Student thesis series INES nr 510

Soil organic carbon prediction using multispectral Sentinel-2 data and the LUCAS topsoil database

Philipp Hansen

2016
Department of
Physical Geography and Ecosystem Science
Lund University
Sölvegatan 12
S-223 62 Lund
Sweden

Philipp Hansen (2020)

Soil organic carbon prediction using multispectral Sentinel-2 data and the LUCAS topsoil database

Bachelor degree thesis, 15 credits in Physical Geography and Ecosystem Analysis

Department of Physical Geography and Ecosystem Science, Lund University

Level: Bachelor of Science (BSc)

Disclaimer

This document describes work undertaken as part of a program of study at the University of Lund. All views and opinions expressed herein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Soil organic carbon prediction using multispectral Sentinel-2 data and the LUCAS topsoil database

Philipp Hansen

Bachelor thesis, 15 credits in Physical Geography and Ecosystem Analysis

Supervisor:
Marko Scholze,
Department of Physical Geography and Ecosystem Science, Lund University

Industry supervisor:
Qiang Wang,
Vultus AB

Exam committee:
Jing Tang,
Department of Physical Geography and Ecosystem Science, Lund University
Stefan Olin,
Department of Physical Geography and Ecosystem Science, Lund University

Abstract

Its carbon sink potential as well as soil fertility benefits make organic carbon a soil variable for which reliable quantification methods are sought. This thesis work aims at investigating the possibility of adapting a large soil spectral library to build models for SOC predictions with remotely sensed, multispectral data. For this purpose, the continental-scale LUCAS topsoil database was spectrally resampled to simulate the reflectance measured by the Sentinel-2 satellite. Multivariate partial least squares regression models were created based on the spectrally resampled LUCAS database (i) for all mineral cropland soil samples and (ii) for a regional subset of the mineral cropland samples relative to location of the validation samples in southern Sweden. The global model was poor (RPD = 1.09) in relation to a comparable model produced with the original spectral information. This outcome was related to the insufficient spectral information of Sentinel-2 type data to account for the variability of soil chromophores within this large dataset. Despite the reduced extent and a sample size (n = 70) that is comparable to moderately successful SOC modelling attempts, the regional model yielded only a slight performance improvement (RPD = 1.12). Reasons for this outcome could be the spatially dispersed sampling strategy used to collect the LUCAS database or the high sand content of the samples. Both models failed to produce reasonable predictions of the validation dataset. The investigation of the difference between spectrally resampled LUCAS reflectance and remotely sensed Sentinel-2 reflectance revealed that the data of these two measurement methods are not readily compatible.

Acknowledgements

The topic and scope for this work was provided by the industry supervisor, Dr. Qiang Wang. The company he works for, Vultus AB, is providing precision agriculture solutions based on Sentinel-2 products. In addition to the numerous constructive discussions, I'd like to thank Dr. Wang for providing the spectrally resampled LUCAS dataset used in this thesis. Many thanks to Dr. Johannes Albertsson, SLU Alnarp, for providing validation soil samples. I am also grateful for the excellent project supervision and advice by the thesis supervisor Dr. Marko Scholze, Lund University.

Table of Contents

List of Abbreviations

SOC: Soil Organic Carbon

SOM: Soil Organic Matter

Remote Sensing and Spectroscopy:

BSI: Bare Soil Index

DRS: Diffuse Reflectance Spectroscopy

FWHM: Full Width Half Maximum

MSI: MultiSpectral Instrument

NDVI: Normalized Vegetation Difference Index

NIR: Near-Infrared

SSL: Soil Spectral Library

SWIR: Short-Wave Infrared

Datasets:

LUCAS: Land Use and Coverage Area frame Survey

Statistics:

LV: Latent Variable

PLSR: Partial Least Squares Regression

RMSE: Root Mean Square Error

RPD: Ratio of Performance to Deviation

std: Standard Deviation

Government Bodies and Organizations:

ESA: European Space Agency

EU: European Union

IPCC: Intergovernmental Panel on Climate Change

1 Introduction

Soils are composed of a multitude of mineral and organic constituents. The fraction of soil consisting of organic materials at various stages of decomposistion is termed soil organic matter (SOM) (Ben-Dor et al., 1997). While its amount and quality has a strong influence on soil properties, the fraction of SOM that is organic carbon (SOC) is increasingly recognized by scientists and policy makers alike for it's significant carbon offset potential (Lal, 2004; Van-Camp et al., 2004). These aspects have led to interdisciplinary attempts to understand and quantify SOC stocks and fluxes within disciplines ranging from global atmospheric modelling to field-scale precision agriculture.

A large share of Earth's land surface has been adapted to meet the production needs of mankind (IPCC, 2007). It has been established that up to 60% of SOC in temperate and 75% in tropical climates is lost due to this conversion process (Lal, 2004). These SOC fluxes are determined by the sequestration gains derived from the humification of organic material and losses due to heterotrphic respiration, erosion and leaching (Lal, 2004). While erosion of SOC from agricultural soils has been suggested to act as a slight carbon sink at 0.12 Pg C yr^{-1} (Oost et al., 2007), the remainder of the lost carbon is ultimately released into the atmosphere.

The managed soils of croplands have SOC pools that are particularly volatile and change according to management practices (Conant et al., 2011; Lal, 2004). In many cases, the natural SOC has been depleted due to long-term agricultural activity, so that there is a great theoretical potential to restore carbon in soils (Nocita et al., 2014). Lal (2004) has estimated that between 50 and 66% of the total cumulative carbon emissions from soils could be offset by the sink capacity of the current agricultural and degraded soils.

The importance of SOC for the global carbon cycle has been mirrored by increasing body of research conducted (Smith et al., 2018), as well as the political attention this topic has attracted in recent decades (Nocita et al., 2014; Van-Camp et al., 2004). In their fourth assessment report, the Intergovernmental Panel on Climate Change (IPCC) listed enhanced SOC sequestration in agricultural ecosystems as a pathway to restore a significant amount of terrestrial carbon that has been emitted to the atmosphere due to land conversions and exploitative land use practices (IPCC, 2007). Likewise, the Soil Thematic Strategy that has been adopted the European Commission highlights the potential of appropriate management practices to stabilize and even increase the amount of carbon sequestered in soil (Van-Camp et al., 2004).

Despite the significance of SOC for the carbon cycle, the most prominent subtopic in research remains the relationship between SOC and of soil quality (Smith et al., 2018). SOC improves the soil physical, chemical and biological properties, making it a prime indicator for soil quality (Ben-Dor et al., 1997; Karlen et al., 1997). It has been found that SOC improves the soil structure and its water and nutrient retention is orders of magnitudes higher than that of mineral soil constituents (Spaccini and Piccolo, 2013). SOC also increases the aggregate stability of the soil and prohibits surface crusting, which facilitates infiltration and decreases the risk of erosion from surface runoff. Soil organisms feed on organic matter, so that a maintained SOC stock will lead to abundant soil life.

The decomposition of soil organic matter caused by soil organisms is accompanied by the mineralization of plant available nutrients, which is of interest for soil amendment calculations (Stenberg et al., 2005). In recent decades, the development of precision agriculture technology has aimed to improve the efficiency of nitrogen fertilizer usage by calculating variable application rates (Stenberg et al., 2010). This is primarily accomplished by analyzing the reflectance of established crops. To improve the amendment strategy, an understanding of the amount of nitrogen that will become plant available from SOM mineralization is needed. This requires a solid and reproducible method to quantify SOC.

The causes of ongoing depletion of SOC stocks in many agricultural soils are the prominent management practices of modern agriculture (Wesemael et al., 2010). Tillage and bare fallow periods encourage decomposition of remaining SOC, allow for leaching, aeolian and water erosion of SOC (Oost et al., 2007; Rochette and Angers, 1999).

A wide array of conservative management practices have been suggested to mitigate further carbon release and increase the carbon sequestration of SOC depleted soils (Lal, 2004; Tola et al., 2019). These include conservation or zero-tillage, cover crops, perennial crops, conservation buffers and improved rotations (Conant et al., 2011; Spaccini and Piccolo, 2013).

While the significance of SOC for ecosystem productivity and the global carbon cycle is well documented, difficulties persist in the effective and spatially contiguous quantification of SOC. Many of the soil maps available today have a low spatial resolution and are based on outdated methods (Stevens et al., 2013). Standard soil tests are a robust and simple method to obtain a precise point measurement of SOC, but the high costs connected to timely data collection and laboratory analysis limit their applicability on larger scales (Castaldi et al., 2019; Ward et al., 2019). In addition, SOC varies spatially and temporally as a function of underlying environmental, climatic and management related variables (Conant et al., 2011; Wesemael et al., 2010). The high spatial heterogeneity of these variables complicate a precise interpolation of SOC point data using geostatistical methods (Conant et al., 2011).

To obtain SOC estimates at an appropriate spatial resolution while being cost-effective and practical, different methods are required (Stevens et al., 2013). Starting in the 1970s, the use of diffuse reflectance spectroscopy (DRS) in the visible to shortwave infrared part of the electro-magnetic spectrum (400-2500 nm) has been developed to infer information from soil (Nocita et al., 2014; Steinberg et al., 2016). This indirect method takes advantage of empirically established relationships between soil constituents and their reflectance (Steinberg et al., 2016). While DRS is nowadays commonly used in laboratories to quantify SOC in soil samples, the increasing sophistication of sensor technology has even made it possible to obtain data at a sufficient spectral and spatial resolution from remote sensing platforms (Ben Dor et al., 2015). The monitoring of SOC in annual agro-ecosystems is a prime application for remote sensing, as exposed bare soil areas are commonly present after harvest. Precise agricultural topsoil SOC quantifications have been performed using hyperspectral data from an airborne sensors (Steinberg et al., 2016), and there have also been promising attempts using multispectral data collected by the ESA Sentinel-2 satellites (Castaldi et al., 2019; Gholizadeh et al., 2018).

To obtain SOC estimates from remotely sensed reflectance data, a soil spectral library (SSL) is required (Nocita et al., 2014). A SSL consists of soil spectral information paired with soil physical or chemical properties (Ben Dor et al., 2015). The correlation between reflectance and soil properties such as SOC is then commonly established with multivariate statistics such as PLSR (Castaldi et al., 2019; Stenberg et al., 2010). An appropriate library requires the collection and standardized analysis of a representative amount of soil samples to account for the spatial variability of soil properties for the whole study area (Guerrero et al., 2016; Nocita et al., 2014; Stevens et al., 2013). With a lack of standardized methods to collect and analyze soil samples, many of the small SSLs around today have a limited applicability (Ben Dor et al., 2015). The surging political and scientific interest in large scale, harmonized soil databases, however, has led to recent efforts to construct country and even continental scale SSLs (Ben Dor et al., 2015; Tóth et al., 2013b). One of the largest consistent soil databases to date has been compiled under the Land Use and Coverage Area frame Survey (LUCAS) program in 2009. The resulting LUCAS topsoil database comprises 19,967 European soil samples (Orgiazzi et al., 2018).

The increasing availability and quality of remotely sensed data and SSLs makes SOC quantifications on larger scales feasible (Castaldi et al., 2019). This provides new opportunities to investigate the effectiveness of policies and land management as well as allowing farmers to consider in-field SOC variability in the calculation of agricultural amendments (Stenberg et al., 2005; Wesemael et al., 2010). It has been shown that the spectral information in the LUCAS topsoil database allows for the development of SOC predictive models (Nocita et al., 2014; Steinberg et al., 2016; Stevens et al., 2013). Other researchers have successfully used multispectral Sentinel-2 data and in situ soil sampling for SOC predictions in croplands (Bhunia et al., 2019; Castaldi et al., 2019; Gholizadeh et al., 2018).

1.1 Aim

The aim of this thesis is to scrutinize the possibility of developing a model for SOC quantification from Sentinel-2 data using the LUCAS topsoil database as the SSL. As both of the LUCAS database and Sentinel-2 data are open access and no field sampling is required for the described method, a successful model would make SOC quantifications more accessible and avoid the effort and costs involved in field sampling.

