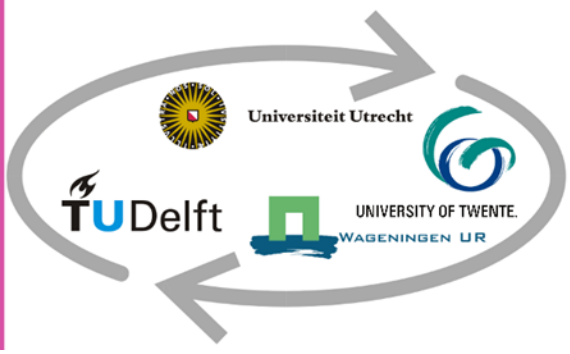
GIMA
Geographical Information Management and Applications

Who can command the Random Forest and make the trees pull Data out of the earth?

Predicting soil types through Random Forest machine learning using open-source data

Master's Thesis

Author: Daan Molleman
Supervisor: Prof.dr.ir. Peter van Oosterom
Responsible Professor: Jesús Balado Frías PhD
External supervisor: Kap. Bas J.H. Ooink MSc
Date: 01-03-2021



This page was intentionally left blank.

Preface & Acknowledgements

This master thesis is written as part of the GIMA programme from September 2020 to February 2021. In the period of writing this thesis I have learned a wide variety of new skills and techniques that will help me progress my career in GIS. In the beginning I was unfamiliar with both machine learning and soil science, and by doing this research, I have gained great insight into the ins and outs of both topics.

I would like to give out a special thanks to Bas Ooink who has been enormously helpful in providing both technical and subject-related assistance. Right from the beginning we had a strong and healthy working environment which we kept going during the whole period.

Next, I would like to thank Peter van Oosterom who provided me with guidance on subjects such as structure and performing research. Jesús Balado Frías was most helpful in assisting me with in depth questions on machine learning methods and always provided quick and detailed feedback.

Abstract

For military terrain analysis, a detailed soil map is needed to assess the terrain accessibility during a mission. Predicting soil classes where no data is available is a difficult task. For this reason, the Random Forest algorithm has been applied to predict the individual soil properties: sand, silt, clay, coarse fragments, organic content, and cation exchange capacity which can be combined into the Unified Soil Classification system (USCS) soil classification. Using open-source data points in combination with explanatory variables that are available for all of Europe, the model is trained to predict soil property values in areas where no soil samples are present at a spatial resolution of 30 meters. The predictors used include satellite imagery, spectral indices, hydrological data, digital elevation models and its derivatives. As the liquid limit and the plasticity index are both needed for the USCS, but are not included in European soil samples, they must be calculated using the other soil properties. From the clay content and the cation exchange capacity, a linear regression model was set up using US-data in order for the two properties to be predicted in Europe. The linear regression model reached an R-squared of 0.842 for the liquid limit and 0.895 for the plasticity index. In this study, an innovative method for model validation is used that ensures consistent validation statistics by generating subsets that each contain points that are well distributed over the entire range of values and are also geographically dispersed. It was found that hydrological predictors scored high in importance when predicting sand, silt, clay and cation exchange capacity. Moreover, variation in coarse fragments was mostly explained by the digital elevation model and its derivatives, and the nitrogen content of a soil reaped the highest importance in predicting organic content in soils. In predicting the individual soil properties, the highest amount of variation was explained for clay content, resulting in 68.5%. This is followed by sand and silt content (48.4% and 57.9% resp.). For cation exchange capacity and organic content, explained variations of 40% and 44.5% were attained. The lowest R-squared statistic was reached for coarse fragments where only 38.1 percent of the variation was explained. The Random Forest algorithm proved effective in predicting soil properties with limited samples available while maintaining a spatial resolution of 30 meters. Additionally, an improved method for determining the Atterberg limits was developed to be used in areas where no data on the limits is available. Furthermore, a validation method was constructed that provided consistent statistics describing explained variation.

Key words: predicting soil properties, USCS, liquid limit, plasticity index, Random Forest, machine learning, remote sensing, open-source data.

Contents

1. Introduction

In military tactical analysis, information on the enemy's potential movements is of vital importance. Knowing what routes are accessible for each type of land vehicle provides insight into possible scenarios and allows for educated decisions on troop and vehicle placement. The Dutch military takes great use out of a tool that shows both the possibilities and impossibilities of movement during an ongoing mission. Such an educated decision regarding movement can be based on a terrain analysis. Part of the terrain analysis is the Environmental Evaluation (EE) that assesses the relevant environmental aspects of a certain region to later be used for the final terrain analysis visualisation of accessible and inaccessible terrain. Traditionally these analyses are done by hand and were therefore prone to error and delays. It is expected that applying GIS to automate the process will reap benefits regarding accuracy and time, both of which are essential in the context of military operations.

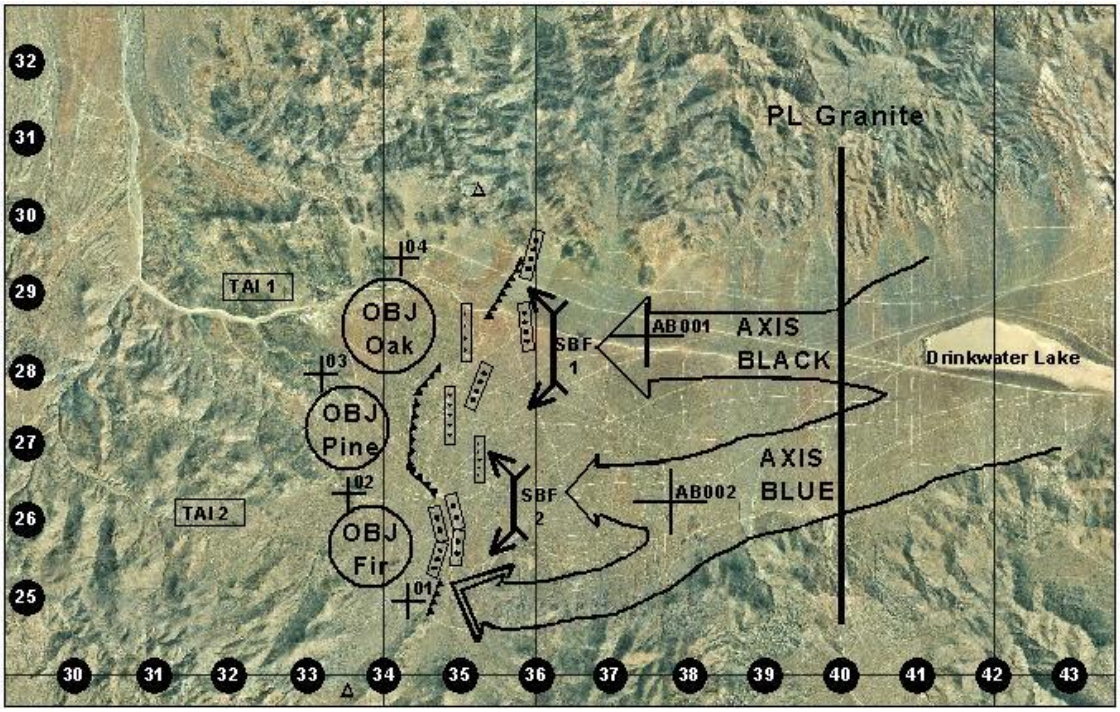


Figure 1.1: *military terrain analysis with (source GlobalSecurity, 2021)*

In order to conduct an automated terrain analysis in a military context, it is necessary to distinguish the different characteristics of the terrain. The elements that are part of the environmental aspects are slope, soil, vegetation infrastructure, and water properties such as drainage and groundwater (see figure 1.2). When these 5 elements are evaluated correctly, it is possible to assess to what extent the terrain will impact troop and vehicle movement (Koch et al., 2012). Additionally, predictions can be made on advantageous locations for both forces, allied and enemy. To illustrate, an example of a terrain analysis provided that also indicates possible moving routes.

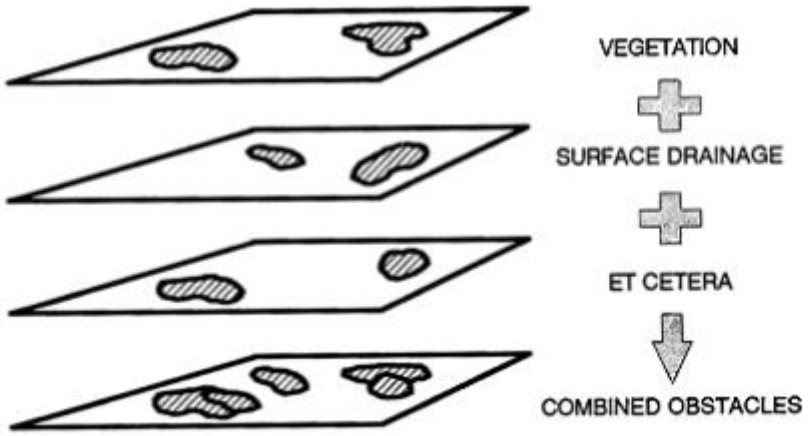


Figure 1.2: *environmental analysis (from Glinton et al., 2004)*

Evaluating terrain characteristics in a military context requires data with a spatial scale of 1:50,000. A scale of this size is necessary for proper assessment of the smallest military unit, a platoon. Any scale coarser than this will not provide enough detail to make sound judgements of the terrain accessibility and will therefore leave room for error (Koch et al, 2012). Data with a scale of 1:50k is available for most environmental aspects in Europe, excluding soil. It is crucial that a soil map with a proper scale is available in order to construct a usable product. Since such a soil map is not readily available, one must be constructed. This can be done through combining both vector and raster data using a soil type classification that is appropriate for assessing soil accessibility. Data used in this research will be outlined later. A widely used soil classification is the Unified Soil Classification System (USCS). This soil classification system is also used in other military operations, making it a suitable classification system for the automated terrain analysis. Across militaries, vehicles have been tested to what degree they can pass certain soil types. Their pressure on the soil has been calculated, which can be related to the soil classes in the USCS. Consequently, given a USCS soil map, a go/no go map can be created for each vehicle.

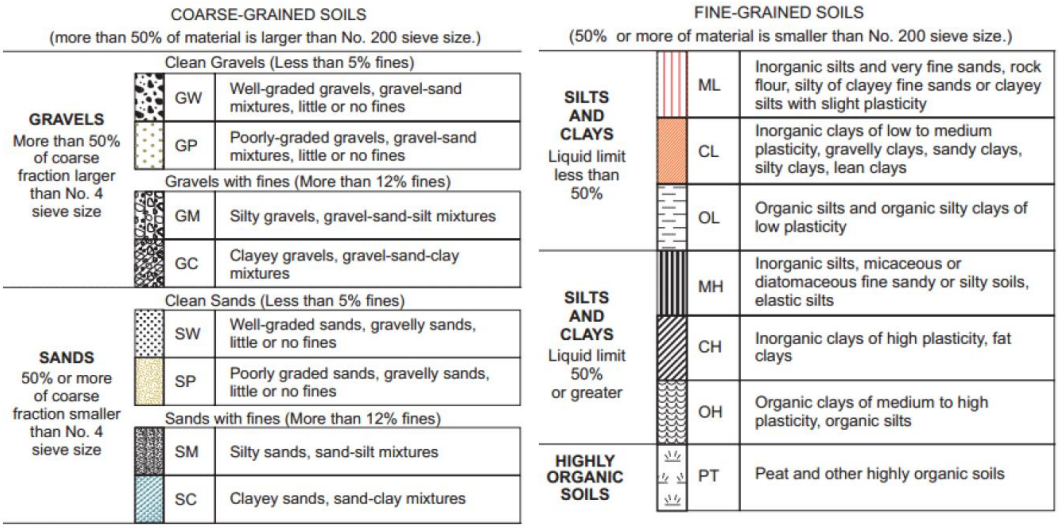


Figure 1.3: *The USCS classification (Lehner & Harmann, 2007)*

While soil data is crucial for vehicle movement and therefore the terrain analysis, little to no data is available on a global scale. When this data is available however, the scale is often too coarse to be of any use in the analysis. Additionally, some data on soil is available within the agricultural sector, but

these datasets often use different classifications as they are meant so assess soil suitability and fertility for crops. When searching for open-source soil datasets using the USCS classification, nothing is available, as this type of data is restricted to foreign militaries (Koch et al., 2012). Collecting soil data can be labour-intensive, time consuming and expensive. More specifically, it is not recommended to perform soil sampling in an area that is under military threat. Moreover, using older, likely more inaccurate maps could result in misinterpretations and uneducated decisions. It is therefore useful to have an updated version of the relevant area available in order to make fast and educated decisions. However, as there are no maps available at the proper spatial resolution, this is not yet possible. To remedy the lack of soil datasets available, mapping technologies can be applied to model the required soil properties at an accurate enough scale. This relates to the scientific specialisation of Digital Soil Mapping (DSM) that encompasses modelling soils for all sorts of purposes (Grunwald et al., 2011). The main intention of this scientific trend is to step away from static soil maps that delineate areas with certain taxonomic classes using strict borders between them and move towards digitally created maps using modern techniques to reduce costs and enhance the spatial resolution. This trend was set in motion by Young (1973) who advocated for soil surveys that included all sorts of different soil characteristics and changes, rather than one static map derived from historic documents. Moreover, old-fashioned soil maps tend to only cover smaller areas of a country or region, and when a larger area is subject to research, the quality of the spatial resolution tends to decrease rapidly (Grunwald et al., 2011). The improved spatial resolution is of most importance as McBratney and Pringle (1999) suggested that spatial autocorrelations for sand, clay and chemical soil properties were occurring at a spatial resolution of less than 300m.

In the last decade, research has been conducted on combining soil samples with machine learning methods to construct a classified soil map of a specific region (Santanello et al., 2007; Da Silva Chagas et al., 2016). When certain truths are combined with assumptions and a machine learning algorithm, the soil types of a region can be predicted with varying accuracy (Gambill et al., 2016). An example of a machine learning method is the Random Forest algorithm that combines multispectral imagery, height models, weather and water data, available soil samples and many other variables to construct a soil map according to a classification (Breiman, 2001; Tesfa et al., 2009). This method caters to the desires of the Joint ISTAR Commando as they require a soil map with a scale of 1:50,000, constructed through open source rasters and freely available soil samples. Such a map will assist them in creating an automated terrain analysis method and therefore improves the speed in which an accessibility map can be computed.

The Random Forest algorithm is a method mainly used in ecology and soil science as it provides high accuracy and is fairly robust (Cutler et al., 2007; Brungard et al., 2015; Heung et al., 2014). Moreover, it is resistant to overfitting and the method is insensitive to value range, meaning that it does not require standardisation (Breiman, 2001). In soil science, the method is used for mapping organic matter present in the soil (Grimm et al., 2008; Wiesmeier et al., 2011). Additionally, Random Forest is used to analyse soil depth (Tesfa et al., 2009) and to update soil survey maps (Rad et al., 2014). An advantage of the Random Forest algorithm is that it is capable of dealing with relatively small sample sizes and unstandardised data inputs (Qi, 2012). Moreover, the Random Forest algorithm allows for an assessment of variable importance, providing the ability to remove unimportant and possibly distorting predictors (Strobl & Zeileis, 2008). The Random Forest algorithm is also proven successful in handling data with high multicollinearity (Belgiu & Drăguț, 2016), meaning that providing datasets that show highly similar patterns will not negatively influence the result.

The structure of this thesis is as follows: first, a short overview of the research objectives is given, stating the research questions and its limitations and scope. Next, the theoretical background for this thesis is outlined. The soil classification systems are discussed and the needed soil properties for classification are explained. The workings of the Random Forest algorithm are also detailed in the theoretical section. In section 4, the methodology for this thesis is provided, containing the geographical delineation of the research area, a summation of the training and comparison data and a workflow for performing the Random Forest algorithm. This section also details the method for verifying the results and generating the final USCS classification. The results section displays the results that were initially generated and follows up with a more in-depth run of the Random Forest algorithm by describing the variable selection process and providing the results for this second iteration. Next is the discussion, that deliberates on the methods and results of this thesis and identifies its shortcomings. Moreover, this section compares the results to similar academic research, and lastly provides topics for further research. Last is the conclusion, summarising the results and answering the research questions posed in section 2. In this section, a compact description of the research aims will be provided to identify the goal of the thesis. From this explanation, one main research question is constructed, encompassing the global direction of this research. To provide more exact objectives of the research, several sub-questions are set up that each tackle one aspect. These aspects then all add up to provide a greater insight into the main research question.

1.1 Research aim and Questions

The aim of this thesis is to assess the effectiveness of machine learning in constructing a soil map to be used in an automated military terrain analysis. It is done by only using open-source data such as multispectral imagery, digital terrain models and soil samples. The machine learning method used is the Random Forest algorithm since it has proven useful in previous research and allows for accurate predictions with a relatively small supply of samples. The goal is to achieve relatively accurate spatial resolution by using ground samples and high-resolution imagery and rasters. This combination will allow this method to be used anywhere in Europe, as all collected data is available on a European scale. For this same reason, the spatial resolution of the final product will also be similar across Europe, only shifting slightly due to the geographic projection of the output raster. The output rasters are classified according to the USCS classification to ultimately be used in a military terrain analysis. The research question for this thesis is as follows:

"To what extent can soil be classified for the use of a terrain analysis through the Random Forest machine learning method using open-source data?"

To answer this main research question more accurately, it has been divided into several sub questions that each answer a part of the main research question. The sub questions are as follows:

1. *With what accuracy can the Random Forest algorithm predict distribution patterns of soil properties?*
2. *To what degree does the algorithm predict the soil property values correctly?*
3. *To what extent can the soil property maps be combined to calculate the liquid limit and plasticity index?*
4. *What predictors are most influential in predicting soil properties?*
5. *In what way does the used data influence the model's outcome?*
6. *What level of uncertainty is present when using the Random Forest algorithm to predict soil properties?*

The first sub question assesses whether the Random Forest algorithm can predict where certain soil properties are relatively high, and where they are relatively low. This does not consider the actual values of the soil properties and only investigates the visual distribution of high and low values. The second sub question does ask whether the predicted soil values are correct and is answered by applying statistical methods to combine the predicted and observed values. The liquid limit and plasticity index are complicated soil properties, which will become evident later. Therefore, methods are explored to determine these soil properties and the best one is chosen. Next, sub question four takes both a qualitative and quantitative approach to assess what independent variables explain variation in the dependent variable, i.e., the soil properties. The next sub question takes a qualitative approach and considers what influence the input data has on the output of the Random Forest algorithm and investigates why this effect occurs. Last, there is the sub question that assess the amount of uncertainty present when using the Random Forest algorithm to predict soil properties. This is done through a statistical method that uses the Random Forest predictions to create a 90% confidence interval that indicates how certain the model is of its predictions.

1.2 Research Scope

The largest limitation to this research is expected to be the use of open-source data. This type of data is often not as detailed as commercially acquired data and is more likely to contain errors. It is therefore necessary to make use of multiple sources of open data in order to calibrate and verify the product of this research. Moreover, the scope of this research does not contain further work on the automated terrain analysis. This thesis is solely meant to explore soil analysis and achieve the best accuracy possible. Additionally, the area in which the research will be conducted is limited to the Netherlands, Flanders and the westmost part of Germany. This is chosen due to its representativeness in regard to soil types and the readily available data for comparing the results. Additionally, the research area is kept to this size to compensate for the limitations in computer processing power. Last, the prediction of soil characteristics is limited to areas without trees and urban land uses.

