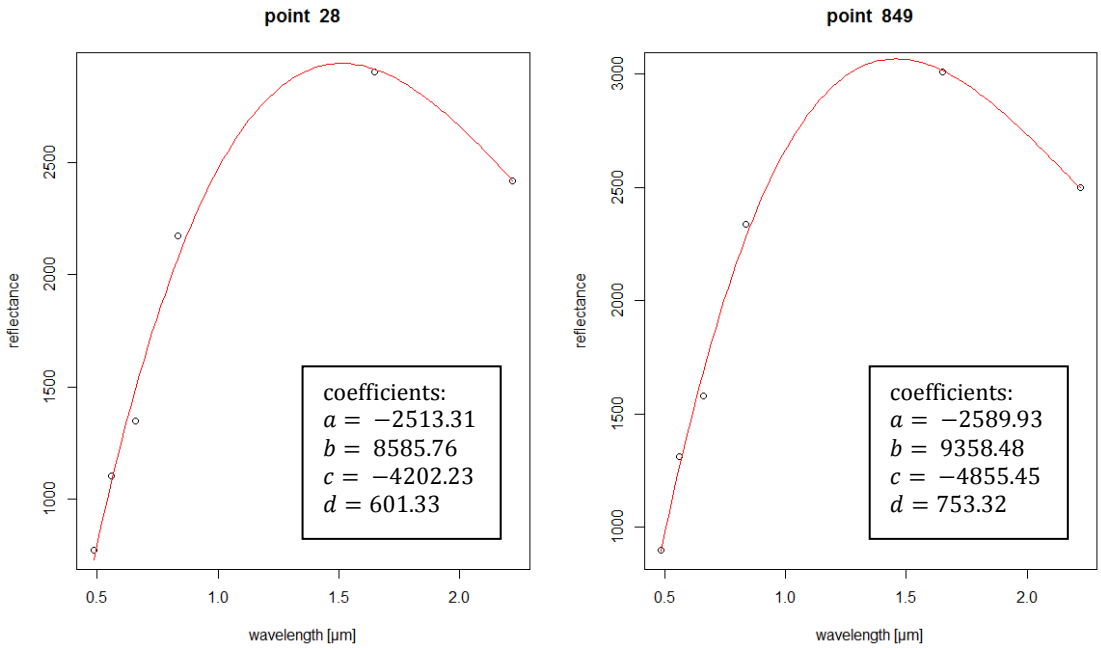


Figure 10: 3rd order polynomial function fitted on point 28 and point 849 based on their spectral reflectance (observed values in Landsat Band 1:5, 7). Dots represent Landsat Band reflectance from Band 1 (left) to Band 7 (right).

This computation is done for each day of acquisition. The results are 16 raster stacks (one respectively for every Landsat image), each containing the four coefficients (a, b, c and d) of the function. This is done according to the formula for a third degree polynomial function:

$$y = a + bx + cx^2 + dx^3$$

To reduce the dimension of the dataset (16 x coefficient a, b, c and d) and to cope with the problem of missing data values, as a last step of this computation the median values over time of each coefficient is calculated. This leads to a final raster stack of four layers representing the median of the coefficients of the polynomial function.

Soil line - relation between Landsat Band 3 and Band 4: The soil line concept was introduced in 1977 (Richardson and Wiegand, 1977). It describes the linear relationship between the Red and Near-Infrared reflectance of bare soil as characterized by slope and intercept parameters (Fox et al., 2004, p. 1326):

$$NIR = \beta_0 + \beta_1 RED$$

where β_0 is the intercept and β_1 the slope of the function. It is considered as a robust concept which is influenced little by varying soil roughness, soil moisture or higher reflectance values due to solar irradiation and angle. Fox et al. (2004) further stress, that it "can be related to site-specific soil conditions within a field" and that "this relationship may provide a means for directing soil sampling" (Fox et al., 2004, p. 1326). The pixels considered for the soil line may change its reflectance (due to soil moisture, solar illumination angle,...) over time and thus move up and down the soil line, but their variability has no effects on the soil line coefficients - respectively the intercept β_0 and the slope β_1. Figure 11 shows that a pixel on the soil line moves upwards if it is dry and moves down the soil line if it is wet. The right figure also illustrates very well the possible migration of a pixel on the soil line over a season. It starts as a wet soil on the lower part of the soil line, moves away from it as it gains biomass and finally ends on the soil line again after harvesting - but this time more on the upper part as the soil is drier than in the beginning.

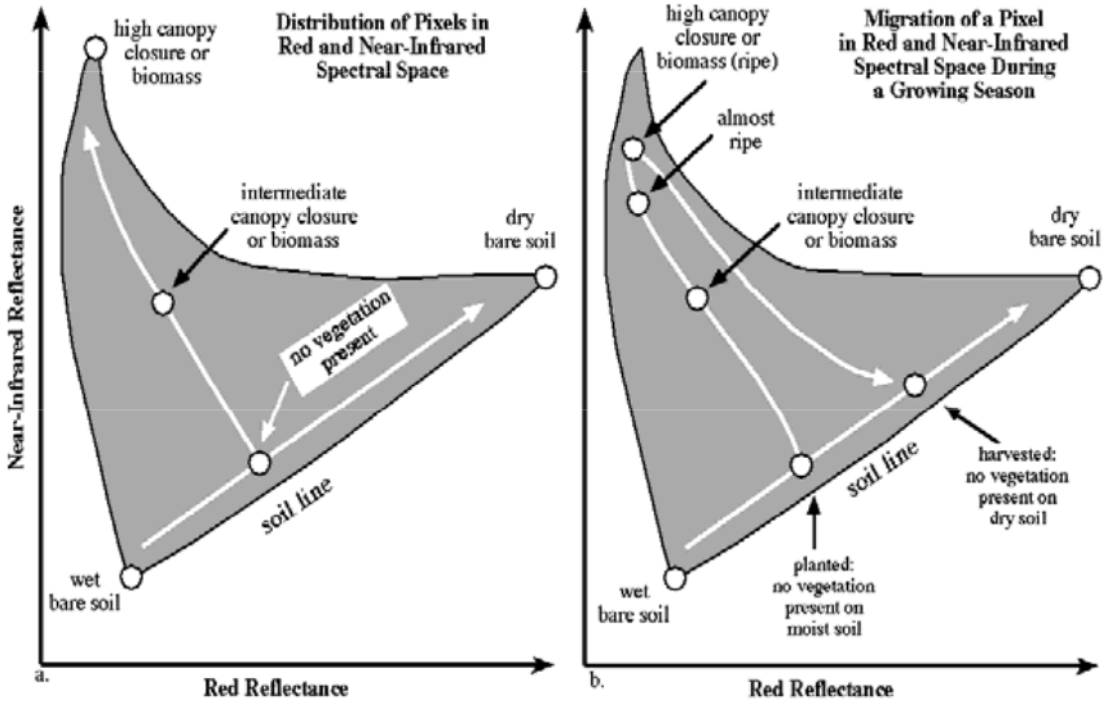


Figure 11: a) The grey shaded area represents a possible distribution of reflectance values in a remote sensing image plotted between the red and the near-infrared band. The greater the amount of photosynthetically active vegetation present, the greater the near-infrared reflectance and the lower the red reflectance. This condition moves a pixel's spectral location in a perpendicular direction away from the soil line. b) The migration of a single vegetated agricultural pixel in red and near-infrared multispectral space during a growing season is shown. After the crop emerges, it departs from the soil line, eventually reaching complete canopy closure. After harvesting, the pixel will be found on the soil line, but perhaps in a drier soil condition (Jensen, 2007, p. 343).

The soil line coefficients - intercept β_0 and the slope β_1 - can be used as input feature for the classification. The soil line will be calculated for each pixel for each Landsat image using a moving window of 3x3 and 7x7 pixels. As at least two data points are needed to calculate the soil line, the moving window approach includes neighbouring pixel values. The differences between the two window sizes will be evaluated. To reduce the dimension of the dataset (16 x coefficient β_0 and β_1) and to cope with the problem of missing data values - as a last step of this computation the median values over time of each coefficient are calculated. This leads to a final raster stack of two layers representing the median intercept and slope coefficients of the soil line.

The generated datasets

- "Median value of spectral reflectance",
- "Fitted polynomial function" and the
- "Soil line slope and intercept"

serve as input variables for the classification algorithms.

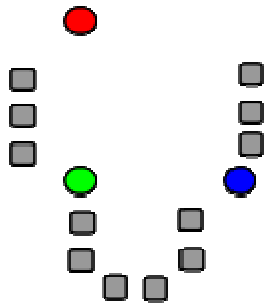
2.4. Unsupervised classification

Classification is a method to look for underlying patterns and structures in datasets and classify or group them accordingly. It identifies the class an observation belongs to and tries to group observations similar to one another. Different classification algorithms have a different design and look for different structures in the dataset, therefore the outcome of a classification is highly dependent on the data structure and the algorithm used. A good classification manages to maximize similarity of observations within a group, and maximizes the dissimilarity between the groups enabling good separability and distinction. Whether a classification is good or the result is satisfying, is a difficult task to determinate. There are several measures - many measuring class separability - giving an impression of the quality of the classification, but nevertheless a project specific, individual evaluation is often required as well (cf. Kaufman and Rousseeuw, 2005).

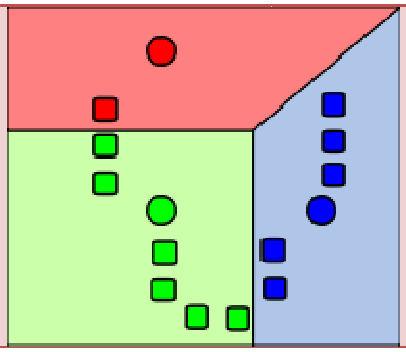
Unsupervised classification is not guided by knowledge of the user. It uses algorithms which structure the dataset according to underlying patterns. The only possibility to influence the outcome is by choosing the desired algorithm, the input features and by defining the number of output classes. Unsupervised classification has the advantage that not much site-specific information is needed and can be applied anywhere without necessary knowledge concerning input variables and training samples - as it is necessary with supervised methods. Above all it should be mentioned, that algorithms are heuristic solutions - so there is no guarantee that the global optimum will be found and the result may be strongly influenced by the initial clusters. The following two unsupervised classification methods were applied:

kMeans algorithm: the kMeans algorithm is a popular method for cluster analysis and data mining. It partitions the observations into *k* clusters in which each observation belongs to the cluster with the nearest mean - the sum of squares from points to the assigned cluster centres is minimized. The process can be divided into four steps (cf. Kwedlo, 2011; "kmeans clustering," 2015) shown in Table 2:

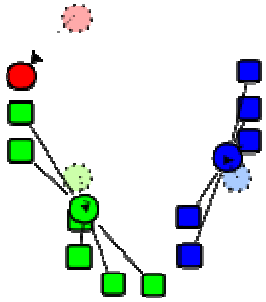
Table 2: Processing steps of the kMeans standard algorithm.

Step 1: The *k* initial means are randomly generated and distributed on the dataset (in gray). In this case *k* = 3: three classes should be differentiated - shown in color.	

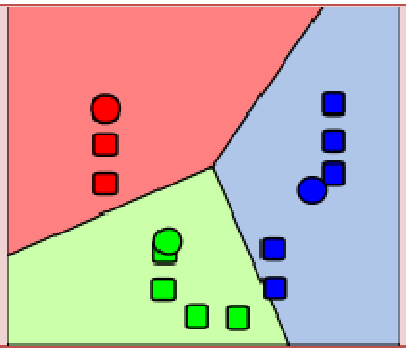
Step 2: *k* clusters are differentiated by assigning each observation with the nearest mean using the euclidean distance measure. (Euclidean distances gives a metric distance between two points and can be applied in 2- dimensional to *n*-dimensional space.



Step 3: The centroid of each cluster becomes the new mean.



Step 4: Step 2 and Step 3 are repeated and iterated until the predefined "convergence threshold" or the "maximum number of iterations" has been reached.



There are various settings besides *k* that can be defined throughout the process. In Step 1 the initial means can either be randomly generated (as shown in the figure) being an artificial data point or randomly chosen from the dataset - being an observation. Concerning this aspect Lu et al. (2008) claim that "for most of the clustering algorithms convergence usually depends highly on the choice of the initial cluster centers. Thus the determination of initial cluster centers is very important for such algorithms" (Lu et al., 2008, p. 788). Pena et al. (1999) conclude in their research on a comparison of four initialization methods for the kMeans algorithm, that the random initialization outperformed the other methods with respect to the effectiveness and the robustness of the kMeans algorithm. In this study, we use random initialization. Furthermore - as mentioned in Step 4 - the maximum number of iterations has to be defined. If the desired convergence is not reached, the algorithm stops after certain amount of iterations.

One of the main disadvantages of the kMeans algorithm are its dependency on parameter *k*, and that it tends to seek clusters of similar size due to its spherical concept.

SOM algorithm: The SOM (self-organizing map) algorithm is one of the most popular artificial neural network models. It is also called "kohonen networks" after their inventor Teuvo Kohonen. The SOM algorithm describes a mapping from a higher-dimensional input space to a lower-dimensional map space. The process is similar to the kMeans algorithm shown in Table 2. The terminology is different: calling the centroids "neurons" and the iterative process where the neurons are adjusted according to their input data points is called "training". Before training the neurons are randomly initialized (like in Step 1 of Table 2) and with each data point presented to all the neurons, the neurons change their position toward the data point (see Step 3). This changing parameter - which is called learning parameter - determines how much to adjust and is also defined initially. Adjusting too much will cause the neuron to not "learn" about other data points, whereas adjusting too little will cause the neuron to not "learn enough" about the data points. SOM has the characteristic that neurons not only adjust themselves to the data, but also adjust the neighboring neurons as well - this is called cooperative learning:

The neuron closest to the data point is the "winning" neuron and gets to move "the most" toward the data point. The neighboring neurons get to move toward the point but with a lesser distance.

Cooperative learning explains why similar neurons in the SOM tend to be grouped together at the end of the process (see Figure 12). Cooperative learning decreases with each iteration until it finally stops - at this point only competitive learning is performed. When a neuron wins in the competitive part, it becomes closer to its data point but doesn't adjust its neighbor neurons anymore (cf. Green, 2010).

The following input parameters for the SOM algorithm can be defined (Wehrens and Buydens, 2007):

rlen	the number of times the complete data set will be presented to the network.
alpha	learning rate, a vector of two numbers indicating the amount of change. Default is to decline linearly from 0.05 to 0.01 over rlen updates.
radius	the radius of the neighborhood, either given as a single number or a vector (start, stop). If it is given as a single number the radius will run from the given number to the negative value of that number; as soon as the neighborhood gets smaller than one only the winning unit will be updated. The default is to start with a value that covers 2/3 of all unit-to-unit distances.
init	the initial representatives, represented as a matrix. If missing, chosen (without replacement) randomly from input data.

While SOM with small number of neurons (clusters) performs similar to kMeans - it also uses the Euclidean distance - the additional advantage is that it preserves information about the similarity between neurons giving the possibility to merge neurons.

Figure 12 shows two example plots of a self organizing map. Each circle represents a neuron (class) - in total the SOM algorithm was run with 25 neurons represented on the grid. By default similar neurons are mapped next to each other. The left plot illustrates the differences between the neuron and its neighbors - where red indicates "strong similarity" between the neuron and its surroundings and white "strong difference". These indicators

can also be delineated in the right plot - showing the characteristics (magnitude and composition) of each neuron regarding its four dimensions.

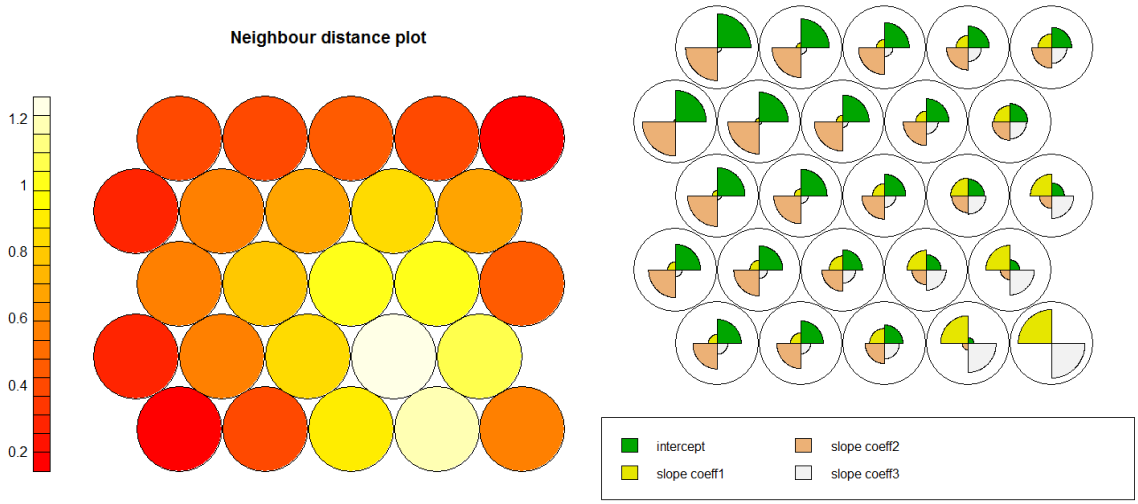


Figure 12: Example plots of self organizing maps. Each circle represents a neuron (class). The left map shows the difference between neighboring circles from red (very similar) to white (very different). The right map shows the four dimensions and its magnitude and composition of each class - the differences from the left plot can be delineated.

2.5. Post-processing and Analysis

The maps produced with both classification algorithms are further processed by using a Majority Filter implemented in QGIS. It reduces local, pixel-wise dissimilarities by looking in the neighborhood for a more general pattern. The radius defining the neighborhood to assess can be defined. Figure 13 shows how the algorithm is implemented in ArcGIS, where "InRas1" is the input and "OutRas" the output after majority filtering. Small local dissimilarities are reduced.

Figure 13: Example Majority Filtering as it is implemented in ArcGIS (ESRI, n.d.).

The resulting classification maps of the area of interest are compared and evaluated analytically. This is divided into internal and external cluster analysis. An overview of the evaluation methods is given in Table 3.

Table 3: Overview of the evaluation methods.

Analytical Evaluation	
Internal Cluster analysis	External Cluster analysis
mean Silhouette widths	adjusted Rand Index (ARI)

2.5.1. Internal Cluster Analysis

The internal cluster analysis measures the homogeneity of the clusters within the classification without any reference to additional external information. It gives high rates to classifications with low similarity between clusters but with high similarity within clusters. The Silhouette Index introduced by Peter Rousseeuw is such a measurement (cf. Rousseeuw, 1987). It is defined as:

$$s(i) = \frac{\big(b(i) - a(i)\big)}{\max\big(a(i), b(i)\big)}$$

where $a(i)$ is the average dissimilarity of the observation with its corresponding cluster (to which i belongs) and $b(i)$ the minimum average dissimilarity of observations to another cluster. So $a(i)$ can be interpreted as how good i is assigned to its cluster and $b(i)$ gives information about neighboring clusters - clusters which are similar to one another. In comparison to other internal evaluation indices like the Dunn Index (cf. Dunn, 1974) or the Davies Bouldin Index (cf. Davies and Bouldin, 1978), the Silhouette Index holds the advantage of normalizing its rating between - 1 and 1. This enables to compare different input data sets. With other indices which may start with 0 and range to infinity this is made difficult as different input data sets itself may have a better "starting score" not regarding its clustering technique. Silhouette widths of 1 indicate a perfect classification, values around 0 a weak classification and negative values a misclassification (cf. Kaufman and Rousseeuw, 2005). A value less than 0.25 is considered as poor by Kaufman and Rousseeuw. For computation the `silhouette` function from the `cluster` package in R was used.

2.5.2. External Cluster Analysis

In the external cluster analysis the classification maps are compared with the soil map shown in Figure 6. The adjusted Rand Index (ARI) - a measure of agreement between two data clusterings - is calculated. It bases on the publication of William M. Rand in 1971 where he introduced the Rand Index. The Rand Index is defined "as the percent of pairs of instances that locate in either the same or different clusters in both [...] clustering" (Bento et al., 2005, p. 311). The ARI - introduced by Hubert and Arabie in 1985, is the corrected-for-agreement-by-chance version of the Rand Index. Its use is recommended if the sizes of the clusters are not uniform - which is usually the case (cf. Hubert and Arabie, 1985). Santos and Embrechts (2009) further recommend it as "the index of choice for measuring

agreement between two partitions in clustering analysis with different number of clusters" (Santos and Embrechts, 2009, p. 4).

To calculate the ARI, the labels of the two clusterings don't have to correspond to one another - an important point for the evaluation of unsupervised classification, where the label can change from one classification run to another. The number of clusters and cluster sizes can differ between the two clusterings - but the number of classified objects has to be the same.