2 Background

2.1 Reflectance of soil organic matter

The soil constituents that interact with electromagentic radiation are called chromophores (Stenberg et al., 2010). SOM is not a single substance, but an array of materials which are divided into fibric, hemic and sapric materials based on their stage of decomposition (Ben-Dor et al., 1997). The decomposition stage is known to have a profound impact on reflectance, with fibric components resembling the reflectance of senescent leaves, while hemic and sapric substances absorb more radiation (Baumgardner et al., 1986). Prior to diffuse reflectance spectrosopy measurements for the LUCAS topsoil database, larger fibric components such as plant roots were removed from soil samples (Tóth et al., 2013b).

There is generally an inverse correlation between soil reflectance and the amount of SOM in soil. Absorption features related to SOM have been identified for many wavelengths of the Visible-NIR spectral range, with Stenberg et al. (2010) citing the importance of bands around 1100, 1600, 1700 to 1800, 2000, and 2200 to 2400 nm. Other research also highlights the activity of SOM in the visible range (Baumgardner et al., 1986; Ben-Dor et al., 1997).

It has been found that organic matter does not play a significant role in soil reflectance if it makes up less than 2 % ($\approx$12 g C kg^{-1}, Baumgardner et al., 1986). Other chromophores, whose absorbance features are then dominant, include mineral components such as clay minerals, iron oxides, carbonates or quartz (Ben-Dor et al., 1997). Due to these multiple soil covariates, the reflectance of soil is highly variable and spatially dependent. Therefore, the geographical scale and sampling density determine the soil reflectance variability within a soil spectral library and thus the possible accuracy of soil property predictions that can be achieved (Stenberg et al., 2010).

2.2 Soil organic carbon in Europe

Topsoil SOC concentrations fluctuate as a function of many interrelated environmental and anthropogenic drivers such as temperature, precipitation, vegetation type, land use and management. The resulting high SOC variability in the EU has been visualized by mapping the SOC of the 19,672 LUCAS soil samples (Fig. 1). Low SOC contents have been recorded for most of the Mediterranean, where the mean SOC of cropland soils is lower than for any other climatic regions in Europe at 12 to 16 g kg^{-1} (Tóth et al., 2013b). The most organic soils are found across peatlands in Ireland, the UK, Sweden and Finland, which make up about 50 % of the total SOC in the EU (Jones et al., 2005). A general south-east to north-west trend in European SOC has been noted by several continent scale studies (Tóth et al., 2013a).

The study site of this thesis is located in southernmost Sweden. This area has been classified into the sub-oceanic to sub-continental climate region together with most of Poland, eastern Germany, parts of the Czech Republic and western Denmark (Tóth et al., 2013b). For this climate region, the second lowest mean cropland SOC was measured at 15 g kg^{-1}. The low SOC in the in the agricultural land in Scania stand in stark contrast to the rest of Sweden with its many organic forest and peat soils.

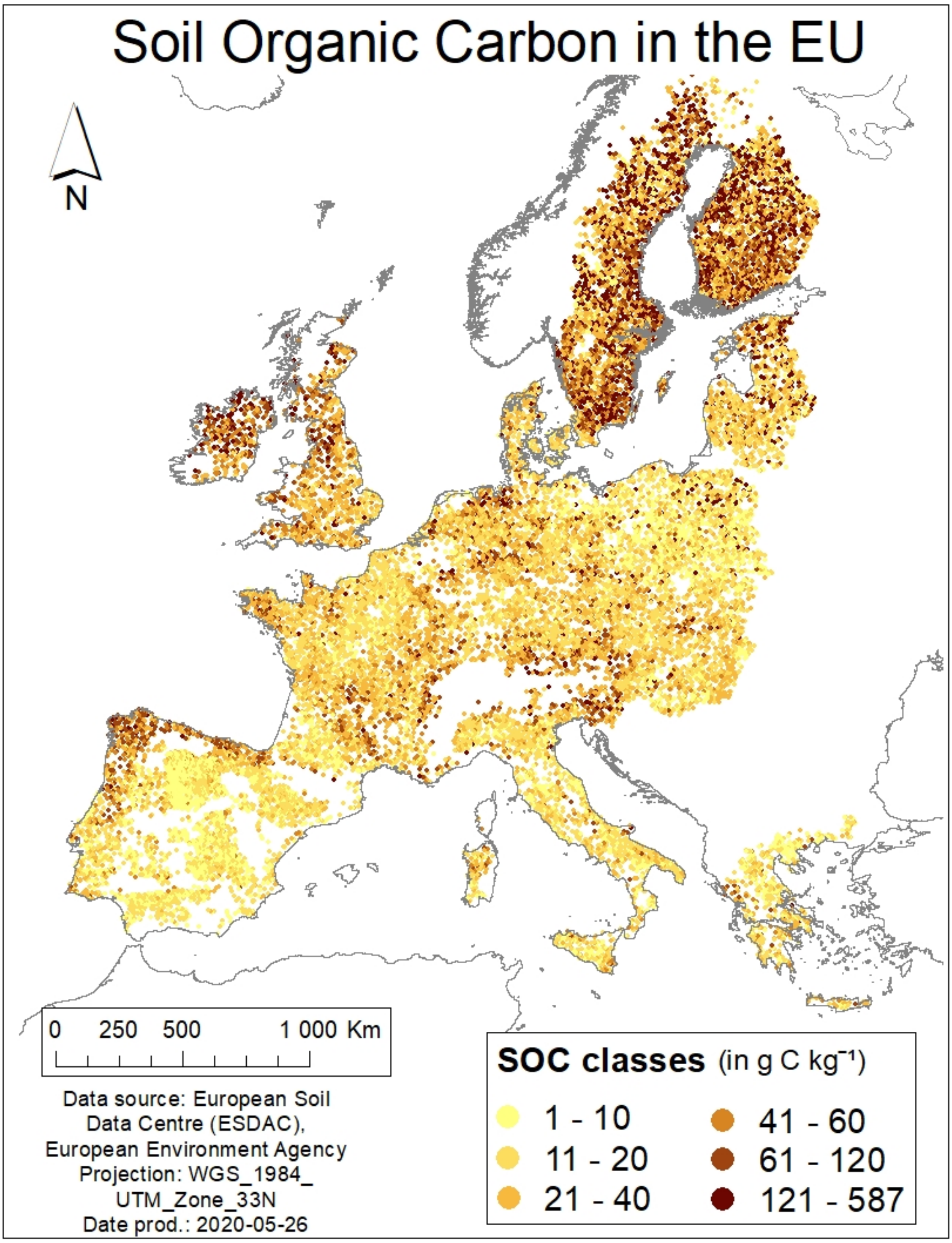

Soil Organic Carbon in the EU
N
0 250 500 1 000 Km
Data source: European Soil Data Centre (ESDAC), European Environment Agency
Projection: WGS_1984_ UTM_Zone_33N
Date prod.: 2020-05-26
SOC classes (in g C kg⁻¹)
● 1 - 10
● 11 - 20
● 21 - 40
● 41 - 60
● 61 - 120
● 121 - 587

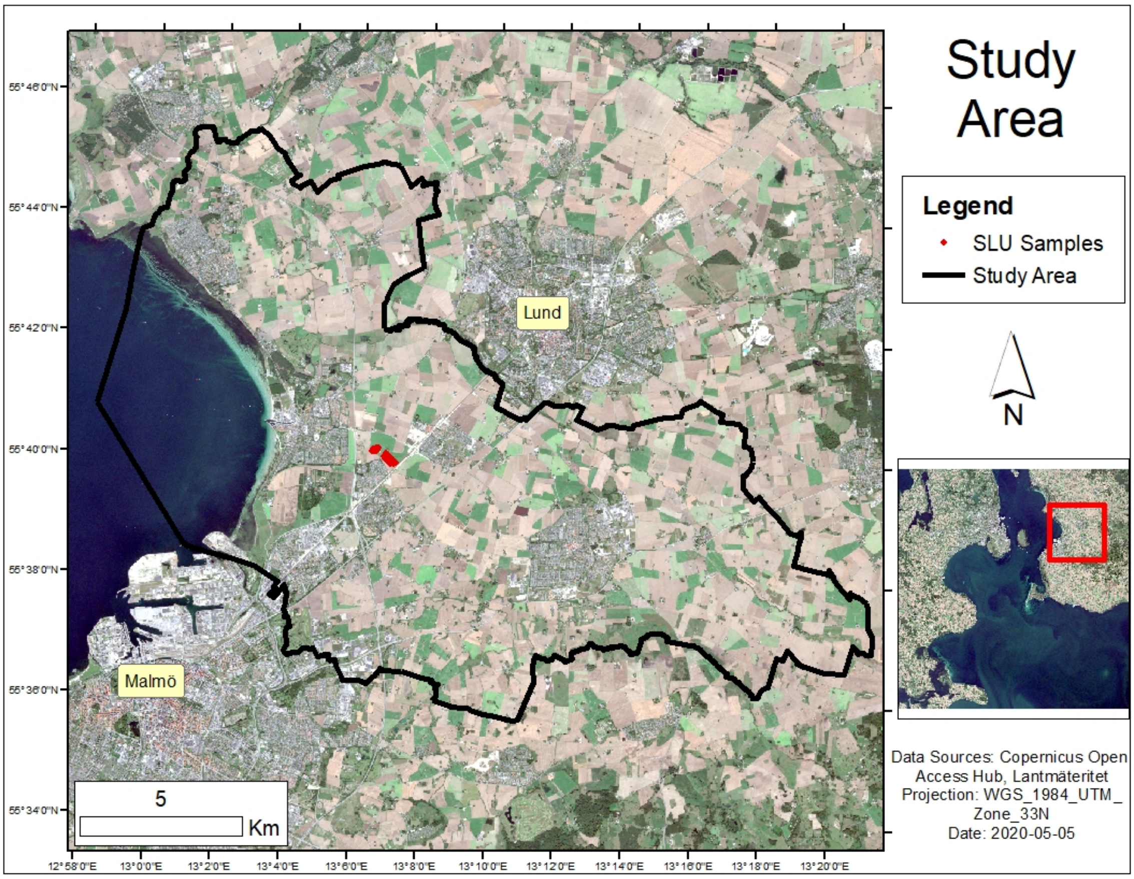
Figure 1: The distribution of organic carbon in the EU based on the complete LUCAS topsoil database. Every point on the map represents the organic carbon measurement of a LUCAS soil sample. Samples were obtained from the top 15 cm of mineral soil in 2009 (Tóth et al., 2013b).

3 Materials and Methods

3.1 Study area

An appropriate study area for the target analysis was chosen according to the geolocation of the validation soil samples in Sweden's southernmost province, Scania. Three municipalities - Lomma, Burlöv and Staffanstorp municipality - have been selected as an initial area of interest (Fig. 2). Despite their location between the two major cities Malmö and Lund, a majority of the land (about 126 km^2) is used for agricultural purposes within these municipalities.

Approximately 50 % of the food that is currently produced in Sweden comes from Scania (Dänhardt et al., 2013). The mild climate in combination with the fertile soils allow for the intensive cultivation of annual crops such as wheat, barley, sugar beet and rape seed (Dänhardt et al., 2013). Upon harvest in late summer, a large share of the fields are tilled and prepared for the next crop. In that time period, the reflectance of the bare soil can be recorded by remote sensing platforms. The Sentinel-2 satellite image depicting the study area in figure 2 was taken on the 25th of August and exemplifies the large share of fields that do not display a vegetative cover at this time of year.



Data Sources: Copernicus Open Access Hub, Lantmäteritet
Projection: WGS_1984_UTM_Zone_33N
Date: 2020-05-05

Figure 2: Location of the study area in Southern Sweden

3.2 Sentinel-2 imagery

The Copernicus Sentinel-2 mission consists of two satellites in the same sun-synchronous orbit, which have been launched in 2015 and 2017. With a swath width of 290 km, the multispectral sensors on board allow for a temporal resolution of 2-3 days in the mid-latitudes, while providing a high spatial resolution across thirteen spectral bands (Table 1). While the higher resolution bands are meant to provide information on surface features, band 1, 9 and 10 are primarily used for atmospherically correction and cloud detection.