1.3 Research Methodology

Now that the objective and scope have been outlined, an overview can be given of the general methods that will be used to complete this objective. The main method for this study is the Random Forest algorithm, which is a machine learning method that allows for prediction of a dependent variable through a set of independent variables. This method will be elaborated on in section 2.6. For running the algorithm, a training, validation and testing phase is performed. These will result in statistics that are used to assess the accuracy of the model. To run the Random Forest algorithm to predict soil properties, soil samples are needed to train the algorithm, the sources of these samples are also outlined in the research design. In addition, the set of independent variables used to predict the soil properties is listed, and their relevance to the soil properties is outlined. To calculate additional soil properties needed for the USCS, a linear regression is performed to create a function that determines the value of these two. The final soil classification is made using a simple script, from which a soil map is created of the research area. This area is compared to an existing soil map to assess the accuracy of the classification.

2 Theoretical Background

In this section, the theoretical foundation of this thesis is outlined, providing insight into the concepts used to conduct the research. First, two soil classification systems are discussed and compared, after which a choice is made for the most suitable one. Next up, the Liquid Limit and the Plasticity Index are explored, and a linear regression model is presented to calculate both soil characteristics using only clay and 'cation exchange capacity', which is also explained. Additionally, several soil moisture characteristics are listed and their relationships to soil characteristics are explained. The same is done for biochemical properties that can explain the required soil characteristics. Last, the core concepts of the Random Forest algorithm are elaborated on.

2.1 Soil Classification Systems

When predicting soil characteristics, it is crucial to translate findings to the proper soil classification systems. Preferably, this should be one that allows for an assessment of vehicle accessibility. For this choice, there are three options that are most suitable: The United States Department of Agriculture (USDA), the Unified Soil Classification System (USCS) (Casagrande, 1948) and the American Association of State Highway and Transportation Officials (AASHTO). The United States Department of Agriculture soil classification (USDA, figure 2.1) is the simpler one of the two, providing a relatively easy classification process when translating raw soil characteristics to the soil classification system (Soil Survey Division Staff, 1993). This soil classification system is primarily used in the agricultural sector and therefore lacks information on other soil characteristics. Moreover, the USDA does not incorporate the presence of gravel, rocks and organic material in the soil. Without information on the presence of these materials, some crucial soil types for assessing vehicle accessibility are left out (Koch et al., 2012). The second system, the Unified Soil Classification System (USCS, figure 2.2) is used for engineering purposes and does incorporate all required soil types in order to perform a terrain analysis (Koch et al., 2012), but can only be accurately measured by a hands-on soil test. When a hands-on soil test is not available, one can use the Liquid Limit and the Plasticity Index to accurately predict the USCS (Casagrande, 1948; Gambill et al., 2016). These two factors are explained in the following section. Moreover, there is no clear-cut method to transform the USDA into the USCS. Using conventional tools, a 40% error margin exists when performing this transformation (Garcías-Gaines & Frankenstein, 2015). Machine learning tools, however, have proven useful in translating the USDA into the USCS with a mere error margin of 2% (Gambill et al., 2016). This method, however, does require detailed information on soil properties such as sand, silt, clay and available water storage at multiple depths.

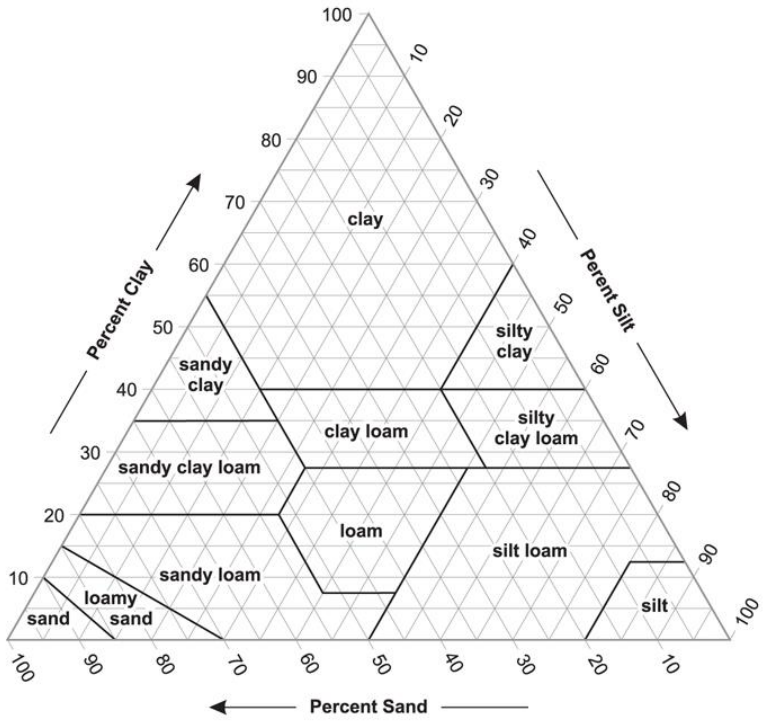


Figure 2.1: *The USDA classification (García Gaines & Frankenstein, 2015)*

To determine sand, silt, clay and gravel content in the soil, the USCS uses a range of sieves in order to define grain size. These are used in the Casagrande (1948) method and are still applied in current-day measurements. The collection of sieves includes the sizes of: 76.2mm, 4.75mm and 0.075mm. All elements that do not pass the 76.2mm sieve are categorised as rocks and boulders and do not take part in determining the USCS classification. Material that passes the 76.2mm sieve and are retained on the 4.75mm sieve are called coarse grained soil or gravel. Soil that passes the 4.75mm sieve but are held back by the 0.075mm sieve are sands. Anything that passes the 0.075mm sieve is either silt or clay. The sieve sizes differ in some regions of the world, but the USCS strictly handles the mentioned sieve sizes.

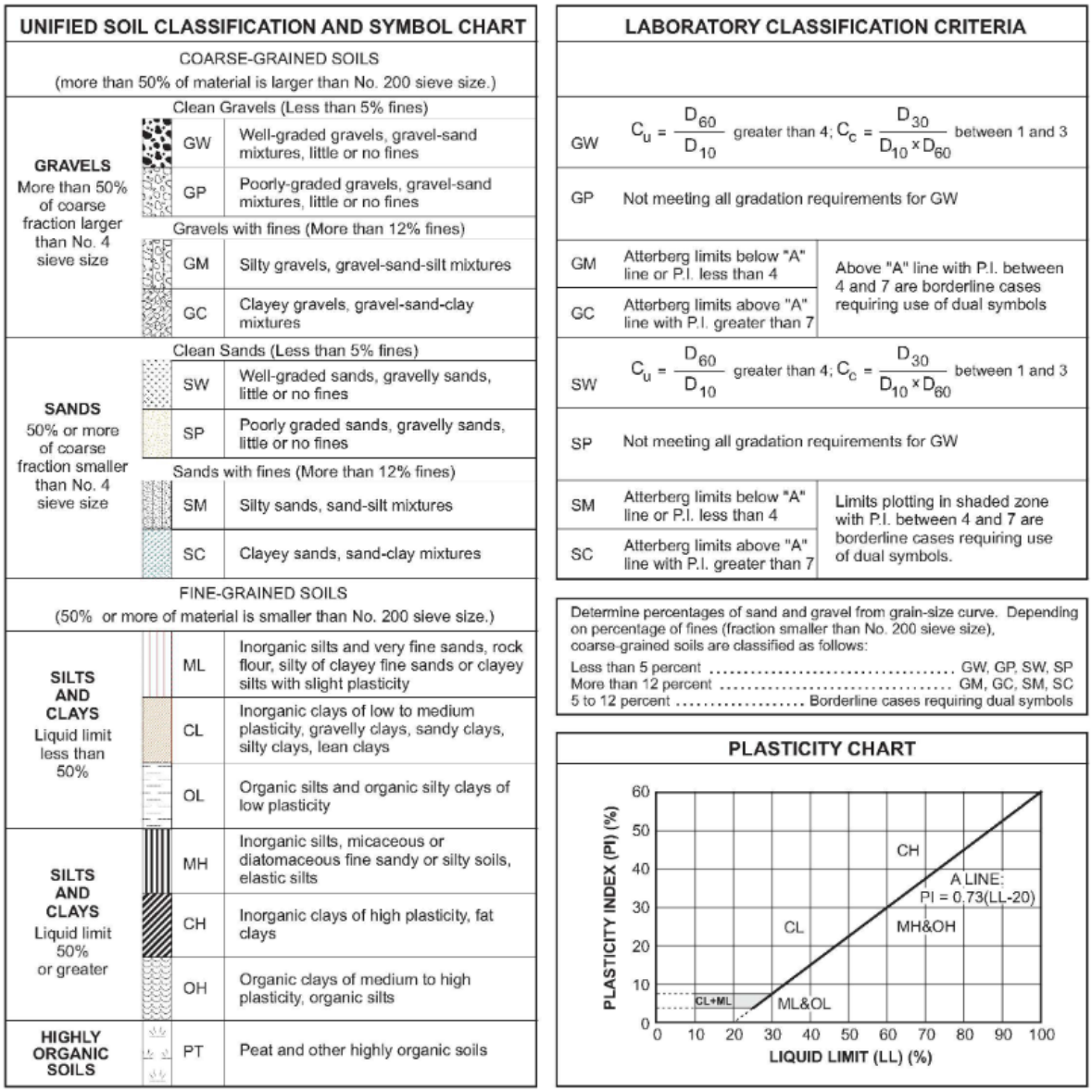


Figure 2.2: *The USCS classification chart including A-line (Caltrans, 2020)*

When using the USCS classification table to classify a soil, one must first check the percentage of coarse materials in the soil, as that divides the soil into either one of two categories: fine- or coarse-grained soils. When a soil falls into the coarse-grained group, it is relevant to determine whether the majority of the grains is gravel or sands. A gravel is either well-graded or poorly graded, based on the spread in the size of the soil particles. When the spread of sizes is even, it is called well-graded, when there is little variation in size, it is called poorly graded. If sands take up the majority of the soil, it is a sandy soil. To further narrow it down, the amount of fine grains in the soil must be examined. When this is less than 5 percent, it is either well or poorly graded. If it is more than 12 percent, the A-line must be consulted to determine whether the sand is clayey or silty. Alternatively, when the share of clays and silts in a soil is 50 percent or more, it is classified as a fine-grained soil. Then, the liquid limit is addressed to determine whether it falls into the low-plasticity or high-plasticity category. For both of the categories, the A-line is used to determine the final classification. For peaty soils, the liquid limit is either 50 or higher and the organic content is more than 20% or, the liquid limit is lower than 50 and the organic content is higher than 15%.

Another soil classification system is the American Association of State Highway and Transportation Officials (AASHTO). This is mostly used for the construction of infrastructure, and therefore covers some soil aspects that relate to the accessibility of terrain by certain vehicles. Similar to the USCS, the AASHTO is generated using particle sizes and plasticity characteristics from the Atterberg limits. It was developed in 1929, and revised multiple times over the last century (Ishibashi & Hazarika, 2010). The AASHTO categorises soils with the purpose to be later used for engineering and infrastructure construction. The classification system is therefore mostly used in organisations that perform these types of operations (US Air Force Engineering Support Agency/Civil Engineering Squad (AFCESA/CES), 1997).

2.2 Cation Exchange Capacity

In calculating both the liquid limit and the plasticity index, the Cation Exchange Capacity (CEC) is used. This refers to the capacity of exchangeable cations a soil can hold. Cations are positively charged ions that can attach to soils (Hazelton & Murphy, 2016). The CEC-property is important in influencing a soil's structural capabilities, nutrients, and pH values. It is expected that soils that contain a large fraction of clay also possess a high cation exchange capacity. Similarly, organic matter tends to have higher CEC values (Dixon, 1990). Clay minerals and organic matter consist of surfaces that have negatively charged spots that can absorb positively charged ions. This is done through electrostatic force which makes positively charged and negatively charged ions attract. Before the cation exchange capacity is determined, salts must be removed from the soil (Hazelton & Murphy, 2016). They are considered as a separate resource of ions and are therefore washed away with water before measurements take place. Otherwise, they saturate the soil and its cation exchange sites preventing proper measurement of the soil property. Before taking any measurements for determining the CEC, a soil is analysed, and the amount and type of exchangeable bases are assessed. The types are calcium, magnesium, sodium, and potassium. Using these, base saturation can be derived by comparing the exchangeable bases and the cation exchange capacity (Sumner & Miller 1996).

2.3 Liquid Limit and Plasticity

When classifying organic soil types such as peat according to the USCS, the factors of Liquid Limit and soil Plasticity are of importance. The Liquid Limit defines the amount of water a soil type can take before becoming liquid (McBride, 2002). This of course differs per soil type and is relevant for vehicle accessibility. Plasticity is a measure to assess the strength of a soil type. It defines the capability of a soil type to suffer deformation at a steady rate without breaking or falling apart (Seybold et al, 2008). Seybold et al. 2008 used a linear regression model to estimate the soil liquid limit and its plasticity index using only basic soil properties.

The liquid limit and the plasticity index are part of the Atterberg limits (Atterberg, 1911). The Atterberg limits are moisture content limits and define a soil's degree of resistance to deformation (McBride, 2002; Seybold et al., 2008). The states are solid, semi-solid, plastic and liquid. The water content at which a soil turns from solid to semi-solid is called the shrinkage limit, from semi-solid to plastic is called the plastic limit, and from plastic to liquid is called the liquid limit. Moreover, the width of the plastic state, being the liquid limit minus the plastic limit, is called the plasticity index. Plasticity can

be defined as the capability of a soil to endure unrecoverable deformation at a stable volume without showing cracks or falling apart. The liquid limit is the amount of water a soil can hold until it starts behaving as a liquid (Atterberg, 1911; Seybold et al., 2008).

Seybold et al (2008) use a soil's clay content and cation exchange capacity to determine both its liquid limit and its plasticity index. Cation Exchange Capacity is discussed in the next subsection. They dictate that clay correlates significantly with plasticity and shrinkage limit. They also state that the organic content of a soil can impact its Atterberg limits, but they find it to be non-significant for all cases. Two soils can have the same PI but can act differently in plasticity over different ranges of soil moisture due to differing limits, making it essential to both assess the liquid limit and plasticity index of a soil. Additional to clay, the cation exchange capacity also greatly explains liquid limit (De La Rosa, 1979; Mbagwa & Abeh, 1998). While De La Rosa (1979) and Mbagwa and Abeh (1998) claim to have great results, it must be noted that they only used 54 and 30 soil samples respectively for their research, meaning that the results might not be as significant as they state. For the liquid limit LL, Seybold et al. propose the following function:

(1) $$LL = 0.655 * Clay + 0.406 * CEC + 12.459$$

For the plasticity index PI they use:

(2) $$PI = 0.408 * Clay + 0.434 * CEC - 1.525$$

For both formulas, the clay is in percentages and the CEC is in cmol(+)/kg. LL and PI result in kg / 100kg. For calculating the linear regression model Seybold et al. (2008) used around 10,000 horizons with measured values for the liquid limit and the plasticity index.

2.4 Soil moisture characteristics

While the liquid limit and the plasticity index are required for classifying soil into the USCS classification, there are also soil properties concerning water that can predict sand, silt and clay values. These are therefore useful for the Random Forest algorithm to be used as explanatory variables. The water properties used are saturated water content, field capacity, permanent wilting point, and available water capacity.

Saturated water content

Some of the mentioned soil moisture properties relate to the saturation of a soil. A soil is saturated when the soil's pores are entirely filled with water, meaning that all air is gone from the pores (Brouwer, 1985). Determining whether a soil is saturated or not is relatively easy when doing field work. When a handful of saturated soil is taken from the ground and squeezed, some water will pour out of it. Soil saturation is important for agricultural practices, as plants require both water and air in the soil to grow and prosper. When a soil is saturated however, no air is available in the soil which has a negative effect on plants when sustained for a long time. Most plants die when their soil is saturated for more than 5 days, with rice being the only one able to withstand exposure to fully saturated soil. Under normal circumstances, saturation does not last longer than a few days since the water moves downward from the topsoil into the lower parts of the soil. The process of moisture moving from the topsoil into lower horizons is called drainage or percolation. When this process occurs, water is replaced by air. Drainage

happens quicker in coarse textured soils, having more sand than fine grains, lasting no longer than a few hours. Soils that have a fine texture usually take longer to drain, ranging from 2 to 5 days.

Field capacity

Relating to the drainage process is the Field Capacity. When the drainage process has come to a stop, the larger pores of a soil contain both water and air, while the smaller pores remain filled with water (Cassel & Nielsen 1986). This stage of the soil's moisture content is called the field capacity (Brouwer, 1985; Veihmeyer & Hendrickson, 1931). This stage of soil saturation is regarded as the ideal state for crop growth and is therefore an important soil property for agriculturalists. It is also used for constructing biophysical models depicting water availability (Güsler & Candemir, 2014). The field capacity concept is widely used among agriculturalists, agronomists, soil experts and conservationists to define the water content of a recently moisturised soil (Cassel & Nielsen 1986). Additionally, the field capacity property is used for modelling water transport in cropping soils (Rab, 2011). Measuring the field capacity of a soil must be done through field work by determining the water content of a recently irrigated and drained soil (Colman, 1947). Measuring the field capacity can be infeasible when large areas are being studied, when water levels are too low or when the terrain is impassable (Mohanty et al., 2015). Due to the impracticability of determining a soil's field capacity, scientists have come up with ways for calculating the field capacity of a soil by using other, already known soil properties. Wilcox and Spilsbury (1941) found that sand was an explanatory variable for the variance in field capacity when assessing Canadian soils. A soil's *moisture equivalent* (Briggs & McLane, 1910) also proved useful in calculating field capacity, as Veihmeyer and Hendrickson (1931) found significant similarities between both soil properties. More recently, Navin et al. (2009) have stated that the calculation of the field capacity has been a complicated matter due to the ambiguity of the soil property's definition. The 'drainage of excess water' is a vague definition for a soil property and research has therefore been done using empirical guidelines which are based on time, pressure, or flux. Navin et al. (2009) propose a flux-based method for mathematically determining field capacity.