An example of application of the ARI is given following the tutorial from BelVecchioUK, (2012) and according to the publication from Rand (1971): Clustering A and B represent two different clustering of same size but with different assignment and labeling.

$$\boldsymbol{Clustering\ A} = [c, d, d, c, c]$$

$$\boldsymbol{Clustering\ B} = [x, z, x, z, z]$$

To calculate the ARI, pairs have to be compared according to the following four options:

- **a)** **≠ and ≠**: pair assigned to different cluster in A and in B.
- **b)** **≠ and =**: pair assigned to different cluster in A and pair assigned to the same cluster in B.
- **c)** **= and ≠**: pair assigned to the same cluster in A and pair assigned to different cluster in B.
- **d)** **= and =**: pair assigned to the same cluster in A and in B.

With a classification of 5 values there are 10 possible pairs:

[1,2], [1,3], [1,4], [1,5], [2,3], [2,4], [2,5], [3,4], [3,5], [4,5]

Starting with the first pair [1,2] leads to the following:

$$\boldsymbol{a} = +1 \begin{cases} \text{Cluster A } [1,2] \rightarrow (c, d) \rightarrow c \neq d \\ \text{Cluster B } [1,2] \rightarrow (x, z) \rightarrow x \neq z \end{cases}$$

The second pair [1,3]:

$$\boldsymbol{b} = +1 \begin{cases} \text{Cluster A } [1,3] \rightarrow (c, d) \rightarrow c \neq d \\ \text{Cluster B } [1,3] \rightarrow (x, x) \rightarrow x = x \end{cases}$$

The third pair [1,4]:

$$\boldsymbol{c} = +1 \begin{cases} \text{Cluster A } [1,4] \rightarrow (c, c) \rightarrow c = c \\ \text{Cluster B } [1,4] \rightarrow (x, z) \rightarrow x \neq z \end{cases}$$

...

The last pair [4,5]:

$$\boldsymbol{d} = +1 \begin{cases} \text{Cluster A } [4,5] \rightarrow (c, c) \rightarrow c = c \\ \text{Cluster B } [4,5] \rightarrow (z, z) \rightarrow z = z \end{cases}$$

If this procedure is done for all the 10 pairs it leads to the following confusion matrix:

Table 4: Confusion matrix.

	Pairs assigned to different cluster (A)	Pairs assigned to the same cluster (A)
Pairs assigned to different cluster (B)	a = 3	c = 3
Pairs assigned to same cluster (B)	b = 3	d = 1

The ARI finally is calculated by:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]}$$

where n is the number of classified objects. It leads to the following result:

$$ARI = \frac{\binom{5}{2}(3+1) - [(3+3)(3+3) + (3+1)(3+1)]}{\binom{5}{2}^2 - [(3+3)(3+3) + (3+1)(3+1)]} = -0.25$$

The ARI is bounded between ± 1 - with expected value 0 and total accordance between the two clusterings as 1. Reference ARI values for map comparison indicating a good, moderate or a bad classification could not be found. Nevertheless the ARI is a good method to compare the different classification results and highlight the best map. In this work the `comPart` function from the `flexclust` package in R was used to retrieve the ARI.

An overview of the methodological workflow is given in Figure 14.

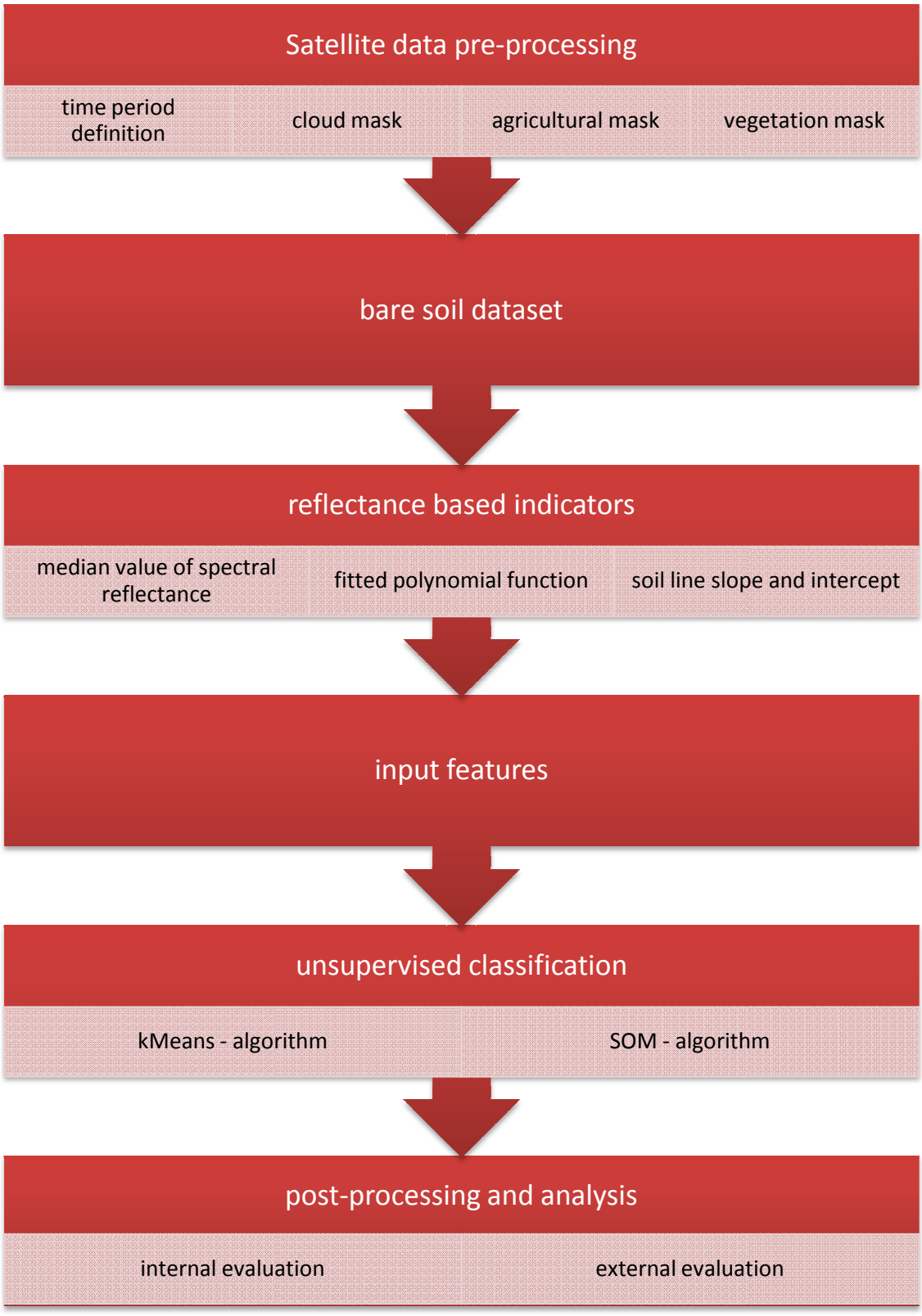


Figure 14: methodological workflow

3. Results and Analysis

In the following chapter the results of the work are described and appear in the same order as listed in the methodological workflow (see Figure 14) - starting with the pre-processing of the Satellite data, moving on with the reflectance based indicators and finishing with the evaluation of the classified maps.

3.1. Satellite data pre-processing

Figure 15 shows NDVI values of four different acquisition dates. The values are grouped in five color categories, where the brown category represents bare soil pixel, with NDVI value ranging from 0.19 to 0.28 as indicated in the legend. Higher NDVI values appear in different green intensities and lower NDVI values in blue. The development of the NDVI values in 2010 can be seen well at the four images covering the time period from the 10th of June 2010 until the 22nd of August 2010. The amount of bare soil increases with time with a first peak on the 21st of July 2010, a week before Landsat images were included into processing (29th of July - 6th of September). It indicates that the time period was well defined.

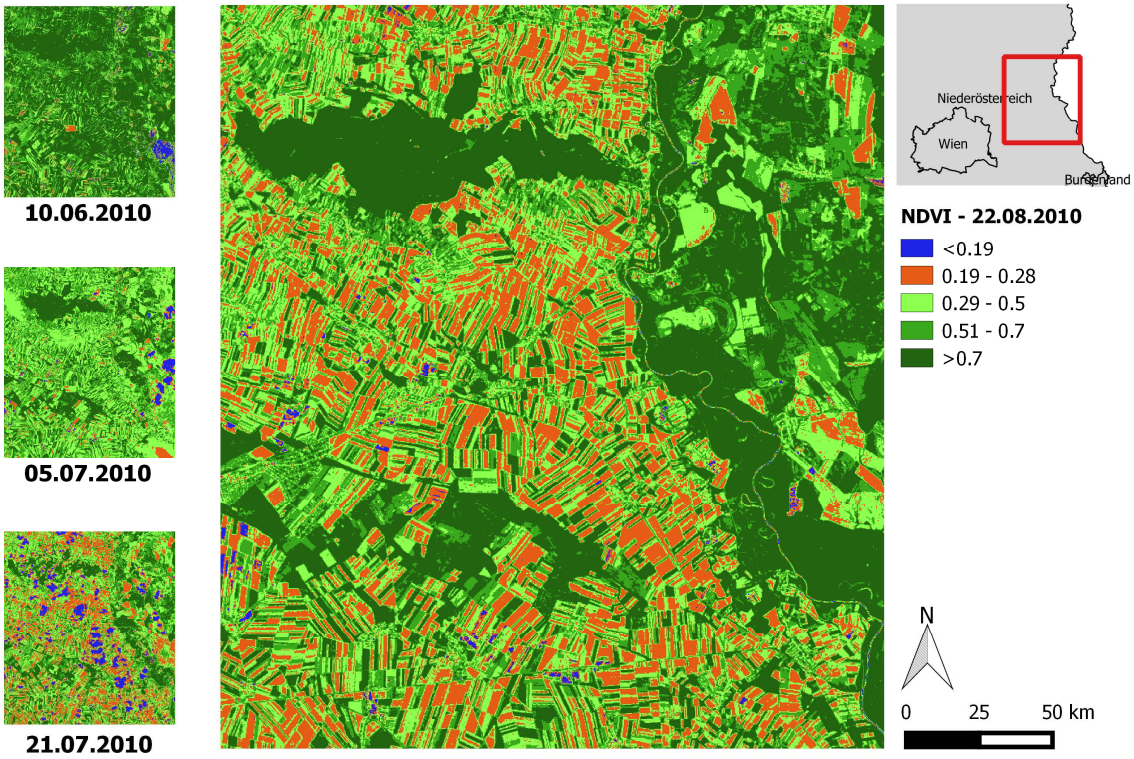


Figure 15: The big map shows NDVI values on the 22.08.2010. The brown area represents bare soil pixels, whereas green and blue areas have higher or lower NDVI values. Slovakian territory, urban spaces, forests and clouds have not yet been masked. The decline of the NDVI from the 10.06.2010 until the 22.08.2010 is shown by the increase of brown areas depicted in the four maps. It matches the natural development of NDVI values over a season with the raise in spring, where crop and biomass development is high and a decline in the late season.

The study was limited to the period 2010 to 2013, where only images between the 29th of July and the 6th of September of each year were included. This led to 16 Landsat-7 images with the following dates of acquisition:

1. 22.08.2010	5. 17.08.2011	9. 03.08.2012	13. 06.08.2013
2. 29.08.2010	6. 24.08.2011	10. 10.08.2012	14. 13.08.2013
3. 09.08.2011	7. 25.08.2011	11. 19.08.2012	15. 22.08.2013
4. 16.08.2011	8. 02.09.2011	12. 04.09.2012	16. 29.08.2013

The selection of the bare soil pixels is illustrated in Figure 16, that shows a scatter plot of an example Landsat image (on the 22.08.2010) before (left) and after (right) the exclusion of non-bare soil pixels. The same resulting bare soil pixels are plotted in red in both plots. The scatter plot represents the feature space obtained with Band 3 (red) and Band 4 (nir). This band combination enables to differentiate between vegetated surfaces and bare soil. As the NDVI is also calculated using this band combination, it is possible to identify higher NDVI values (> 0.28) representing shrub land, grass land and forests above the red path. Lower NDVI values (< 0.19) such as water are found on the data points beyond the red path.

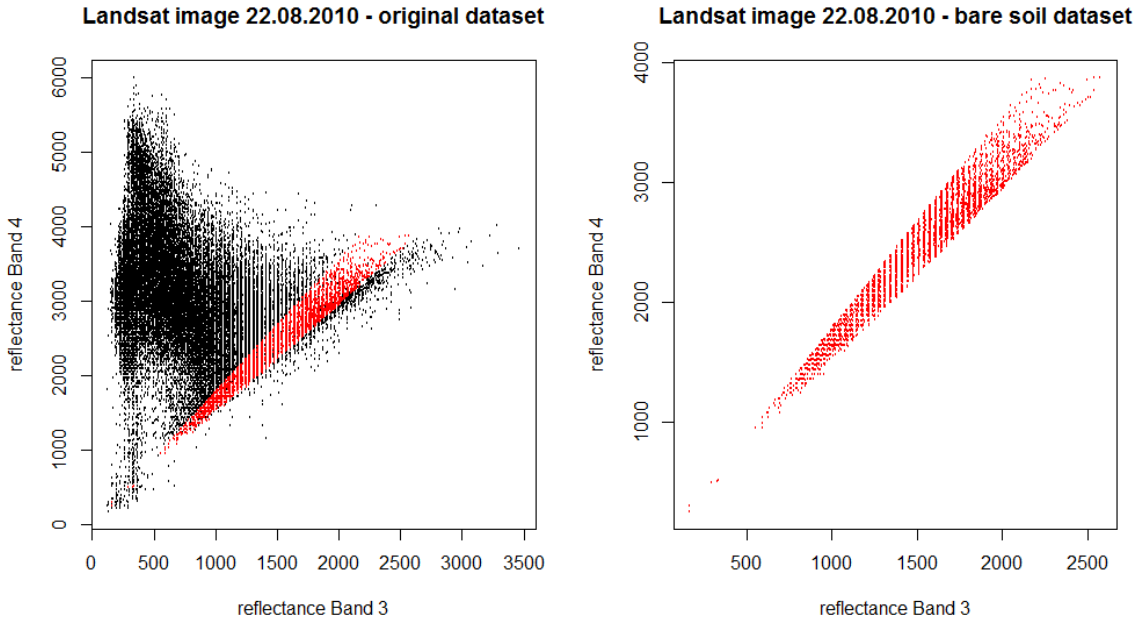


Figure 16:Relationship between Band 3 and Band 4. Bare soil pixels are plotted in red. Black data points are excluded from the original dataset (left) leading to the right plot representing the bare soil dataset.

The 16 images of the resulting dataset have different spatial distribution and amount of bare soil pixels: Landsat image 1 is differing from Landsat image 2, 3,..., 16 regarding cloud coverage and reflectance values. Figure 17 shows the variable distribution of bare soil pixels (mapped in green) and the excluded pixels (in grey) over four different Landsat images.

Figure 17: Landsat images on different dates of acquisition represent bare soil pixels in green and excluded pixels in grey. The four different maps give an impression of the spatial and temporal variability of bare soil pixels in the dataset.

To illustrate the amount of bare soil observations available for each pixel - Figure 18 shows the spatial divergence of data density. Dark red areas provide high data density and light red areas have low data density. Slovakian territory, urban spaces and forests are shown in grey. Geometrical shapes that can be seen on the map are caused by agricultural field-wise management. Fields are managed individually causing a recently harvested field to lie next to a field being at the late season stage. The main explication for these geometrical patterns is that some fields were just more often harvested (and though represent bare soil) during end of June and beginning of September than other fields lying next to them. The average number of observations of bare soil pixels at the test site over all 16 images is 3.37.

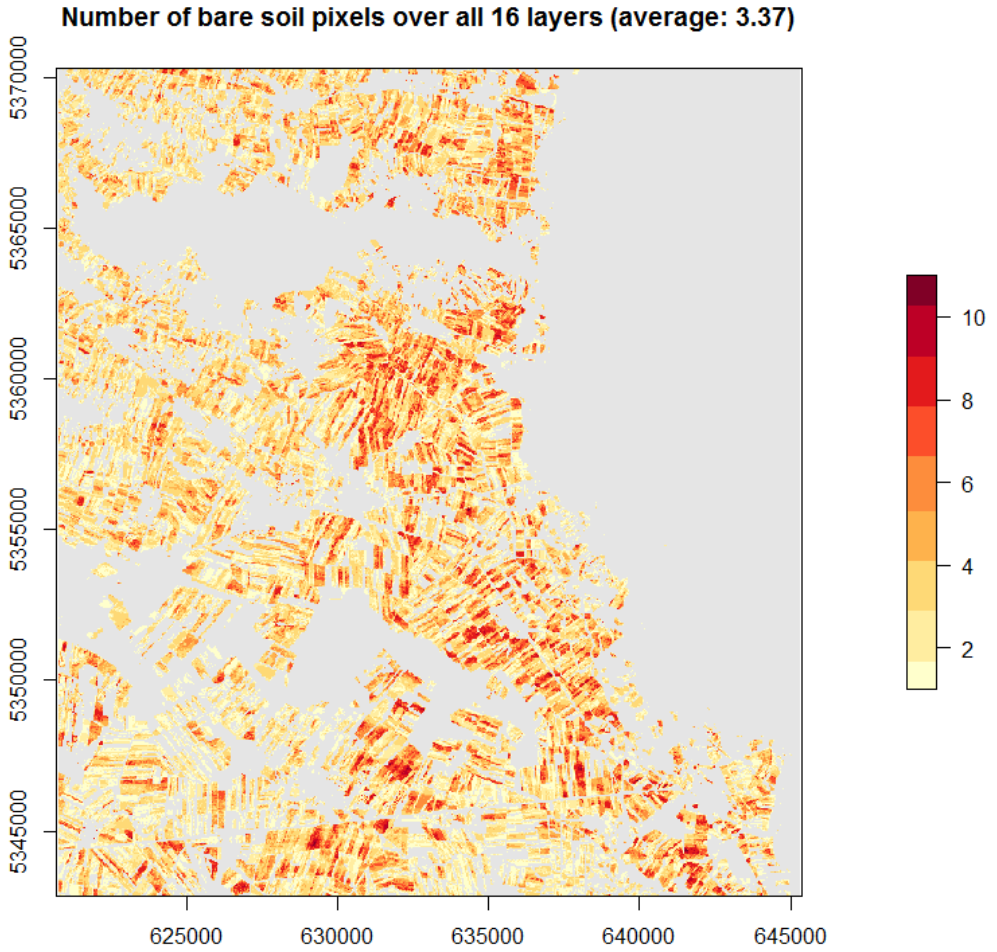


Figure 18: Number of bare soil pixels over all 16 images (layers). Spatial divergence of data density is visualized in this map with the average number of values over the whole area of interest being 3.37.

This heterogeneity of the dataset - concerning the number of bare soil pixels over time - implicates certain difficulties with respect to the clustering algorithm. Many clustering algorithms are sensitive to missing values (NA values) - datasets with data gaps are mostly reduced to some observations having full data coverage over time. Part of the dataset is represented as a matrix in Figure 19. Each pixel (23 pixels - represented as rows) containing NA values in only one of the 16 Landsat images (3 images / days - represented as columns) will be excluded from clustering. For some of the pixels (urban spaces, forests or Slovakian territory) this is the intention, but for the rest of the dataset it implies, that a major part of the area is excluded from clustering due to maybe one NA value over those 16 Landsat images. In the case of the example matrix this would result in pixel (row) 6, 7, 13, 14 and 15 being used as an input for clustering. To deal with this characteristics of the dataset, the median value of the pixel for the 16 days is calculated. In the example matrix this would lead to an exclusion of only 4 pixels (1, 2, 3 and 17) from the clustering.

	day1	day2	day3
1	NA	NA	NA
2	NA	NA	NA
3	NA	NA	NA
4	NA	NA	-3830.884
5	NA	-2900.245	NA
6	-3349.362	-1881.713	-2202.569
7	-3350.536	-2060.874	-2215.704
8	NA	-2405.785	-2285.859
9	NA	-2362.570	-3232.490
10	-3761.727	NA	NA
11	NA	-2601.969	NA
12	-3697.270	NA	NA
13	-3650.018	-2092.401	-3156.622
14	-3850.920	-2473.281	-3274.942
15	-4377.297	-2940.014	-2958.054
16	-4363.226	-2754.735	NA
17	NA	NA	NA
18	NA	NA	-2093.612
19	NA	NA	-1916.844
20	-1731.598	NA	-1675.074
21	-2945.628	NA	-2318.470
22	-3895.346	NA	-2529.523
23	-3996.527	NA	-2265.462

Figure 19: Example matrix of the dataset, where each row represents a pixel and each column a different Landsat 7 image.

Table 5 shows how many pixels are excluded from the original Landsat 7 image each date. The second column shows the percentage of the excluded pixels of the total amount of pixels (750204 pixels per Landsat image). The third column indicates the absolute amount of excluded pixels. The high number mostly result from the agricultural-area mask which already excludes 452628 pixels (60,3 %) itself - cloud coverage and vegetation contribute the rest.