The Copernicus Open Access Hub website[1] was used to search for cloud-free Sentinel-2 imagery. Additional criteria were the amount of precipitation occurring prior to sensing and the time interval between sensing date and the field sampling done on the September 12^{th}, 2019. A cloud-free Sentinel-2 scene was found for the 25^{th} of August, 2019. No precipitation was recorded in the study area the week before sensing according to the Swedish Meteorological and Hydrological Institute (SMHI)[2]. A second Sentinel-2 scene from the 29^{th} of October was downloaded as a backup dataset. The cloud cover of this scene was 0.23 % and only minimal rainfall (<3 mm) occurred three days prior to sensing.

The Sentinel-2 datasets were obtained as Level-2A products, which provide atmospherically corrected Bottom of Atmosphere reflectance data. Band 10, which is used for image correction, is not included in these datasets.

Table 1: Summary of Sentinel-2A spectral bands

Band	**Name**	**Spatial Resolution (m)**	**Central wavelength (nm)**	**Band-width (nm)**[3]
1	Coastal Aerosol	60	443	21
2	Blue	10	492	66
3	Green	10	560	36
4	Red	10	665	31
5	Vegetation Red Edge 5	20	704	15
6	Vegetation Red Edge 6	20	741	15
7	Vegetation Red Edge 7	20	783	20
8	Near-Infrared (NIR)	10	833	106
8a	Narrow NIR	20	865	21
9	Water Vapor Absorption Window	60	945	20
10	Shortwave Infrared - Cirrus	60	1374	31
11	Shortwave Infrared 1 (SWIR 1)	20	1614	91
12	Shortwave Infrared 1 (SWIR 2)	20	2202	175

[3] at Full Width Half Maximum (FWHM)

3.2.1 Image pre-processing

The downloaded Sentinel-2 scene was processed for modelling purposes using ArcMap (ESRI 2017. ArcGIS Desktop 10.5.1. Redlands, CA: Environmental Systems Research Institute) and the Sentinel Application Platform (SNAP) tool. Information on administrative boundaries was included in a vector terrain layer that was obtained from the Swedish Surveying and Cadastral Agency ("Lantmäteriet"). In addition to administrative districts, this vector layer provided information on the spatial extent of agricultural land. A mask of the three municipalities that comprise the study area and all agricultural land was created in ArcMap accordingly.

To maximize the spatial resolution for the analysis, all coarser bands were resampled to 10m in SNAP using the nearest neighbor upsampling method. Of the agricultural land within the mask area, only the areas with bare soil at the time of satellite data collection were of interest. Two indices have been tested for the identification of bare soil pixels: the Bare Soil Index (BSI) and the Normalized Vegetation Difference Index (NDVI). Both of these indices are normalized and thus have a value range of -1 to 1. Greater BSI values indicate bare soil, as the reflectance of soil in the

[1] Retrieved from: `https://scihub.copernicus.eu/dhus/#/home`

[2] Retrieved from: `https://www.smhi.se/data/meteorologi/kartor/dagliga/nederbord/2019/augusti`

visible red and short infrared bands is higher compared to vegetation (Bhunia et al., 2019; Eq. 1)

$$BSI = \frac{(Band_{11} + Band_4) - (Band_8 + Band_2)}{(Band_{11} + Band_4) + (Band_8 + Band_2)} \tag{1}$$

The NDVI is commonly used to distinguish vegetation from non-vegetation and has found application in other soil property estimation attempts using Sentinel-2 (Castaldi et al., 2019; Gomez et al., 2019). A low NDVI hints at the absence of vegetation.

$$NDVI = \frac{Band_8 - Band_4}{Band_8 + Band_4} \tag{2}$$

By visually inspecting the results, appropriate thresholds for bare soil identification were tested. With regard to the validation samples (see section 3.3), a NDVI < 0.26 threshold was ultimately chosen and the mask polygons updated accordingly.

For the modelling of SOC, three bare soil fields within the study area were selected (Fig. 3). The reflectance for each pixel within these fields was extracted using SNAP.

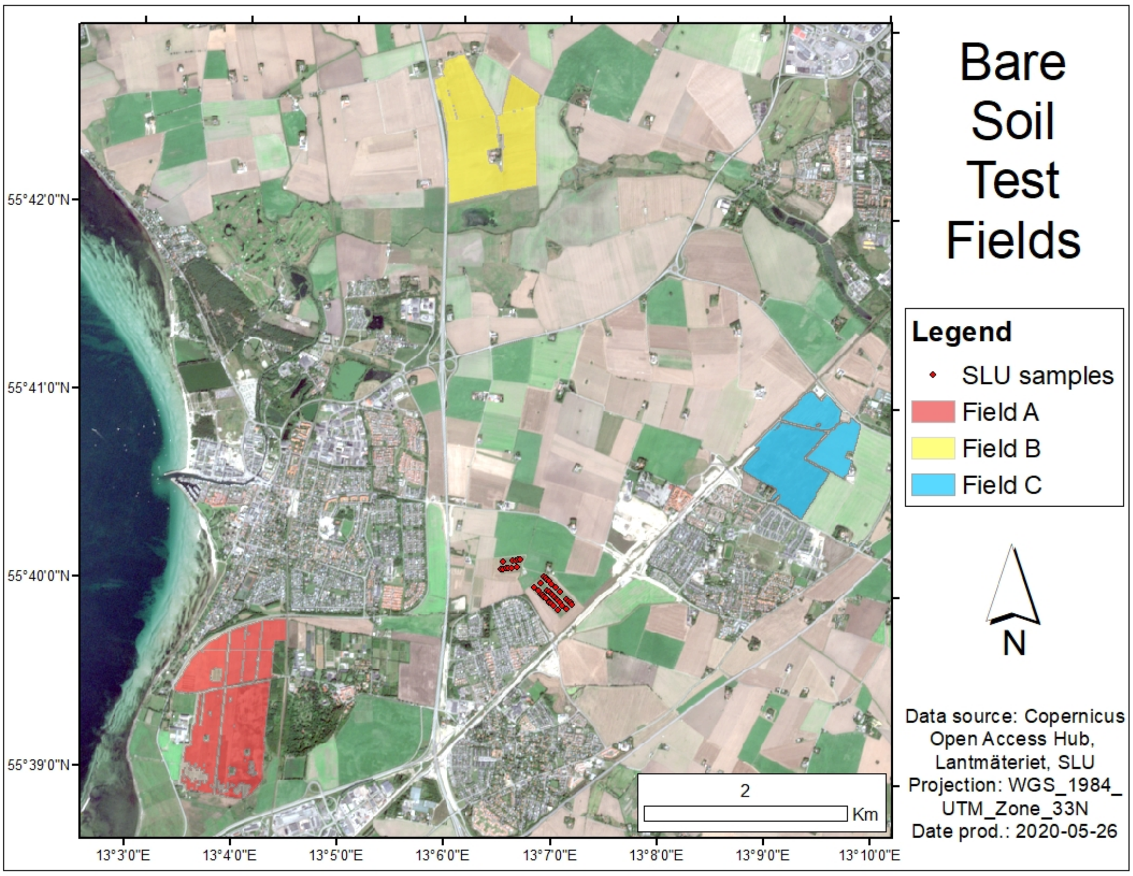


Figure 3: Validation sample location and selected bare soil fields.

3.3 Validation soil samples

A campus of the Swedish University of Agricultural Sciences (SLU) is located within the study area. On experimental fields of the University near Åkarp soil sampling has been conducted by Farid Jan and Elin Lund on the 12^{th} of September, 2019 (Fig. 3). A total of 44 samples were collected and pH, macro nutrients and SOC measured for three depth intervals (0-20 cm, 20-60 cm, 60-90 cm).

As the vegetative cover of the fields was very heterogeneous at the time of Sentinel-2 data collection, most of the pixels that the soil sampling points were located in did not fall within the in the initial BSI and NDVI range. Through visual inspection of different band combinations, the most convincing result could be produced by increasing the NDVI threshold to 0.26, which was then used for all pixel extractions in SNAP. Eight soil sampling points were located within Sentinel-2 pixels that fell below this threshold. An identity mask was created accordingly, the pixel reflectance data were extracted and then paired with the topsoil SOC sampling measurement (0-20 cm).

3.4 LUCAS topsoil database

The Land Use and Coverage Area frame Survey (LUCAS) was established by the Statistical Office of the European Union (EUROSTAT) in 2001 to create a pan-European database on landscape parameters that are relevant for agricultural and environmental policy development and evaluation (Tóth et al., 2013b). Since 2006, this survey has been periodically performed every third year for 2x2 km grid cells of all EU member states (Orgiazzi et al., 2018). The land cover is classified using satellite and airborne imagery.

In 2009, an extension to the periodic LUCAS was granted to produce a consistent, coherent and harmonized topsoil database for the EU (Tóth et al., 2013b). In this soil sampling campaign about 20,000 soil samples were collected using a multi-stage stratified random sampling approach that aimed at representing the proportion of different land use types in the EU (Tóth et al., 2013a). For each sampling point five topsoil samples (0-20 cm) were taken and combined to a composite sample. All of these samples were then analyzed for physical, chemical and reflectance properties using a standardized procedure in the same laboratory (Orgiazzi et al., 2018).

After air drying, crushing and sieving of each sample, the Visible-NIR absorbance from 400-2500 nm was recorded using the FOSS XDS Rapid Content Analyzer (FOSS NIRSystems Inc., Denmark) (Nocita et al., 2014). A total of 4200 absorbance bands were recorded at a 0.5 nm measurement interval. SOC was assessed by subtracting the carbonate content from total carbon, which was measured in a VarioMax CN Analyzer (Elementar Analysis, Germany) (Nocita et al., 2014).

The LUCAS topsoil database is available to researchers, public administrations and private companies for non-commercial purposes through the European Soil Data Centre (ESDAC) website[4].

3.4.1 Resampling of LUCAS

As the soil sample surface structure causes non-linear light scattering, not all radiation that does not return to the sensor is absorbed (Stenberg et al., 2010). Therefore, the measured absorbance (A) for each band and sample was transformed into reflectance (R) according to equation 3 (Nocita et al., 2014).

$$R = \frac{1}{10^A} \tag{3}$$

To use the hyperspectral LUCAS SSL for the prediction of SOC with Sentinel-2 multispectral reflectance, the 4200 LUCAS bands had to be resampled to simulate reflectance recorded by the MultiSpectral Instrument (MSI) of Sentinel-2. First, half of the LUCAS bands were removed, leaving one band for every nm of the recorded spectrum. While the (MSI) on board of Sentinel-2 measures reflectance within the given bandwidths, it is not equally sensitive for all wavelengths. The bandwidths provided in Table 1 represent the Full Width Half Maximum (FWHM) of every band, representing the spectral distance between the two wavelengths for which half of the amplitude of the maximum reflectance wavelength is recorded. To obtain information on the amplitudes recorded by MSI at 1 nm resolution, the spectral library of ENVI (L3Harris Geospatial 2015. ENVI 5.3. Broomfield, CO) was assessed. The reflectance amplitudes for each band were visualized in ENVI (Fig. 11, see Appendix). A text file containing the required information was downloaded and used to resample the LUCAS SSL in MATLAB (The MathWorks Inc. 2020. MATLAB R2020a. Natick, MA). The resampled LUCAS data was provided by the industry supervisor for this thesis, Dr. Qiang Wang.