Permanent Wilting Point

Water is not stored in soil permanently. Rather, it evaporates and is absorbed by plant roots to grow. Though when no additional water is added to the soil, the plants eventually take up all the water until now moisture is left in the soil. When the level of water decreases in the soil, it becomes harder for the plants to extract it and at a certain point, the amount of water is insufficient for the plant to survive and it dies. The soil water content at this exact time is the 'low' or 'dry' end of water and is called the permanent wilting point (Brouwer, 1985; Cassel & Nielsen, 1986). The permanent wilting point is measured at a kPa of 1500 (Slatyer, 1967). There is some water left in the soil, but not enough for plants to extract and survive. Another definition is given by Tolk (2003) as "the largest water content of a soil at which indicator plants, growing in that soil, will fail to recover when placed in a humid chamber". This definition focuses on the recovery of plants, rather than their death. Moreover, while field capacity is mostly determined by soil characteristics, the permanent wilting point is constructed through a combination of factors including soil, plants, atmosphere, and hydraulics. Due to the variety of explanatory variables for the permanent wilting point of a soil, the property differs greatly across soils (Ghorbani et al., 2017). For instance, the distribution of plant roots is influenced by grain size and grain composition. When a plant's roots are well distributed in a soil, the water content of a soil at wilting point will be lower (Tolk, 2003). This example explains the diversity in factors influencing the permanent wilting point. Similar to the Field Capacity, the permanent wilting point is also influenced by grain size (Tolk, 2003). The laboratory measurement of the permanent wilting point incorporates sunflower plants that are planted in containers only affected by transpiration, as they are sealed off from other factors. The plants are grown until they have three sets of leaves. Then, water supply is stopped,

and the plants are left in an environment with low evaporation until all three layers of leaves wilt. To check whether the plants have actually wilted, they are placed in a moist chamber during the night. When they are still wilted, then the permanent wilting point has been attained, after which the soil's moisture content can be examined (Tolk, 2003). While the sunflower method provides a consistent result for the permanent wilting point, Richards and Weaver (1943) state that measuring the statistic at -1.5 MPa matric potential yields better, more accurate results. A sieved soil sample is placed in a moist chamber and 1.5MPa pressure is applied until water content equilibrium is reached between the plate and the soil. Measuring the soil property in the field is the most ideal situation, as it provides an in-situ observation of the real-world environment where realistic plant-soil-interactions are allowed to occur (Tolk, 2003).

Available Water Content

Connecting both the field capacity and the permanent wilting point results in the Available Water Content (AWC). From the perspective of plants, the soil can be seen as a source of water. A saturated soil means that the source is filled to its capacity, but situations occur in which water drains below the reach of a plant's roots before any water can be obtained. At the end of the drainage, the soil is at field capacity and the plant's roots use the rest of the water source (Brouwer, 1985). The source has emptied when the permanent wilting point has been reached, meaning there is no water left to reach for the plant. Using this knowledge, it can be stated that the amount of water that a plant can actually use is the water content of a soil at field capacity minus the water level at the permanent wilting point (Brouwer, 1985; Cassel & Nielsen, 1986). This relationship between the two statistics is illustrated in the figure below.

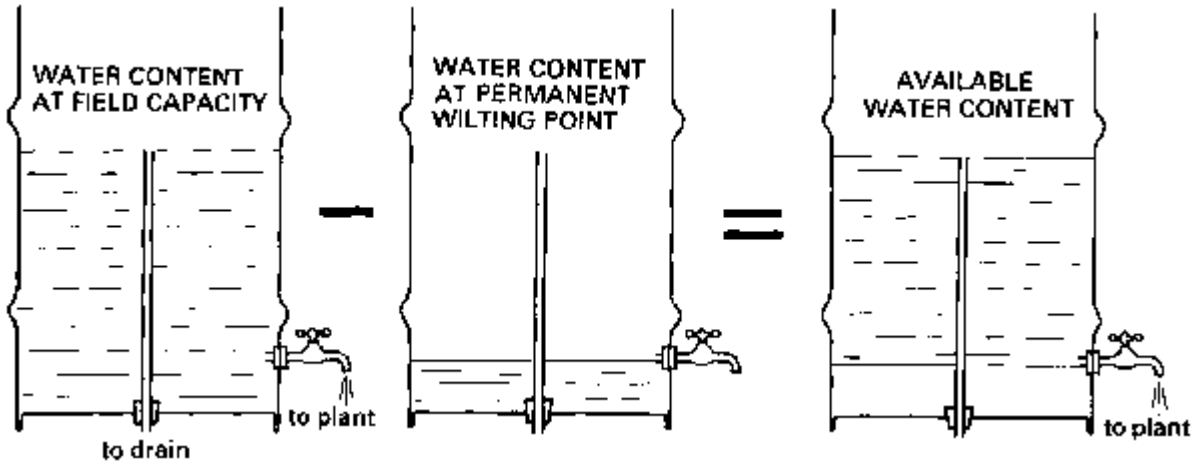


***Figure 2.3**: AWC as a function of field capacity and permanent wilting point (from Brouwer, 1985)*

The available water content of a given soil is constant but can vary greatly across different soil types and compositions. It is heavily reliant on soil texture and structure with sand having the lowest available water content, and clay having the highest. This is depicted in the table below.

***Table 2.1**: AWC per grain type in mm/m (from Brouwer, 1985)*

Soil	Available water content in mm water depth per m soil depth (mm/m)
Sand	25 to 100
Loam	100 to 175
Clay	175 to 250

In addition to grain size having an impact on the available water content of a soil, the organic matter content can also be influenced by this soil moisture property (Hudson, 1994). It is suggested that with

an increasing amount of organic matter, the soil can hold additional water, equal to multiple times the amount of extra organic content. Furthermore, the notion of a static value for available water content is questioned (Logsdon, 2019). Laboratory measurements of field capacity and permanent wilting point are determined by exact pressure points that are found to be too arbitrary for reliable, real world measurements (Gardner, 1971). Just like Navin et al. (2009) did for the field capacity, Gardner (1971) questions the definition for the field capacity provided by Cassel and Nielsen (1986). By rethinking the way field capacity is measured, the calculation of available water content is also challenged. It is stated that dynamics in plant roots disturbs the laboratory measurements of the field capacity (Gardner, 1971). The same goes for the difference betweens soils that do and do not have plants. Moreover, drainage of a soil and water uptake by plants do not occur sequentially, rather, it happens simultaneously meaning that the actual available water content might be higher than the one measured in the laboratory, since plants take up a part of water that is expected to drain. Furthermore, plant roots can block pores in the soil that would act as pathways for drainage water (Scholl et al., 2014).

2.5 Additional biochemical properties

The cation exchange capacity is a relevant chemical soil property due to its importance in calculating the liquid limit and plasticity index. Further soil biological and chemical soil properties are the bulk density, soil organic carbon stock and nitrogen. These soil properties are expected to assist in finding organic and peaty soils.

A widely used measure to assess the amount of plant material and other biological materials in a soil is the organic content of a soil. It is often expressed as gram per kilogram, therefore having a potential range of 0 to 1000. The maximum value is however never reached, as a soil cannot consist of organic matter alone. The percentage of organic matter that occupies a soil determines whether a soil can be characterised as mineral or organic. The dividing content is 30 percent (300g/kg) (FAO, 2009). A soil's organic content consists of any material that originated from a living organism but is separated from those organisms and has started to decompose in the soil (Melillo et al., 1989). The decay process of plants starts at fresh plant litter that over a span of time eventually transforms into soil organic matter. At both the early and the late stages of the decomposition process, nitrogen and carbon is either released or contained (stocked) (Aber & Melillo, 1980; Berg & Staaf, 1981). Related to the soil organic carbon is therefore the nitrogen in a soil. When organic matter dissolves, dissolved organic matter is generated (DOM), heavily influencing the biochemistry of nitrogen in a soil (Kalbitz & Geyer, 2002). Nitrogen is heavily present in organic matter and therefore can assist in detecting peaty soils (Hemond, 1983).

Bulk density is a statistic of the compactness of a soil. It can be defined as the ratio of the mass of oven dried material to the bulk volume. A soil consists of solid materials and pores filled with either water or air. As the bulk density property is a ratio of the mass against the volume, a higher value for bulk density means that a soil contains relatively few pores, while a low bulk density value signifies relatively large and many pores, since the actual mass of a soil sample is a relatively small chunk of the total volume (Chaudhari et al., 2013). Bulk density plays a role in influencing multiple different other soil properties such as structural strength (Blake & Hartge, 1986), flowability of powders (Abdullah & Geldart, 1999), soil moisture retention (Gupta & Larson, 1979; Vereecken et al., 1989) and degree of compactness (Hakansson, 1990; Hakansson & Lipiec, 2000). Measuring the bulk density of a soil is most desirably performed at constant and standardised moisture conditions, to avert issues generated by swelling and shrinking caused by water content in the soil (Hakansson & Lipiec, 2000). To solve the

swelling issue caused by soil moisture, the soil sample is oven-dried at 105 degrees Celsius for more than 48 hours (Agus et al., 2010). More importantly, bulk density is affected by the amount of organic content in the soil and its texture (Chaudhari et al., 2013). The bulk density soil property is often used to examine and anticipate soil processes and the aggregation of soil data when measuring horizon mass. Regarding its relation to soil texture and grain size, bulk density ranges from 1.0 to 1.6 mg/m3 for clay and 1.2 to 1.8 mg/m3 for sand (Chaudhari et al., 2013). Furthermore, bulk density is used to calculate soil organic carbon stock, which can be useful in determining where organic and peaty soils are located (Post et al., 1982).

As stated above, bulk density is used to calculate soil organic carbon stock, which can be used to determine to what extent soils contain organic carbon. Soil organic carbon is the content of carbon held within the organic components of a soil (FAO, 2009; Zhu et al., 2010). Usually, research into soil organic carbon stock is performed due to environmental reasons (Martín et al., 2016; VandenBygaart, 2006), since it plays a significant role in keeping greenhouse gasses from the air. Additionally, soil organic carbon is crucial for the enhancement of soil quality, management of clean water and maintaining sustainability in food production (Singh et al., 2007). The total organic carbon impacts soil characteristics relevant to the USCS classification and to characteristics that are needed to calculate the liquid limit and the plasticity index. The most notable is the cation exchange capacity. When a soil has a high degree of clay, the organic carbon fraction contribution to the CEC is marginal, while the organic carbon fraction in sandy soils has a relatively high contribution to the CEC (Pluske et al., 2021).

2.6 Random Forest Algorithm

Random forest is a machine learning method used for data classification and regression (Gambill et al., 2016). The algorithm consists of a set of decision trees also called a forest or an ensemble. The decision trees are randomly generated using predictor values, hence the name of the algorithm (Breiman, 2001; Segal, 2004). While overfitting is usually a problem for decision trees, the multitude of trees in the Random Forest algorithm compensate for overfitting by having a voting mechanism in place that selects the majority of votes, rather than the most specific outcome (Segal, 2004).

2.6.1 Random Forest uses

The algorithm falls into the regression based voting approach (Cootes et al., 2012). This approach originated when the Generalised Hough Transform voting method was utilised to identify shapes in imagery (Ballard, 1981). Since then, many variants were created, each for their own purpose. Object position voting was created using the Implicit Shape Model (Leibe et al., 2004) that recognises patches on an object. Furthermore, the regression based voting approach was used to detect human body parts in the Polesets model (Bourdev & Malik, 2009). Hough Forests use the regression approach to locate an object by dividing the dataset up into multiple sub-regions (Gall & Lempitsky, 2009). The contemporary training method involved combining both forest regression and classification to separate irrelevant background imagery, and only counting votes that originated from the objects relevant to the purpose of the research (Cootes et al., 2012). Random forests are also used in medical research to identify joint centres in the human body using depth images (Girshick et al., 2011). Similarly, Criminisi et al., (2010) used regression forest to detect anatomy by using bounding boxes situated around organs.

2.6.2 Remote sensing uses

In the realm of remote sensing, the Random Forest algorithm has been applied for both classification and regression. Ismail & Mutanga (2010) use the Random Forest algorithm to model water stress through the use of hyperspectral data. They state that traditionally, regression trees are sensitive to outliers and prone to overfitting. Additionally, it is concluded that the Random Forest regression method produces better results than both bagging (Breiman, 1996) and boosting (Friedman, 2002) (Ismail & Mutanga, 2010). Abdel-Rahman et al., (2013) make use of the random forest algorithm to determine nitrogen concentration in sugarcane leaf samples reaching 61% prediction rate. Their best predictors were the visible, red edge and near-infrared wavelengths. More soil related is the study performed by Millard & Richardson (2015) who used the Random Forest classifier to identify peatland ecosystems. Furthermore, Fassnacht et al. (2014) utilise the powerful Random Forest regression algorithm to estimate aboveground forest biomass using LiDAR and multispectral imagery from multiple satellite platform sources.

2.6.3 Random forest core workings

Classical regression techniques that have a pre-specified relationship between response variables and predictors. The regression analysis is then used to either prove the relationship as significant or disprove it as coincidental. The Random Forest regression analysis is devoid of such assumption of relationship. The algorithm's main workings are setting up certain decision rules using the predictor variables (Verbyla, 1987; Clark & Pregibon, 1992). The decision rules are located in decision trees, constructed through recursively dividing the input data into smaller and smaller batches. The division is done using binary splits using one predictive variable. To decide at what value a split should be made, the algorithm performs an exhaustive procedure to determine what split is most beneficial. In the case of forest-based regression, the most desirable split is the one that has the highest degree of homogeneity inside the two resulting groups when compared to the dependent variable (Prasat et al., 2006). Assessing this degree of homogeneity is done through examining the between-groups sum of squares, where the highest value is chosen. The result of running the Random Forest algorithm is a tree ensemble, or forest, in which the decision nodes, or branches, are determined by decision parameters and a range of leaf nodes that state the mean response. First, a maximal number of trees is generated, and then the algorithm uses cross-validation techniques to shorten the trees in order to prevent overfitting (Therneau & Atkinson, 1997).

2.6.4 Trees

In most applications of the Random Forest algorithm, the model takes two sets of data as input: the training data (dependent variable) and the prediction data (independent/explanatory variables). The algorithm trains a model based on values that are already known through a training dataset. Once trained, the model can be utilised to predict another dataset that contains the same predictors (also known as explanatory variables) but of which the dependent variable is not known. The core of the algorithm consists of a wide range of trees, which together are called a forest or ensemble (Breiman, 2001). The trees are fed with data and each tree results in a prediction, of which the majority vote decides what the final outcome will be. Since multiple trees are used rather than a single decision tree, the Random Forest algorithm prevents overfitting, a situation where a model too closely tries to mimic reality but ends up modelling only outliers. The randomness of the model prevents this from happening. Another advantage of the Random Forest is that it is able to estimate missing data based on other datasets while still maintaining high accuracy (Gambill et al., 2016). More technically, pseudocode for the tree creation in the Random Forest algorithm is shown below.

For b=1 to B:

draw a bootstrapped sample of size N from the training data

grow a random forest tree Tb to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree:

select m variables at random for the p variables

pick the best split

split the node

2.6.5 Bootstrapping

Each individual tree does not contain all explanatory variables, rather it contains a subset of them. This omission, called bootstrapping, caters to the randomness of the algorithm reducing overfitting. Similarly, each individual tree does not receive each piece of a data set. Traditionally, the total forest also does not receive all the data, instead, a randomly selected 90% of the dataset containing explanatory variables is used as training data, and the other 10% is used for testing the model (Breiman, 2001). When a subset of data is chosen randomly, it is called a Bootstrapped dataset in which one data entry or sample can occur more than once. The data that is not included in the bootstrapped dataset is called 'out of bag data' (Breiman, 1996; 2001).

2.6.6 Splitting trees

In essence, the trees are decision trees where certain 'if' statements (figure 2.4) are made to split the trees into branches. An example of such a statement could be whether the clay content is higher than 55%. Each branch then splits up into other branches using new 'if' statements until ideally one value is left. The decision where only one value is left is called the leaf node, and if there are still multiple decisions, it is called a decision node. The Random Forest algorithm is trained by analysing the properties of known dependent variables. Using these properties, decision trees are set up aiming to split the data as efficiently as possible in order to have the highest information gain. How well each variable does at splitting the decision trees is also called variable performance. If a node in the decision tree is able to fully isolate one class, called a pure subset, then the highest amount of information gain is achieved.

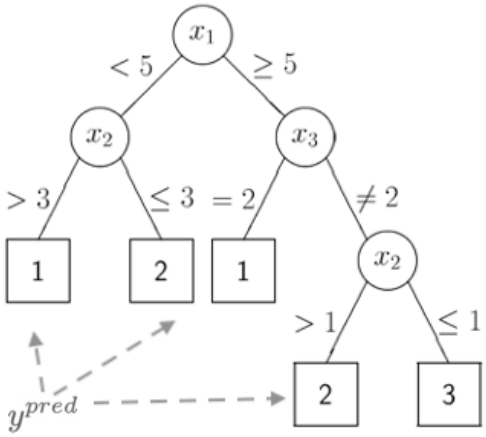


Figure 2.4: *random forest decision tree where each node represents an 'if statement' (from: Bookdown, 2020)*

Growing a tree is done based on a bootstrapped sample from the input training data. The tree is grown to full size without any pruning being done. For each tree, the Random Forest algorithm uses a randomly selected subset of variables that each determine the split at a node. Similarly, the trees get fed data that is not included in the creation of the tree, called out-of-bag data (Breiman, 2001), meaning that the tree is processing data it has not seen before to prevent a confirmation bias (Prasat et al., 2006). Then, the difference in mean square errors between the out-of-bag data and the data that was given to the tree to grow (Maindonald & Braun, 2006). This is done for each variable, resulting in an estimate depicting the prediction error for each variable. This is called the out-of-bag error and measures a variable's capability to predict values that it has not been trained with (Breiman, 2001; Palmer et al., 2007). This error measure can also be used to measure a variable's importance in predicting the dependent variable. Variable importance is evaluated based on "how much worse the prediction would be if the data for that predictor were permuted randomly" (Prasad et al., 2006). This very ability to create statistics of the workings of the algorithm reduces the impact of it being a "black-box" process like other machine learning algorithms (Prasad et al., 2006).

2.6.7 Predictors

In order to perform the supervised learning Random Forest algorithm, one requires predictors to feed into the algorithm. Predictors are data sources that explain or 'predict' something about an unknown attribute. For instance, when looking only at the colours of a satellite image, one can deduct something is water from it being the colour blue. Similarly, it is possible to deduct soil characteristics when looking at terrain models, hydrological data, climate and weather data and multispectral imagery.

3. Research Design

This section presents an outline of the research design that will be used in the thesis. First, the research area is defined and arguments for this choice are given. In the research area, some areas of interest are highlighted that either possess key soil features or play a role in the results section. Next, an overview of the general workflow is given which is elaborated on in the following subsections. In subsection 4.3, the training data for the model is listed, accompanied with solid assumptions after which their relevance to the project is assessed. Section 4.3 also contains a linear regression analysis that explores the possibility of generating the Atterberg limits using clay content and cation exchange capacity. This section is followed up by an exploration of the independent variables needed, the collection of this data and their relation to the soil characteristics. Fourth is the general workflow for running the Random Forest algorithm where algorithm settings are discussed. Additionally, an investigation is performed to assess whether enough sample points are provided for the Random Forest algorithm. Next, the validation and testing methods are outlined, which are divided up into phases. Lastly, the method for generating the final product is outlined, a raster map containing the USCS classification at a 30m scale.

3.1 Research Area

The initial aim of the terrain analysis model is to get the best prediction of soil types possible for the continent of Europe. It is likely that areas that already have a lot of detailed information on them will be the most useful for this research. The Netherlands-Belgium area is such a region with readily available data. While satellite data is available at a global scale, data on soil properties is mostly available within Europe. When looking for information on specific subjects, such as the USCS, data is significantly more limited. The two countries that have detailed open-source data of their soils are the Netherlands and Belgium. They provide descriptions of their soils which can be qualitatively reclassed into the USCS classification using an unvalidated method. When using any machine learning method, it is of crucial importance to have a method of verifying the results of the algorithm. Therefore, the choice for the research area has fallen on the Netherlands, the westmost part of Germany, and the Flemish part of Belgium. As stated in the research scope, the analysis of soil characteristics is limited to bare soil, grasslands, agricultural areas, and marshes. To delineate this area, the Sentinel 2 Global Land Cover map (S2GLC) is used (Malinowski et al., 2019). From this land cover map, certain land covers are selected for analysis. The set of land covers used is listed in appendix 7, and visible in figure 3.1.