Table 5: Excluded pixels per date. Total amount of pixels per image is 750204.

day of acquisition	% of excluded pixels	amount of excluded pixels
22.08.2010	83,4	625636
29.08.2010	98,0	735212
09.08.2011	94,8	711376
16.08.2011	94,4	708160
17.08.2011	88,6	664662
24.08.2011	89,1	668625
25.08.2011	81,0	607300
02.09.2011	87,6	657158
03.08.2012	99,4	745727
10.08.2012	99,0	742589
19.08.2012	90,9	682016
04.09.2012	98,2	736354
06.08.2013	84,6	634341
13.08.2013	99,2	744094
22.08.2013	87,9	659179
29.08.2013	90,5	679237

The structure of the reflectance based indicator datasets - which serve as an input for classification - is described in Table 6.

Table 6: Structure of the input datasets.

Input datasets	Structure of the dataset
Band medians	6 raster: each raster represents one Landsat band and its pixels median values of the 16 images (- the median value over time is calculated). Each Landsat band is computed separately: Band 1, Band 2, Band 3, Band 4, Band 5, Band 7.
Fitted polynomial function	4 raster representing the median values of the 16 images of four polynomial coefficients: intercept, slope coefficient 1, slope coefficient 2, slope coefficient 3
Soil line slope and intercept	• 2 raster representing the median values of the 16 images of the intercept and the slope computed on a 3x3 window • 2 raster representing the median values of the 16 images of the intercept and the slope computed on a 7x7 window

3.2. Input features for classification

The following three figures show the density plot for the three different input features used for classification. Density plots are calculated using the existing soil map and each of the input features - the different lines represent the five soil types present in the area (see Figure 6). In Figure 20 the density plot is calculated using the "Band medians - dataset". It helps to study the distribution of the dataset and delineate how many classes can be expected. Major visual distinctions between the soil classes can mostly be delineated in Band 4 and Band 5.

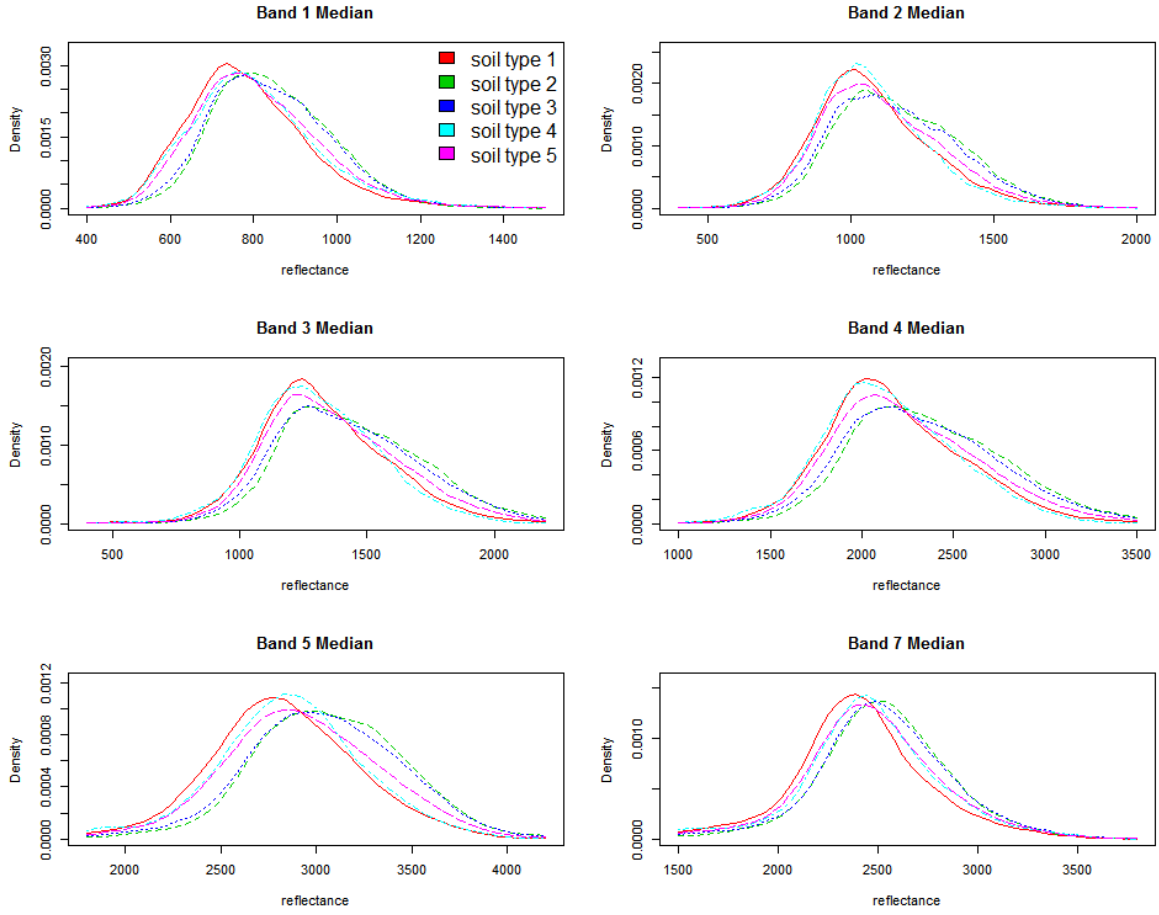


Figure 20: Density plot of the five soil classes mapped in the existing soil map and the "Band medians - dataset".

In Figure 21 - the density plot is computed with the "Fitted polynomial function - dataset" - a similar result is shown. The difference - compared with the "Band medians - dataset" lies in the distinction of the classes, which seems better than in Figure 20.

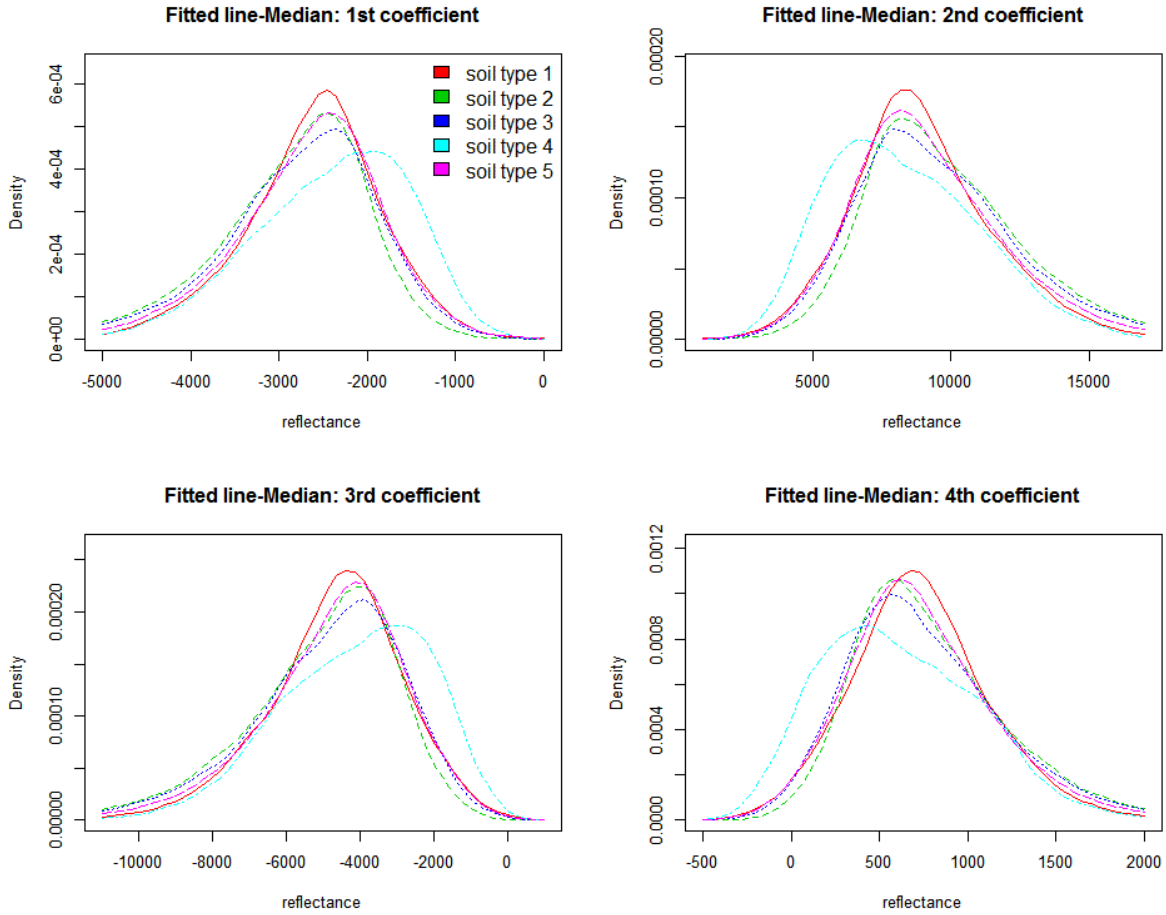


Figure 21: Density plot of the five soil classes mapped in the existing soil map and the "Fitted polynomial function - dataset".

A totally different density plot is drawn when the "Soil line - dataset" is used. A good separability between classes would be indicated by the plot but with further analysis and as Figure 23 - Figure 25 show, the values have very little variation and do not show clear shapes as the other datasets. Nevertheless the differences between the 3x3 and the 7x7 window are noticeable. In the upper two plots (3x3 window) all classes show volatile density distributions, whereas in the 7x7 window only soil type 4 and 5 show the same pattern. The volatile density distributions of the four plots, where each class forms a peak mainly without other overlapping classes, would be promising for classification.

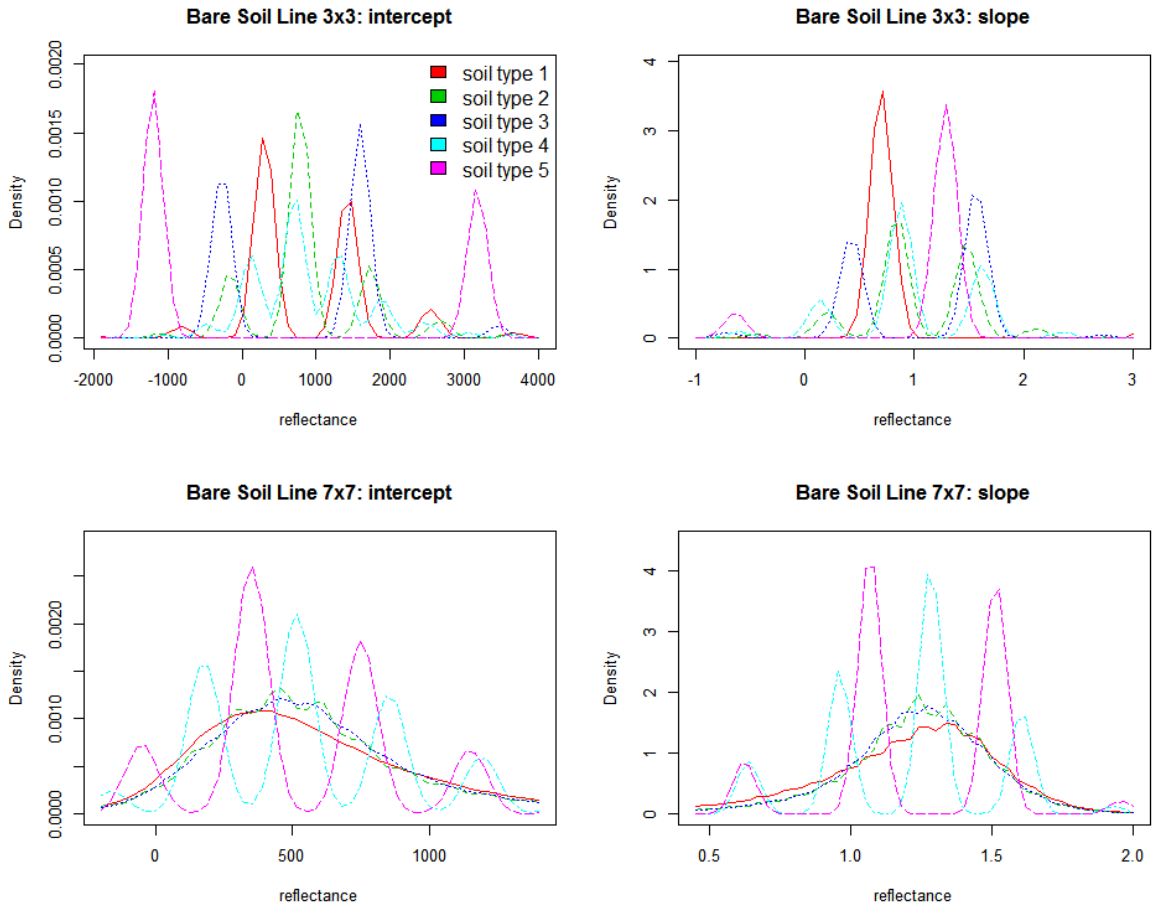


Figure 22: Density plot of the five soil classes mapped in the existing soil map and the "Soil line - dataset".

In order to assess the differences shown in Figure 22 a look was taken on the soil line slope and intercept dataset. It clearly seems that the 3x3 window was too small as it has a strong "salt and pepper" effect and no clear patterns can be observed over the area. In Figure 23 a comparison between the soil line parameter slope is shown, using a 3x3 window and using a 7x7 window for calculation.

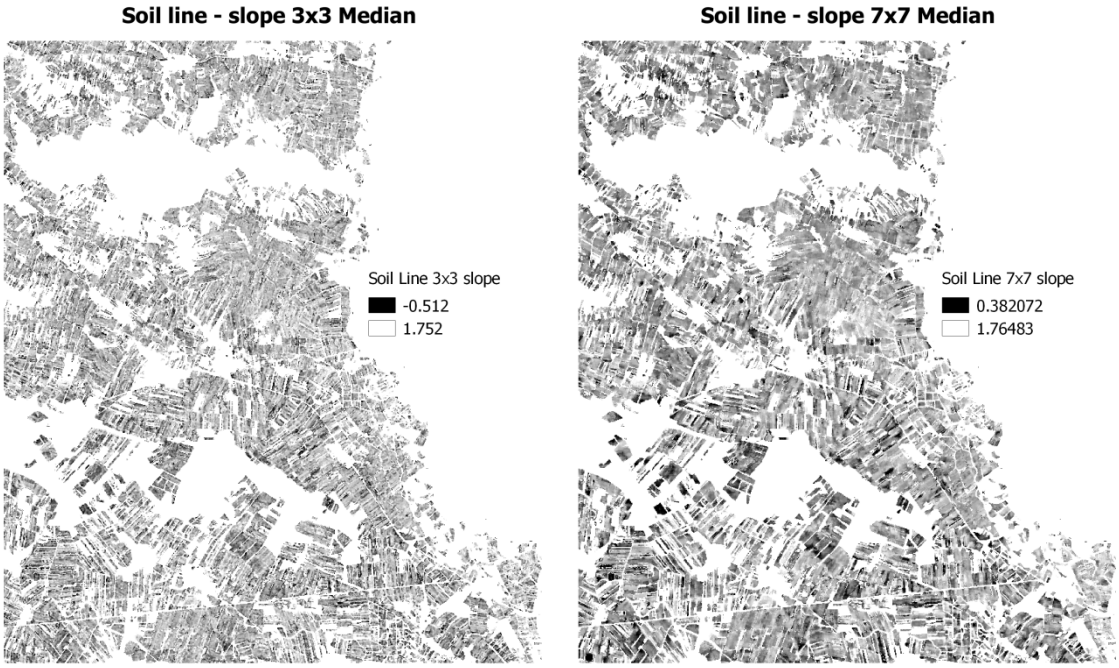


Figure 23: Comparison of the same coefficient (slope of the Soil line) calculated in a 3x3 window (left) and a 7x7 window (right).

To illustrate the homogeneity of the 3x3 dataset see Figure 24. The number above the x-axis indicate the number of data points falling into this bar (as the numbers are very low, just one main bar is visible in each plot).

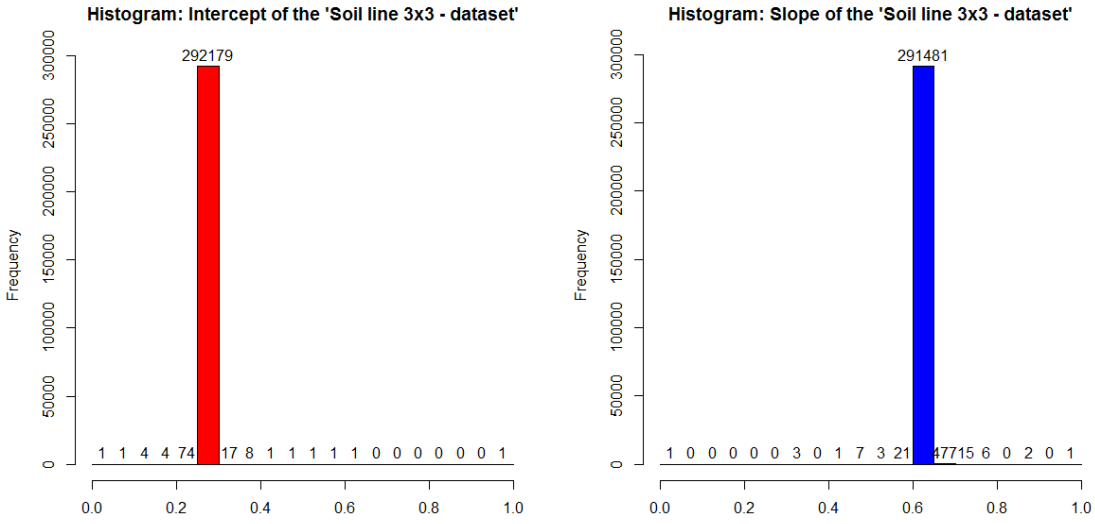


Figure 24: Histogram of the two coefficients (intercept and slope) of the "Soil line 3x3 - dataset". It shows very well the homogeneity of the dataset.

"Soil line 7x7 - dataset" shows little less homogeneity - see Figure 25. Due to their homogeneity both "Soil line - datasets" don't offer good basis for clustering.

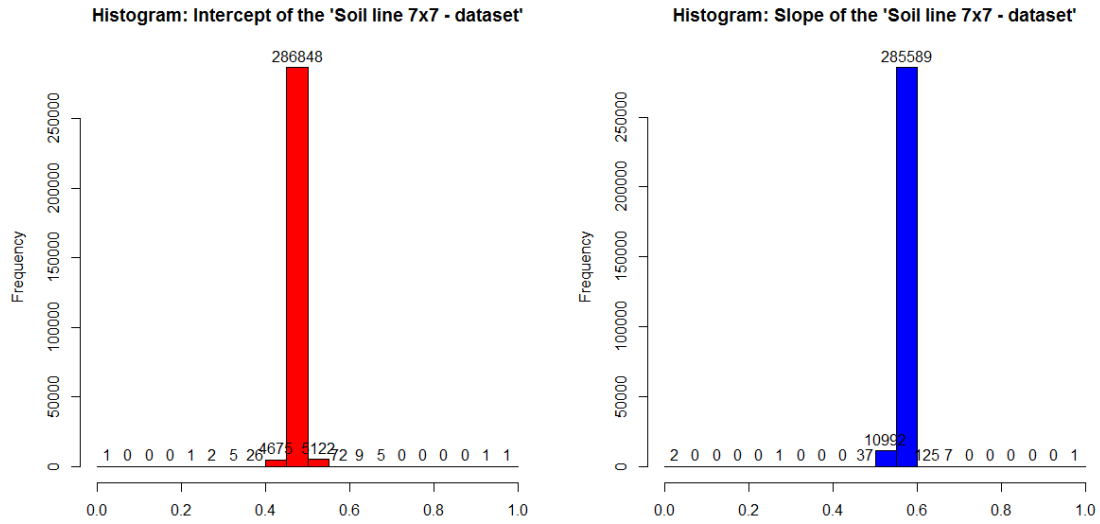


Figure 25: Histogram of the two coefficients (intercept and slope) of the "Soil line 7x7 - dataset".

In order to illustrate strong local patterns for comparison, two raster of the other datasets ("Band medians - dataset" and the "Fitted polynomial function - dataset") are depicted in Figure 26. In contrast to the "Soil line - dataset" (see Figure 23) shapes and patterns can visually be detected above all in the right plot - both datasets offer a better input feature for clustering.