[4] Retrieved from: `https://esdac.jrc.ec.europa.eu/`

3.5 Soil organic carbon preditive models

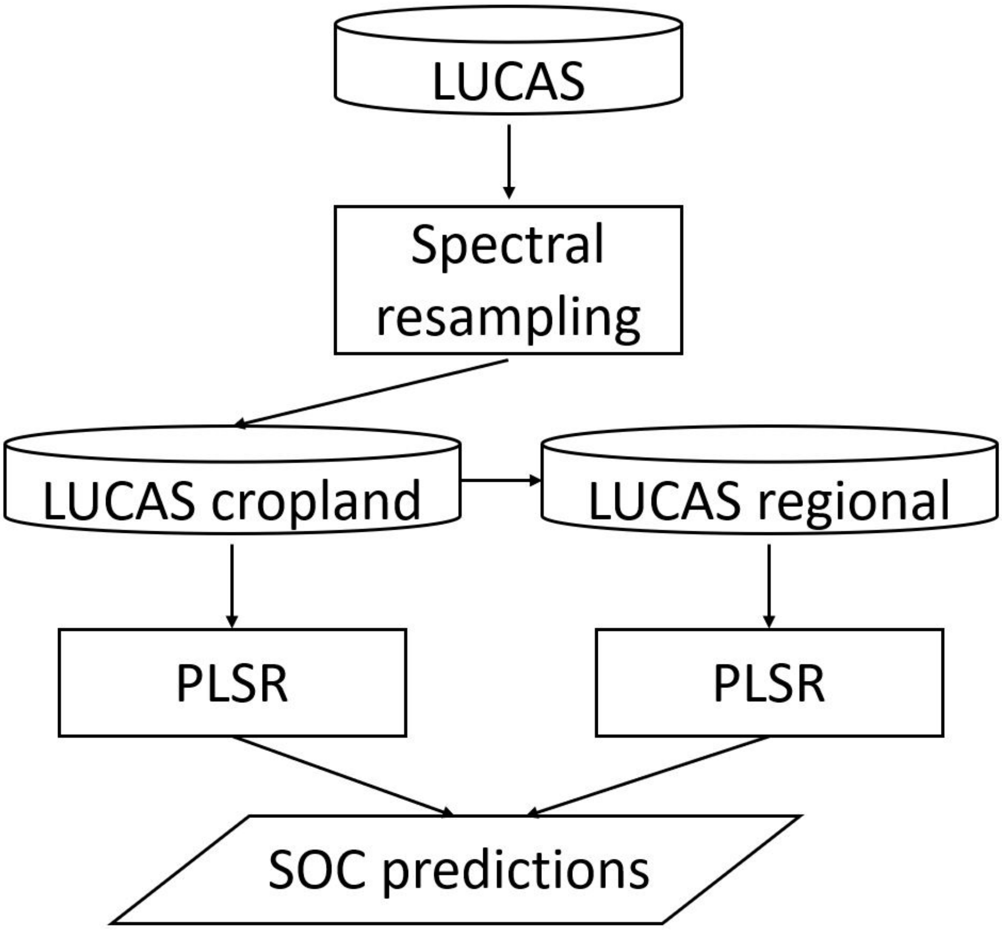


Figure 4: Overview of the general workflow

The LUCAS database is comprised of soil samples that are representative of the land use and land cover types in the European Union. As the target analysis concerns only agricultural land, an effort was made to obtain meaningful LUCAS subsets for statistical modelling.

Information on land use at each sample location is included in the LUCAS database. In addition, the samples have been classified as mineral or organic soil depending on their SOC content. In accordance to the soils present in our study area, all LUCAS samples that were classified as mineral and taken on cropland were extracted as first subset ($n = 8332$, Table 2).

From this, a regional subset of LUCAS cropland samples was obtained with respect to the geographical location of the study area. One column within the LUCAS database concerns the European NUTS2 regions, which are subdivisions of countries for statistical purposes (Panagos et al., 2013). The geographical subset was attained by selecting the LUCAS samples that fell within three NUTS2 regions adjacent to the study area that encompass southern Sweden and the Danish island Zealand.

Predictions of soil properties based on soil reflectance measurements are commonly achieved by creating multivariate statistical models (Ward et al., 2019). The numerous bands resulting from diffuse reflectance spectroscopy are often highly collinear and noisy (Nocita et al., 2014). Partial Least Squares Regression (PLSR) is a multivariate regression analysis that has been developed to deal with the multiple, correlated and noisy predictor variables in chemometrics (Wold et al., 2001) and has become the preferred statistical method within Visible-NIR spectroscopy studies (Stenberg et al., 2010). This regression approach has recently also been employed to explain and model SOC based on multispectral Sentinel-2 reflectance (Castaldi et al., 2019).

In contrast to the related Principal Component Analysis, PLSR takes into account both the predictor variables X and response variable Y in the model development process (Stenberg et al., 2010). The PLSR algorithm produces a set of new X variables, also called latent variables (LV), with the aim of maximizing the covariance between the predictor and response variables (De Jong, 1993). From the LV, the predictor variable is approximated, resulting in weights for each predictor and response variable.

Several LV are produced iteratively, i.e. after the first set of LV has been calculated, the X matrix of the model calibration dataset is deflated by subtracting the predicted X matrix (Wold et al., 2001). Every PLSR iteration is referred to as a component. The final linear model consists of regression coefficients that are dependent on the amount of components (Eq. 4).

$$b_k = \sum_{a=1}^{A} c_a * w_{ka} \tag{4}$$

where c are the PLSR Y-weight and w the PLSR X-weight for each component a and X variable k (Wold et al., 2001). The resulting linear regression coefficients b_k can then be used to derive Y predictions ($\hat{Y}$) based on the predictor variable observations (Eq. 5).

$$\hat{Y} = X_k * b_k \tag{5}$$

The created LUCAS subsets were used to build predictive PLSR models with the reflectance of the 12 simulated Sentinel-2 bands as predictor variables X and the measured SOC as response variable Y. While an increasing number of components generally improves the fit of the model for the input data, it grows prone to overfitting for new data (Wold et al., 2001). Therefore it is necessary to select an appropriate set of components, which is achieved with cross validation (Nocita et al., 2014). This method splits the input dataset randomly into training and validation data, builds a PLS model for the training data and predicts Y of the validation data. The number of components was chosen according to the lowest averaged Root Mean Square Error (RSME$_{cv}$) resulting from 10 cross validation iterations (Castaldi et al., 2019).

For both models the SOC values of the input datasets were predicted using the chosen set of components and compared to the measured SOC. To evaluate the fit, the Root Mean Squared Error of Prediction (RMSE$_p$; Eq. 6) and the Ratio of Performance to Deviation (RPD; Eq. 7) were calculated.

$$RMSE_p = \sqrt{\frac{\sum_{i=1}^{n}(y - y_{pred})^2}{n}} \tag{6}$$

$$RPD = \frac{std_Y}{RMSE_p} \tag{7}$$

The RMSE$_p$ reflects on the mean predicted residual, where y are the measured SOC values, y_{pred} the predicted SOC and n the number of samples. For an improved fit comparison between different datasets, the RPD takes into account the underlying standard deviation of the measured SOC values (std_Y) (Castaldi et al., 2019). While a RPD below one stands for a very poor model, 1 < RPD < 1.4 represents a poor model, 1.4 < RPD < 1.8 an ok model and everyting above 1.8 a good to very good model (Gholizadeh et al., 2018).

3.6 Model validation

To test the applicability of the model for Sentinel-2 data, the SOC of the SLU validation soil samples was predicted using the paired reflectance observation of the satellite. The outcome was evaluated by calculating SOC statistics and RMSE$_p$.

4 Results

4.1 LUCAS resampling

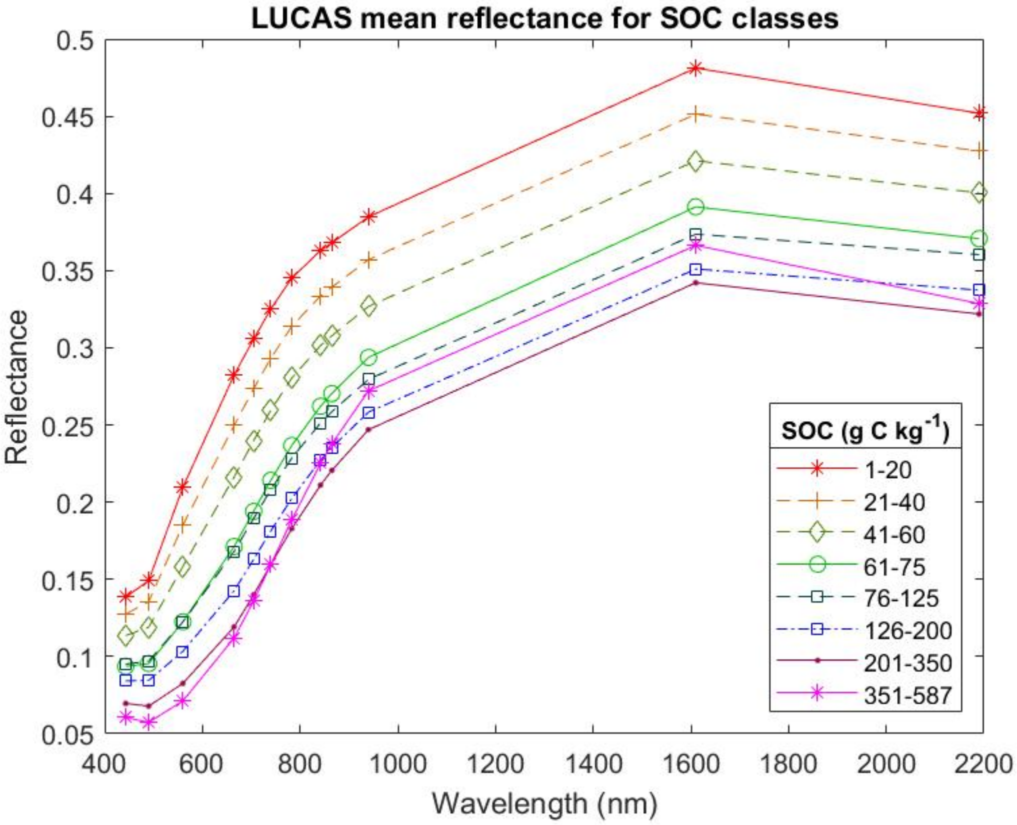


Figure 5: Mean reflectance of LUCAS soil organic carbon classes following figure 3 by Nocita et al. (2014). The center of each of the twelve simulated Sentinel-2 band is represented by a marker (table 1). The plot is based on the reflectance of all of the nearly 20,000 soil samples in the LUCAS topsoil database.

With the resampling of the LUCAS SSL to simulate Sentinel-2 reflectance, plenty of initial detail and thus spectral information has been excluded or agglomerated. Nocita et al. (2014) gauged the capability of predicting SOC using the hyperspectral LUCAS SSL and visualized the spectral response to SOC content by plotting the mean reflectance of different SOC classes across the measured spectral range. This figure has been recreated for the complete, spectrally resampled SSL (Fig. 5). The inverse correlation between SOC and reflectance becomes apparent here, as the reflectance decreases consistently with increasing SOC for most of the defined classes (Baumgardner et al., 1986). Only for the highest SOC class, which represents organic peat and forest soils, a greater mean reflectance can be observed for simulated NIR and SWIR bands (800-2200 nm; Table 1). Nocita et al. (2014) found that differences between the lower SOC classes was most evident, which is also the case for the spectrally resampled LUCAS database in figure 5. All of the mineral cropland soil samples in the regional LUCAS subset and the majority of the Eu-wide LUCAS cropland samples used in the analysis fall within 1-40 g C kg^{-1}, or the first two SOC classes (Table 2). The difference in mean reflectance between these two classes increases from 400-600 nm and is then nearly constant at about 0.03 for the remaining spectral interval.

4.2 Soil datasets

Table 2: Dataset overview and soil organic carbon statistical summary for the two LUCAS calibration datasets and the validation samples collected by SLU.