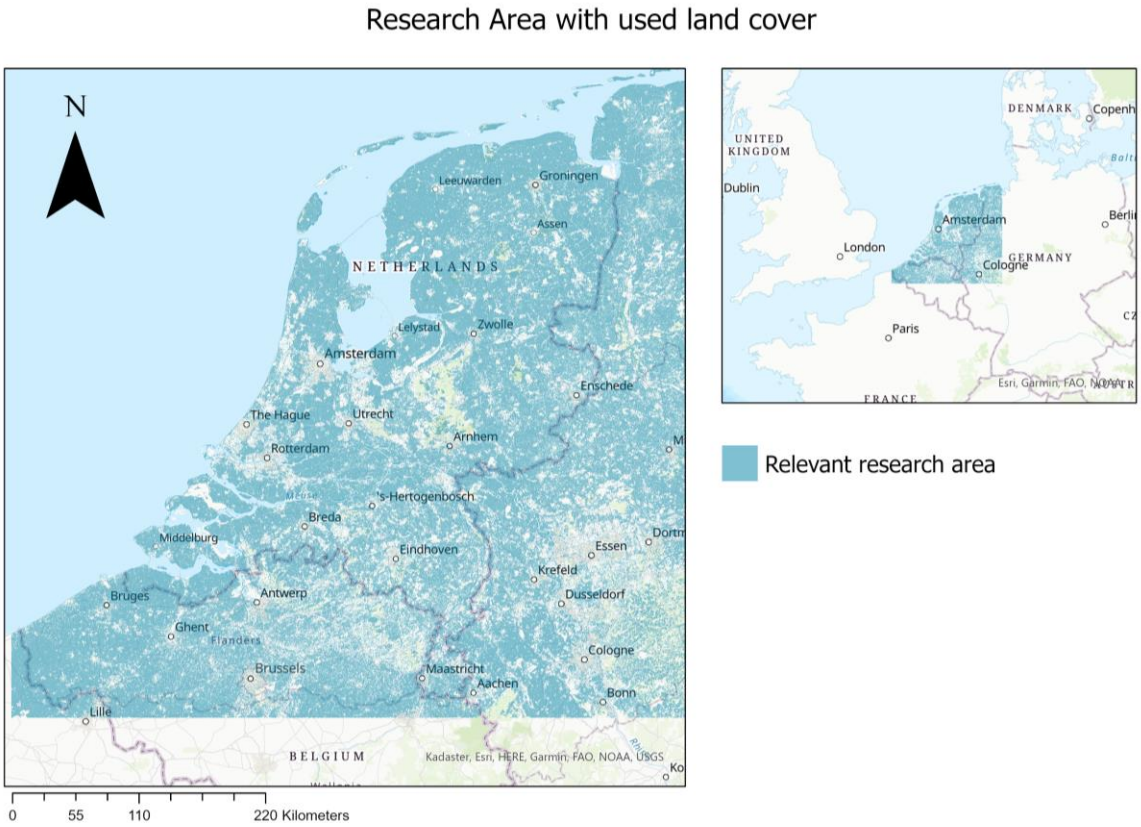


Figure 3.1: *the research area for this study with the relevant land cover areas*

In the research area, there are some areas of interest where some key patterns are expected, or areas that will be referred to later in the thesis. The areas are displayed below in coloured polygons with their names listed in the legend. The areas are relevant due to their distinct soil properties. Most areas indicated in the figure are expected to have significantly large areas of peaty soils, which are crucial for the automated terrain analysis. These areas include Het Groene Hart, the Hondsrug and the Noordoostpolder. Other areas are relevant due to the dense concentration of high values of predicted soil properties in the results section. Other areas are mentioned in the results section as well to indicate certain patterns in the distribution of the soil properties. These include the Dutch River Area, the Beemster, Het Groene Hart, and the Gent-Antwerpen-Terneuzen-Brugge area.

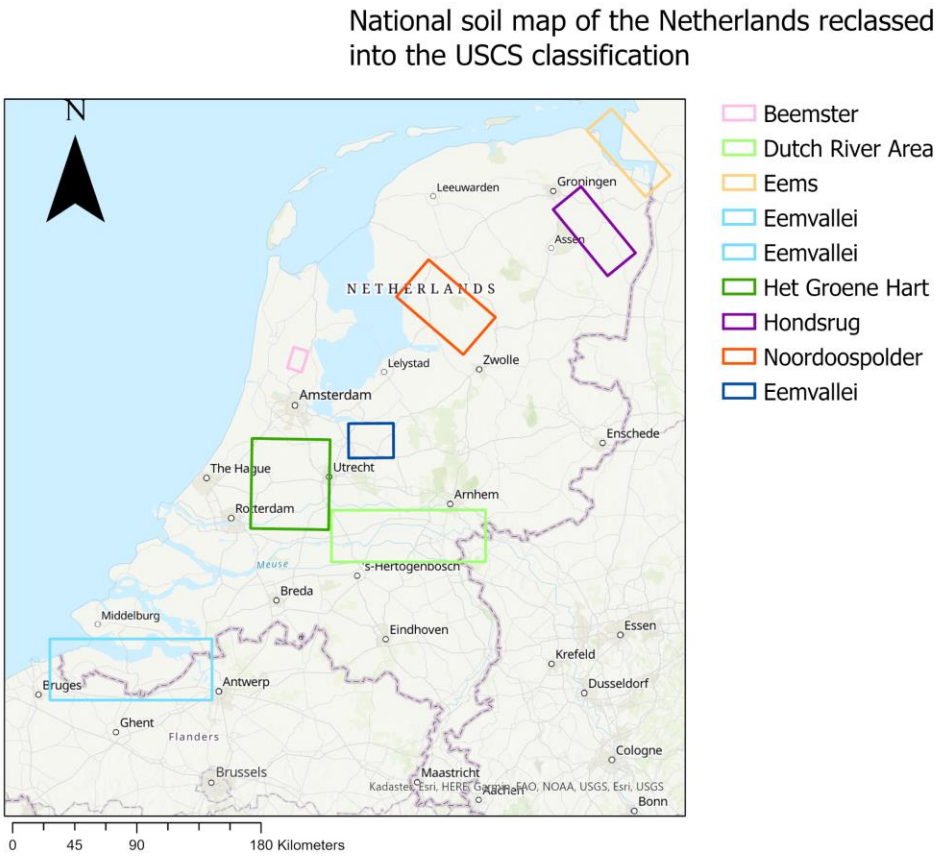


Figure 3.2: *areas of interest for the research area*

3.2 Workflow identification

Before performing any actions towards the final product of this thesis, it is necessary to identify what is required for this end goal in terms of steps and information. As the goal is to classify soil into the USCS classification using soil samples in combination with high resolution rasters, it is first necessary to assess how the USCS classification can be attained. As discussed in the theory, the data needed for this classification system is percentage sand, silt clay, organic content, coarse materials, and the value for CEC. Additionally, information on the liquid limit and the plasticity index is required. Getting this data on a 1:50k (30m) scale is not possible by searching online sources. The best resolution available for data on soil properties is 250x250m, but this data, however, is calculated using an algorithm and can therefore not be seen as actual true information. Data that is freely available on a 1:50k scale is satellite imagery, height data and derivatives describing vegetation and moisture. On its own, satellite, elevation and derived data cannot be used to classify soil but using the Random Forest algorithm as described in the theory, the entire collection of data can be combined and be put to use in predicting soil properties. Creating a map for each individual soil type allows them to be combined into the USCS classification. For executing the Random Forest algorithm, training data is needed. Data on individual soil properties (sand, silt, clay etc.) is needed to train the algorithm on points where the data is known in order for it to predict the properties in areas where it is not already known.

3.3 Training and Comparison data

For training the random forest algorithm, training data is needed in the form of true information. Only when a soil's properties are known to be true, can the algorithm be correctly. It is therefore crucial to get as much data as possible that resembles the true situation on the ground. For comparing the Random Forest algorithm results, this data also comes in useful. It allows for quick reference in order to see if the predicted soil class matches the soil class indicated by the ground truth data. Data of which the accuracy is expected to be less than true can also be used. While it cannot assist in assessing accuracy regarding USCS classifications, it can however provide indirect information that assists in determining the correct soil classification.

3.3.1 Ground truths

Not all data surrounding soil requires predicting, some is already known. This known data comes in the form of soil profiles and is called ground-truth data. This type of data is used to train the Random Forest algorithm in order to predict the six soil properties in areas where no data on them is available. It comes in the form of point feature vector maps and represents actual dig and test sites of soil. It can therefore be stated that this information is suitable for both training and testing machine learning algorithms. The soil profiles contain data on the percentages sand, clay, silt, organic matter, coarse materials, and cation exchange capacity. The main source for these soil profiles is the ISRIC-database (ISRIC, 2020). A subset of this database is the WoSIS database. It contains soil profile sets, each with different standardisations. Comparing these sets is likely to result in some commonalities between them and will therefore provide a single large database containing real soil data. Due to the fact that the WoSIS databases are offered in separate files, each containing data on only one soil property, it is necessary to combine the datafiles in order to have all soil properties in one dataset. A problem encountered with this operation is that the individual datasets do not contain the same profiles. While there are some locations that contain data on all needed soil properties, the majority of points only contain one or two soil properties and are therefore inadequate for comparing model results. The individual datasets, however, are useful for training the random forest algorithm, as individual rasters are created for the individual soil properties. Additionally, the Land Use/Cover Area frame statistical Survey Soil (LUCAS) dataset (Orgiazzi, 2018) provides additional point data on dig sites. The points contain data on all needed soil properties: percentage sand, silt clay, organic matter and coarse materials, as well as cation exchange capacity. Again, from clay and CEC, the liquid limit and the plasticity index can be calculated. When the LUCAS is combined with the WoSIS, a larger and more spread-out point vector dataset is created for training the algorithm. If the Random Forest algorithm is to be applied elsewhere in Europe, ground truth data points can be sourced from that particular region on the continent.

Samples from multiple depths are collected in the soil profile databases, an example of this can be found in figure 3.3. This study however is focused on the topsoil characteristics, meaning that lower depth measurements are irrelevant to the research and will likely distort the predictions if the collected dataset remains untouched. To remedy this, only samples between 0- and 30-centimetre depth are used. When multiple points are present at this depth, for instance between 0 and 15, and between 15 and 30, their mean value is used. When this filtering operation is performed, the following number of soil sample points remain.

Table 3.1: *Number of soil samples per soil property per source*

Soil property	Nr. of WOSIS points	Nr. of LUCAS points	Total
Sand	417	582	999
Silt	547	582	1129
Clay	5334	582	5916
Coarse fragments	297	582	879
Organic content	234	582	916
Cation exchange capacity	23	582	605

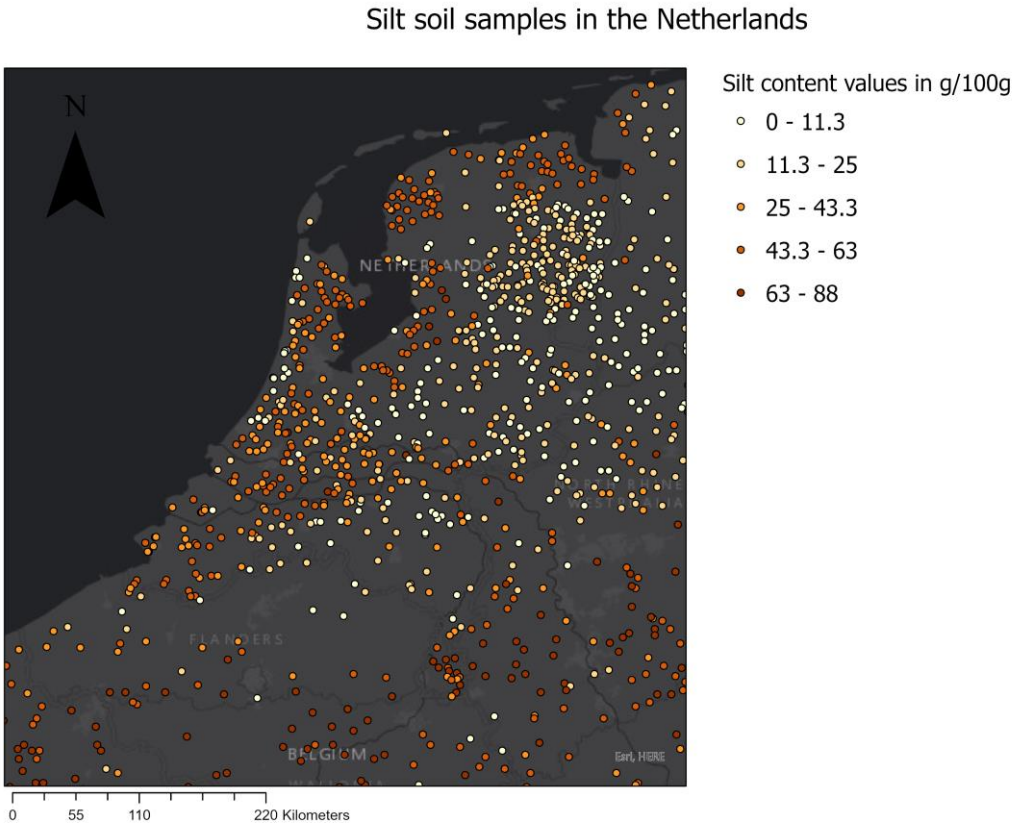


Figure 3.3: *silt soil samples in the research area when WoSIS and LUCAS are combined*

Both datasets are combined and thereby are instantly useful for training the Random Forest algorithm, as it requires information on individual soil properties as input. For verifying the final result of the algorithm, being a USCS-map of a certain region, some more operations are required to get the LUCAS and WoSIS points from having only data on soil property, to displaying the correct USCS classification. While soil profiles provide a quick overview of the soil properties of different point locations, they do not yet function as ground control points for assessing the accuracy of the final product. The soil profile points are transformed into data that can be used to verify whether the final soil classification has assigned the correct USCS classes to the locations of the points. Hence, they are translated into the USCS classification using their soil properties. Of each point, the percentages of sand, silt and clay are known. Additionally, information on the organic content, coarse materials and cation exchange capacity is provided. Given both the clay content and the CEC, both the liquid limit and the plasticity index can be calculated through a regression analysis improving upon the method of Seybold et al. (2008), described in section 4.3.3. Combining this information according to the USCS classification method, all points can be categorised. How this classification is performed is described in section 3.7.

3.3.2 Solid assumptions

Similarly, large scale vector data containing features that present the exact soil type are available in both the Netherlands and Belgium from national soil maps. While they do not depict the exact classification according to the USCS, they can be qualitatively reclassed into the correct classification system using the existing soil descriptions. Using this reclassed map, a qualitative comparison can be made between the final classification result of this study, and the national soil maps. The national soil map of the Netherlands is displayed below to provide an indication of what this looks like.

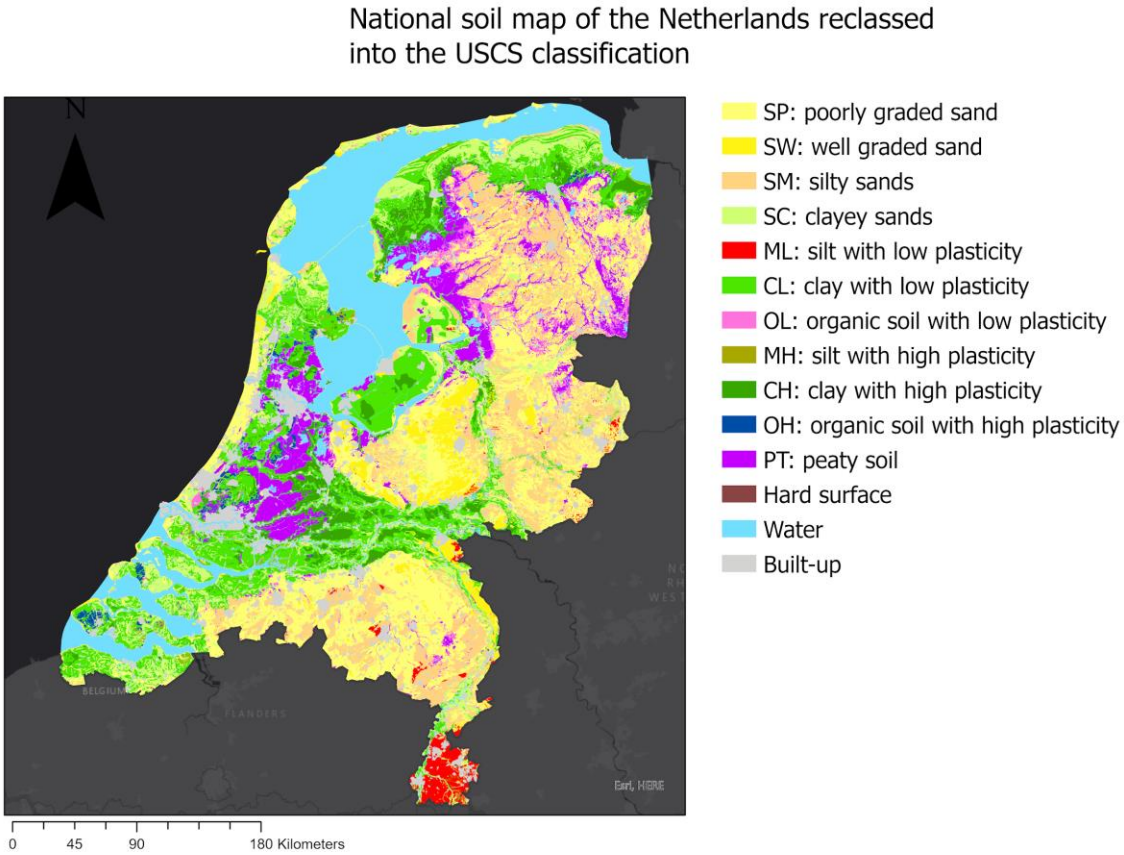


Figure 3.4: *national soil map of the Netherlands reclassed into the USCS classification using the description of the original labels.*

Another form of solid assumption comes in the form of interpolated raster data containing individual soil properties at a coarse scale. Data like this is obtained from the SoilGrids portal by ISRIC and is called SoilGrids data. Using an algorithm, ISRIC compiled their point based WoSIS data into their own rasters, resulting in raster datasets containing information on individual soil properties at a 250m scale. While this spatial resolution is significantly coarser than the spatial resolution of the final product, the maps do bear some resemblance to reality and can therefore act as solid assumptions to indirectly verify the results of the Random Forest algorithm. As the SoilGrids data is also generated using an algorithm, but they probably do contain patterns in the spatial distribution of said soil properties.

While SoilGrids provides a good indication of the distribution of individual soil properties, the soil map of the Netherlands provides information on actual soil properties. While this dataset has a plethora of different classes, these classes can be reclassified into the USCS classification according to their description in the product documentation. The result of this reclassification then provides an indication of what soil types should be where. This dataset should not be taken as ground truth, rather it is a solid

assumption, as the spatial resolution is not as precise as desired and there are some classification differences between local governments that result in some sharp borders between soil type regions.

To summarise, the table below lists information on the data that acts as ground-truth data and solid assumptions.