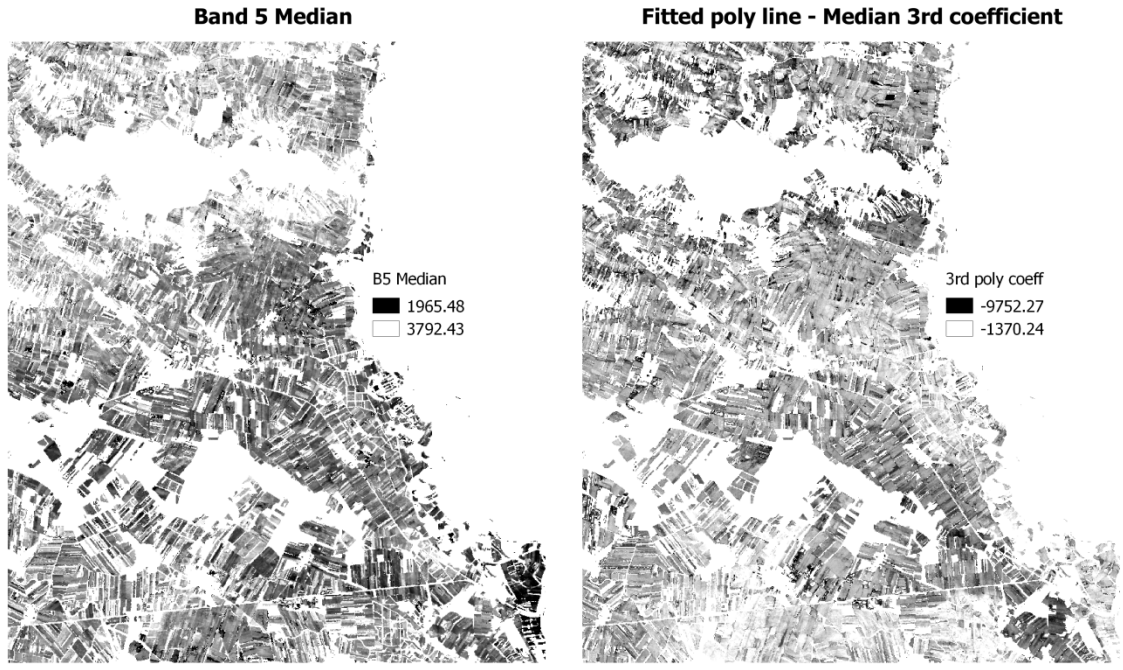


Figure 26: Comparison of the Band 5 median raster and the median of the 3rd coefficient of the fitted polynomial function. Both layers show strong local patterns.

In order to guarantee equal weighting in the classification between the datasets, they are normalized according to the following formula:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

where x represents the whole dataset and x_i one individual observation. z_i is the normalized value. After normalization each dataset ranges from 0 to 1. This is done to take account of the different range of the parameters.

The two classification algorithms are performed using following R packages and parameters:

kMeans: For computation the function `kmeans` from the R package `stats` was used.

```
kmeans_classification <- kmeans(data, algorithm="Lloyd", 3, iter.max=1
000, nstart=500)
```

`algorithm="Lloyd"`: the Lloyd algorithm was used as it does not specify the initial placement of centers (cf. Kanungo et al., 2002). Pena et al. (1999) conclude in their research that the random initialization outperformed the other methods, motivating the choice of this algorithm.

`iter.max=1000`: the maximum number of iterations is set to 1000, which is by far above the number of iterations the algorithm needed.

`nstart=500`: the number of initial random sets chosen is set to 500 (- this parameter has high influence on computation time). The number was chosen in order to ensure a good result and to keep computation time on an acceptable level.

SOM: For computation the function `som` from the R package `kohonen` was used.

```
som_classification <- som(data, grid=somgrid(xdim=2, ydim=2, topo="hex
agonal"), rlen=500, keep.data=TRUE, n.hood="circular")
```

`grid=somgrid(xdim=2, ydim=2, topo="hexagonal")`: specifies the grid, i.e. the number of output classes. A grid of 2x2 gives 4 nodes (output classes). `topo="hexagonal"` indicates the shape of the nodes, having influence on the number of neighboring nodes. An hexagonal or a square shape can be chosen. The hexagonal shape causes up to 6 neighbouring nodes.

`rlen=500`: the number of times the complete dataset will be presented to the network. The number is equal to the `nstart` parameter of the `kmeans` algorithm to ensure comparability between the two algorithms.

`keep.data=TRUE`: saves the information of the neigbouring nodes.

`n.hood="circular"`: indicates the neighbourhood, based on the shape of the nodes. `"circular"` is default for hexagonal maps.

As the learning rate `alpha` was not specified in the command above it was kept on the default value: Default is to decline linearly from 0.05 to 0.01 over `rlen` updates (cf. Wehrens and Buydens, 2007).

3.3. Maps and Cluster Evaluation

The following paragraph reports the output for the unsupervised classification algorithms after applying the majority filter with radius=1. The order of the maps will be dataset-wise as this has greater influence on the maps than the algorithm applied. After each map the corresponding table will show the results of the analytical evaluation. The internal evaluation criteria - Silhouette Index - bases on a distance matrix which is a very computer intensive process. In order to deal with this, a bootstrapping method was applied. 500 times a sample of 10.000 observations was taken to calculate the distance matrix and the corresponding Silhouette width. The mean value of those 500 Silhouette widths is reported in the analytical evaluation table.

3.3.1. "Band medians - dataset"

The classification in Figure 27 shows three classes and was computed using the kMeans algorithm. The red area delineates very well Feuchtschwarzerde in the east but has also included Tschernosem and Paratschernosem in the south. For a better distinction of Kolluvisol and Kulturrohböden an extra class would be needed as both are classified now within one class: blue. The green area is mostly representing Tschernosem. Coarse patterns and shapes can be delineated from the map and spatial variability is a lot higher in the classification compared to the reference map.

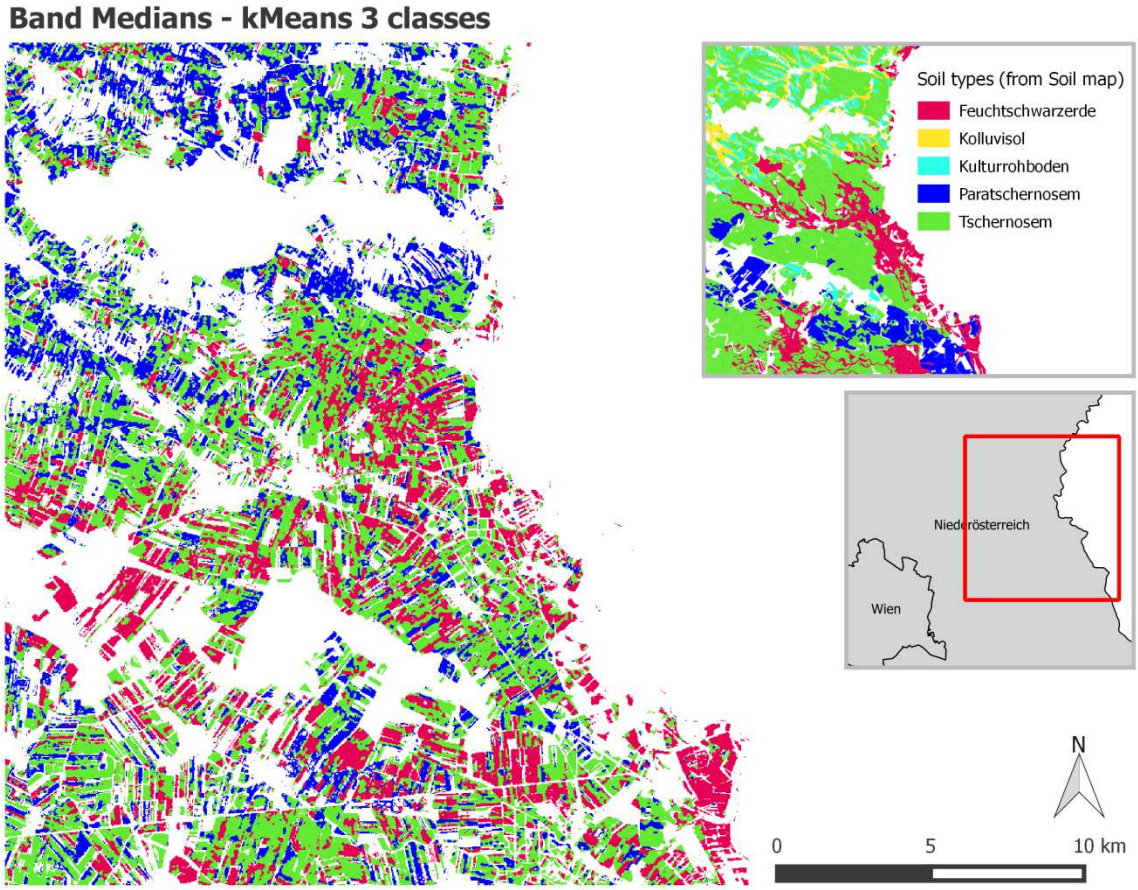


Figure 27: kMeans classification using "Band medians - dataset" - 3 classes.

Band Medians - kMeans 3 classes		
Internal analysis	mean Silhouette width	0,356
External analysis	adjusted Rand Index	0,057

The classification in Figure 28 shows four classes and was computed using the kMeans algorithm. The green area represents well Tschernosem and Paratschernosem, which is no problem as the two soil types are very similar to one another. It also includes big areas of Feuchtschwarzerde, where the green area lacks better separability. In comparison to Figure 27 an extra class led to a better separability between the classes in the northern third of the region. Coarse patterns and shapes can be delineated from the map, although the north matches better the reference map than the south. Spatial variability is a lot higher in the classification compared to the reference map.

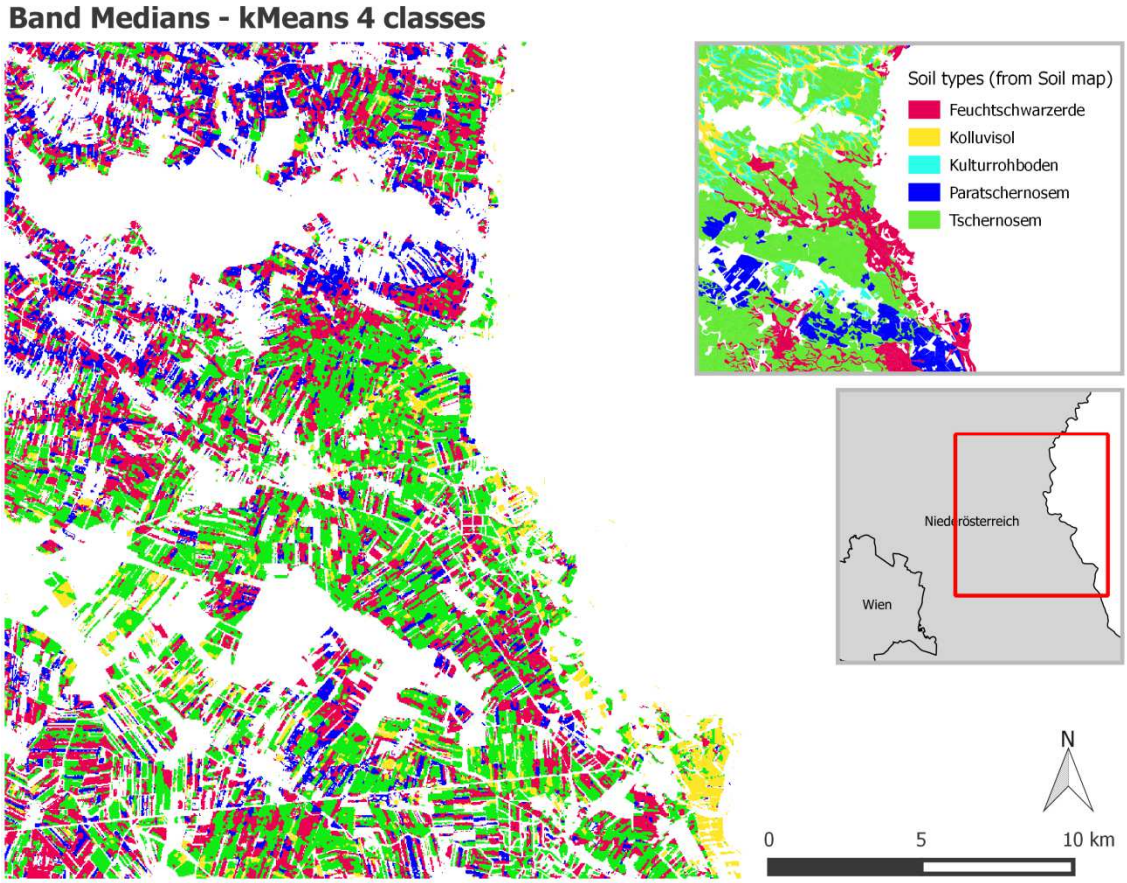


Figure 28: kMeans classification using "Band medians - dataset" - 4 classes.

Band Medians - kMeans 4 classes		
Internal analysis	mean Silhouette width	0,345
External analysis	adjusted Rand Index	0,048

The classification in Figure 29 shows four classes and was computed using the SOM algorithm. Compared with Figure 28, it shows that the two classification algorithms produce very similar results. The two maps are nearly equal and do not have major differences, which is also indicated by their close internal and external evaluation values.

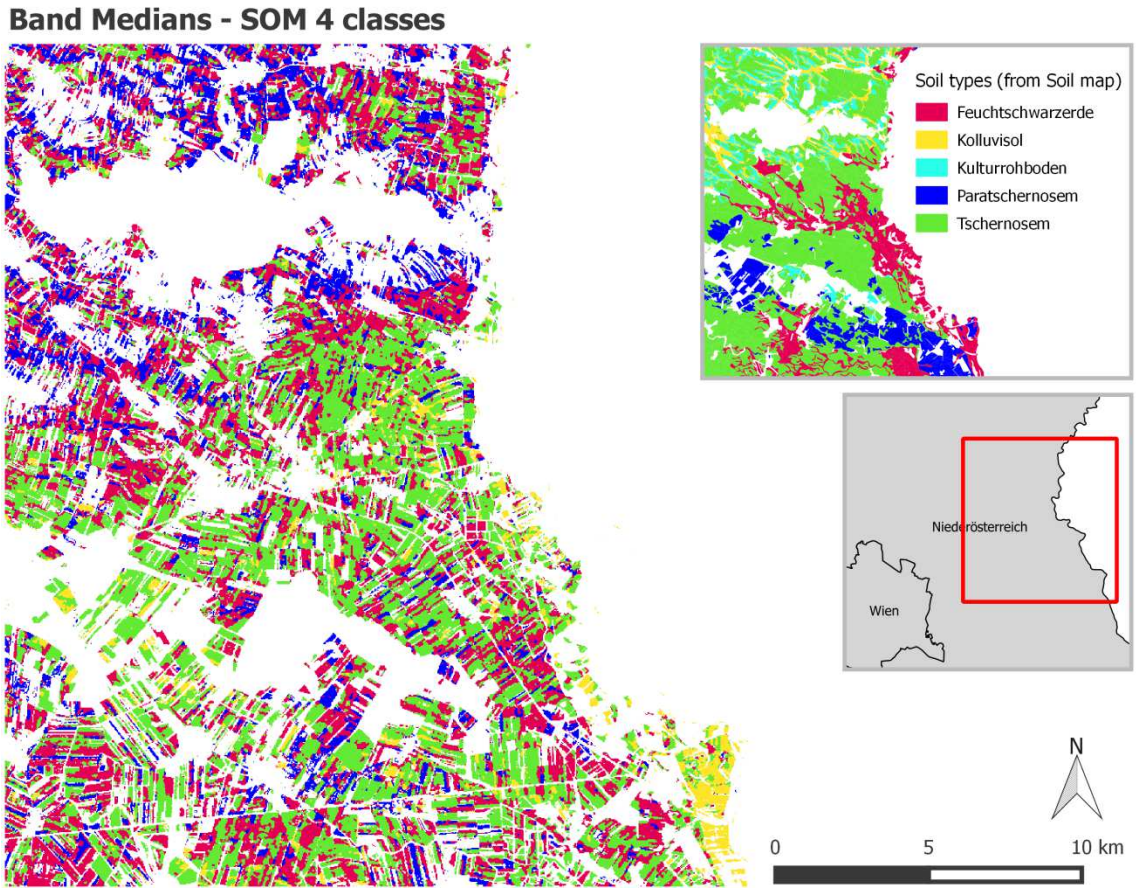


Figure 29: SOM classification using "Band medians - dataset" - 4 classes.

Band Medians - SOM 4 classes		
Internal analysis	mean Silhouette width	0,348
External analysis	adjusted Rand Index	0,045

3.3.2. "Fitted polynomial function - dataset"

The classification in Figure 30 shows three classes and was computed using the kMeans algorithm. The red area indicates very well Feuchtschwarzerde in the east but does not differentiate it with Tschernosem and Paratschernosem in the south. Both soil types - Tschernosem and Paratschernosem - are classified green in the classification. The northern part indicates well the diversity as it is shown in the reference map. Spatial variability is higher in the classification compared to the reference.

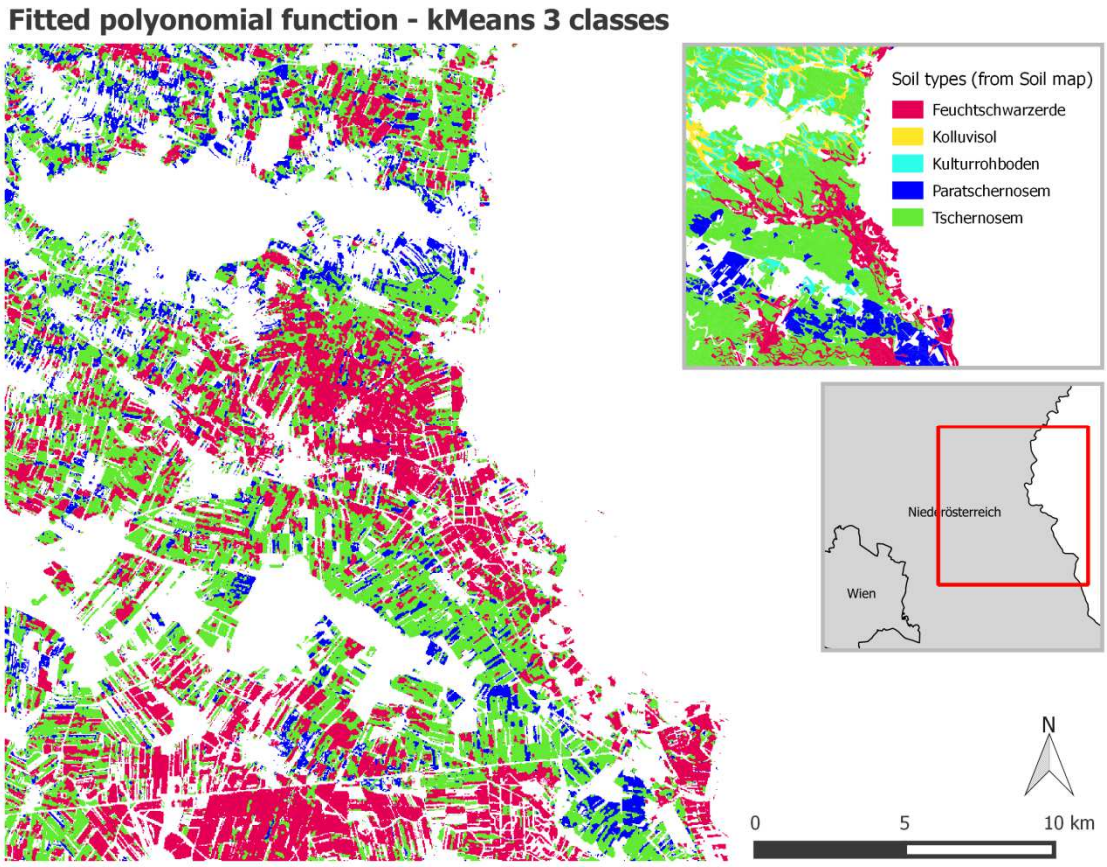


Figure 30: kMeans classification using "Fitted polynomial function - dataset" - 3 classes.

Fitted polynomial function - kMeans 3 classes		
Internal analysis	mean Silhouette width	0,484
External analysis	adjusted Rand Index	0,059

The classification in Figure 31 shows four classes and was computed using the kMeans algorithm. In comparison with Figure 30 the blue class expanded and the red area declined - the latter matching quite well with Feuchtschwarzerde from the reference map. The fourth class (yellow) is mostly present in the north, contributing to higher class variability in that region. Shapes and coarse patterns can be delineated very well from the classification. Higher spatial variability is visible in the classification compared with the reference map.