			SOC g kg^{-1}			
Dataset	**Extent**	**n**	**min**	**mean**	**max**	**std**
LUCAS cropland	EU-25 states	8332	2.0	17.1	160.3	10.87
LUCAS regional	Skåne, Sjaelland	70	7.7	17.0	37.0	5.32
Validation samples	2 fields in Skåne, ~13 ha	8	12.6	16.4	18.5	1.9

An overview of the LUCAS subsets and the SLU validation soil samples is provided in table 2. A total of 8332 LUCAS samples were taken on croplands from mineral soils across the EU-25 member states. The climate systems within the EU range from boreal to mediterranean and oceanic to continental, with an accordingly large array of natural ecosystems (Tóth et al., 2013b). The associated range of pedogenic processes in addition to the natural variability of parent materials have led to the many, spatial heterogeneous soil types found in the EU. The regional LUCAS subset consists of 70 samples and does not only have a minor geographic extent, but also less variation in soil types. While both datasets have the same mean SOC, the range and standard deviation of the complete LUCAS cropland dataset is greater.

The SLU validation dataset is crucial for assessing the model's applicability to reflectance measurements by the Sentinel-2 MSI. This small dataset (n = 8) has a minor geographic extent compared to the calibration datasets (Table 2). The SOC range of this validation dataset is just 5.9 g C kg^{-1}, but the mean of 16.4 g C kg^{-1} is similar to the mean of the two LUCAS datasets.

Figure 6 shows the mean reflectance of the datasets used in this thesis, which markers indicating the center of each Sentinel-2 band. The first two lines represent the LUCAS datasets, whose reflectance is derived from spectrally resampled laboratory measurements. The reflectance of the validation samples are the paired Sentinel-2 measurements from the satellite image taken on August 25th. Due to the small sample size of the validation dataset (n = 8), the mean reflectance of two of the selected fields were added (Fig. 3). Field A consists of 11,111 Sentinel-2 pixels and Field B of 9,837 pixels. All datasets with actual Sentinel-2 measurements show a lower mean reflectance than the resampled LUCAS subsets. If compared to figure 5, the Sentinel-2 mean reflectance graphs resemble the mean reflectance of higher SOC classes. This is the case despite the validation samples having a slighly lower SOC mean than the two LUCAS calibration datasets (Table 2).

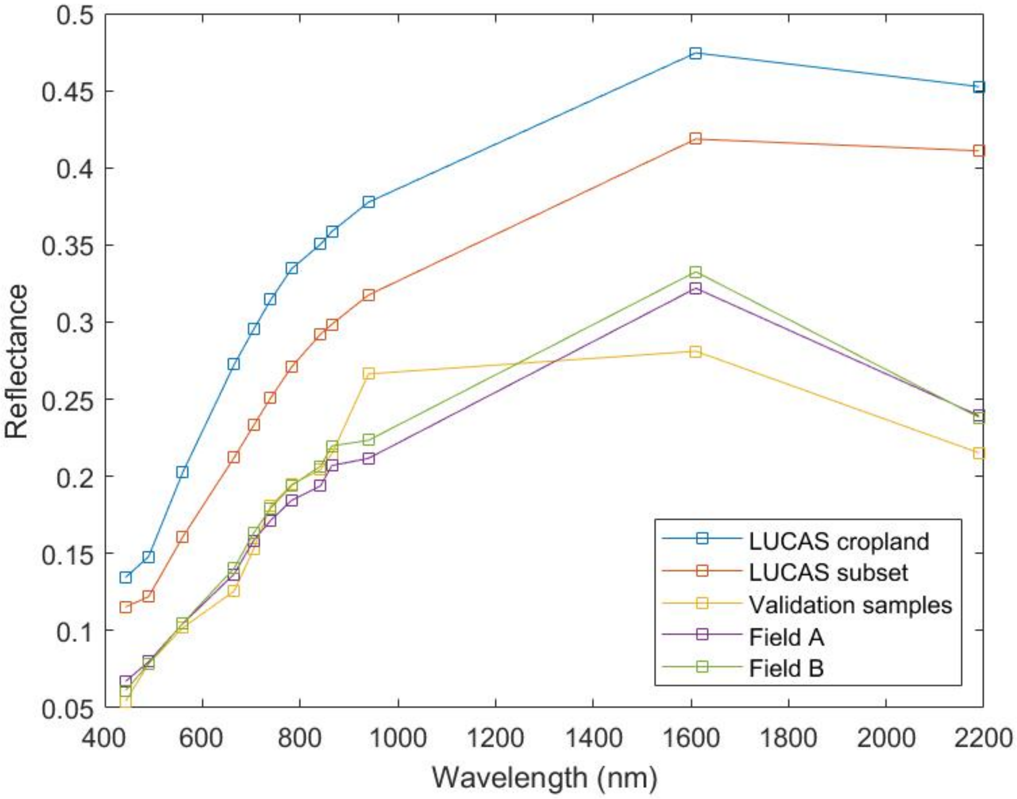


Figure 6: Mean reflectance of datasets used in this thesis. The first two datasets consist of spectrally resampled LUCAS data and the latter three stem from the Sentinel-2 scene taken on the 25^{th} of August.

4.3 Partial least squares regression models

The PLS model computed for the combined LUCAS cropland samples could not explain more than 18.5% of the SOC variance within the dataset ($R^2 = 0.185$; Fig. 7). For the regional subset, the maximum coefficient of determination (R^2) using all of the possible 12 components was 0.388.

The appropriate number of components was chosen according to the lowest RMSE$_{cv}$. For the LUCAS cropland model, the lowest RMSE$_{cv}$ was given by 12 components (Fig. 12, see Appendix). As the fit of the model improved less than 0.1 g C kg^{-1} for more then four components (RMSE$_{cv}$ = 9.98 g kg^{-1}), a total of four components were ultimately chosen to decrease the risk of overfitting.

Due to the small calibration sample size (n = 70) of the regional subset, this model is prone to overfitting using many components (Wold et al., 2001). While the RMSE$_{cv}$ of the regional subset was more variable for different amount of components, the lowest RMSE$_{cv}$ of 5.12 g C kg^{-1} was given for 3 components (Fig. 13, see Appendix).

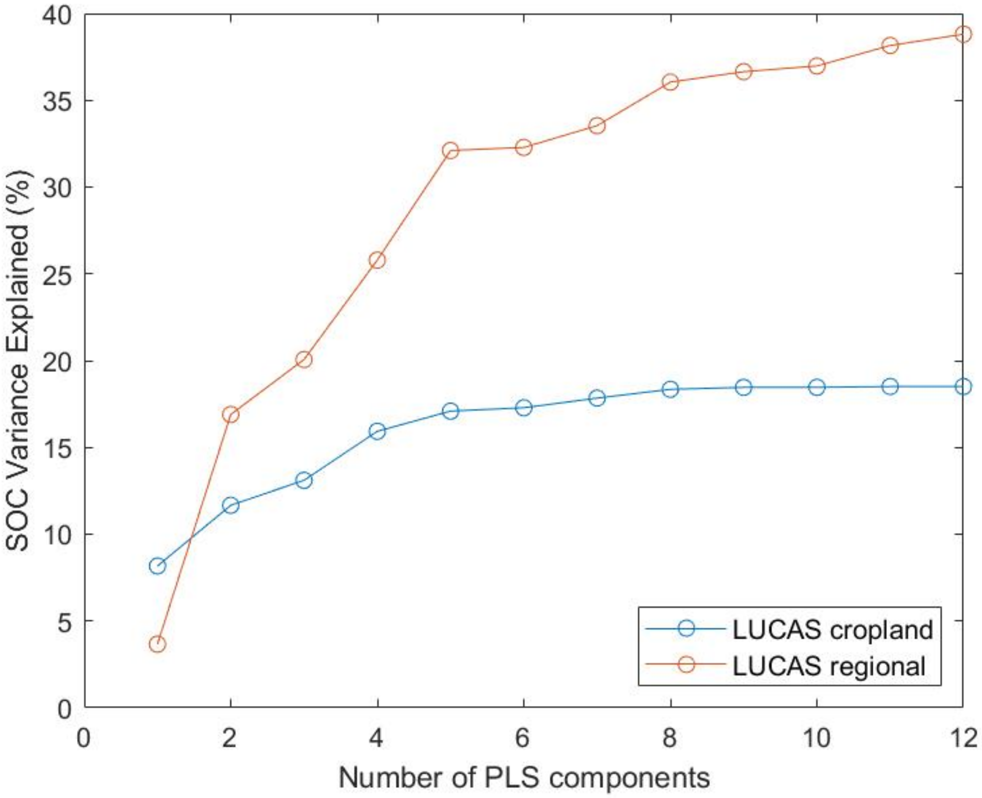


Figure 7: Cumulative explained variance for the components of both PLSR models.

Based on the chosen components, PLS regression coefficients were computed, which provide a linear model for the reflectance of 12 Sentinel-2 type bands to approximate soil organic carbon. These coefficients and the intercept of each linear model are given in table 3.

Table 3: PLS regression coefficients for the pan-European LUCAS cropland and the LUCAS regional model calibration dataset. The LUCAS cropland coefficients are the result of four PLS components and the LUCAS regional coefficients are based on three components.

	PLS regression coefficients	
	LUCAS cropland	**LUCAS regional**
Intercept	26.32	30.24
Band 1	66.44	-35.96
Band 2	44.87	-54.05
Band 3	-41.63	-66.64
Band 4	-114.17	-76.09
Band 5	-74.11	-43.42
Band 6	-39.74	-8.71
Band 7	-2.63	31.16
Band 8	44.13	56.93
Band 8a	71.89	69.29
Band 9	92.89	80.39
Band 11	-12.02	78.92
Band 12	-45.57	-162.82

SOC predictions for the calibration datasets were derived from a linear combination of sample reflectance and the regression coefficients. The prediction of the LUCAS cropland dataset resulted in a high RMSE$_p$ of 9.97 g kg^{-1}(Fig. 8). While the mean of the predicted and observed SOC are equal, the standard deviation of the predictions is less than 40 % (Table 2, 4). The R^2 of 0.16 and an RPD of 1.09 indicate that this model is has poor predictive capabilities and does not reliably distinguish the reflectance alterations caused by different SOC contents of the soil.

Table 4: Statistical summary of the soil organic carbon predictions of the two calibration datasets with the produced linear model.

		Predicted SOC g kg^{-1}			
Dataset	n	min	mean	max	std
LUCAS cropland	8332	-3.3	17.1	34.5	4.3
LUCAS regional	70	11.3	17.0	22.5	2.4

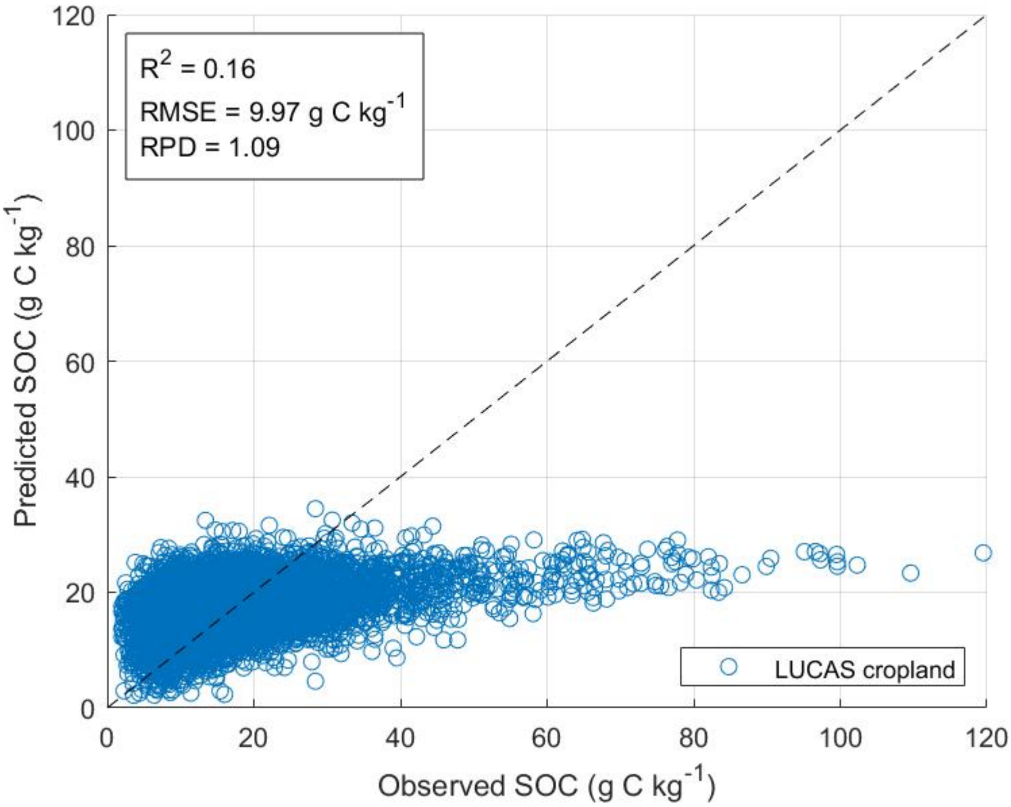


Figure 8: SOC prediction for LUCAS cropland subset based on a PLSR model with 4 components.