Table 3.2: *list of sources of ground truth data and solid assumption data*

Data Name	Info	Category	Type	Resolution	Date	Source
WOSIS	Soil profiles containing %clay, sand silt and coarse fragments, organic matter and CEC	Ground-truth	Vector	Point	2019	ISRIC
LUCAS	Soil profiles containing %clay, sand silt and coarse fragments, organic matter and CEC	Ground-truth	Vector	Point	2015	ESDAC
Soil grids	Rasters containing %clay, sand silt and coarse fragments, organic matter and CEC	Solid Assumption	Raster	250m	2020	ISRIC
National soil map Netherlands and Belgium	Reclassified into USCS classification according to soil type description in original product	Solid Assumption	Vector	Polygon	2018	WUR

3.3.3 Liquid Limit and Plasticity Index

In contrast to De La Rosa (1979) and Mbagwa and Abeh (1998), Seybold et al. (2008) used significantly more soil samples to generate their linear regression formulas for the liquid limit and the plasticity index. Even though the N increased more than a 100-fold when compared to earlier research, the linear regression model can still be improved by adding more soil samples to the statistical analysis. For this purpose, the United States SSURGO (Soil Survey Geographic database) is used to collect data on soil samples. It is published by the Natural Resources Conservation Service (NRCS) and contains soil data from all states of the United States of America. The advantage of this dataset over data from LUCAS or ISRIC is that the SSURGO contains field test data on the liquid limit and the plasticity index, making it possible to perform a valid regression analysis on the data, rather than guessing what the result should be.

For a representative result of the relationship between clay and CEC, and liquid limit and plasticity index, data of each state was used, except for Hawaii and Alaska, as they differ too much from the European climates. From each state's database, only one county's data was taken to reduce processing time needed to parse the data and to perform the statistical analysis on it. This selection still resulted in more than 54.000 valid soil samples, meaning that the samples contained data on all needed soil characteristics: clay content, cation exchange capacity, liquid limit, and plasticity index. Moreover, only

soil samples that appear in the standard USDA classification (see figure 1.3) are selected, as the SSURGO database also contains soil data on bedrock and other divergent soil types.

Figure 3.5: *scatter plots of liquid limit and plasticity index as a function of clay and cation exchange capacity*

The linear regression analysis was performed in Microsoft Excel using the analysis package. In the analysis, liquid limit was selected as the dependent variable, and clay and cation exchange capacity were selected as the independent variables. Second, the plasticity index was selected as the dependent variable. An initial assumption was that there might be a higher correlation between clay, CEC and liquid limit and plasticity index within the soil's USDA classes (figure 2.1). This hypothesis was however disproven when the R-squared value for the regression analyses performed on the individual USDA classes was significantly lower than when all soil samples were used in the regression analysis. This might be explained by the fact that the values for both clay content and CEC have a very limited range within a USDA class. Therefore, only a part of the entire regression plot is used, resulting in more of a circle than a line. The performed regression analysis resulted in the following formulas for the liquid limit and the plasticity index, accompanied with their statistics:

(3) $$LL = 0.773 * clay + 0.373 * CEC + 10.921$$

This linear regression formula for the liquid limit has an R-squared value of 0.842 and a p-value of 0, meaning that the results are significant. This proportion of variation explained improves upon the statistic generated by Seybold et al. (2008) by 3 percent, as their study reached 81 percent of variation explained for the liquid limit using clay and CEC.

(4) $$PI = 0.588 * clay + 0.149 * CEC - 1.884$$

The formula for the plasticity index has an R-squared value of 0.895 with a p-value of 0 meaning that this regression analysis is significant as well. This regression analysis improves greatly upon the regression suggested by Seybold et al. (2008) who reached a formula that explained 71 percent of the variation in the plasticity index.

3.4 Required Data and Data Collection

For the running the Random Forest algorithm for predicting individual soil properties, data on multispectral imagery, radar, elevation, and hydrology is used. In addition, indices and derivatives are calculated from these datasets. The data was collected from different sources and pre-processed in different ways. The datasets used all come from open access sources and therefore do not require any payment or subscription to be downloaded. The only data source that needs a subscription is the EU Hydro dataset, albeit that they do allow for a free download of their data. The caveat with open-source data is that accessing it can sometimes be a more time-consuming process, as opposed to commercial data where customer service is more important. Moreover, most open-source raster datasets are coarser than the required spatial resolution of 30m, therefore the selection of datasets is limited.

3.4.1 Satellite Data collection

Table 3.3: *list of satellite bands used.*

Band name	**Band width in µm**	**Sensor choice**	**Band Nr.**	**Original resolution**
Coastal	0.435-0.451	Sentinel 2	1	30
Blue	0.439-0.535	Sentinel 2	2	10
Green	0.537-0.582	Sentinel 2	3	10
Red	0.646-0.685	Sentinel 2	4	10
Red Edge 1	0.694-0.714	Sentinel 2	5	20
Red Edge 2	0.731-0.749	Sentinel 2	6	20
Red Edge 3	0.768-0.796	Sentinel 2	7	20
NIR	0.767-0.908	Sentinel 2	8	10
NIR small	0.848-0.881	Sentinel 2	8A	20
Water vapor	0.931-0.958	Sentinel 2	9	60

SWIR 1	1.539-1.681	Sentinel 2	11	20
SWIR 2	2.072-2.312	Sentinel 2	12	20
Radar VV		Sentinel 1	VV	10
Radar VH		Sentinel 1	VH	10

The required data from Sentinel 1 and Sentinel 2 is collected through the Google Earth Engine. The aim was to retrieve data layers that have similar weather and atmospheric conditions resulting in a rather similar image. Since both sensors do not have swaths covering the entirety of the research area, multiple images had to be stitched together to form a single image covering the entire relevant region. Collecting remotely sensed satellite data is done through Google Earth Engine. The choice is made for Google Earth Engine due to three reasons. First, the service has all relevant satellite data in one place, making collecting and downloading the data needed a more efficient and streamlined process. The second reason for choosing GEE over other services is the code editor that allows for exact filtering of data, as opposed to other satellite downloading services (EarthExplorer, Glovis, CopernicusHub). The code editor allows for the selection of geometry, date, cloud cover, satellite bands and spatial resolution. Additionally, products coming from the Google Earth Engine are pre-processed and therefore do not need further colour correction.

Before using the code editor of GEE, some preparation needs to be done. The script for downloading the Google Earth Engine data can be found in the appendix 5. The choice is made for less than 5% cloud cover. Regarding the temporal considerations when collecting Sentinel 2 multispectral data, the median is taken of satellite imagery that dates between April 1st and May 31st of the years 2015-2020 and contains less than 5 percent cloud cover. These dates are chosen due to the absence of snow, and the limited presence of vegetation. The Sentinel 1 imagery is collected in a similar way, but it was collected over the span of the spring season instead of over two months. From the Sentinel 1 platform, the VV and VH bands are collected. By collecting the average of the reflectance over a span of five years, the influence of weather is minimised.

3.4.2 Sentinel 2 indices

Using the Sentinel 2 imagery extracted from the Google Earth Engine, several indices can be calculated that each use a different combination of Sentinel 2 bands and modifiers to create a derived property. The first and most straightforward one is the Normalised Difference Vegetation Index (NDVI) depicting the relative greenness per cell. This NDVI uses the green, red and near infrared band to calculate the greenness index. The NDVI is commonly used to determine organic matter in the soil and can also be applied as an indicator for other soil properties (Zhou et al., 2020). An example for the NDVI is presented in figure 3.6. Furthermore, there is the Normalised Difference Moisture Index (NDMI) used to assess the water content in vegetation (Beucher et al., 2019). For this, the ratio between the NIR and the SWIR bands is calculated. To gain another perspective on vegetation, the Enhanced Vegetation Index (EVI) is used (Zhou et al., 2020). This index is more sensitive to small differences in vegetation in areas where vegetation count is already high. It also adjusts better to atmospheric haze and corrects for surface beneath the canopy. The Soil Adjusted Total Vegetation Index (SATVI) originated from a

need to minimise the influence of soil brightness on the vegetation index (Huete, 1988). Lastly, the Normalised Soil Moisture Index (NSMI) acts as a proxy to soil moisture (Hong et al., 2018). It is a non-dimensional measure of reflectance bands calculating the ratio between the reflectance at two different wavelengths.

Along with the various indices that can be derived from the Sentinel 2 platform, there is also the Tasselled Cap Wetness index that is derived from the same platform. The tasselled cap is mostly used in tandem with the NDVI and is therefore used simultaneously with the NDVI. The tasselled cap is a wetness index formed by reflection bands from multispectral imagery. It measures soil and canopy moisture and is created through multiplying satellite band values with a set of constants. It is an effective method of data transformation as information loss is kept to a minimal and spectral imagery can directly be connected to physical properties of the soil (Crist & Cicone, 1984; Zhang et al., 2002). It was first introduced by Kauth and Thomas (1976) to perform agricultural assessments of soil and canopy information. They used it to detect changes in the brightness, vegetation greenness and soil wetness. The tasselled cap wetness index has mostly been used in ecology and environmental monitoring to detect changes in surface properties. It is therefore common practice to only calculate the tasselled cap using spectral imagery from a single day and comparing this to another day to then detect any changes (Zhang et al., 2002). Additionally, the tasselled cap is frequently used in change detection in forests in combination with the NDVI (Franklin et al., 2002). When attempting to detect the average situation of an area regarding the tasselled cap, it is more relevant to use the average values for the multispectral reflectance (Raynold & Walker, 2016). Lastly, the land surface temperature and average precipitation derived from Sentinel 2 are used as potential predictors, as they might add some insights into the influence of weather.

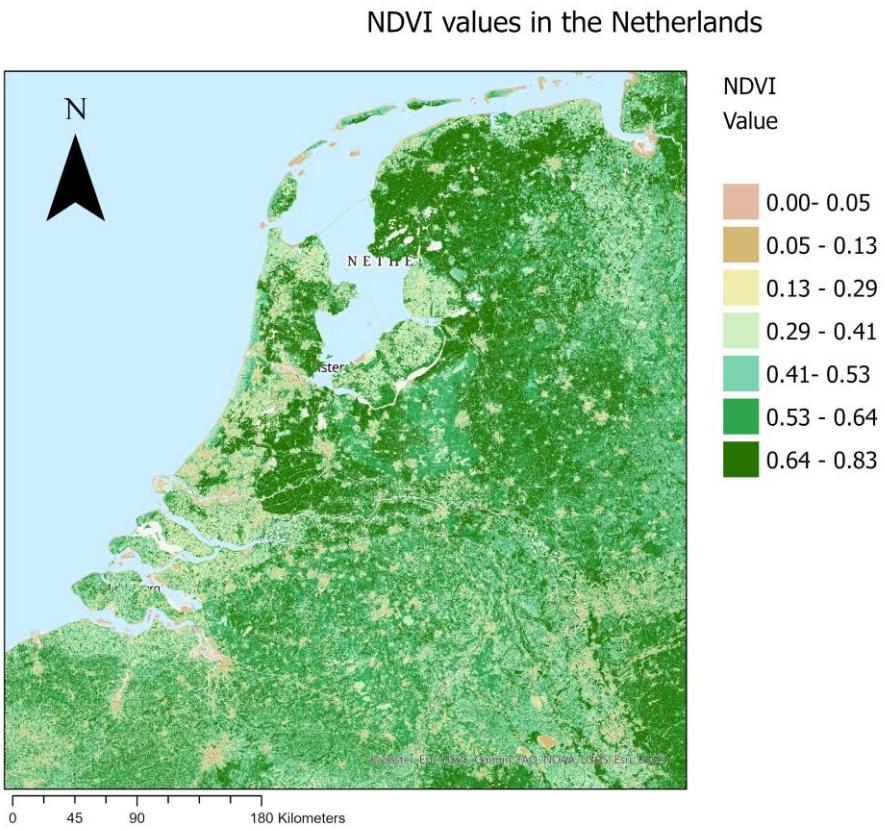


Figure 3.6: *an example of the NDVI values for the Netherlands*

3.4.3 EU DEM derivatives

Additionally, elevation data was used in combination with other predictors when running the Random Forest algorithm, given the fact that elevation indeed is a suitable predictor for soil attributes (Odeha et al., 1994; Thompson et al., 2001). The elevation data is retrieved from the EU DEM (European Environment Agency, 2020) in two batches, both containing one half of the Netherlands' elevation. These are put into a mosaic to form one map containing the digital elevation model. From this model, slope is derived as has influence on water runoff, chemical properties, and plant behaviour (Rezaei & Gilkes, 2005; Thompson et al., 2001). As the slope product is derived from the already pre-processed DEM, no further alterations are required. In addition to the slope is the curvature. This property can be used to analyse hydrological processes by describing the physical properties of a drainage basin which can be used to understand the erosion and drainage processes (Moore et al., 1991). Furthermore, the average height of a pixel when compared to all the cells within a 100-meter radius is calculated to indicate anomalies in height.

From this digital elevation model, further properties can be derived using terrain analysis tools. For the calculation of the following factors, SAGA GIS is used. First is the Valley Depth (VD) which can be calculated as the height difference between the elevation model and an interpolated ridge (Conrad & Olaya, 2012). This derived property explains to a certain degree Organic Carbon in the soil and might therefore assist in predicting organic content (Zhou et al., 2020). Additionally, the Modified Catchment Area is used as this was also expected to assist the Random Forest algorithm in predicting organic content (Zhou et al., 2020). Length Slope Factor (LSF) is also used as a predictor that is derived from the elevation map. It was originally designed to predict soil loss and was the most influential factor in the RUSLE for the continent of Europe (Panagos et al., 2015). The LSF takes into account both the steepness of a slope, as well as the length of a slope, which can influence soil erosion. Using the elevation model, a derived property can be calculated that provides an indication of the wetness, called the Topological Wetness Index (TWI). It quantifies topological influence on hydrological processes by taking the natural logarithm of the upslope local drainage area divided by the slope in radians. While this factor is expected not to produce a diverse range of numbers in a flat area like the Netherlands, it is still added to the list of potentially useful predictors (Radula et al., 2018). Lastly, a relatively niche factor is used called the catchment slope, which indicates the slope of the catchment area (Gericke & Plessis, 2012). This factor is considered niche since it is only relevant for catchment areas.

3.4.4 Hydrology data

There are also datasets used that are not derived directly from remotely sensed data using a mathematical function, rather, they were derived from the European based SoilGrids maps. The maps from the European Soil Hydraulic Database (Tóth et al., 2017) are generated using pedotransfer functions and were trained using European Hydro-pedological data. These datasets include information on the already discussed field capacity (FC), saturated water content (SWC or THS) and permanent wilting point (WP). The datasets from the EU Hydro dataset are generated using climatological data and therefore are not influenced by weather anomalies. Moreover, an interpolated version of the Soil Organic Carbon Stock (SOCS) from the SoilGrids database is also used as a predictive raster. Adding to the soil moisture characteristics is the Available Water Content (AWC) and the Water Wilting Point (WWP), also obtained from SoilGrids. Even though the WP and WWP are expected to measure the same soil property, they are still added simultaneously to assess the difference between two data sources, and to see whether one predictor has the advantage over the other. The predictors mentioned above originally

have a spatial resolution of 250 meters, which is too coarse to be used in the Random Forest algorithm. For this reason, they are each resampled into 30 by 30-meter pixel rasters using cubic convolution interpolation. While this assumes the fact of spatial autocorrelation, the decision for the resampling has been made since it is expected that the properties mentioned above (water-related properties and soil organic carbon stock) do in fact have spatial autocorrelation. Additionally, the choice for interpolation was made to prevent the original coarse pixel size from causing a pixelated image resulting from the Random Forest algorithm.

To sum up, a table is presented listing additional datasets used in running the Random Forest algorithm and thereby predicting individual soil properties.

Table 3.4: *Derived properties and extra predictors listed by source*

Dataset name	Information	Derived from
NDVI	(NIR - RED) / (NIR + RED)	Sentinel 2
NDMI	(NIR - SWIR) / (NIR + SWIR)	
EVI	2.5 * (NIR-RED) / (NIR+6*RED-7.5*BLUE+1)	
SATVI	(SWIR1-RED) / (SWIR1+RED+1) * 2 - (SWIR2/2)	
NSMI	(Ri−Rj) / (Ri+Rj)	
Tasselled Cap	0.1509*BLUE + 0.1973*GREEN + 0.3279*RED + 0.3406*NIR − 0.7112*SWIR1 − 0.4572*SWIR2	
Valley Depth	Height difference between the elevation and ridges	EU-DEM
Modified Catchment Area	Delineation of river areas	
LSF	Length Slope Factor	
TWI	Topological Wetness Index	
CS	Catchment slope	
Curvature	Description of basins	
Slope	Slope of surfaces	
Average height	Height difference between pixel and the average of the surrounding pixels in a 100m radius	
Field Capacity	Water content at field capacity of a soil	EU Hydro
WP	Water content at wilting point	
Saturated content	Saturated water content	
SOCS	Soil organic carbon stock	Soil Grids
AWC	Available Water Content	
WWP	Water content at permanent wilting point	

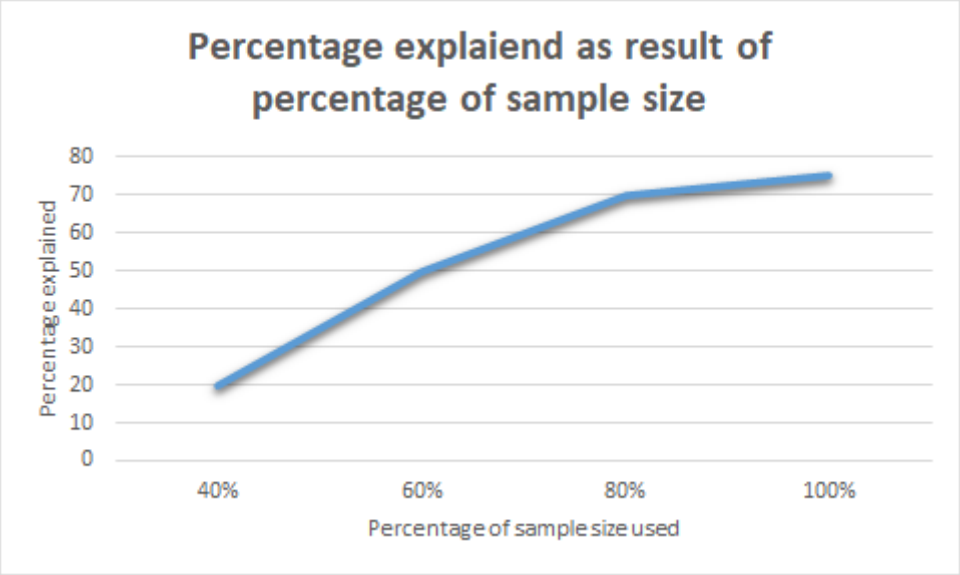
3.5 Running the Random Forest Algorithm

For running the Random Forest algorithm, the spatial statistics toolbox in ArcGis Pro is used. The toolbox contains the “Forest-based Classification and Regression” tool, allowing for a geospatial use of the algorithm. While the name differs from the one mentioned in Breiman (2001), it is the same in essence. The tool simplifies the user experience greatly, nevertheless it still allows for the adjustment of many different settings, including verification methods and algorithm specific variables.