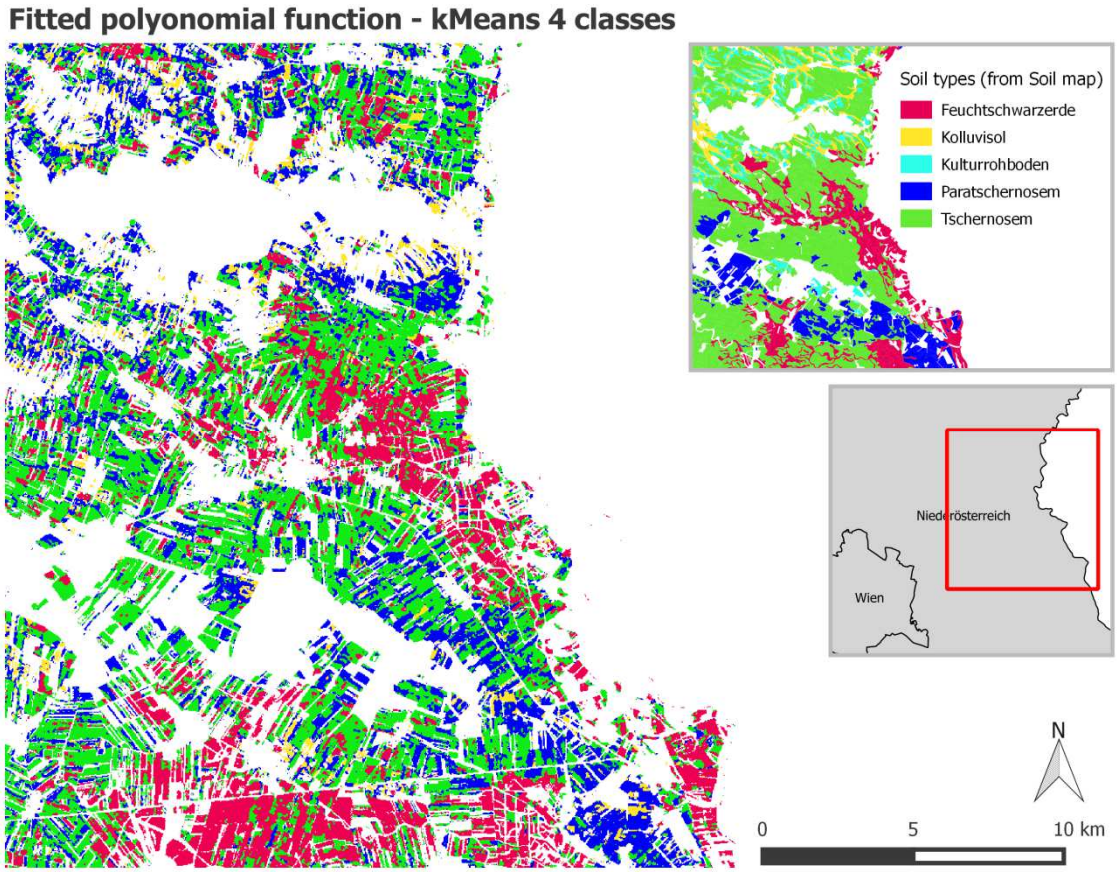


Figure 31: kMeans classification using "Fitted polynomial function - dataset" - 4 classes.

Fitted polynomial function - kMeans 4 classes		
Internal analysis	mean Silhouette width	0,468
External analysis	adjusted Rand Index	0,075

The classification in Figure 32 shows four classes and was computed using the SOM algorithm. In comparison with the classification in Figure 31 no major differences between the two classifications can be delineated. This is proven by the almost equal output values of the analysis.

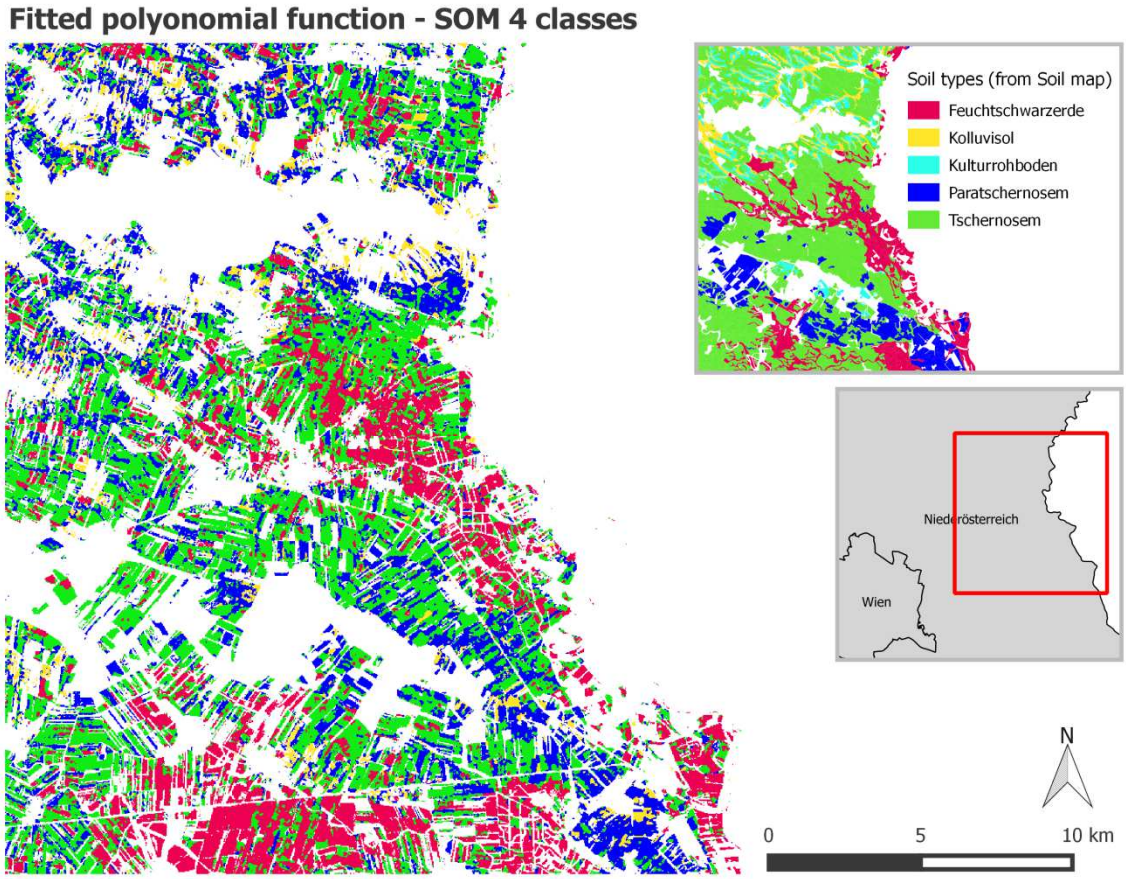


Figure 32: SOM classification using "Fitted polynomial function - dataset" - 4 classes.

Fitted polynomial function - SOM 4 classes		
Internal analysis	mean Silhouette width	0,467
External analysis	adjusted Rand Index	0,075

3.3.3. "Soil line - dataset"

The classification in Figure 33 shows three classes and was computed using the kMeans algorithm. Visually there are just two classes distinguishable, as the algorithm just assigned very few separated observations to the third class - therefore being invisible on the map. The algorithm apparently does not find more spatial patterns and differentiations in the homogeneous dataset to classify them extra. Hence this classification does not provide good information on spatial shapes and patterns - indicated as well by a low Rand Index value of 0,031.

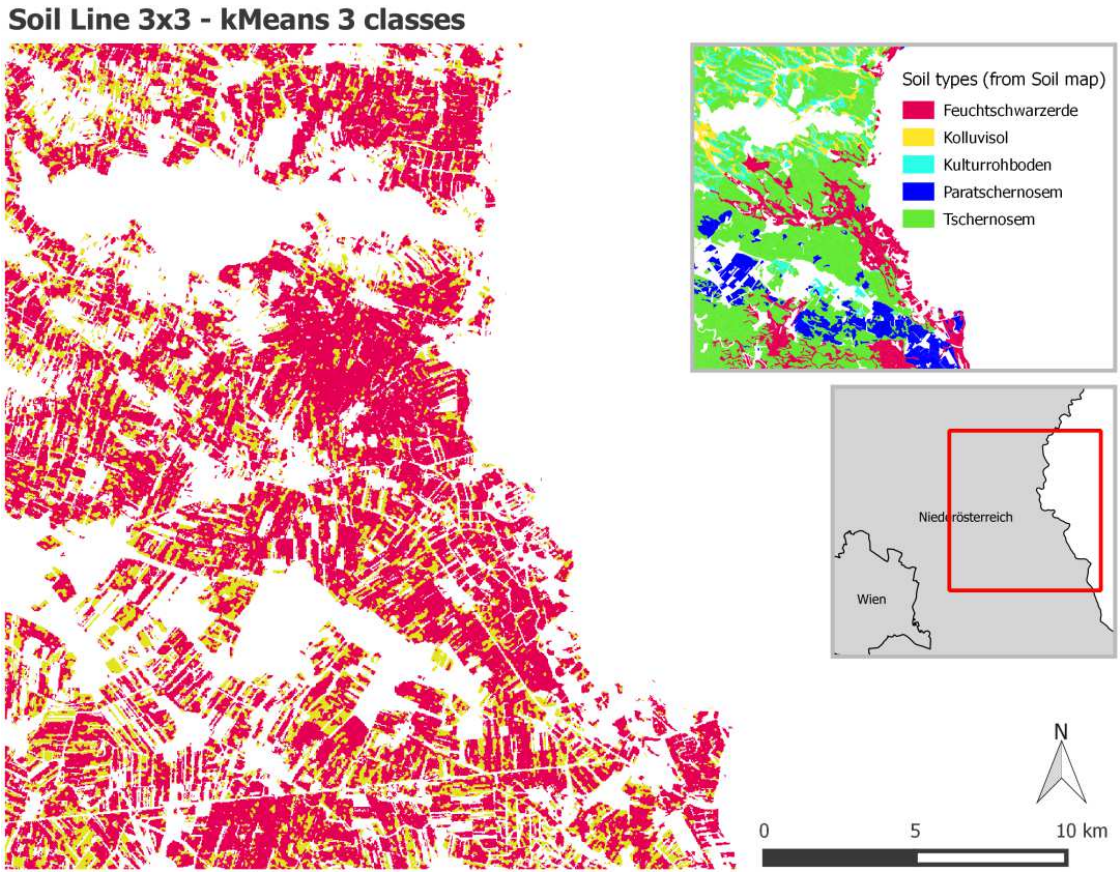


Figure 33: kMeans classification using "Soil Line 3x3 - dataset" - 3 classes.

Soil Line 3x3 - kMeans 3 classes		
Internal analysis	mean Silhouette width	0,538
External analysis	adjusted Rand Index	0,031

The classification in Figure 34 shows three classes and was computed using the kMeans algorithm. In comparison with the classification in Figure 33 the third class gains importance but according to the reference map, spatial shapes are still not represented well. This is proven by a low Rand Index value of 0,015.

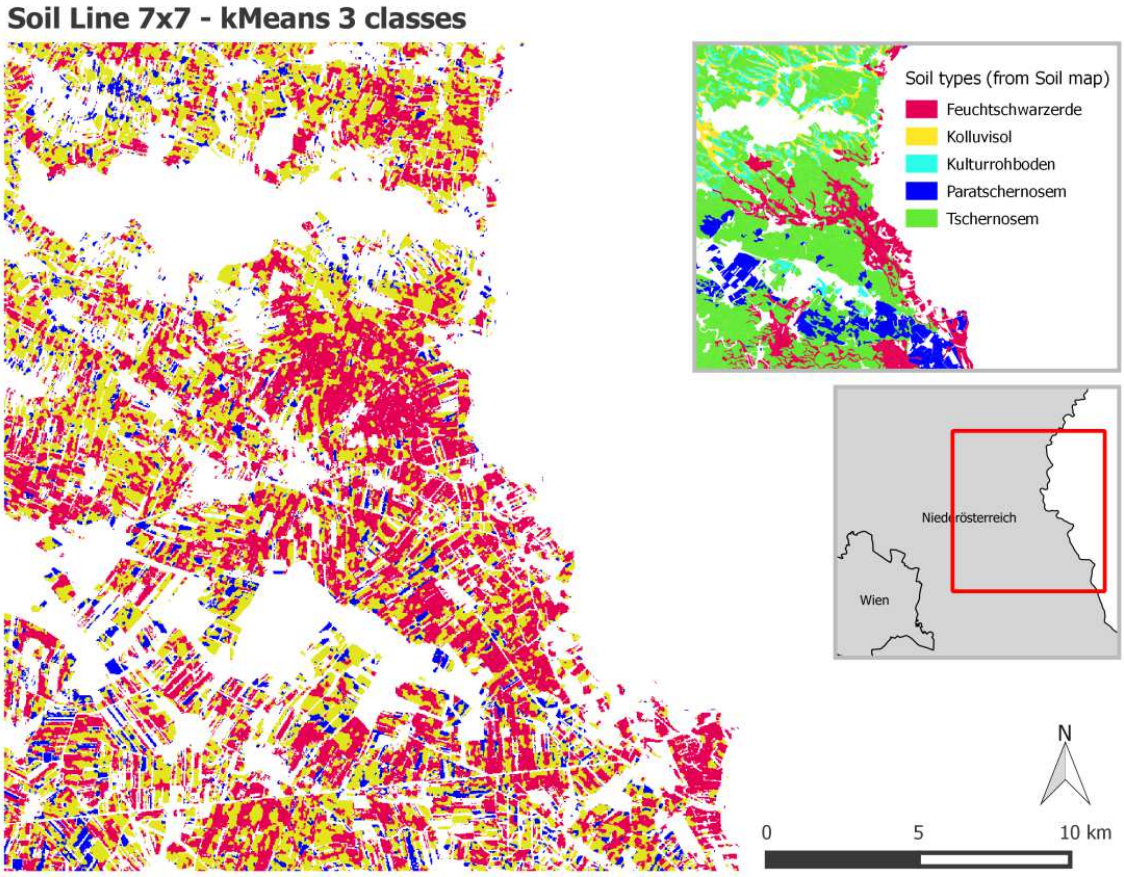


Figure 34: kMeans classification using "Soil Line 7x7 - dataset" - 3 classes.

Soil Line 7x7 - kMeans 3 classes		
Internal analysis	mean Silhouette width	0,502
External analysis	adjusted Rand Index	0,015

The classification in Figure 35 shows four classes and was computed using the kMeans algorithm. Analog to Figure 33 the homogeneous dataset apparently does not allow a better split of observations on the four classes as again just three classes are visible. The fourth - invisible class - has very few and separated observations assigned, which makes it impossible to delineate on the map. The low Rand Index indicates that the classification has no reliable information regarding spatial shapes and patterns.

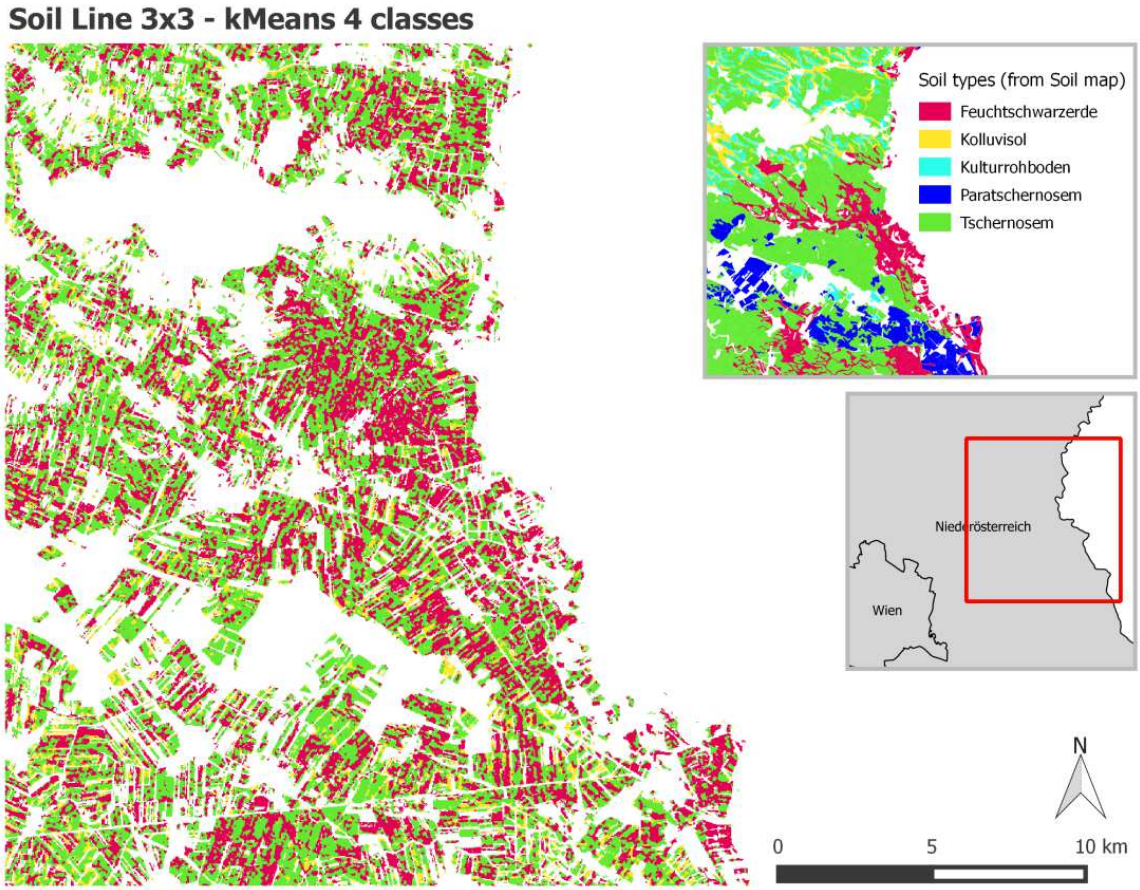


Figure 35: kMeans classification using "Soil Line 3x3 - dataset" - 4 classes.

Soil Line 3x3 - kMeans 4 classes		
Internal analysis	mean Silhouette width	0,469
External analysis	adjusted Rand Index	0,008

The classification in Figure 36 shows four classes and was computed using the kMeans algorithm. Spatial shapes cannot be delineated well, except for the northern part of the region, where spatial variability in the reference map is high as well. The Rand Index underlines the low quality of the classification concerning the representation of the spatial shapes of the soil map.

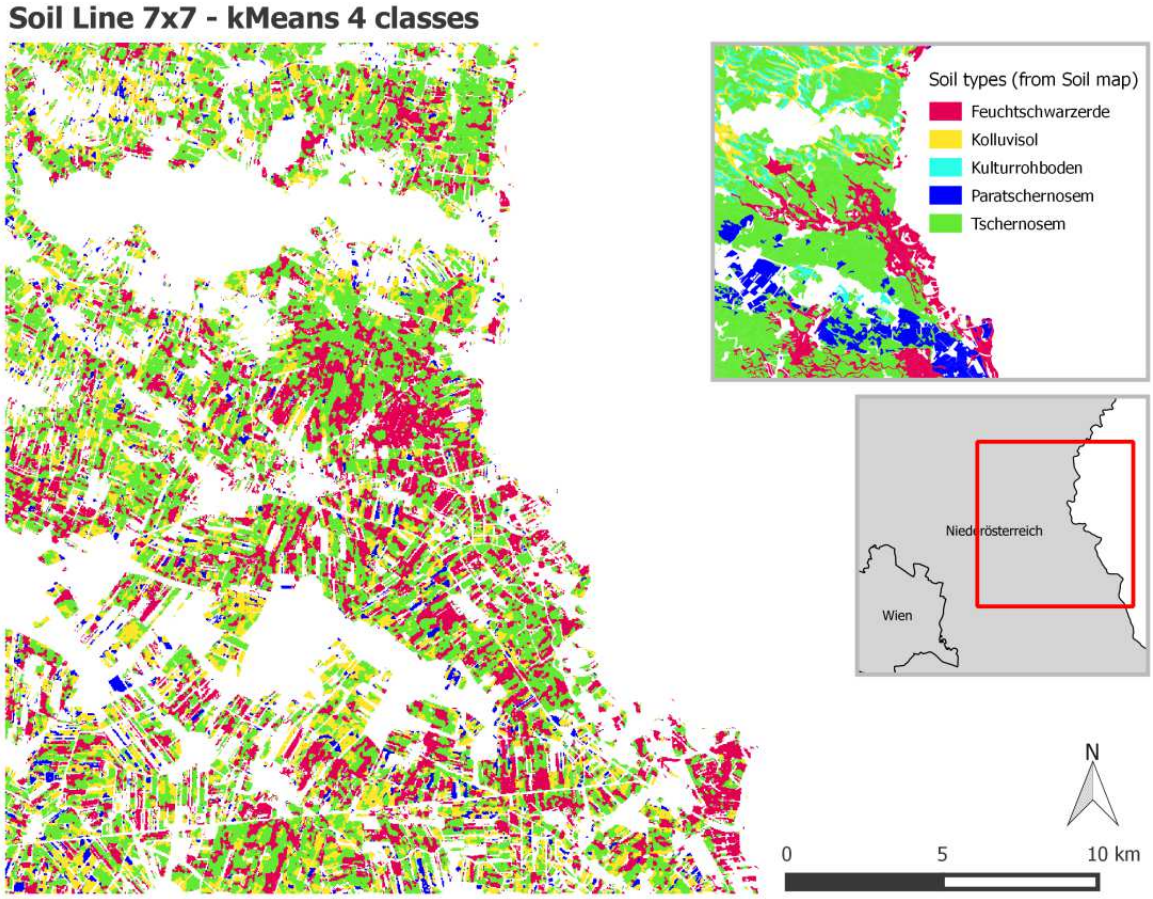


Figure 36: kMeans classification using "Soil Line 7x7 - dataset" - 4 classes.