The SOC prediction of regional LUCAS calibration dataset resulted in a R^2 of 0.26 (Fig. 9). The RPD of 1.12 indicates that this model also has poor predictive power.

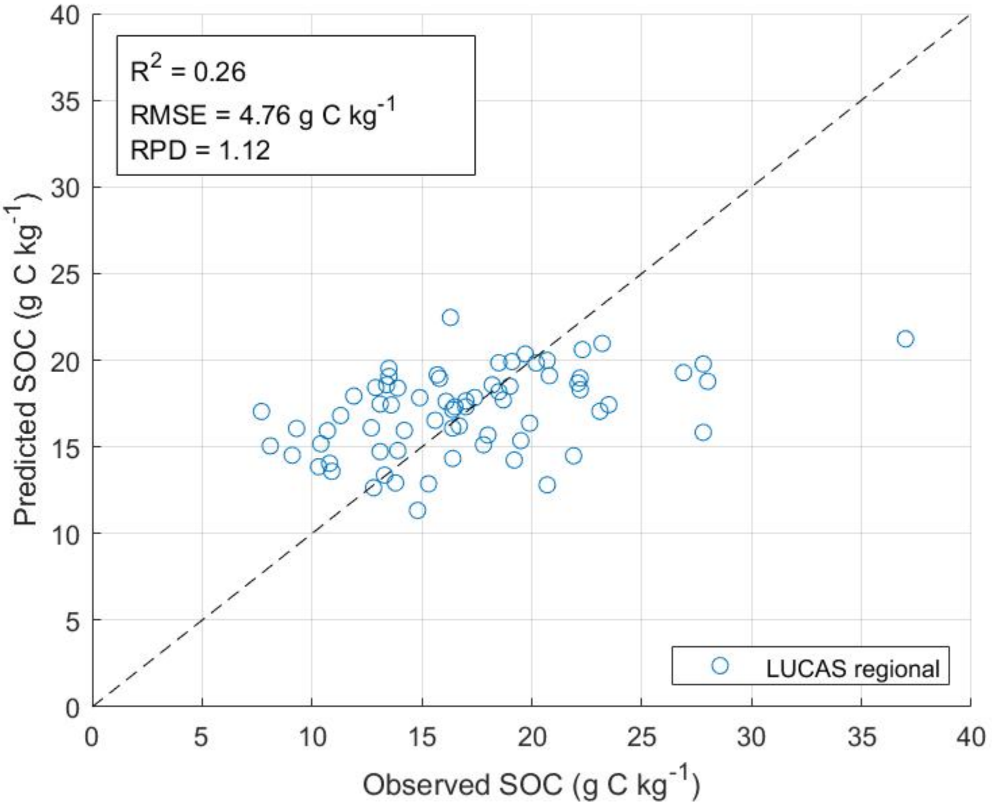


Figure 9: SOC prediction for LUCAS regional subset based on a PLSR model with 3 components.

4.4 Forward modelling

Both models were used to predict the SOC of the validation samples (n = 8) by a linear combination with the paired Sentinel-2 reflectance for each validation soil sample. The mean predicted SOC was 33.1 g kg^{-1} for the LUCAS cropland coefficients and 42.3 g kg^{-1} for the regional model coefficients (Fig. 10). These mean SOC values are twice and two and a half times higher than the measured values, respectively (Table 2).

Validation sample three, with a measured SOC of 15.1 g kg^{-1}, is a positive outlier for both predictive models, with a predicted 42.1 g C kg^{-1} for the cropland and 53.9 g C kg^{-1} for the regional model (Fig. 10). A comparison of the paired Sentinel-2 reflectance between all validation samples revealed that the reflectance for sample three is higher for four bands between 740 and 950 nm (Fig. 14, see Appendix). These reflectance properties lead to this sample being a prediction outlier.

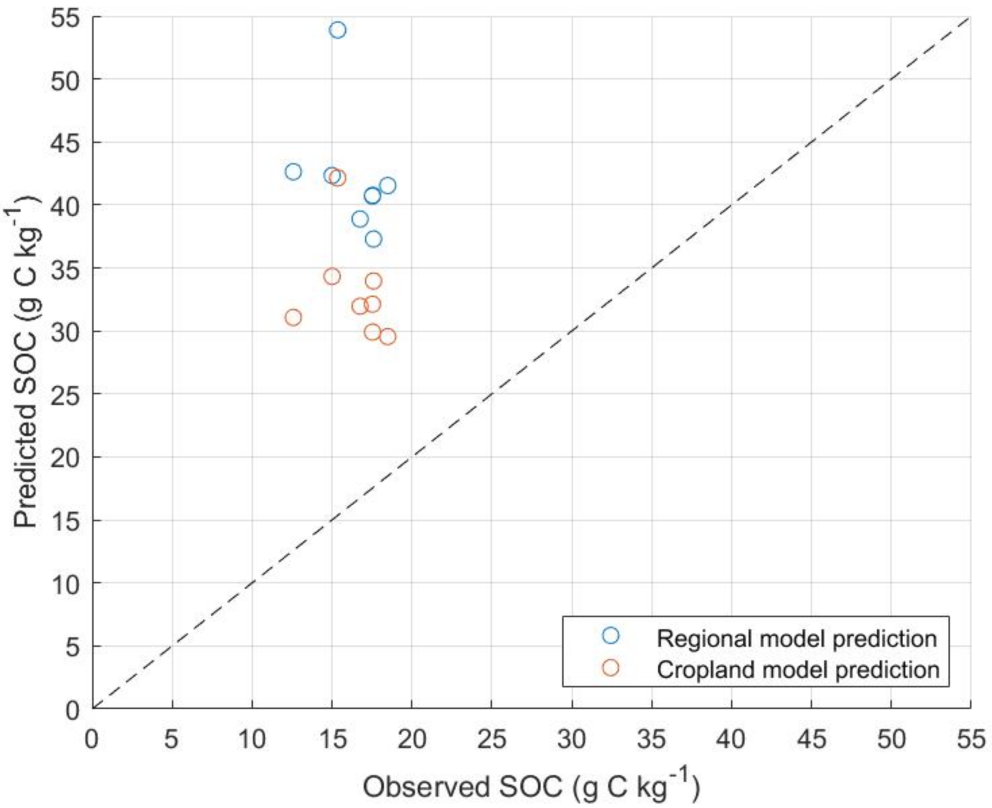


Figure 10: Predicted versus observed SOC of the validation soil samples. A set of predictions was computed for each of the PLSR linear models.

5 Discussion

5.1 Soil organic carbon prediction in research

The interest in large-scale, low-cost SOC quantification is reflected in the ever increasing body of research on the topic (Smith et al., 2018). Numerous SOC prediction papers have been published using spectral data derived from laboratory spectroscopy as well as hyperspectral and multispectral remote sensing platforms (Castaldi et al., 2019; Gholizadeh et al., 2018). Laboratory spectroscopy provides a high spectral resolution and consistent and controllable sampling conditions, which has allowed for good predictive models ($R^2 > 0.8$) using multivariate statistics (Stenberg et al., 2010). The multispectral data provided by the Sentinel-2 mission represents the lower end of the spectral and spatial resolution range that have been used for SOC predictions, but the high revisiting time and open access of the data make it attractive for modelling purposes (Castaldi et al., 2019).

5.1.1 Modelling soil organic carbon with the LUCAS

Both Ward et al. (2019) and Nocita et al. (2014) have established different methods to quantify SOC with the LUCAS dataset using forms of PLSR. The reference model developed by Ward et al. (2019) has been produced for all cropland samples of the LUCAS database (n = 8294), thus nearly the same calibration dataset than for the LUCAS cropland model in this thesis (n = 8332). Ward et al. (2019), however, included 3800 of the 4200 spectroscopy bands measured for each LUCAS sample, compared to the 12 simulated Sentinel-2 bands in my model. In contrast to the poor performance of the LUCAS cropland model presented above ($\mathrm{R}^2 = 0.159$, $\mathrm{RMSE}_p = 9.97$ g kg^{-1}), their model was proved to be more accurate ($\mathrm{R}^2 = 0.59$, $\mathrm{RMSE}_p = 7.37$ g kg^{-1}). This shows that the loss of spectral information through the resampling done in this thesis has a strong influence on the performance of the PLSR model. A PLSR produces regression coefficients for every predictor variable, which form a linear combination to approximate the response variables (Wold et al., 2001). Thus, the amount of variation in a dataset that can possibly be accounted for by the PLSR model decreases for less predictor variables or spectral bands, which is reflected by the different outcome of my and Ward et al. (2019) analyses.

5.1.2 Modelling soil organic carbon with Sentinel-2

While the poor model performance for the entire LUCAS cropland dataset is not evidence enough to conclude that the resampled SSL does not contain enough information to predict SOC with a PLSR, there is certainly too much spectral variation within this large dataset for accurate modelling. This might not be surprising considering the geographic extent of the soil samples, which leads to a diversity of soil types in the dataset (Orgiazzi et al., 2018). Thus, the mineral and organic chromophores determining the soil reflectance properties vary greatly within the large SSL.

Most attempts of SOC quantification with Sentinel-2 data have therefore been conducted on a much smaller scale and with a different sampling strategy compared to the LUCAS database (Castaldi et al., 2019; Gholizadeh et al., 2018). For that purpose, case specific SSLs are produced in an analogous manner to our validation dataset: Georeferenced SOC soil samples are paired with the reflectance for Sentinel-2 pixels that these samples are located in. The reduced geographical extent decreases the amount of natural soil variability that has to be accounted for by the model.

As the LUCAS dataset has a low spatial sampling density relative to the extent of my analysis, no LUCAS samples are located within the target study area. Instead, a regional SSL was built based on land use and the geographical distance to the study area in southern Sweden. With n = 70 samples, this regional SSL is comparable in sample size to the SSLs created by Castaldi et al. (2019), who also computed partial least squares regression models to approximate SOC. Castaldi et al. (2019) created one SSL for a larger study area comparable in size to my analysis, located in the Belgian loam belt. For the loam belt area, the SOC content was generally lower and less variant (mean = 10.7, std = 2.8 g kg^{-1}) than in the LUCAS regional SSL, but the the PLSR model yielded an equally poor RPD of 1.1. For the remaining SSLs, RPDs between 1.0 and 2.6 were achieved. Most notable differences between these SSLs and my dataset is the higher sampling density and the different sampling strategy. Several samples are taken from each sampled field in the study area, leading to a more spatially clustered sampling that accounts for in field variability. This stands in stark contrast to the dispersed LUCAS regional subset, where several kilometers lie in between each sample.

Moderate to good modelling results (RPD = 1.6 - 1.92) were also obtained by Gholizadeh et al. (2018), who created individual SSLs (n = 50) for four fields in the Czech Republic. This field scale approach functions well despite the minimal standard deviation of 1.5 to 3.9 g C kg^{-1} of the spectral libraries.

Both Castaldi et al. (2019) and Gholizadeh et al. (2018) compared their Sentinel-2 models to models derived from hyperspectral airborne data. They found that the hyperspectral data generally only led to a slight increase in model performance. This led them to conclude that the spectral resolution of Sentinel-2 is good enough for SOC modelling, if the size of the study area and the sampling strategy are appropriate. Too much variation in the dataset due to the large extent and dispersed sampling thus remains a possible explanation for the poor outcome of the LUCAS regional model.