The execution of the Random Forest algorithm using the forest-based regression tool in ArcGis is performed in combination with an optimisation of the composition of predictive rasters for each soil property and thereby reaching relatively peak accuracy. To determine what set of explanatory variables are most suitable for predicting a certain soil property, some steps must be taken to assess each variable’s influence on the soil property in question. First, a qualitative assessment of each explanatory raster must be performed to indicate what their possible addition would be in explaining the variation in soil property values. When this process is complete, the rasters that are expected to perform well with a soil property are used as explanatory rasters in the forest-based regression tool and their results are interpreted and noted. Next, other rasters that were not expected to correlate with the dependent variable are added to investigate whether they might reveal unexpected or hidden patterns in their values. Again, the results of the training are noted. Explanatory variables that have a low relative importance are removed from the predictive set. The final set of explanatory variables will partly consist of rasters that are expected to correlate with the dependent variable and will also contain predictive variables that provide an unexpected edge to the algorithm. In addition to selecting the set of explanatory variables based on their performance in predicting the value of an individual soil property, the number of trees is set to 200.

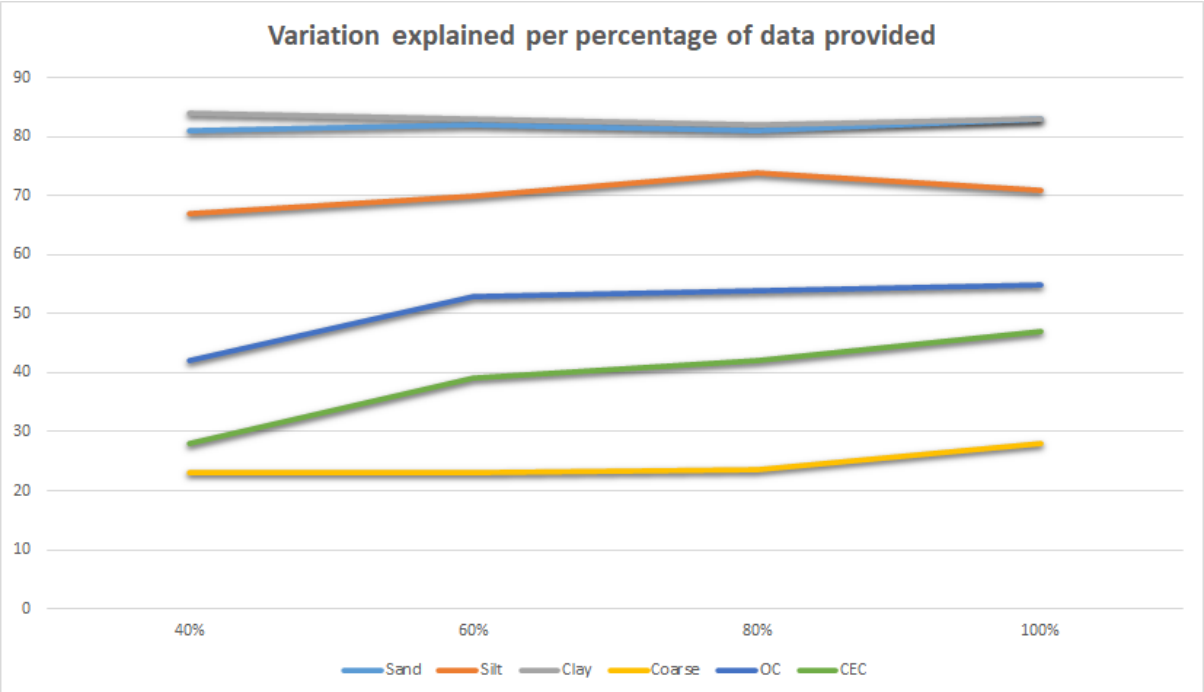
3.4.3 Sample Size Check

In machine learning, it is uncertain whether enough training data is provided to the algorithm. Before any results are produced, it is of importance to assess whether enough training data is available for a proper use of the machine learning method. It is common practice to plot a learning curve that shows the percentage of variance explained against the percentage of total data used. If the percentage of variation explained increases rapidly when more data is used, then it is likely not enough data is provided. Otherwise, when for instance 40% of the data reaps the same amount of quality as 100%, enough data is collected. Usually, when the explained variation is plotted against the percentage of sample size, the graph looks as follows (adapted from Byrd et al., 2012 and Ng et al., 2020).



Graph 3.1: *example of percentage of explained variation as a result of sample size to assess if enough data is provided to the machine learning algorithm.*

For each soil property, such a test is performed using 40, 60, 80 and 100 percent of the sample points to assess whether the Random Forest algorithm is supplied with enough points to provide consistent outputs.



Graph 3.2: *Variance explained per percentage of data provided for each relevant soil property*

From this graph, it is visible that the percentage of variation explained is constant for each of the percentages of the sample dataset used. It can therefore be concluded that the number of training points is sufficient for the Random Forest algorithm, as the percentage of variation explained decreases for a lower number of sample points, e.g., silt, OC, and CEC, while it does not rapidly increase when more points are provided. It seems that the percentage of variation explained has hit an asymptote where adding more values does not increase the percentage of variation explained.

3.4.4 Calculating Liquid Limit and Plasticity Index

Both the liquid limit and the plasticity index are difficult to predict by using solely remotely sensed data as input. Moreover, both soil properties do not appear in either the LUCAS or WoSIS dataset. Due to the lack of ground truth data, Random Forest cannot be applied to predict liquid limit and plasticity index. Therefore, another method of calculating both soil properties must be used. This method that uses both clay content and cation exchange capacity is outlined before and the formula is applied using map algebra. Verifying the results of this method can be done through assessing whether the liquid limit outcomes from the generated formulas have a lower or higher value than 50, and then comparing this to the solid-assumptions USCS map. The liquid limit in areas with OH, CH and MH should be higher than 50, and in other areas it should be lower.

3.5 Validation and Testing

When the rasters of individual soil properties are generated, it is vital to the quality of the final product to assess the accuracy of the individual rasters. In other words, does the raster correctly predict the values of the soil properties? The first and most simple method is comparing the ground truth points to the raster value at their location. It is expected that these values should be similar, as the ground truth points are used to train the algorithm. Therefore, the algorithm should know that at this location, given the provided predictors, the value should be the same as the ground truth point. However, performing this comparison on the same points that the algorithm is trained on will result in a biased assessment of the model's accuracy. When generating results using the ArcGIS forest-based regression, one cannot take the resulting soil property maps as a fact. The model must be validated and tested. Validation data is used during the training of the model, but the model does not use that data to ‘learn’. The validation data is used to evaluate the model in an unbiased way measuring the model fit. If this dataset is included into the model configuration however, the model’s bias increases. The test dataset is used after the model is completed to provide a model evaluation. In summary, the test set is locked away during the training process, while the validation set is used during the training without influencing the training process.

To perform a scientifically based validation and test, it must be assured that the testing data is not in any way present in the training and validation of the model. For this reason, a randomly selected 20 percent of the ground truth dataset is kept separate from the rest before any training or validation is performed. This set of data points is used at the end of the training-validation-test pipeline. The pipeline used for training, validating, and testing the model is illustrated below.

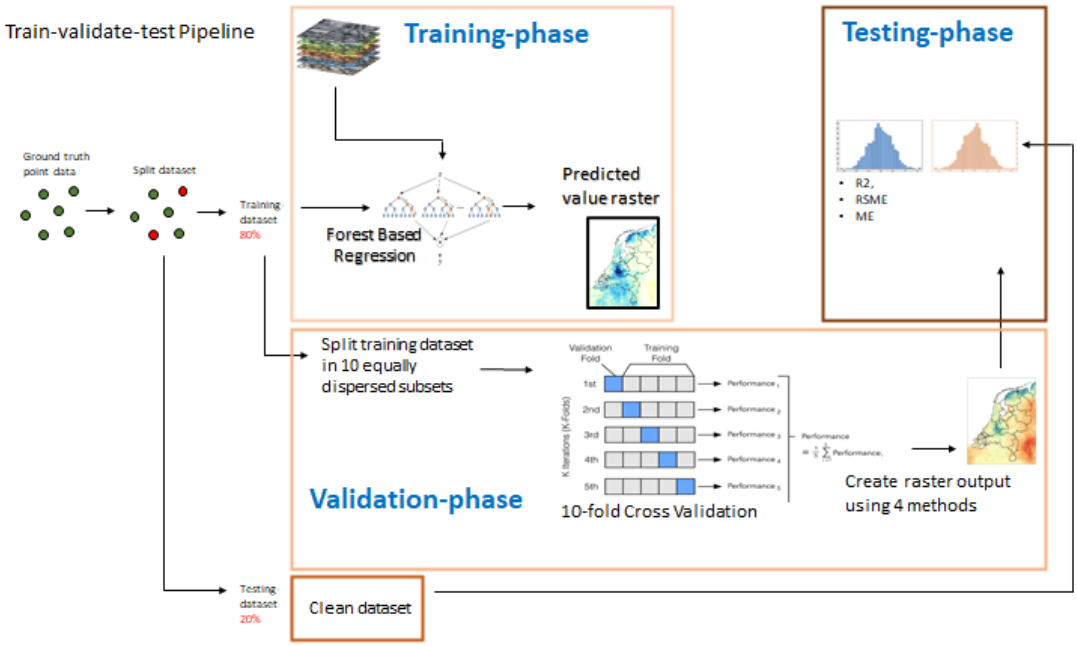


Figure 3.5: *train-validation-test pipeline*

As seen in the pipeline, the data split is done before any other operation. 80 percent of the data is used as training data, and the other 20 percent is reserved for testing. This division is done both geographically and value-wise. First, the entire dataset is split up into 5 subsets based on their values, i.e. subset 1 has the lowest 20% of the values and subset 5 has the highest 20% of the observed values. From these 5 subsets, the testing subset is created by taking a geographically spread 20% from each of the 5 subsets, thereby accounting for 20% of the entire sample set. The 10 validation subsets are generated in a similar way, only for these sets, a geographically spread 8% is taken from each of the 5 subsets.

The training phase incorporates selecting the best scoring predictive rasters. This is done through running multiple combinations of predictive rasters through the Forest Based Regression tool with 10 percent selected for testing. For the process of selecting the best scoring variables, the entire sample set is used, from which 10% is kept separate for testing. This testing set generates an R-squared value and RMSE accompanied with a P-value. When the R-squared consistently results in a high value and the lowest RMSE is achieved, the set of rasters used will be selected for the final model. This, however, is only done when the P-value is below 0.05. When the best combination of rasters is selected, a predicted value raster is generated once. This raster will be used to clean data points from both the training and testing subset, as some points fall outside of the relevant research area.

After the highest scoring predictive rasters have been selected and a single output has been created, the validation phase commences. For the purpose of proper validation, a k-fold-cross-validation is performed. A cross-validation is performed to gain a statistical insight into the model's workings and accuracy. This type of validation calculates three statistics that each say something different about the model's precision: the R-squared, Root Mean Square Error (RMSE), and Mean Error (ME). To improve the estimate of the model's performance, a k-fold validation is applied, meaning that the training dataset is randomly divided into 10 subsets. Then, the cross-validation is run 10 times, with each time another subset selected as the validation fold, and the rest of the subsets being the training folds. This process results in 10 cross-validation performances (10 times a calculation of R-squared, RMSE, ME), of which

the mean can be calculated in order to depict the average performance of the model for an individual soil property. Additionally, when performing the k-fold-cross-validation, the Forest Based regression tool in ArcGIS also outputs a Q-high and a Q-low raster, each depicting the upper and lower end of the 90% confidence interval, respectively. A feature of the ArcGIS forest-based regression tool is the option to create uncertainty rasters. One depicts the lower bound of the 90% confidence interval, and the other shows the upper bound. Subtracting the lower from the upper bound therefore creates the range in which 90% of the trees in the Random Forest resulted in. If this range is large, the model is more uncertain than when this range is low.

It is expected that an average of the values from the 10 output rasters generated during the k-fold-cross-validation will result in the most optimised output of the model. If one of the 10 k-fold-cross-validation rasters predicts outliers, then these are nullified when a mean is created of all 10 rasters, further adding to the reduction of overfitting that the Random Forest already offers. The average value per pixel of the 10 generated rasters can also insert a form of distortion in the result. Since the training dataset is divided into 10 subsets, which are used to run the Random Forest algorithm 10 times, each time leaving out one of the subsets, this means that a single ground truth point is present in 9 out of 10 runs. From this, it can be deducted that in 9 out of 10 rasters, the pixel residing at the same spot as the ground truth point is predicted correctly, while it is likely that 1 out of 10 times the value differs. For this reason, several different methods for generating the validated raster map are used including the mean method, the median method, standard deviation method and restructured method. For all four methods of generating the validated raster (mean, median, standard deviation and restructured method), the test is performed during the testing phase. Each method is outlined below.

- **Mean Method**
  The mean method uses the results from each k-fold-cross-validation test and calculates the mean from the 10 resulting rasters.

- **Median Method**
  The median method is performed in the same way as the mean method but uses the median of all 10 rasters instead of the mean.

- **Standard deviation method**
  The standard deviation method calculates the standard deviation of the 10 values per pixel and discards the values that deviate two standard deviations from the average. Then, the mean is calculated from the remaining pixel values.

- **Re-Structured method**
  The restructured method discards the notion of using the resulting rasters from the k-fold-cross-validation to create the validated raster map. Rather, it performs the k-fold-cross-validation and calculates its statistics but uses all the training data (80% of total samples) to create the validated prediction map. This result is therefore not ‘optimised’ like the methods described before.

Last, there is the testing phase. The values in the training dataset are compared to the predicted values, and using their differences, the same statistics from the k-fold-cross-validation are calculated. These statistics provide the final verdict regarding the model’s performance. When these statistics are calculated for each individual soil property, the average can be calculated again, indicating the performance of Random Forest in predicting the soil properties needed for classifying soil into the USCS classification.

3.6 Additional comparative methods

Another method of comparing the results is performing a map algebra calculation that adds the values of sand, silt and clay together into one raster. In theory, the value for each pixel should then be 100. In spite of the seeming simplicity and straightforwardness of this method, there are some caveats. First, when the combination of the three soil properties ends up being 100, the possibility remains that there is still some error in the prediction, where one value is estimated too low and one other is estimated too high. Furthermore, when the summed-up value is either too high or too low, there is no way of telling in which of the three soil properties the error occurred. Therefore, this method should only be used to provide a quick indication of the algorithm's output and should not be considered as the factual truth when a value does result in 100.

3.7 Generating the USCS classification

When rasters are calculated for all 8 needed soil properties (sand, silt, clay, CEC, OC, coarse fragments, LL, PI), the proper USCS classification can be determined. To achieve this, the same python script is used when classifying the LUCAS and WoSIS datasets. The rasters of the predicted soil properties are converted to points, which are combined using a spatial join. Fields are added to calculate the liquid limit and the plasticity index, and all 8 soil properties are then used to generate the USCS classification. Then, the point layer is translated to a raster dataset with the USCS classification as band value. This results in a product that contains the USCS classification with a spatial resolution of 30x30m.

A caveat with classifying soils into the USCS classification is that the classification table (figure 2.2) is not 100% inclusive of all soils. This might result in some 'unknown' soil classifications, which is not desirable when the goal is to use the generated map in practice. For instance, when a soil has over 50% sand, it can end up being either SW or SP, or SC or SM. When a sand soil has less than 5% fines, it is either SW or SP, but when it has more than 12% fines, it falls in either the SM or SC category, depending on its position on the A line. Therefore, when classifying a soil that has over 50% sands and a fine value between 5 and 12, it traditionally reaches a sort of middle category. In real life however, there is no middle category. For this reason, the dividing value is set on 12, meaning that anything with less than 12% fines also falls into the SP/W category. Additionally, the peat category is not defined by numbers, rather, it is described as "highly organic with a dark colour and an organic smell". Since the peat category is not defined using exact numbers, some estimation needed to be made on what properties the soil type would have. The requirements a soil should meet to become classified as peat are: have a liquid limit of 50 or higher and have an organic content value of 200 or above, *or* have a liquid limit of lower than 50 and have an organic content value of 150 or higher.

4 Results

The results section starts with the qualitative and qualitative selection of predictors for each soil property to create the most optimised result. Next, the variable importance for each soil property are depicted in graphs, along with their source. This allows for a quick overview of what explanatory variables are useful for predicting the six soil characteristics. This is followed by the validation results that were gathered during the validation phase of the pipeline. Using the training and validation data, map results are produced, from which an additional map is created using the sum of sand, silt and clay values. Next, the test results and uncertainty maps are presented to provide an insight into the independent accuracy of the prediction model. Last, the maps of the individual soil properties are combined and translated into the USCS classification.

4.1 Selecting Predictive Variables

For the second iteration of the Random Forest algorithm, a selection of rasters was made for each individual soil property. The final selection of rasters for each property reaped the highest test score from multiple tests. To come to this final set of rasters, some steps were taken to select, add and filter out certain predictive variables. First, a qualitative assessment was made on which rasters were expected to perform well in predicting a soil property. This is done through analysing literature and by linking predictive variables to physical properties of the soil property in question. As discussed in the theoretical section of this thesis, there are indeed some independent variables that help explain certain soil properties. Then, after an initial R-squared is generated, additional rasters are added to assess whether they might have an unexpected correlation with the dependent soil property. Below, the independent variable selection process is outlined for each dependent variable (sand, silt, clay, coarse fragments, OC, and CEC).

4.1.1 Sand Predictors

For sand, the predictors expected to correlate well were water-related variables, elevation and slope variables, and Sentinel 2 bands from the visible spectrum (red, green, blue). Water variables were expected to do well since sand is relatively dry when compared to clay and silt. The grains are larger and therefore let through more water, relating to the Field Capacity and Available Water Capacity (Brouwer, 1985; Cassel & Nielsen, 1986). Due to the grains being larger, the soil can store more water than other soils, which might lead to a higher saturated water content. Elevation and slope variables were expected to do well as sandy soils reside mostly in higher up areas. The slope factor has to do with sedimentation in rivers. When a river's slope is steep, all sedimentation soils remain in the water, but when slope decreases and the water in the river slows down, sand grains are the first to sink to the bottom of the river (Rezaei & Gilkes, 2005; Thompson et al., 2001). Additionally, the valley depth and modified catchment area were implemented to support the notion that rivers might influence sand values (Moore et al., 1991). The red, green and blue bands from Sentinel 2 were used as initial predictive variables as they might detect the distinct yellow colour of sand.

After performing the forest-based regression training-only function in ArcGIS, the results indicated that all water related variables performed well, as the output of the tool each presented them with a relatively high variable importance. The EU DEM and its derived predictors each had a different importance in predicting sand. The elevation map and valley depth explained some of the variance in sand, while the

slope, curvature, average height, and modified catchment area all performed sub-par. Additionally, the red, green and blue band from Sentinel 2 also added some importance to the explanation of the soil's sand content. The seeming irrelevance of the optical bands is likely due to the fact that sandy soils are not bare most of the time, rather, they are covered with herbaceous vegetation or shrubs which conceal the yellow colour. To remedy this, more predictors were added to unveil possible correlations that were not initially expected. Another test was done using the initial variables, and all bands from Sentinel 2, vegetation and moisture indices derived from Sentinel 2, land surface temperature and the soil organic carbon stock. The results again presented the multispectral bands to be unimportant in predicting a soil's sand content. The same result was produced for the Sentinel 2 derived vegetation and moisture indices (NDVI, NDMI, EVI, SATVI, NSMI). The Tasselled Cap Wetness index, however, did explain some of the variation in sand. Moreover, the land surface temperature and soil organic carbon stock also assisted in predicting sand values. Lastly, all other Sentinel 2 bands were added to assess their importance in predicting sand, and only band 1 and 9 were important in predicting sand in addition to the already present red and blue bands.