Soil Line 7x7 - kMeans 4 classes		
Internal analysis	mean Silhouette width	0,480
External analysis	adjusted Rand Index	0,022

The classification in Figure 37 shows four classes and was computed using the SOM algorithm. The classifications in Figure 33 and Figure 35 already have shown that the "Soil Line 3x3 - dataset" is too homogeneous to successfully delineate various classes. In this classification just two major classes are visible. As the other two classes have again very few and spatially separated observations assigned, it is hard to visually detect them.

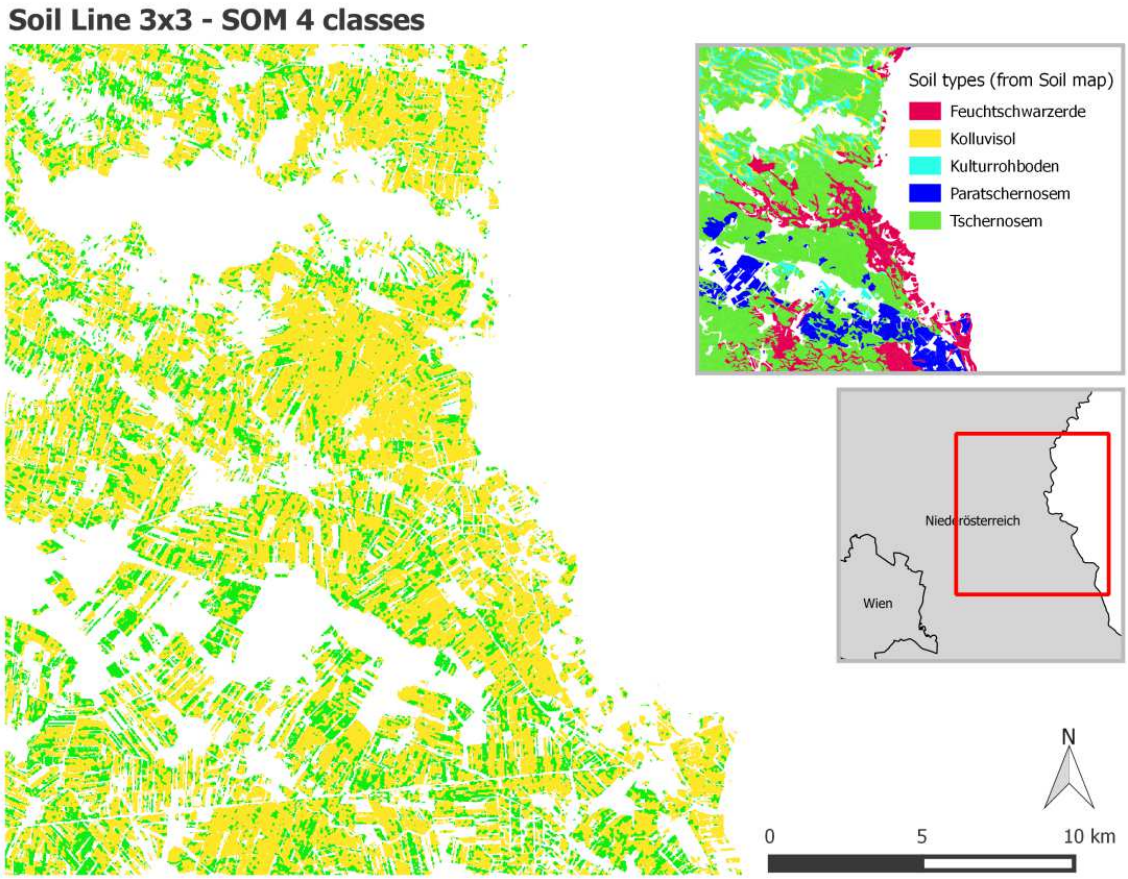


Figure 37: SOM classification using "Soil Line 3x3 - dataset" - 4 classes.

Soil Line 3x3 -SOM 4 classes		
Internal analysis	mean Silhouette width	0,538
External analysis	adjusted Rand Index	0,028

The classification in Figure 38 shows four classes and was computed using the SOM algorithm. Its classes are distributed equally over the area without major agglomerations and hence do not represent well the coarse spatial patterns of the reference map. A low Rand Index proves this point.

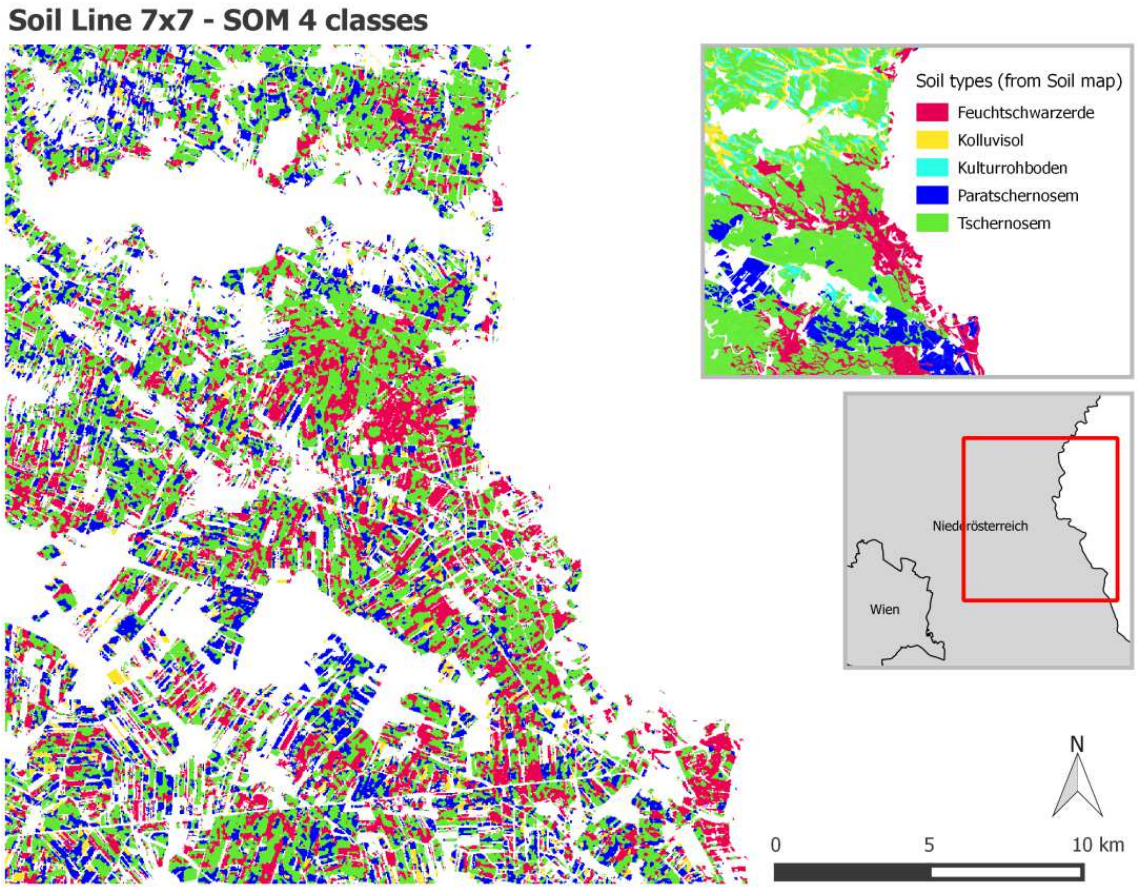


Figure 38: SOM classification using "Soil Line 7x7 - dataset" - 4 classes.

Soil Line 7x7 -SOM 4 classes		
Internal analysis	mean Silhouette width	0,479
External analysis	adjusted Rand Index	0,024

3.3.4. "Fitted polynomial function - dataset" - automatic / manual merge

As the "Fitted polynomial function - dataset" outperformed the other two datasets, it was used for further SOM classification. In this approach 25 classes (neurons) were calculated and merged to 4 or 5 classes. The merging was done either manually by visual assessment of the attributes of the neurons, or automatically using a hierarchical clustering approach. The latter was done by clustering the attributes of the 25 neurons and plotting them in a dendrogram (see Figure 40), enabling to delineate 4 or 5 different classes. A first assessment of the SOM clustering can be seen in Figure 39. Beside the number of pixels assigned to each neuron (upper left plot) an assessment of the quality of the 25 neurons is

illustrated (upper right plot). The upper right plot shows the mean distance of objects mapped to their corresponding neuron. The smaller the distances, the better the objects are represented by the neurons. The lower plot indicates the similarity between neighboring neurons - a low value indicates high similarity to one another.

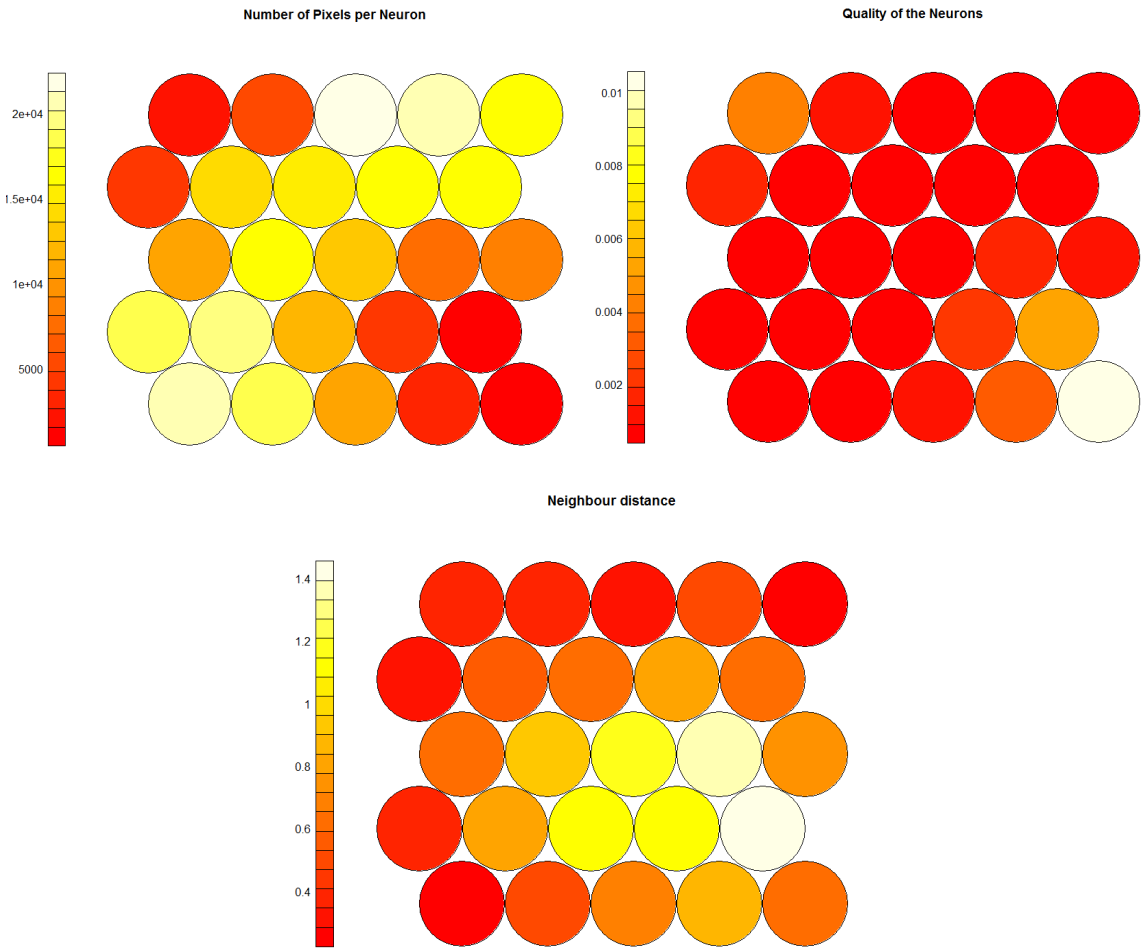


Figure 39: Upper left plot illustrates the number of pixels assigned to each neuron. The upper right plot shows the quality (homogeneity) of each neuron - where the lower the value the better the quality. The lower plot indicates the similarity between neighboring neurons.

The automatic merging approach - via hierarchical clustering - led to the following dendrogram - with the red line indicating the cut of the 4-class clustering (see Figure 42) and the blue line the cut of the 5-class clustering (see Figure 43).

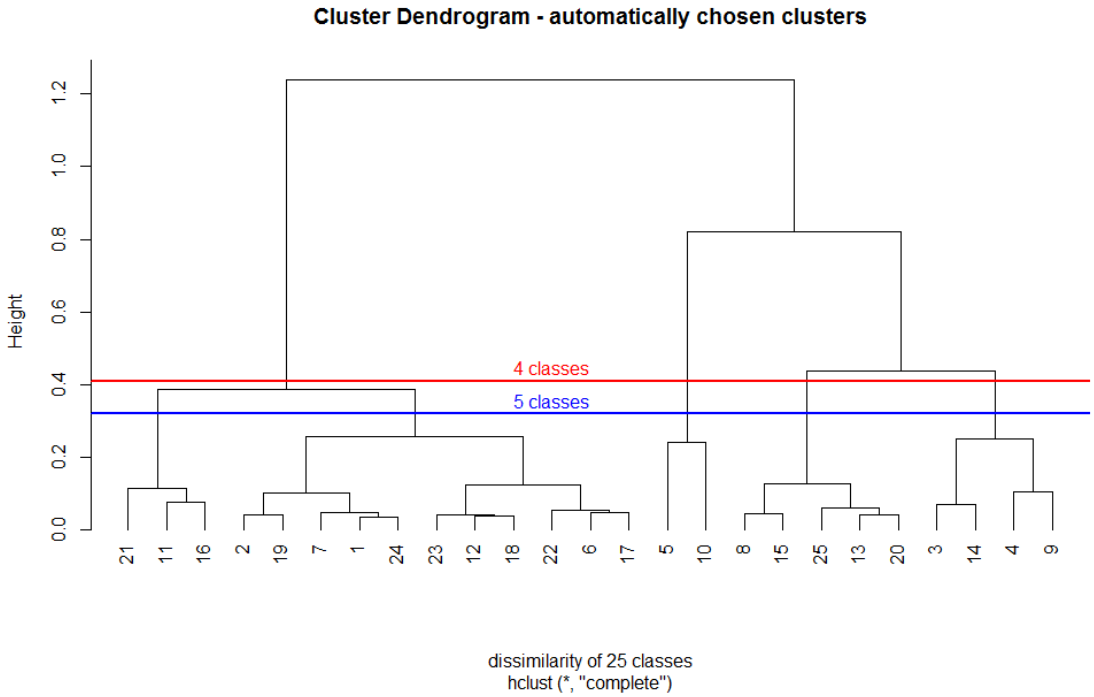


Figure 40: Dendrogram of the hierarchical clustering of the 25 neurons.

The merged neurons of the 4-class, and the 5-class clustering are shown in Figure 41. The merged classes are illustrated by the colour of the circles, which change from 4- to 5-class clustering. The neurons' attributes depicted in the pie charts are the same between the left and the right plot (they represent the same dataset and SOM clustering).

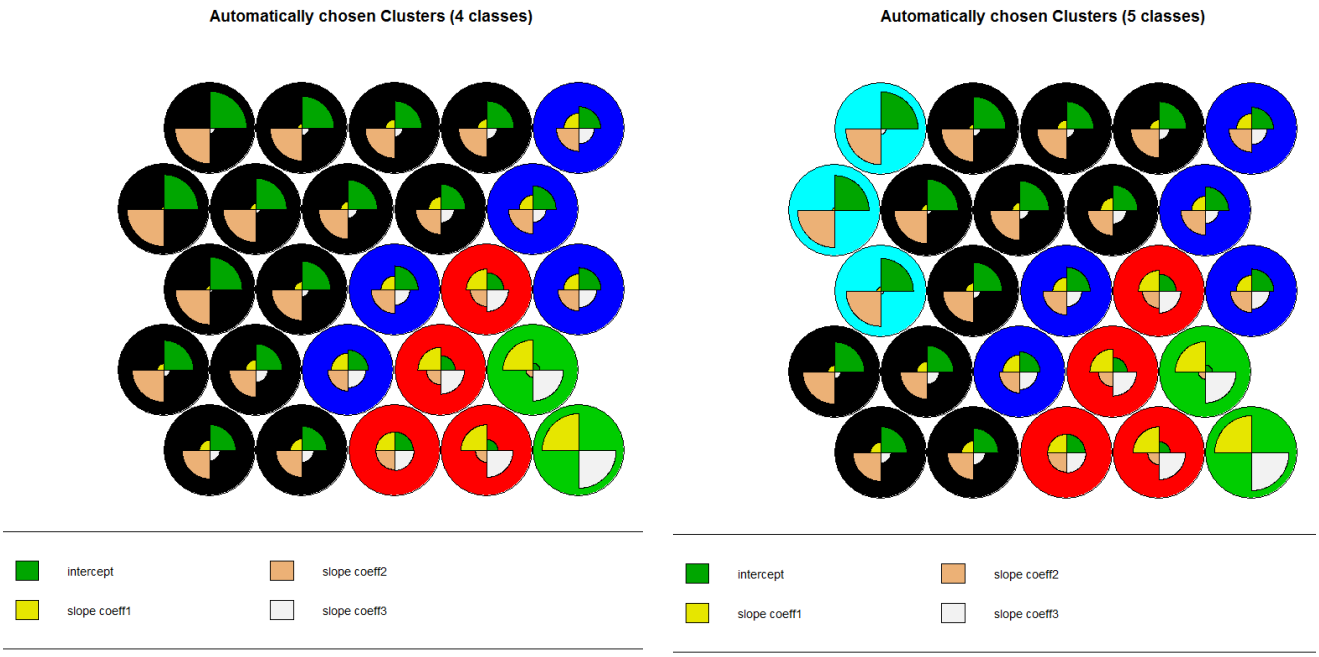


Figure 41: Automatically classified neurons and their attributes shown as pie charts inside the circles.

The automatic approach led to the maps presented in Figure 42 and Figure 43. The classification in Figure 42 shows four classes and was computed by the SOM algorithm. Four classes (red, blue, green and yellow) are visible but the yellow and green class have very few observations assigned to it. Visually the classification does not seem very promising - due to the fact that the red class spans over the whole area and represents every soil type shown in the reference map. A Rand Index of 0,087 is one of the best results in comparison with the other classifications and therefore does not prove this point.

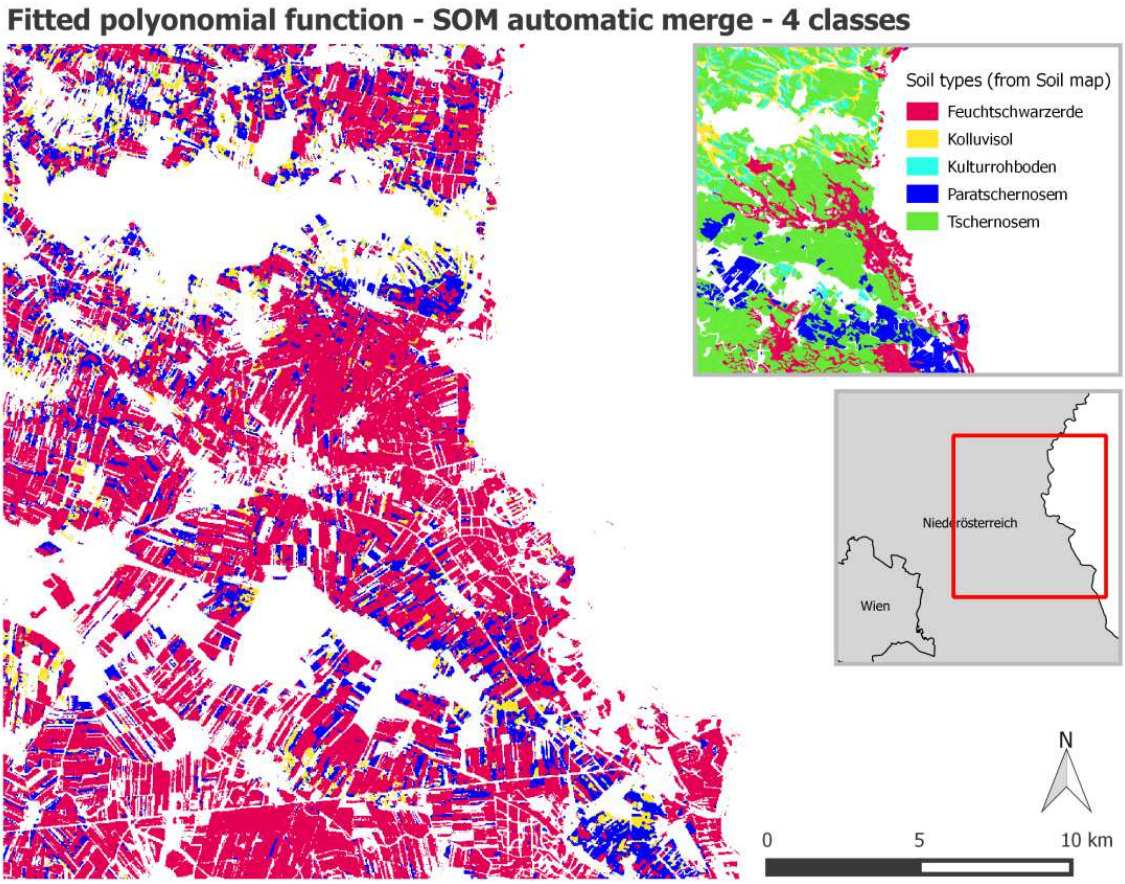


Figure 42: SOM automatic merge classification using "Fitted polynomial function - dataset" - 4 classes.