5.1.3 Soil covariates and prediction accuracy

In addition to the geographical distribution of samples, there are prediction accuracy defining soil covariates that have been assessed in relevant research. Soil reflectance is not only determined by organic matter, but also mineral components and soil moisture (Baumgardner et al., 1986). A frequently discussed soil covariate for SOC detection is the sand content of the sampled soil. Stenberg et al. (2010) found that their multivariate model to predict SOC for Swedish agricultural soils improved significantly when the sandiest soil samples were removed from the calibration dataset. This was related to the quartz minerals in sand, which scatter light and thus obscure the absorbtion features of soil chromophores (Stenberg et al., 2010). Castaldi et al. (2019) found that of the 7 models created in their study, the one with the weakest performance (RPD = 1.0) occurred on very sandy soil. This study site was on morainal soil in the Demmin region of eastern Germany. The SOC range and standard deviation for sampled for that region is similar to the LUCAS regional SSL, but the mean is 4.2 g kg^{-1} lower (Castaldi et al., 2019). Nocita et al. (2014) also detected a slight increase in prediction accuracy upon considering the sand content as a covariate in their analysis. Their model, however, concerned the LUCAS database at hyperspectral resolution.

The mean sand content of the LUCAS regional dataset is 58 %, which indicates that there are predominantly as sandy soil types present. The exclusion of the sandiest samples (sand > 80 %) in this dataset did not improve the model performance, but it remains unclear if the generally high sand content has had an impact on the performance of the LUCAS regional model. This issue should be further investigated by creating regional models for areas with a finer textures.

5.1.4 Model refinement strategies

As shown in figure 5, there is a general reflectance difference between soil organic carbon classes after resampling the LUCAS dataset. This poses the question if good predictive models could be obtained from the resampled LUCAS dataset with an improved methodology.

One determining criterion for model performance is the selection of calibration samples (Stenberg et al., 2010). A meaningful calibration dataset should be representative of the range of SOC and soil properties while not including more variability than can be accounted for by the limited spectral information of the 12 bands. Several adaptations to the calibration sample selection have been found to improve the PLSR model accuracy for spectroscopy data.

Guerrero et al. (2016) investigated PLSR predictive SOC models created with SSLs varying in sample size and scale. They applied so called 'spiking' for all created models, which was achieved by statistically selecting eight representative samples of the validation dataset and adding them to the calibration dataset. The prediction accuracy of all spiked models ($0.835 < \mathrm{R}^2 < 0.962$) increased compared to unspiked models ($0.00 < \mathrm{R}^2 < 0.913$). Spiked calibration datasets have also been reviewed as a promising method in other relevant literature (Stenberg et al., 2010). The process of spiking, however, has only been applied for laboratory spectroscopy data in the reviewed literature for this thesis and not for multispectral datasets. To test the performance of spiked models compared to unspiked models with Sentinel-2 data, a large enough validation sample dataset

is required. This was not the case in this thesis (n = 8).
The physical distance to the target study area was considered for the selection of LUCAS regional dataset by including samples from adjacent administrative districts. Another conceivable method to produce appropriate calibration datasets are spectral distance algorithms, which was explored by Nocita et al. (2014) for the LUCAS database. For their local PLSR approach, an individual PLSR was performed for each validation sample based on calibration samples that were selected by their spectral similarity with the validation sample in question. The best modelling results were obtained when both the spectral and the geographic distance were considered for calibration sample selection (Nocita et al., 2014). This approach is computationally intensive, as an individual model is produced for every modelled unit. In contrast to the spiking method, however, no in situ SOC measurements are necessary to build these models.

Neither spiking nor local PLSR could be attempted in this thesis, with the critical issue being the profound spectral differences between simulated and remotely sensed data discussed in the next section.

5.2 Combining laboratory and satellite measured reflectance

Both the EU-wide and the regional model overestimated the SOC of the validation samples by a mean factor of 2 and 2.5 respectively (Fig. 10). Asides from the poor predictive power of the derived models, a possible reason for the observed shift in the predicted values lays in the difference between simulated and remotely sensed Sentinel-2 reflectance. As can be seen in figure 6, the mean reflectance of the remotely sensed validation samples is lower compared to the reflectance of the resampled or simulated Sentinel-2 reflectance of the two LUCAS calibration datasets, despite all datasets having similar SOC means (Table 2). According to the inverse relationship between SOC content and reflectance, the PLSR model resulting from the two LUCAS datasets lead to higher SOC predictions when they are applied for the reflectance of the SLU validation dataset. That this is the case despite the fact both models barely distinguish between high and low SOC of the calibration dataset (R^2=0.16 and R^2=0.26) speaks for the stark difference between laboratory and remotely sensed reflectance.

The lower reflectance of spaceborne sensors compared to laboratory measurements has also been noted by Gholizadeh et al. (2018). Their samples for the laboratory measurements and the Sentinel-2 reflectance measurements stemmed from the same fields, which increases the evidence of the general difference of reflectance spectra using different measurement methodologies.

This finding puts the feasibility of the initial aim in question. Even if an improved model could be derived from the resampled LUCAS database, this model would not be applicable for actual Sentinel-2 data. An empirical transformation of the resampled LUCAS SSL would be needed to approximate the reflectance that is actually recorded by the spaceborne sensor. A calibration dataset to detect and mathematically relate the observed reflectance difference for each band is thereby hardly conceivable, as a satellite cannot measure a laboratory soil sample and a satellite pixel cannot be measured in the lab. In addition, none of the examined scientific articles concerning soil property predictions with remotely sensed data makes use of a laboratory measured SSL.

6 Conclusion

The objective of this thesis was to investigate the possibility of adapting a large soil spectral library to build models for soil organic carbon predictions with multispectral Sentinel-2 data. Two models were created based on (i) all cropland samples in the LUCAS topsoil database and (ii) a regional subset of the LUCAS cropland samples. The prediction accuracy of the models was low with a RPD of 1.09 and 1.12, respectively. The poor model performance was related to the variability of soil reflectance in the calibration datasets, which is caused by their large spatial extent and dispersed sampling strategy. Both models overestimated SOC when they were applied for Sentinel-2 reflectance. This outcome was investigated, revealing that the laboratory reflectance data and remotely sensed Sentinel-2 reflectance are not readily compatible.

This leads to the conclusion that successful soil organic carbon predictions with Sentinel-2 data using the presented method are unlikely. Instead of combining reflectance data stemming from inherently different measurements techniques, future efforts could focus on establishing a comprehensive soil spectral library with Sentinel-2 data. Additionally, the calibration sample selection needs to be optimized to produce models capable of predicting soil organic carbon with a reasonable accuracy. Promising methods such as spiking and spectral distance algorithms have been established for spectroscopy data. These calibration sample selection methods could be tested for soil organic carbon modelling with Sentinel-2.

7 Appendix

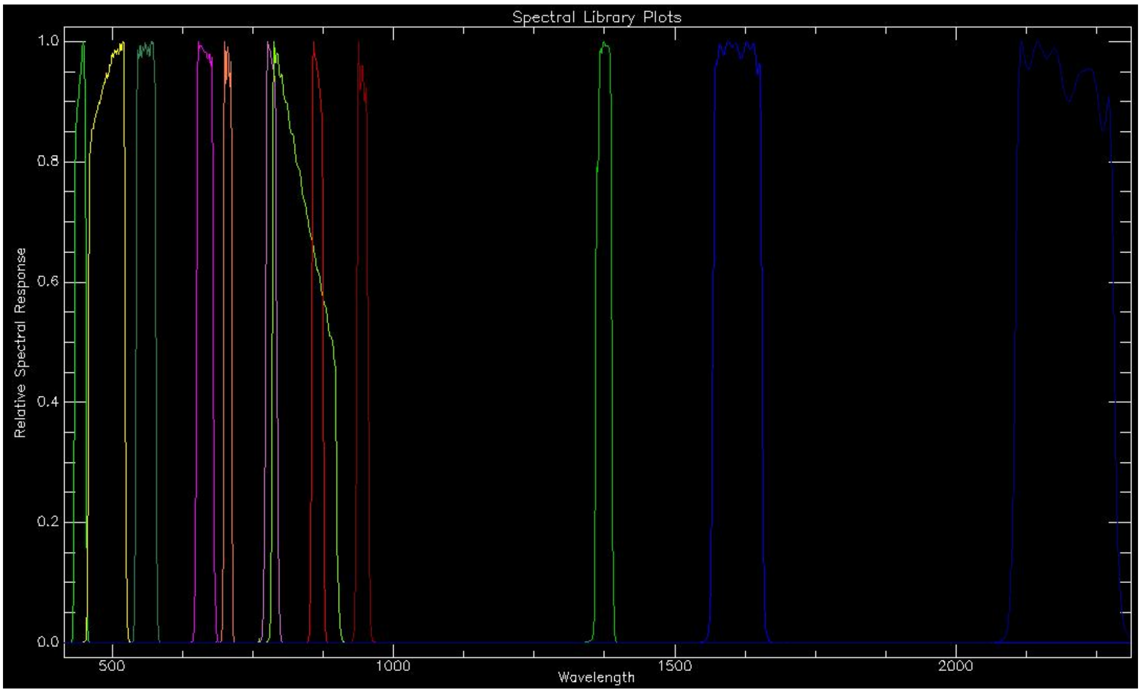


Figure 11: Sentinel-2 spectral band responses according to ENVI (v5.3). The spectral resampling of the LUCAS database was based on this information.

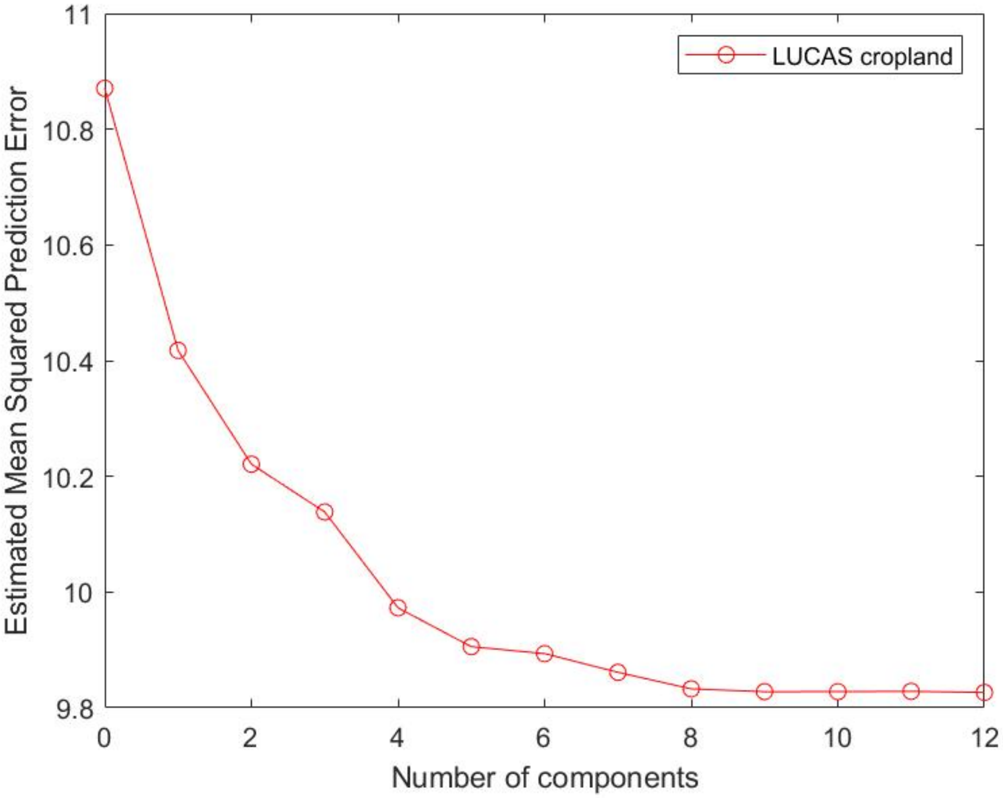


Figure 12: RMSE$_{cv}$ for the range of possible components of the LUCAS cropland PLSR.