For the final set of predictors, the unimportant variables are removed from the selection, and the unexpected important ones are added, creating the final set of independent variables for the dependent variable: sand. This set is shown in table 4.1.

Table 4.1: *predictors for sand*

Sand predictors	
Permanent wilting point	NDVI
Field capacity	NDMI
Saturated water content	EVI
Available water capacity	SATVI
Tasselled cap wetness index	NSMI
EU DEM elevation	Valley depth
Soil organic carbon stock	Sentinel 2 band 1, 2, 4 & 9
Land surface temperature	

4.1.2 Silt Predictors

Predicting the silt content of a soil closely resembles predicting sand. For example, silt is expected to be wetter than sand as it has smaller pores for water to flow out of, therefore the water-related rasters are also assumed as important for predicting silt (Brouwer, 1985; Cassel & Nielsen, 1986). Especially the indices derived from the EU DEM, valley depth and modified catchment area, were expected to be influential in predicting silt as the mechanics of rivers might play a role in silt distribution (Rezaei & Gilkes, 2005). In addition to the water-related variables, all elevation related rasters are added as well to incorporate the height and slope variables into the equation (Thompson et al., 2001). While the Sentinel 2 bands did not reap any benefits for sand, bands 1 through 8 were added to not exclude the possibility that these still might have an impact for predicting a soil's silt value.

The results for silt are somewhat different to those for sand. To start, another set of Sentinel 2 bands play a role in predicting silt values, as bands 1 through 4 explain some degree of silt value. Especially band 3, the green band, explains the most out of all Sentinel 2 bands. The rest of the bands, however, did not provide any assistance to the Random Forest algorithm as their importance in predicting silt was negligible. The most influential predictive variables are the permanent wilting point, field capacity and

the EU DEM elevation map. Similarly, to the sand predictors, the rest of the water-related variables derived from the EU Hydro dataset perform well in predicting silt. Differing from sand, the slope map did assist in predicting silt, and both valley depth and the modified catchment area did prove useful. To gain insight into potentially more relevant independent variables, the other rasters are added to see and assess their importance to the Random Forest algorithm. The Sentinel 2 derived indices did not prove important, nor did Sentinel bands 9 through 12. Predictors that unexpectedly did assist in the algorithm were soil organic carbon stock and the land surface temperature.

When keeping the relevant predictors and removing the unimportant ones, the following set of independent variables is created to be used to predict silt.

Table 4.2: *predictors for silt*

Silt predictors	
Permanent wilting point	MCA
Field capacity	VD
Saturated water content	TWI
Available water capacity	Tasselled cap wetness index
EU DEM slope	Land surface temperature
EU DEM elevation	Sentinel 2 band 1, 2, 3 & 4
Soil organic carbon stock	

4.1.3 Clay Predictors

According to the literature discussed in the theoretical section of this framework, both the available water capacity and soil organic carbon stock should correlate well with the fraction of clay in a soil. For this reason, the available water capacity and other independent variables from the EU Hydro dataset are added along with the soil organic carbon stock (Brouwer, 1985). Extending upon this assumption, the moisture indices derived from Sentinel 2 are also used in the first test run for the predictive rasters for clay (Cassel & Nielsen, 1986). Additionally, the radar imagery from Sentinel 1 is added to the list of initial predictors as clay is also expected to correlate with satellite imagery from this platform. Radar is sensitive to water, and clay usually does not allow perfect drainage of water and will therefore be highlighted on the Sentinel 1 dataset (Zhou et al., 2020). Lastly, the variables from the EU DEM dataset are added, including elevation, slope, curvature, and average height. These independent variables are again expected to correlate with clay, as the elevation variables correlate with clay according to previous research as discussed in the research (Thompson et al., 2001).

After running the first test for selecting predictors to predict clay, the best performances came from the permanent wilting point, field capacity and the elevation model. The soil organic carbon stock also performed well as expected. Similarly, the radar imagery from Sentinel 1 did help in predicting the values for clay, which is surprising, since this variable did not prove important in predicting both sand and silt. Furthermore, while the elevation model performed well in explaining variance in clay, its derived variables (slope, curvature, average height, valley depth and modified catchment area) performed slightly worse than the DEM. Nevertheless, the slope, curvature and average height did explain some variation in clay and were therefore deemed as useful for predicting this soil property along with valley depth, topographic wetness index and catchment slope. Moreover, the moisture indices derived from Sentinel 2 performed poorly, scoring the lowest variable importance out of all variables included. For the second test, all Sentinel 2 bands were added to see whether there might be

some correlation between clay and satellite reflectance. The derived indices related to vegetation are added as well, along with the land surface temperature. Lastly, the tasselled cap wetness index is added to the list of predictors to estimate clay values.

From the variables added in the second test run only the tasselled cap wetness index and the land surface temperature added information gain in predicting a soil's clay content. The other predictors that were added did not help to predict clay content values. By doing both test runs, the following set of predictors was constructed that are used to predict clay using the Random Forest algorithm.

Table 4.3: *predictors for clay*

Clay predictors	
Permanent wilting point	Tasselled Cap Wetness index
Field capacity	EVI
Saturated water content	SATVI
Available water capacity	VD
EU DEM slope	CS
EU DEM elevation	TWI
EU DEM curvature	Sentinel 1 SAR VV
EU DEM average height	Land surface temperature
Soil organic carbon stock	

4.1.4 Coarse Fragments Predictors

Predicting the percentage of coarse fragments in a soil is significantly more complicated than predicting sand, silt, and clay. This is partly due to the fact that less attention is paid to this soil property in literature compared to the other soil properties that are included in this research. Another reason that selecting the most important predictive variables for coarse fragments is less straightforward than the others, is that the selected research area does not lend itself to the analysis of coarse fragments. In spite of this complication, the selection process for the predictors of coarse fragments was performed in a similar way to the other variables. The elevation model and its derivatives are the predictors that are assumed to be most important in estimating coarse fragment content in the soil (Vaysse & Lagacherie 2015). Additionally, the water-related variables were also expected to be of influence in predicting the variance of coarse fragments, since water is allowed to flow freely through coarse fragment particles. Furthermore, the Sentinel 1 radar imagery was added as coarser and more rugged terrain shows up differently on radar than smooth terrain (Hengl et al., 2017).

Most of the qualitative assumptions that were made produced relatively solid results. For predicting coarse fragments in soils, the EU DEM elevation model was most important. The slope, curvature and average height were also influential in predicting the dependent variable. Similarly, the expectations regarding the water-related variables were also met since they also played a role in the information gain of the algorithm. From this dataset, the saturated water content variable performed best, while the available water content had relatively little importance. Unlike the previously mentioned explanatory variables, the Sentinel 1 radar imagery performed less of the variance in coarse fragments than expected, but it still remained important enough to not be discarded from the set of predictors. For the second test phase in selecting the predictors for coarse fragments, the bands from Sentinel 2 were added, along with the Sentinel 2-derived vegetation and moisture indices. Additionally, the soil organic carbon stock

variable was added to see if chemical properties might shed light on the variance in the dependent variable.

From this test, the Sentinel 2 bands proved useful in explaining coarse fragments values. Bands 1, 2, 3, 4, 5, 9 ,11 and 12 were deemed important enough to be incorporated into the predictor set for coarse fragments. Furthermore, only the tasselled cap wetness index explained coarse fragments distribution to some degree, while the other indices were valued as unimportant. Surprisingly, the soil organic carbon stock performed relatively well in predicting the dependent variable. Combining these findings, a predictive variable set can be constructed for coarse fragments and is as follows.

Table 4.4: *predictors for coarse fragments*

Coarse Fragments predictors	
EU DEM slope	VD
EU DEM elevation	CS
EU DEM curvature	TWI
EU DEM average height	MCA
Field capacity	Land surface temperature
Saturated water content	Sentinel 1 SAR
Available water capacity	Tasselled Cap Wetness index
Permanent wilting point	Sentinel 2 bands 1 through 5
Soil organic carbon stock	Sentinel 2 bands 9, 11, 12

4.1.5 Organic Content Predictors

Organic content in the soil can, according to previously performed literature discussed in the theoretical section, be predicted by vegetation and moisture indices, height variables and radar imagery. For this reason, a first test was done assessing the importance of the NDVI, NDMI, SATVI, EVI, NSMI, tasselled cap wetness index, the digital elevation model, slope map, curvature, average height and Sentinel 1 radar imagery. Since the indices used in this first test are derived from the Sentinel 2 satellite platform, its bands are not yet included in order to see whether the indices might explain a sufficient percentage of the variation in organic content in the soil. The vegetation indices are expected to perform the best out of all selected independent variables as organic content is assumed to have a green-ish colour (Zhou et al., 2020). Therefore, the NDVI, EVI and SATVI should score a high importance in the random forest algorithm test (Heute, 1988). Moisture should also be of influence, as other researchers have stated that organic content in the soil holds higher concentrations of moisture than soils with little or no organic content (Hong et al., 2018).

The early assumptions made on predictors for organic content were confirmed by the results of the test assigned each variable a percentage representing their relative importance. Out of the first set of explanatory variables, the normalised difference vegetation index, normalised difference moisture index and digital elevation map scored the highest degree of importance. Additionally, the other predictors that were selected for the initial test also performed relatively well, but the total percentage of variation explained left something to be desired. Especially the Sentinel 1 radar imagery explained a fair percentage of the variation in organic content compared to the previous soil characteristics. To remedy the subpar performance of the set of variables from the first test, more explanatory rasters were added. Valley depth, topographic wetness index and modified catchment area were added to possibly increase the variation in elevation variables in the explanatory set. Furthermore, the nitrogen content and soil

organic carbon stock were added. Initially, these were not added to assess whether organic content could be predicted without the assistance from explanatory variables from SoilGrids, but as the first test did not reap the desired outcome, the choice was made to add both the nitrogen and the soil organic carbon stock. As the moisture indices performed well in predicting organic content, the EU Hydro rasters were also added to the second iteration of the test to assess if these also explained variation in the soil property. Lastly, the bands from Sentinel 2 were added to the second test to see what variables are important in predicting a soil's organic content.

From this follow up test, it seemed that nitrogen content explains by far the largest amount of variation in soil organic content. This is not surprising, as the decomposition process of organic matter releases large amounts of nitrogen into the soil and air. Similarly, the soil organic carbon stock also added a great deal of explanation to the test, having the second greatest variable importance in the algorithm. In contrast, the bands from the Sentinel 2 satellite did not add a vast amount of explanation to the variation in soil organic content, but some were still added to the final set of explanatory variables for organic content. More notable was the small variable importance for all EU Hydro rasters, meaning that soil water properties do not predict organic matter, but topsoil moisture indices do. The final set of explanatory variables for organic matter is as follows.

Table 4.5: *predictors for organic content*

Organic Content predictors	
Nitrogen	VD
Soil organic carbon stock	MCA
NDVI	TWI
NDMI	Sentinel 2 bands 1, 2, 3, 4 & 5
EVI	EU DEM elevation
SATVI	EU DEM curvature
NSMI	EU DEM average height
EU DEM slope	Tasselled Cap Wetness index
Sentinel 1 SAR VV & VH	

4.1.6 Cation Exchange Capacity Predictors

Cation exchange capacity is the measure that defines the capacity of exchangeable cations a soil can hold. This is therefore a chemical soil property, rather than a physical one. For this reason, it is expected that predictive rasters that contain information on chemical soil properties will perform best in the test of the variable importance in the Random Forest algorithm. The available rasters that describe chemical soil characteristics are the EU Hydro datasets and the soil organic carbon stock. The hydrological rasters are expected to have the highest impact, as water directly influences the exchange of cations in a soil. Additionally, the EU Hydro rasters provide information that concerns the inner workings of a soil, rather than only the topsoil reflectance for instance. Another chemical property that might add to the prediction of the cation exchange capacity is the nitrogen content in the soil. While this chemical property does not directly influence the exchange of cations in a soil, it might indirectly explain the dependent variable to some degree (Zhou et al., 2020). The same assumption is made for the soil organic carbon stock, as it also has no direct causal relationship to cation exchange capacity but might eventually prove useful in predicting the soil property. The vegetation indices are included as it is in no way expected that they will explain the variation in the cation exchange capacity (da Silva Chagas et al., 2019).

This set of predictive rasters performed relatively well, as all of the included predictors explained a fair percentage of the variation in cation exchange capacity. The most important variable in predicting the dependent variable using the Random Forest algorithm was the permanent wilting point, followed by the nitrogen content, field capacity and the soil organic carbon stock. The assumptions towards the EU Hydro rasters and the chemical soil properties were therefore confirmed. Nevertheless, there are still physical soil properties that might to some degree explain the variation in cation exchange capacity. To test this hypothesis, the EU DEM and its derived properties are added to the second test, and all the bands from the Sentinel 2 satellite, along with the land surface temperature.

From this second iteration of the test to see what predictors the highest importance have in explaining cation exchange capacity, the initial set of independent variables remained the most important ones. However, the elevation map and its derived properties also added some explanation to the algorithm. This might be caused by the fact that cation exchange capacity correlates with clay content, and soils in low-elevation areas tend to have a higher percentage of clay. More surprisingly, bands 1, 5 and 9 of the Sentinel 2 platform also had a moderate degree of importance and were therefore added in the final set of predictors for cation exchange capacity, which is listed below.

Table 4.6: *CEC predictors*

CEC predictors	
Permanent wilting point	NSMI
Field capacity	NDMI
Saturated water content	Slope
Available water capacity	Curvature
Nitrogen	Average height
EU DEM elevation	LSF
Soil organic carbon stock	MCS
Land surface temperature	CS
Sentinel 2 band 1, 5 & 9	

4.2 Variable importance

Variable importance provides a good insight into what predictors and what sources have the most influence in predicting the soil properties. From these statistics, an overview can be generated that presents the average influence per predictor in predicting the required soil properties for classifying soil according to the USCS classification. The variable importance for each of the soil properties is discussed and summarised to their data source to have an indication of what data source performs best in predicting soil properties.

4.2.1 Sand Variable Importance

Below, the variable importance for the soil's sand content prediction is displayed in percentages. Additionally, each predictor is coloured according to its source. Each source's total share to the variable importance is displayed in the pie chart of which the colours correspond to the bar chart.

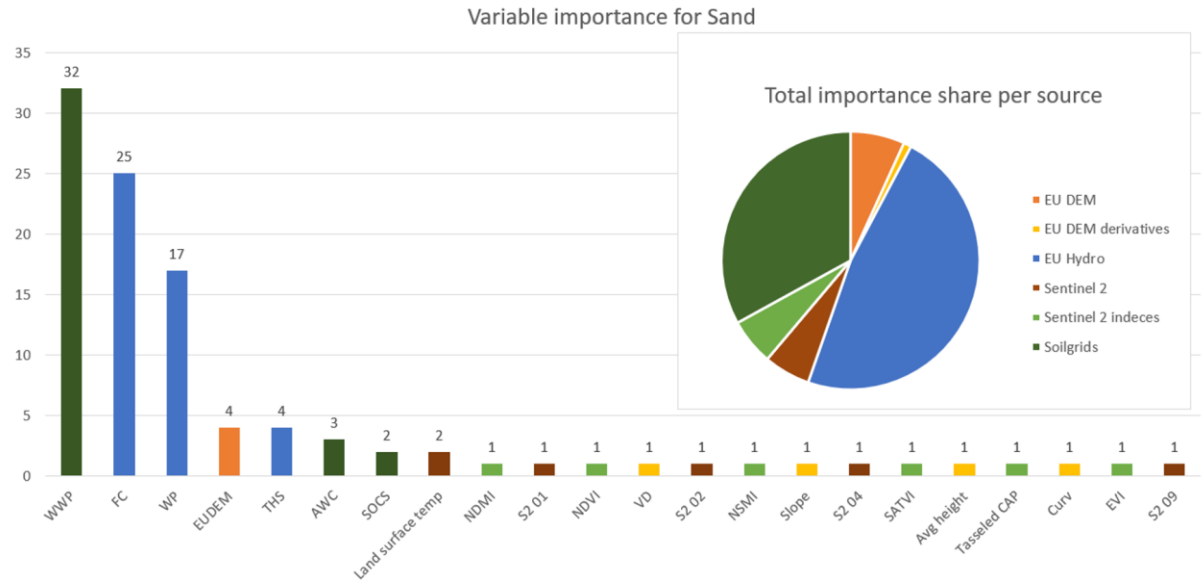


Figure 4.1: *variable importance in predicting sand.*

From this graph, it is discernible that the Water Wilting Point (SoilGrids), Field Capacity (EU Hydro) and Permanent Wilting Point (SoilGrids) together have the most importance in predicting sand properties. Less important but still influential are the EU DEM, Saturated Water Content (THS from EU Hydro) and Available Water Capacity (SoilGrids). It is important to note that four out of the six most important rasters describe water-related soil properties, meaning that water properties predict a significant amount of variation in sand content values. Additionally, the EU DEM and its derivatives also have some influence in the predictions. It should be noted that from the physical soil properties sand, silt and clay, sand is the only one where the Sentinel 2 bands add at least 1 percent of importance. Lastly, the Sentinel 2 derived indices together add 6 percent of importance to the algorithm in predicting sand.

4.2.2 Silt Variable Importance

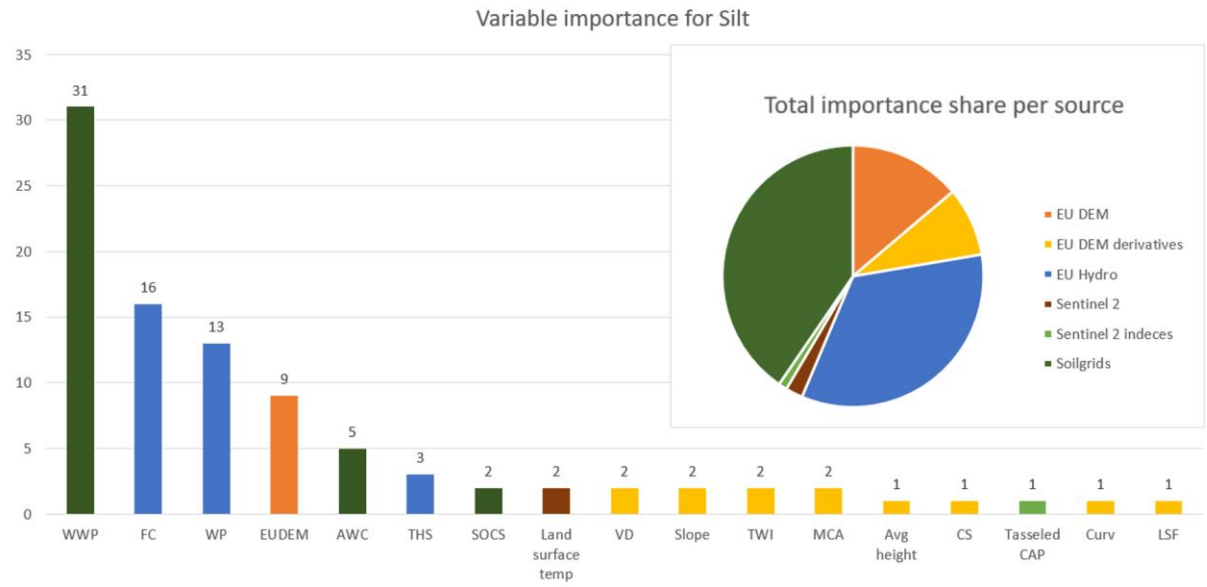


Figure 4.2: *variable importance in predicting silt*

In the graph above, the greatest importance can again be accounted to the WWP, FC and WP, of which two originate from the EU hydro Dataset. Compared to the sand prediction however, the importance in predicting silt content values differs somewhat, as the importance of the EU DEM has more than doubled. Moreover, the EU DEM derivatives also have a higher combined importance when compared to the sand content predictions. From this, it can be assumed that for predicting silt, digital elevation map derivatives are of greater importance than predicting sand content values. Moreover, the Sentinel 2 indices have almost no influence. Only the tasselled cap wetness index has a minor influence in predicting silt content values using the Random Forest algorithm.