Fitted polynomial function - SOM automatic merge 4 classes		
Internal analysis	mean Silhouette width	0,452
External analysis	adjusted Rand Index	0,087

The classification in Figure 43 shows five classes and was computed using the SOM algorithm. Analog to the red area in Figure 42, the green area is very prominent, covering all soil types of the reference map. The northern part matches well with the variability shown in the reference map, whereas in the south coarse shapes visually do not agree with those depicted in the reference map. On the other hand a Rand Index of 0,104 is the best result and indicates that the best analytical agreement between the reference soil map and the classification is obtained with this classification.

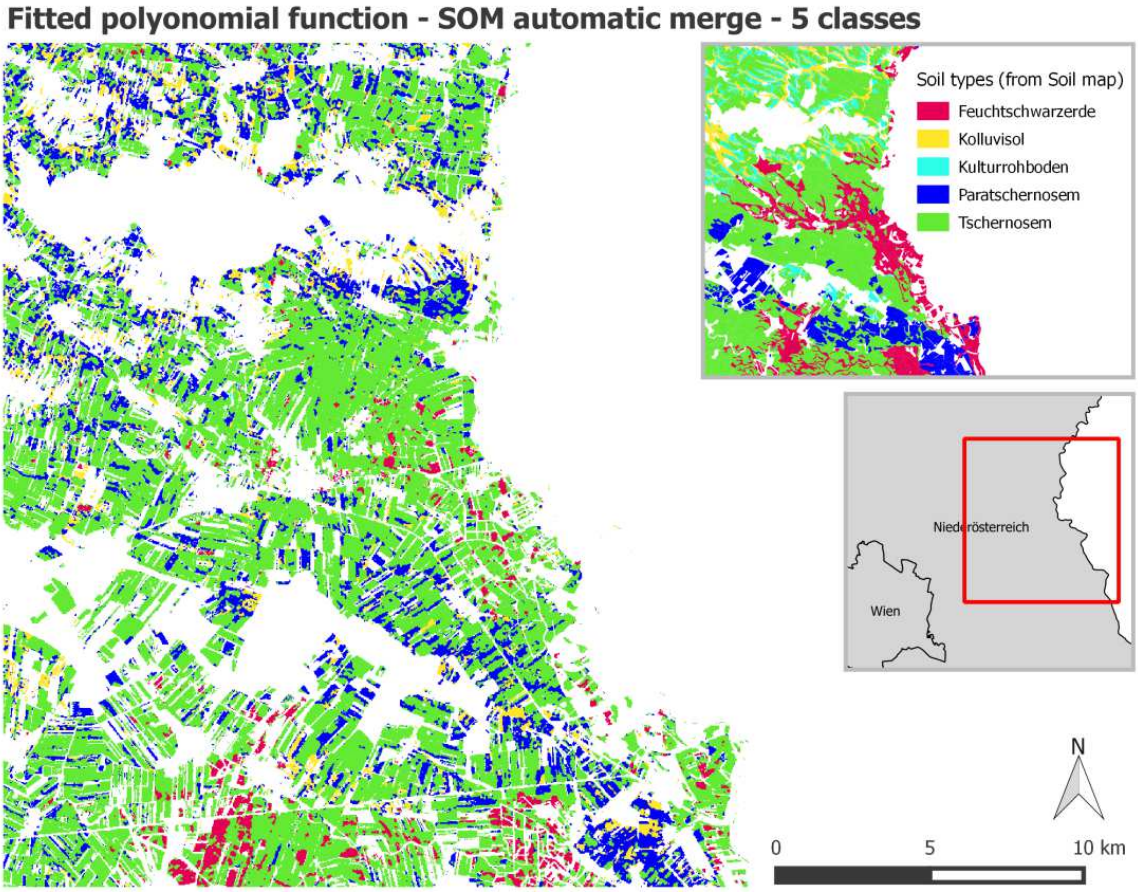


Figure 43: SOM automatic merge classification using "Fitted polynomial function - dataset" - 5 classes.

Fitted polynomial function - SOM automatic merge 5 classes		
Internal analysis	mean Silhouette width	0,390
External analysis	adjusted Rand Index	0,104

The manual merging of the neurons to 4 or respectively 5 classes was done visually by assessing the neurons attributes. Figure 44 shows the classification.

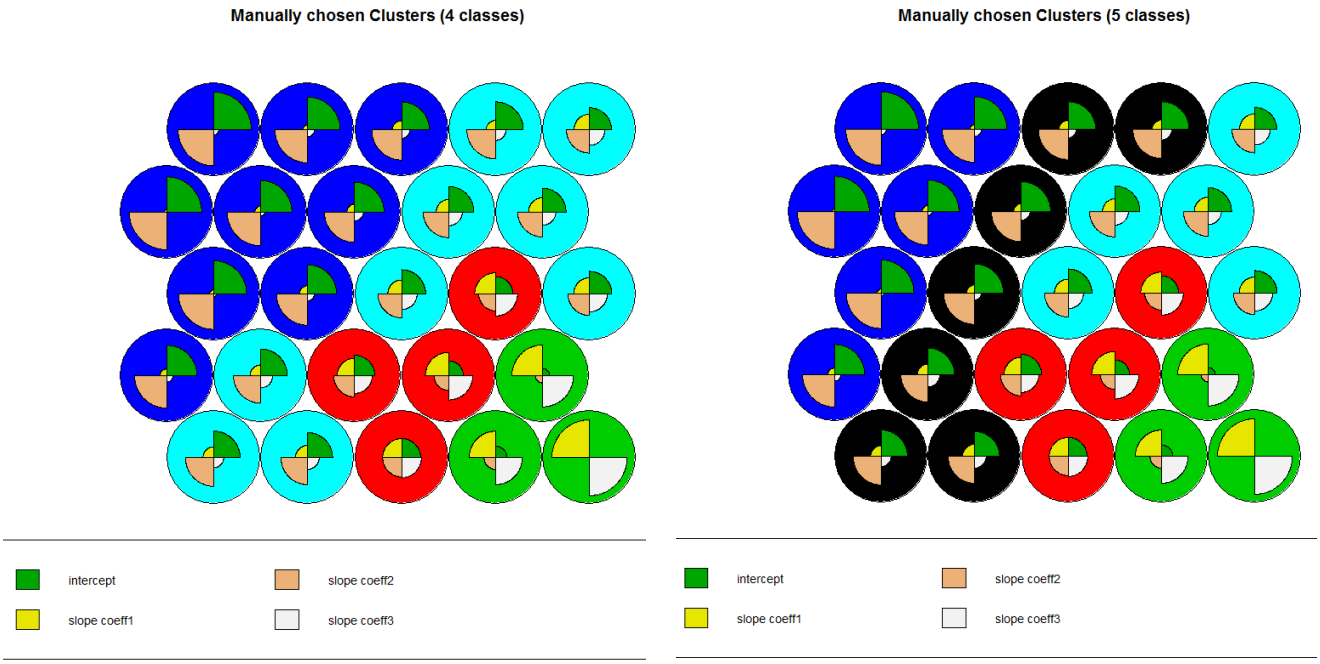


Figure 44: Manually classified neurons and their attributes shown as pie charts inside the circles.

The manual approach led to the following maps: Figure 45 and Figure 46.

The classification in Figure 45 shows four classes and was computed using the SOM algorithm. Coarse shapes can be delineated better than in the classification in Figure 42 although the Rand Index is nearly the half of it. Again the classification is lacking a balanced distribution of observations over all four classes - as one class (blue) only counts with few assigned pixels. The green area does a good job in delineating Paratschernosem and Tschernosem.

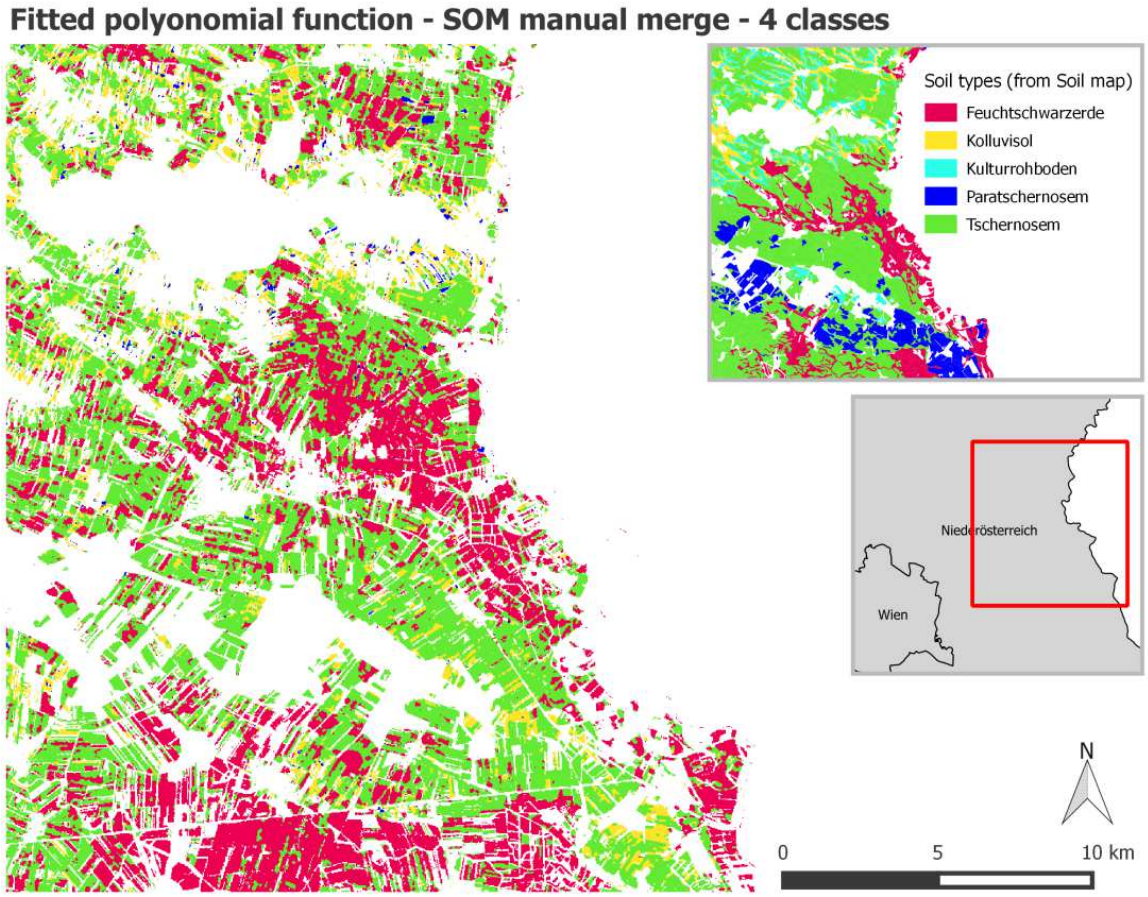


Figure 45: SOM manual merge classification using "Fitted polynomial function - dataset" - 4 classes.

Fitted polynomial function - SOM manual merge 4 classes		
Internal analysis	mean Silhouette width	0,461
External analysis	adjusted Rand Index	0,063

The classification in Figure 46 shows five classes and was computed using the SOM algorithm. It depicts very well coarse shapes and patterns of the soil map and shows higher spatial variability. The red area mostly represents Feuchtschwarzerde from the soil map, which visually seems very accurate. As well the northern area and its spatial variability is depicted well in the classification. The soil types Paratschernosem and Tschernosem are hardly to distinguish from the classification. The second best Rand Index among the classifications, proves the good result.

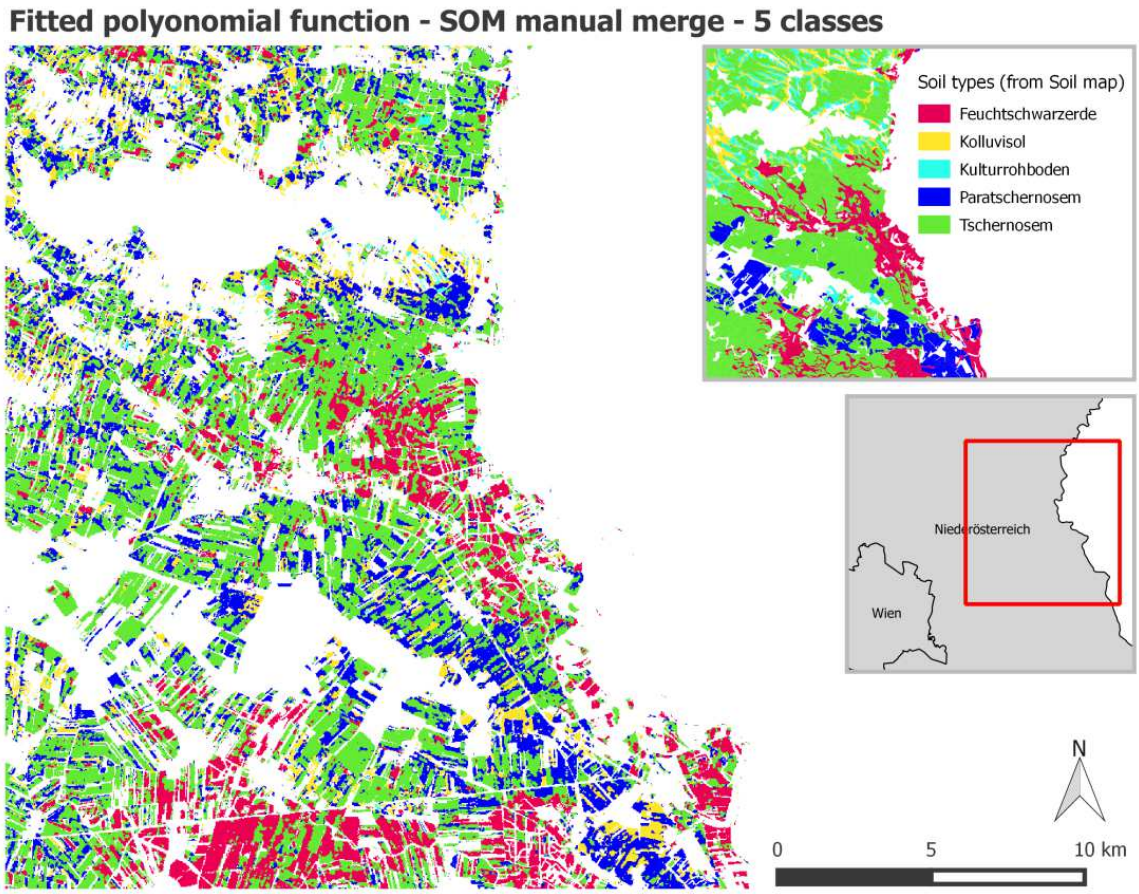


Figure 46: SOM manual merge classification using "Fitted polynomial function - dataset" - 5 classes.

Fitted polynomial function - SOM automatic manual 5 classes		
Internal analysis	mean Silhouette width	0,442
External analysis	adjusted Rand Index	0,094

An overview of the evaluation results of all classification maps is given in Table 7.

Table 7: Overview of the Cluster analysis.

Input dataset - algorithm & number of classes	Internal analysis	External analysis
	mean Silhouette width	adjusted Rand Index
Band Medians - kMeans 3 classes	0,356	0,057
Band Medians - kMeans 4 classes	0,345	0,048
Band Medians - SOM 4 classes	0,348	0,045
Fitted polynomial function - kMeans 3 classes	0,484	0,059
Fitted polynomial function - kMeans 4 classes	0,468	0,075
Fitted polynomial function - SOM 4 classes	0,467	0,075
Soil Line 3x3 - kMeans 3 classes	0,538	0,031
Soil Line 7x7 - kMeans 3 classes	0,502	0,015
Soil Line 3x3 -kMeans 4 classes	0,469	0,008
Soil Line 7x7 -kMeans 4 classes	0,480	0,022
Soil Line 3x3 -SOM 4 classes	0,538	0,028
Soil Line 7x7 -SOM 4 classes	0,479	0,024
Fitted polynomial function - SOM automatic 4 classes	0,452	0,087
Fitted polynomial function - SOM automatic 5 classes	0,390	0,104
Fitted polynomial function - SOM manual 4 classes	0,461	0,063
Fitted polynomial function - SOM manual 5 classes	0,442	0,094

At this point three maps were chosen to be investigated more in detail. The two best performing classifications according to the adjusted Rand Index and one more map ("Fitted polynomial function - SOM 4 classes") which visually, from the authors point of view, seems to be a good classification as well. Zonal statistics were computed for those three maps (see Figure 32, Figure 45 and Figure 46), where histograms show if the soil types from the reference soil map mostly belong to one clustering class or not. Figure 47 shows the histogram of the "Fitted polynomial function - SOM 4 classes" - classification. As the bars indicate, it is not a very clear classification, as most of the class numbers from the

classification cannot be assigned to one soil type and vice versa no soil type can be assigned to one class number.

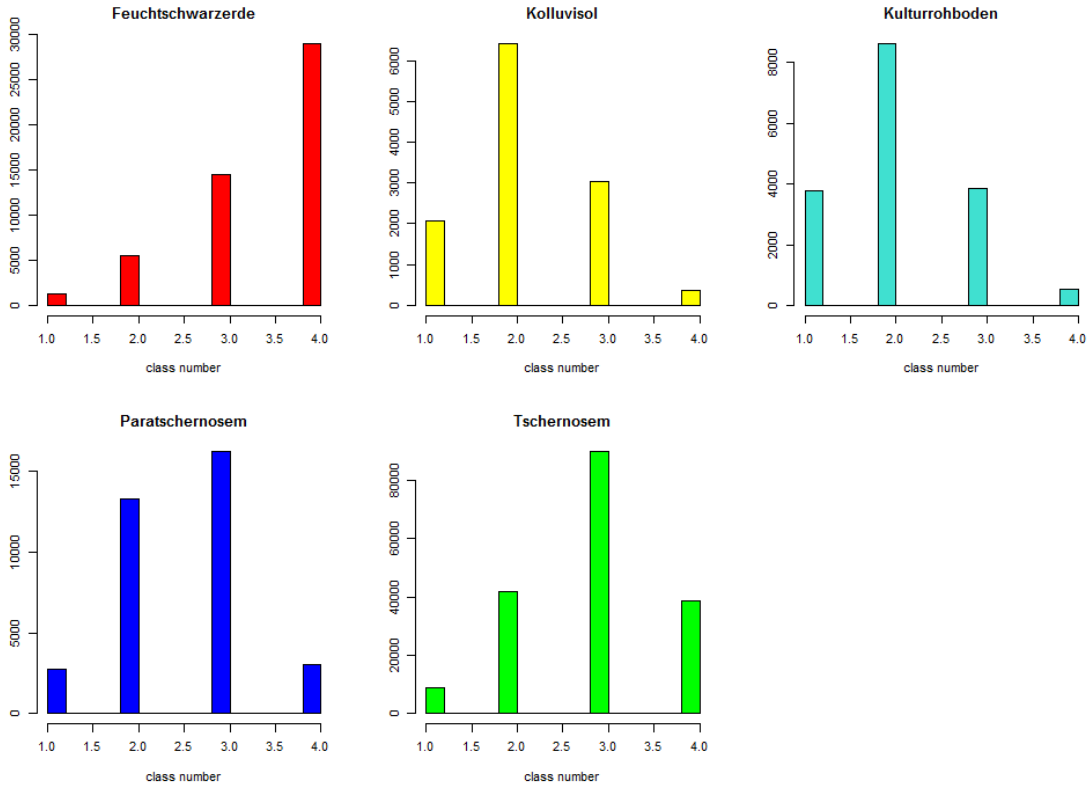


Figure 47: Zonal statistics of the "Fitted polynomial function - SOM 4 classes" - classification, colored according to the legend in Figure 6.

Figure 48 shows the histograms computed for the "Fitted polynomial function - SOM manual 5 classes" - classification. A similar result is drawn like in Figure 47, with no clear separability, when the classification is compared with the reference soil map.