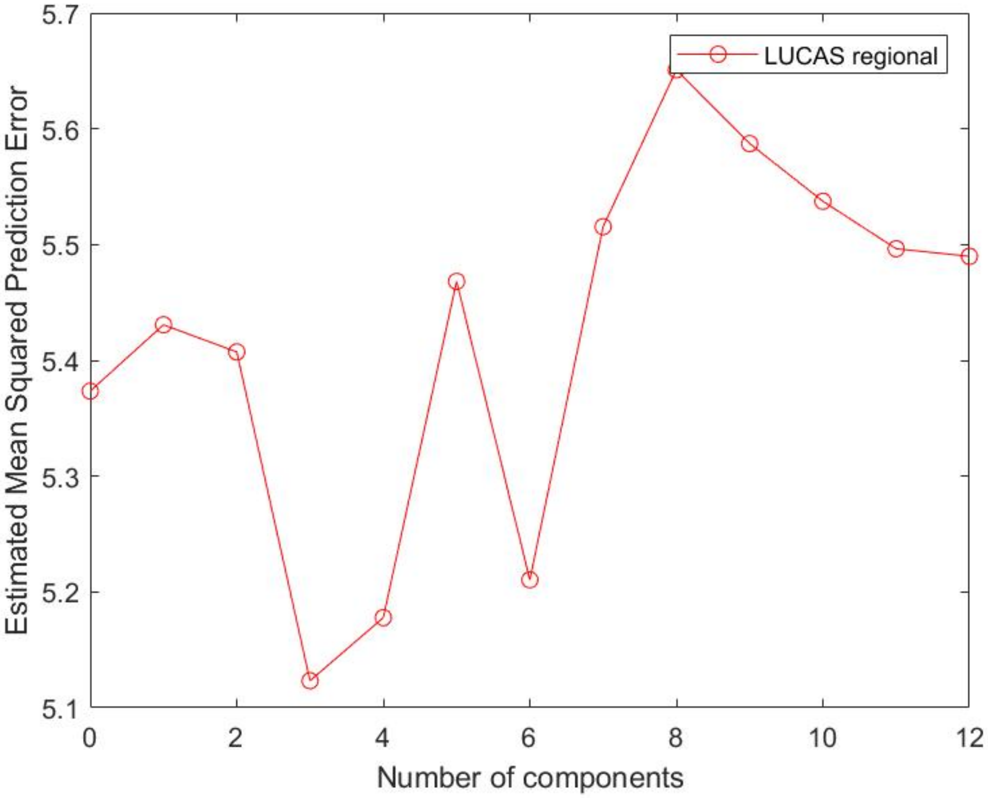


Figure 13: RMSE$_{cv}$ for the range of possible components of the LUCAS regional PLSR.

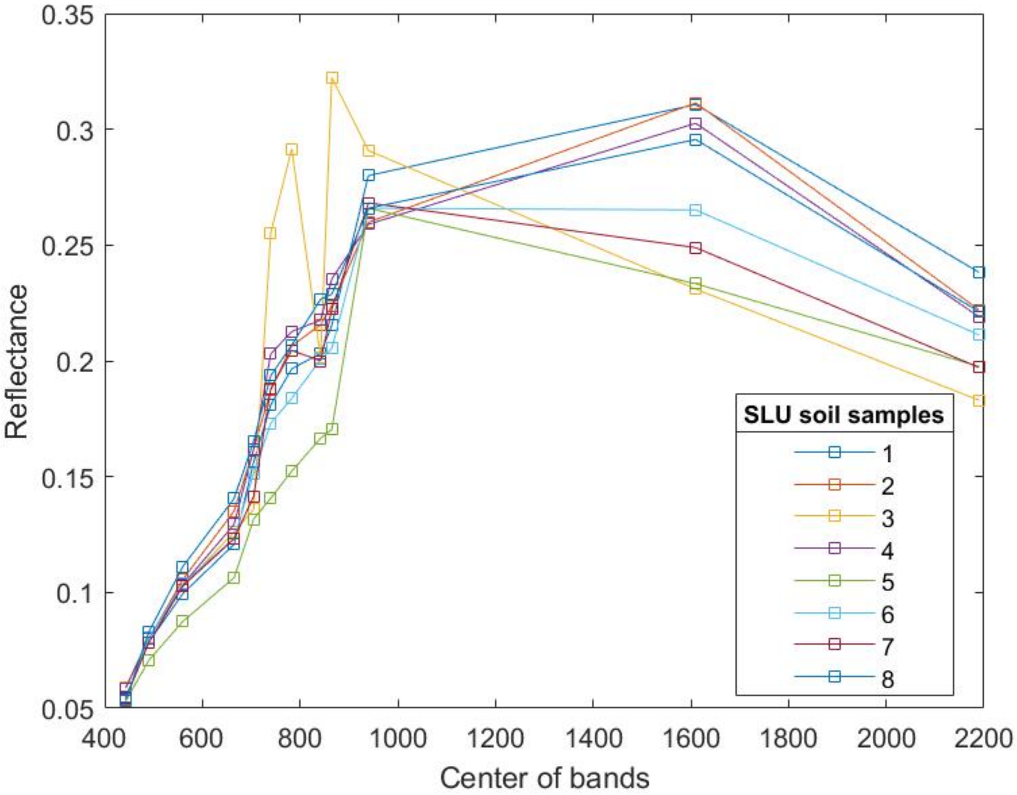


Figure 14: Sentinel-2 paired reflectance for each validation soil sample collected by SLU. Sample 3 has some positively outlying bands in the visible to NIR section of the spectrum (740-950 nm).

References

Baumgardner, M. F., Silva, L. R. F., Biehl, L. L., and Stoner, E. R. (1986). Reflectance properties of soils. *Advances in Agronomy*, 38:1–44.

Ben-Dor, E., Inbar, Y., and Chen, Y. (1997). The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sensing of Environment*, 61(1):1–15.

Ben Dor, E., Ong, C., and Lau, I. C. (2015). Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma*, 245-246:112–124.

Bhunia, G. S., Kumar Shit, P., and Pourghasemi, H. R. (2019). Soil organic carbon mapping using remote sensing techniques and multivariate regression model. *Geocarto International*, 34(2):215–226.

Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., and van Wesemael, B. (2019). Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147:267–282.

Conant, R. T., Ogle, S. M., Paul, E. A., and Paustian, K. (2011). Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. *Frontiers in Ecology and the Environment*, 9(3):169–173.

Dänhardt, J., Hedlund, K., Birkhofer, K., Bracht Jørgensen, H., Brady, M., Brönmark, C., Lindström, S., Nilsson, L., Olsson, O., Rundlöf, M., Stjernman, M., and Smith, H. G. (2013). *Ekosystemtjänster i det skånska jordbrukslandskapett. CEC Syntes Nr 01.* Centrum för miljö- och klimatforskning, Lunds Universitet, Lund.

De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263.

Gholizadeh, A., Žižala, D., Saberioon, M., and Borůvka, L. (2018). Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment*, 218:89–103.

Gomez, C., Dharumarajan, S., Féret, J. B., Lagacherie, P., Ruiz, L., and Sekhar, M. (2019). Use of sentinel-2 time-series images for classification and uncertainty analysis of inherent biophysical property: Case of soil texture mapping. *Remote Sensing*, 11(5):565–585.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., Zornoza, R., and Viscarra Rossel, R. A. (2016). Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil and Tillage Research*, 155:501–509.

IPCC (2007). *Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* [B. Metz, O.R. Davidson, P.R. Bosch, R. Dave, L.A. Meyer (eds)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jones, R. J. A., Hiederer, R., Rusco, E., and Montanarella, L. (2005). Estimating organic carbon in the soils of Europe for policy support. *European Journal of Soil Science*, 56(5):655–671.

Karlen, D. L., Mausbach, M. J., Doran, J. W., Cline, R. G., Harris, R. F., and Schuman, G. E. (1997). Soil Quality: A Concept, Definition, and Framework for Evaluation (A Guest Editorial) .

Lal, R. (2004). Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. *Science*, 304:1623–1627.

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

Oost, K. V., Quine, T. A., Govers, G., Gryze, S. D., Six, J., Harden, J. W., Ritchie, J. C., McCarty, G. W., Heckrath, G., Kosmas, C., Giraldez, J. V., Silva, J. R. M. d., and Merckx, R. (2007). The Impact of Agricultural Soil Erosion on the Global Carbon Cycle. *Science*, 318:626–629.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., and Fernández-Ugalde, O. (2018). LUCAS Soil, the largest expandable soil dataset for Europe: a review. *European Journal of Soil Science*, 69(1):140–153.

Panagos, P., Ballabio, C., Yigini, Y., and Dunbar, M. B. (2013). Estimating the soil organic carbon content for European NUTS2 regions based on LUCAS data collection. *Science of the Total Environment.*

Rochette, P. and Angers, D. A. (1999). Soil Surface Carbon Dioxide Fluxes Induced by Spring, Summer, and Fall Moldboard Plowing in a Sandy Loam. *Soil Science Society of America Journal*, 63(3):621–628.

Smith, P., Lutfalla, S., Riley, W. J., Torn, M. S., Schmidt, M. W., and Soussana, J. F. (2018). The changing faces of soil organic matter research. *European Journal of Soil Science*, 69(1):23–30.

Spaccini, R. and Piccolo, A. (2013). Effects of field managements for soil organic matter stabilization on water-stable aggregate distribution and aggregate stability in three agricultural soils. *Journal of Geochemical Exploration*, 129:45–51.

Steinberg, A., Chabrillat, S., Stevens, A., Segl, K., and Foerster, S. (2016). Prediction of common surface soil properties based on Vis-NIR airborne and simulated EnMAP imaging spectroscopy data: Prediction accuracy and influence of spatial resolution. *Remote Sensing*, 8:613–633.

Stenberg, B., Jonsson, A., and Börjesson, T. (2005). Use of near infrared reflectance spectroscopy to predict nitrogen uptake by winter wheat within fields with high variability in organic matter. *Plant and Soil*, 269:251–258.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., and Wetterlind, J. (2010). Visible and Near Infrared Spectroscopy in Soil Science. *Advances in Agronomy*, 107:163–215.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., and van Wesemael, B. (2013). Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. *PLoS ONE*, 8(6).

Tola, E. K., Al-Gaadi, K. A., and Madugundu, R. (2019). Employment of GIS techniques to assess the long-term impact of tillage on the soil organic carbon of agricultural fields under hyper-arid conditions. *PLoS ONE*, 14(2).

Tóth, G., Jones, A., and Montanarella, L. (2013a). The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environmental Monitoring and Assessment*, 185(9):7409–7425.

Tóth, G., Jones, A., and Montanarella, L. e. (2013b). *LUCAS Topsoil Survey: Methodology, Data, and Results.* Publications Office of the European Union, Luxembourg.

Van-Camp, L., Bujarrabal, B., Gentile, A.-R., Jones, R. J. A., Montanarella, L., Olazabal, C., and Selvaradjou, S. K. (2004). *Reports of the Technical Working Groups Established under the Thematic Strategy for Soil Protection. EUR 21319 EN/3.* Office for Official Publications of the European Communities, Luxembourg.

Ward, K. J., Chabrillat, S., Neumann, C., and Foerster, S. (2019). A remote sensing adapted approach for soil organic carbon prediction based on the spectrally clustered LUCAS soil database. *Geoderma*, 353:297–307.

Wesemael, B. v., Paustian, K., Meersmans, J., Goidts, E., Barancikova, G., Easter, M., and Schlesinger, W. H. (2010). Agricultural management explains historic changes in regional soil carbon stocks. *Proceedings of the National Academy of Sciences of the United States of America*, 107(33):14926–14930.

Wold, S., Sjostrom, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130.