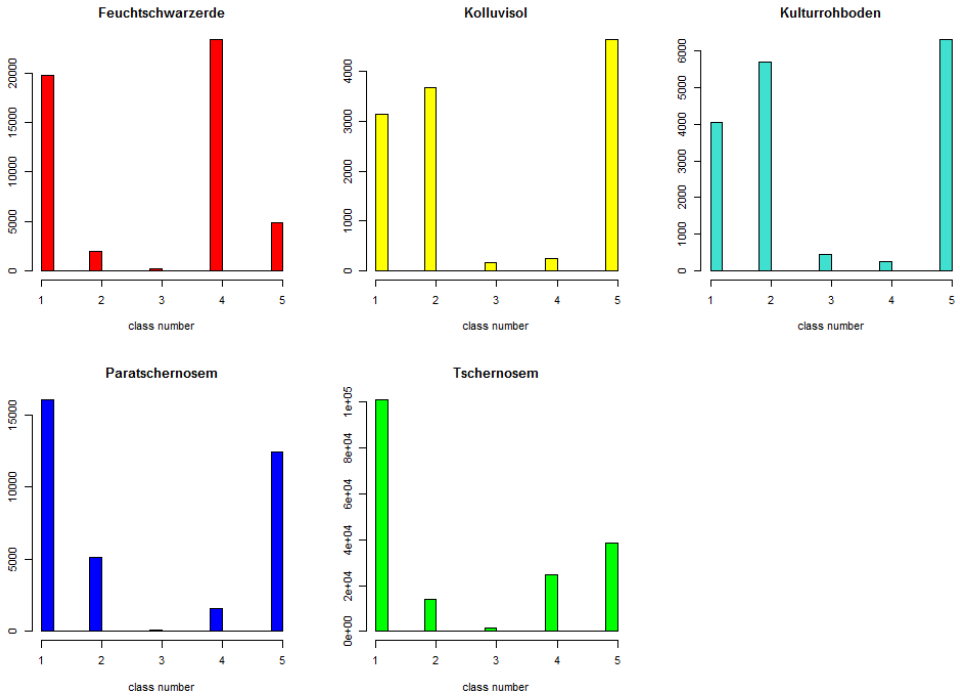


Figure 48: Zonal statistics of the "Fitted polynomial function - SOM manual 5 classes" - classification, colored according to the legend in Figure 6.

Figure 49 shows the histograms computed for the "Fitted polynomial function - SOM automatic 5 classes" - classification. Soil types Feuchtschwarzerde and Tschernosem can mostly be found in class number 1, keeping in mind that class 1 is present in each soil type.

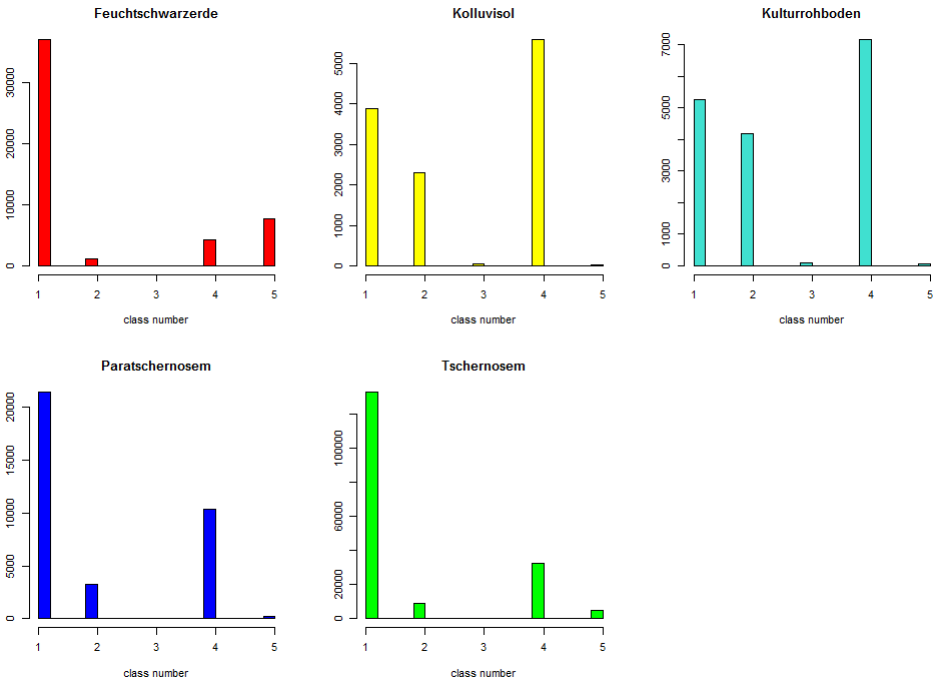


Figure 49: Zonal statistics of the "Fitted polynomial function - SOM automatic 5 classes" - classification, colored according to the legend in Figure 6.

4. Discussion and Conclusions

A ranking of both evaluation indices (mean Silhoutte width, adjusted Rand Index) shows that the different clustering outputs are mostly ranked according to their input dataset - the algorithm used has less influence (see Table 8 and Table 9).

Table 8: Ranking according to the mean Silhouette width.

Ranking according to the mean Silhouette width	Internal analysis	External analysis
	mean Silhouette width	adjusted Rand Index
Band Medians - kMeans 4 classes	0,345	0,048
Band Medians - SOM 4 classes	0,348	0,045
Band Medians - kMeans 3 classes	0,356	0,057
Fitted polynomial function - SOM automatic 5 classes	0,390	0,104
Fitted polynomial function - SOM manual 5 classes	0,442	0,094
Fitted polynomial function - SOM automatic 4 classes	0,452	0,087
Fitted polynomial function - SOM manual 4 classes	0,461	0,063
Fitted polynomial function - SOM 4 classes	0,467	0,075
Fitted polynomial function - kMeans 4 classes	0,468	0,075
Soil Line 3x3 -kMeans 4 classes	0,469	0,008
Soil Line 7x7 -SOM 4 classes	0,479	0,024
Soil Line 7x7 -kMeans 4 classes	0,480	0,022
Fitted polynomial function - kMeans 3 classes	0,484	0,059
Soil Line 7x7 - kMeans 3 classes	0,502	0,015
Soil Line 3x3 - kMeans 3 classes	0,538	0,031
Soil Line 3x3 -SOM 4 classes	0,538	0,028

In the ranking according to the best mean Silhouette width the "Soil line - dataset" performed best, second the "Fitted polynomial function - dataset" and last the "Band medians - dataset". Regarding the adjusted Rand Index a different result is drawn: the "Fitted polynomial function - dataset" outperformed the "Band medians - dataset" as the second and the "Soil line - dataset" as the last - as it can be seen in Table 9.

Table 9: Ranking according to the adjusted Rand Index.

Ranking according to the Rand Index	Internal analysis	External analysis
	mean Silhouette width	adjusted Rand Index
Soil Line 3x3 -kMeans 4 classes	0,469	0,008
Soil Line 7x7 - kMeans 3 classes	0,502	0,015
Soil Line 7x7 -kMeans 4 classes	0,480	0,022
Soil Line 7x7 -SOM 4 classes	0,479	0,024
Soil Line 3x3 -SOM 4 classes	0,538	0,028
Soil Line 3x3 - kMeans 3 classes	0,538	0,031
Band Medians - SOM 4 classes	0,348	0,045
Band Medians - kMeans 4 classes	0,345	0,048
Band Medians - kMeans 3 classes	0,356	0,057
Fitted polynomial function - kMeans 3 classes	0,484	0,059
Fitted polynomial function - SOM manual 4 classes	0,461	0,063
Fitted polynomial function - SOM 4 classes	0,467	0,075
Fitted polynomial function - kMeans 4 classes	0,468	0,075
Fitted polynomial function - SOM automatic 4 classes	0,452	0,087
Fitted polynomial function - SOM manual 5 classes	0,442	0,094
Fitted polynomial function - SOM automatic 5 classes	0,390	0,104

We can claim that most of the classification result (different soil maps) depends on the input dataset used and not on the clustering algorithm applied nor its number of classes. In some cases the similar clustering technique between the kMeans algorithm and the SOM algorithm can be delineated: Two nearly identical maps - "Band Medians - SOM 4 classes" (shown in Figure 28) and "Band Medians - kMeans 4 classes" (shown in Figure 29) - base on the same input dataset but have different algorithms applied. Their similarity is proven by their almost equal evaluation values. But there are still some major differences that can be shown in the performance of the two algorithms. The SOM algorithm, above all in the homogeneous "Soil line 3x3 - dataset" (see Figure 38), seeks to group its data points in mostly two major classes (although four are available). With the same input feature the kMeans algorithm (see Figure 35 for comparison) instead provides a slightly better distribution of data points over the four classes - still nearly leaving out one class. This performance of both algorithms is caused by the composition of the dataset which is very homogeneous (as shown in Figure 23, Figure 24 and Figure 25).

According to Kaufman and Rousseeuw (2005) none of the classifications has a poor internal evaluation value (mean Silhouette width <0.25). In order to select the best classification the focus can be put on the ARI, which is more relevant. The ARI evaluates the agreement between the existing soil map (considered as the truth) and the clustering output maps. The best obtained classifications according to the ARI were the merged SOM classifications (input feature: "Fitted polynomial function - dataset"), where the automatic

(hierarchical clustering) approach performed best. Although the ARI classifies Figure 43 as the best map, from the authors point of view it lacks two important points. First: in contrast to other maps the shapes on a coarse scale are not very well delineated. Second: the "green" class gained too much importance and other classes are under-represented. The author agrees that the "Fitted polynomial function - dataset" performed best, but wants to highlight the "SOM manual merge - 5 classes" classification (see Figure 46) and the 4-class kMeans and SOM classification (Figure 31 and Figure 32), as they visually make the best result. Those three maps show the shapes of the different soil types best, although all have problems with overestimating the "blue" area. Regarding this overestimation the "SOM manual merge - 5 classes" does the best job - having in mind that the other two maps have just 4 output classes and though automatically more data points fall into the "blue" class. The soil type "Feuchtschwarzerde" (see "red" area in Figure 6) was delineated best. "Kulturrohböden" wasn't very successful in its separation to other soil types. As well the "Paratschernosem" and the "Tschernosem" classes were difficult to distinguish. This last case is not very problematic as those two soil types are very similar to one another and are even merged together in different soil maps of the region (as it is done in other layers of the reference shape file for example). Regarding the separability of the soil types the zonal histograms (Figure 47 - Figure 49) didn't show a very successful result, with the "Fitted polynomial function - SOM automatic merge - 5 classes" classification (having the best ARI as well) leaving the best impression.

For unsupervised classification of soil maps, the use of bare soil reflectance based indicators, such as the fit of a polynomial function on the spectral bands, is recommended. This indicator obtains the best results and hence proves the point that it represents soil specific information better than other indicators. A classification algorithm - like the SOM algorithm - which offers the possibility to assess class-characteristics and merge similar classes, enables more sophisticated ways of cluster assessment and though can be recommended. Coarse patterns of soil types can be delineated from the resulting maps. All the classifications show more spatial variability than the reference soil map, which is a good sign that soil classification using remote sensing techniques contributes more detailed spatial information than traditional in-situ soil mapping techniques. This factor is important above all in the Marchfeld region due to its small scale soil variability. The use or additional integration of remote sensing information into traditional in-situ soil surveys, therefore offers the possibility to enhance the quality and spatial resolution of soil maps.

4.1. Outlook

The applied methodology enables to gain an overview of the distribution, the amount of soil types and their shapes. On its basis, maps can be delineated on a coarse scale at low costs. The use of additional information and combinations with other methods - such as regression trees and generalized linear models - are recommended to improve map accuracy (Mulder et al., 2011, p. 12; cf. Wulf et al., 2015). As there always has to be a trade-off between cost-effectiveness and map-accuracy, the presented methodology will be adequate for some purposes. For other uses, more detailed and accurate needs, additional information (such as digital elevation models or higher spectral resolution data - which might not be for free) will be necessary to fulfill this purpose. In this case the presented methodology will be insufficient. Overall remote sensing techniques will gain importance

in the future, being a crucial technology for a lot of environmental issues and agricultural uses - a lot of development can be expected in this field of science.

4.2. Lessons learned

All of the work was done with the open source software "R" and "QGIS". The first program requires coding skills. Having no coding experience before, this was quite a challenging task in the beginning, but with time more and more insights were gained and it became easier. A very important factor was the fast and good help of several internet platforms and tutorials. The author really learned to appreciate R as a very stable and powerful tool for everything dealing with any kind of data. Concerning this aspect QGIS also promises to fulfill a lot of needs (above all with its open-source toolboxes) and as it has an user-friendlier interface than R, some of the processings were initially intended to be done there. Unfortunately over the time there were several functions which weren't working as they should and help or solutions were not found easily. That is why everything - except for the visualization of the maps - was done in the R environment in the end.

BIBLIOGRAPHY

AG Boden, 1994. Bodenkundliche Kartieranleitung. Hannover.

ArcGIS, 2015. ArcGIS [WWW Document]. URL https://www.arcgis.com/features/ (accessed 20.9.15).

ArcGIS - Using the NDVI process [WWW Document], n.d. URL http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Using_the_NDVI_process (accessed 24.7.15).

BelVecchioUK, 2012. Rand Index in Statistics - A Worked Example - Cluster Analysis.

Bento, C., Cardoso, A., Dias, G., 2005. Progress in Artificial Intelligence: 12th Portuguese Conference on Artificial Intelligence, EPIA 2005, Covilha, Portugal, December 5-8, 2005, Proceedings. Springer Berlin Heidelberg.

bmlfuw - Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, 2015. Getreideanbau und Getreidearten in Österreich [WWW Document]. URL http://www.bmlfuw.gv.at/land/produktion-maerkte/pflanzliche-produktion/getreide/Getreide.html (accessed 21.7.15).

Bockheim, J.G., Gennadiyev, A.N., Hammer, R.D., Tandarich, J.P., 2005. Historical development of key concepts in pedology. Geoderma 124, 23–36. doi:10.1016/j.geoderma.2004.03.004

BOKU, n.d. MUbiL [WWW Document]. URL http://mubil.boku.ac.at/?page_id=5 (accessed 22.7.15).

Bouma, J., Broll, G., Crane, T.A., Dewitte, O., Gardi, C., Schulte, R.P., Towers, W., 2012. Soil information in support of policy making and awareness raising. Curr. Opin. Environ. Sustain. 4, 552–558. doi:10.1016/j.cosust.2012.07.001

Davies, D.L., Bouldin, D.W., 1978. Cluster separation measure 1, 224–227.

Dematte, J.A.M., Huete, A.R., Ferreira Jr, G., Nanni, M.R., Alves, M.C., Fiorio, P.R., others, 2009. Methodology for bare soil detection and discrimination by Landsat TM image. Open Remote Sens. J. 2.

Dunn, J.C., 1974. Well-Separated Clusters and Optimal Fuzzy Partitions. J. Cybern. 4, 95–104. doi:10.1080/01969727408546059

ERDAS IMAGINE, 2015. ERDAS IMAGINE [WWW Document]. URL http://www.hexagongeospatial.com/products/producer-suite/erdas-imagine (accessed 20.9.15).

ESA, 2015a. Sentinel-1 - Overview - Sentinel Online [WWW Document]. URL https://sentinel.esa.int/web/sentinel/missions/sentinel-1/overview (accessed 17.7.15).

ESA, 2015b. Sentinel Data Access Overview - Sentinel Online [WWW Document]. URL https://sentinel.esa.int/web/sentinel/sentinel-data-access (accessed 17.7.15).

ESA, 2015c. Sentinel-2 - Missions - Sentinel Online [WWW Document]. URL https://sentinel.esa.int/web/sentinel/missions/sentinel-2 (accessed 11.5.15).

ESRI, n.d. Majority Filter [WWW Document]. URL http://resources.esri.com/help/9.3/arcgisengine/java/Gp_ToolRef/spatial_analyst_tools/majority_filter.htm (accessed 3.8.15).

Estes, J., 2005. The History of Remote Sensing (By John E. Estes 1999, Last Updated 2005) [WWW Document]. URL http://www.geog.ucsb.edu/~jeff/115a/remotesensinghistory.html (accessed 17.7.15).

European Commission - Joint Research Center, 2014. Soil Awareness Raising [WWW Document]. URL http://eusoils.jrc.ec.europa.eu/Awareness/ (accessed 16.7.15).

FAO, 2015. Soil is a non-renewable resource 4.

Fox, G.A., Sabbagh, G.J., Searcy, S.W., Yang, C., 2004. An automated soil line identification routine for remotely sensed images. Soil Sci. Soc. Am. J. 68, 1326–1331.

Green, P., 2010. Self Organizing Maps (Part 1).

Hartemink, A.E., Krasilnikov, P., Bockheim, J.G., 2013. Soil maps of the world. Geoderma 207-208, 256–267. doi:10.1016/j.geoderma.2013.05.003

Hartemink, A.E., McBratney, A.B., Cattle, J.A., 2001. Developments and trends in soil science: 100 volumes of Geoderma (1967–2001). Geoderma 100, 217–268.

Holben, B.N., 1986. Characteristics of maximum-value composite images from temporal AVHRR data. Int. J. Remote Sens. 7, 1417–1434. doi:10.1080/01431168608948945

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218. doi:10.1007/BF01908075

Institut Cartogràfic i Geològic de Catalunya [WWW Document], n.d. URL http://www.icc.cat/eng/Home-ICC/Mapes-escolars-i-divulgacio/Preguntes-frequeents/Que-es-NDVI (accessed 24.7.15).

Jenny, H., 1941. Factors of soil formation. N. Y. McGraw-Hill.

Jensen, J., 2007. Remote Sensing of the Environment: An Earth Resource Perspective., 2nd ed. Prentice Hall, New Jersey.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C., Silverman, R., Wu, A.Y., 2002. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. Pattern Anal. Machiine Intell. 881–892.

Kaufman, L., Rousseeuw, P.J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Series in Probability and Statistics.

kmeans clustering, 2015. Wikipedia Free Encyclopedia.

Kromp-Kolb, H., Formayer, H., Eitzinger, J., 2007. Potentielle Auswirkungen und Anpassungsmaßnahmen der Landwirtschaft an den Klimawandel im Nordosten Österreichs (Weinviertel-Marchfeld Region).

Kwedlo, W., 2011. A clustering method combining differential evolution with the K-means algorithm. Pattern Recognit. Lett. 32, 1613–1621. doi:10.1016/j.patrec.2011.05.010

Lu, J.F., Tang, J.B., Tang, Z.M., Yang, J.Y., 2008. Hierarchical initialization approach for K-Means clustering. Pattern Recognit. Lett. 29, 787–795. doi:10.1016/j.patrec.2007.12.009

Marchfeldkanal [WWW Document], n.d. URL http://www.marchfeldkanal.at/home.htm (accessed 30.9.15).

Miehlich, G., 2009. Bodenbewusstsein - ein Schlüssel zur Förderung des Bodenschutzes (No. 22), NNA-Berichte. Alfred Toepfer Akademie für Naturschutz.

Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping — A review. Geoderma 162, 1–19. doi:10.1016/j.geoderma.2010.12.018

Pena, J.M., Lozano, J.A., Larranaga, P., 1999. An empirical comparison of four initialization methods for the K-Means algorithm.

QGIS, 2015. QGIS [WWW Document]. URL http://www.qgis.org/de/site/ (accessed 20.9.15).

Rand, W.M., 1971. Objective Criteria for the Evaluation of Clustering Methods. J. Am. Stat. Assoc. 66, 846. doi:10.2307/2284239

Richardson, A.J., Wiegand, C.L., 1977. Distinguishing Vegetation from Soil Background Information. Photogramm. Eng. Remote Sens. 43.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation.

Santos, J.M., Embrechts, M., 2009. On the use of the adjusted rand index as a metric for evaluating supervised classification, in: Artificial Neural networks–ICANN 2009. Springer, pp. 175–184.

Schowengerdt, R.A., 2007. Remote Sensing: Models and Methods for Image Processing. Academic Press.

SEOS Project, n.d. Spectral signatures of soil, vegetation and water, and spectral bands of LANDSAT 7.

Sommer, E., Reinthaler, D., Höbaus, E., 2009. Marchfeld Gemüse.

Spectral Response of Landsat 7, n.d.

Stolbovoy, V., Maréchal, B., Jones, A., Rusco, E., Montanarella, L., 2008. Climate change - soil can make a difference! Presented at the Climate change - can soil make a difference?, Brussels.

Thaler, S., Eitzinger, J., Dubrovsky, M., Trnka, M., n.d. Climate change impacts on selected crops in Marchfeld, Eastern Austria.

The International Union of Soil Sciences - IUSS [WWW Document], n.d. URL http://www.iuss.org/index.php?article_id=22 (accessed 14.7.15).

USGS, 2015. Remote Sensing Phenology [WWW Document]. URL http://phenology.cr.usgs.gov/ndvi_foundation.php (accessed 1.10.15).

UW-Madison Satellite Meteorology [WWW Document], n.d. URL http://profhorn.meteor.wisc.edu/wxwise/satmet/lesson3/ndvi.html (accessed 24.7.15).

Wehrens, R., Buydens, L.M.C., 2007. Self- and Super-organising Maps in R: the kohonen package.

Wulf, H., Mulder, T., Schaepman, M.E., Keller, A., Jörg, P.C., 2015. Remote Sensing of Soils.