4.2.3 Clay Variable Importance

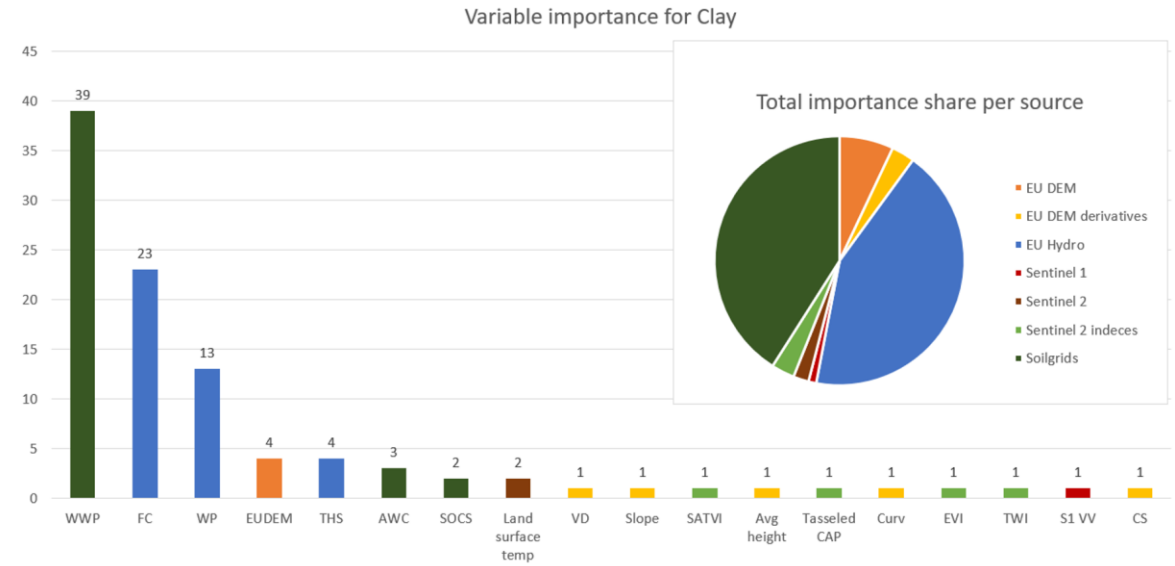


Figure 4.3: *variable importance in predicting clay*

The distribution of importance across variables for clay content prediction closely resembles the distribution for sand content. Again, the WWP and FC have the greatest importance in this prediction, with WP and the EU DEM, THS and AWC coming in second. The EU hydro dataset rasters and SoilGrids datasets are continuing to be dominant in predicting the physical soil properties: sand, silt and clay. Moreover, the EU DEM derivatives add up to 6 percent of importance. Additionally, the Sentinel 2 derivatives add 3 percent of importance in predicting variation in clay content. It is notable that only the Enhanced Vegetation Index and the Soil Adjusted Transformed Vegetation Index have importance in the algorithm, while the NDVI is absent in the importance graph. This might mean that the enhancement and adjustments to the vegetation indices have actual benefits in predicting clay values. Lastly, clay is the only one of the fine-grained soil properties where the Sentinel 1 band has some importance.

4.2.4 Coarse Fragments Variable Importance

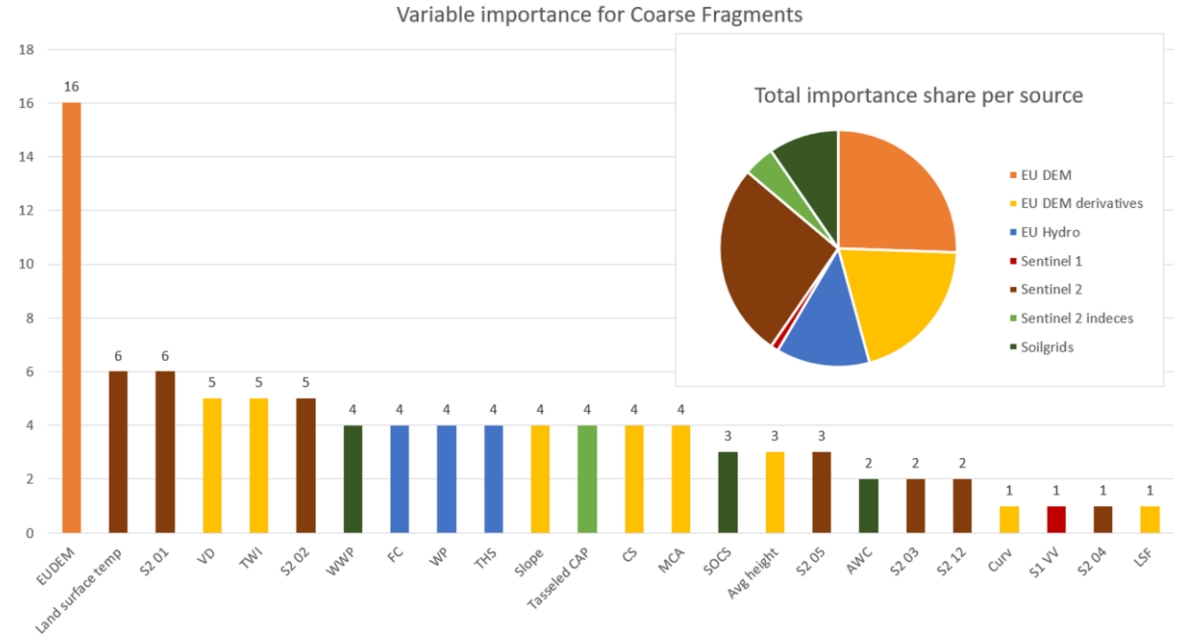


Figure 4.4: *variable importance in predicting coarse fragments*

The variable importance table of coarse fragments paints an entirely different picture when compared to the other physical soil properties. From the graph it is visible that the EU DEM has the highest importance across all predictor variables. The remaining explanatory rasters all have a relatively comparable importance in predicting coarse fragments in a soil. Notable is the Sentinel 2 bands that have a relatively high importance in the algorithm. Combined they even have more importance than only the digital elevation model as seen in the pie chart. The EU DEM derivatives also occupy a significant portion of the pie chart. When the EU DEM and its indices are summed up, their importance is almost half of all the variables combined. The EU hydro datasets play a significantly smaller role in predicting coarse fragments than they did in predicting sand, silt and clay. Moreover, the only Sentinel 2 index that has some importance in coarse fragments prediction is the Tasselled Cap wetness index, which is remarkable, since the NSMI and the NDMI also measure moisture and have an insignificant importance in predicting coarse fragments.

4.2.5 Organic Content Variable Importance

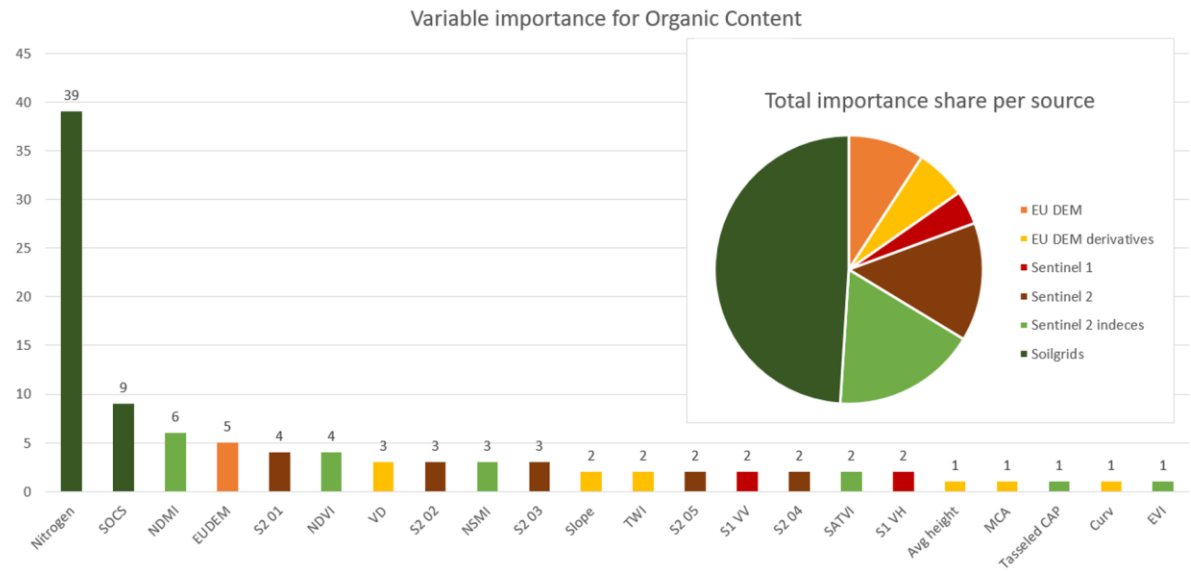


Figure 4.5: *variable importance in predicting organic content*

In predicting the organic content of a soil, nitrogen content is by far the most important predictor variable. This result does not come as unexpected, since decomposing organic content in the soil produces relatively high amounts of nitrogen. Together with the Soil Organic Carbon Stock, Nitrogen takes up almost half of all the importance in predicting organic content. Both the SOCS and the Nitrogen content come from the Soil grids dataset making it the data source with the highest importance. Furthermore, the NDMI and the NSMI moisture indices also have some importance in the Random Forest algorithm. Additionally, the Sentinel 2 bands 1 through 4 are also important. Both the VV and the VH band from the Sentinel 1 Radar satellite are relatively important when compared to the previously discussed soil properties. The EU DEM and its derived properties have a comparable importance to the other soil properties. It is remarkable that the EU hydro datasets have absolutely no importance at all in predicting organic content values, while they had significant importance in predicting sand, silt and clay.

4.2.6 Cation Exchange Capacity Variable Importance

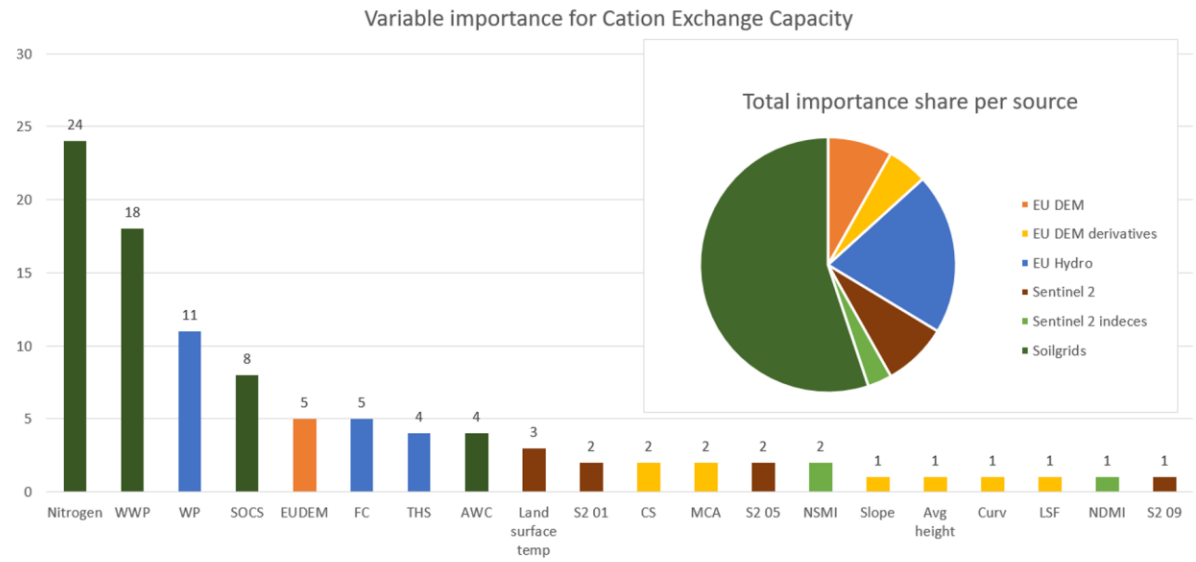


Figure 4.6: *variable importance in predicting cation exchange capacity*

Cation exchange capacity is mostly explained by Nitrogen, WWP and WP. Again, the highest importance comes from the SoilGrids and EU hydro datasets. More than half of the entire importance in the algorithm can be accounted to the SoilGrids data source. The EU hydro dataset has around one fifth of all the importance. The rest of the explanatory variables play a marginal role in explaining cation exchange capacity. It comes as no surprise that nitrogen and the water-related properties have a high importance in predicting cation exchange capacity. It has been mentioned that the cation exchange capacity is a chemical soil property, and it was therefore expected that chemical predictor variables would explain a fair portion of this soil property. While water is not directly a chemical property, it does influence many other chemical properties, which is likely the reason that the water-related rasters score high in importance.

4.2.7 Combined Variable Importance

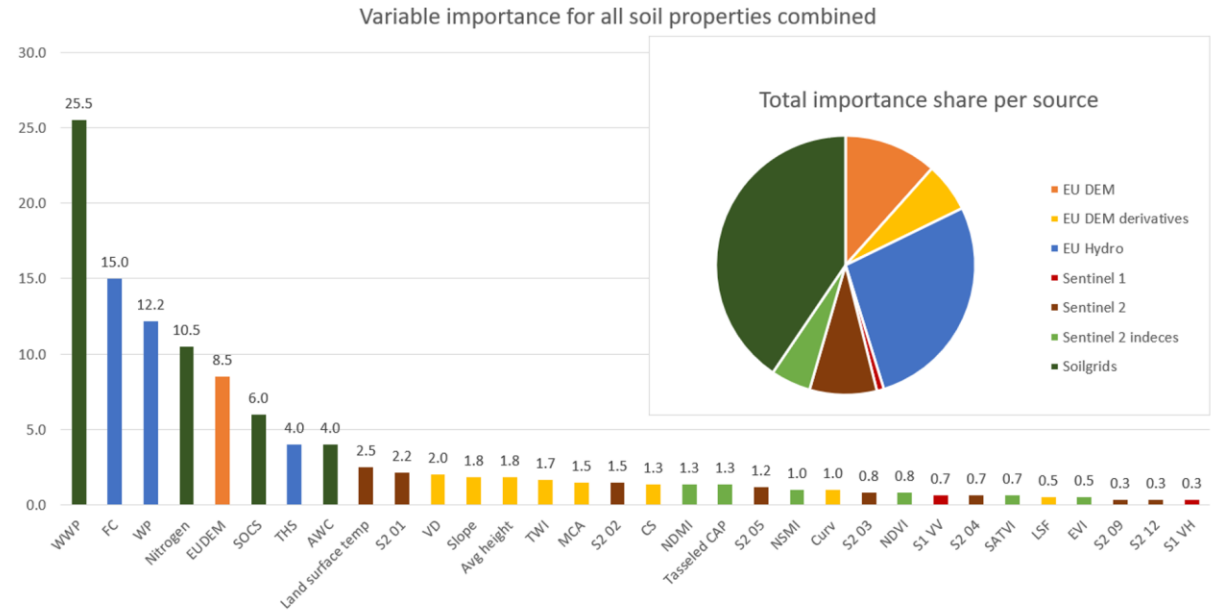


Figure 4.7: *combined variable importance in predicting each soil property*

When combining the importance of all six soil properties into one, the graph shown above is generated. It is clear that the most important variable is the water wilting point. For this reason, the SoilGrids dataset is the source with the highest importance accounted for since it also delivers the influential Nitrogen and Available water capacity variables. Next is the EU Hydro dataset, which provides the wilting point, field capacity and saturated water content predictors. It is notable that the EU DEM predictor is also relatively important in the combined importance of the soil variables. It is one of the few explanatory variables that is comparatively important for each of the soil properties. Similarly, the EU DEM derived properties also have a relatively high combined importance for all the soil properties. The unexpected result is that the Sentinel 1 bands have little to no importance in predicting soil properties.

4.3 Validation Results

For each soil property, a validation phase is performed, resulting in statistics that describe the model's accuracy in predicting the dependent variable using the independent variables. Below, the averages for each statistic is displayed in table 4.2. The R-squared statistic indicates the proportion of the dependent variable that is explained by the independent variables when put into a regression model. The RMSE depicts the average of the differences between the expected value and the observed value, also called the error. It can be read as the average deviation from the observed value. The mean error also depicts the average of errors, but does not correct for negative values, and therefore depicts whether the expected values are too high or too low on average. The full tables for the K-fold-cross-validation are in appendix 3.

Table 4.2: *average validation statistics for each soil property: R2, RMSE, ME*

Soil Property	R-Squared	RMSE	ME
Sand	0.524976	12.43061	0.523636
Silt	0.444924	9.699791	-0.21374
Clay	0.652402	4.429596	-0.09874
Coarse	0.221709	4.02621	-0.46143
OC	0.421685	37.14936	-1.35698
CEC	0.557557	6.657696	0.130412

Of all soil properties, clay evidently has the highest R-squared statistic compared to the other soil properties. This can likely be accounted to the effectiveness of the predictive rasters in explaining variation in clay content. Additionally, clay has the highest amount of ground truth points, making the training of the model for this soil property slightly more effective and the testing more representative. The other R-squared statistics, with the exception of organic content, are in the range of 0.42 to 0.56. This means that given the predictor variables, the model explains 42 to 56 percent of the variation in sand, silt, organic content, and cation exchange capacity. The lowest R-squared is reached by the coarse fragments prediction model with a value of 0.22, meaning that only 22 percent of the variation in coarse fragments. This result is not surprising, as the coarse fragments values in the research area are very small to non-existent, leading to the model having great trouble in predicting the explanatory variable's values.

The Root Mean Square Error depicting the average difference between the expected and the observed value for a soil property was the highest, and therefore the worst-scoring, for the organic content. This large average deviation of the expected values from the observed ones can probably be explained by the enormous outliers present in the organic content dataset. Most organic content points are between 0 and 50, while some values in the dataset reach up to 500, therefore generating a high RMSE when a lower value than 500 is predicted, while 500 is observed. The lowest two RMSE values were attained by the coarse fragments model and the clay model. The low RMSE for the coarse fragments model is unexpected since the R-squared statistic for this soil property scores significantly lower than for the others. It can however be explained by the very low values of coarse fragments in the Netherlands and Belgium, which leads to minor deviations in absolute terms, but large discrepancies in relative terms which leads to a low R-squared value. The second-best RMSE is attained by the clay model, which again is likely due to the high amount of training points and relatively high R-squared value.

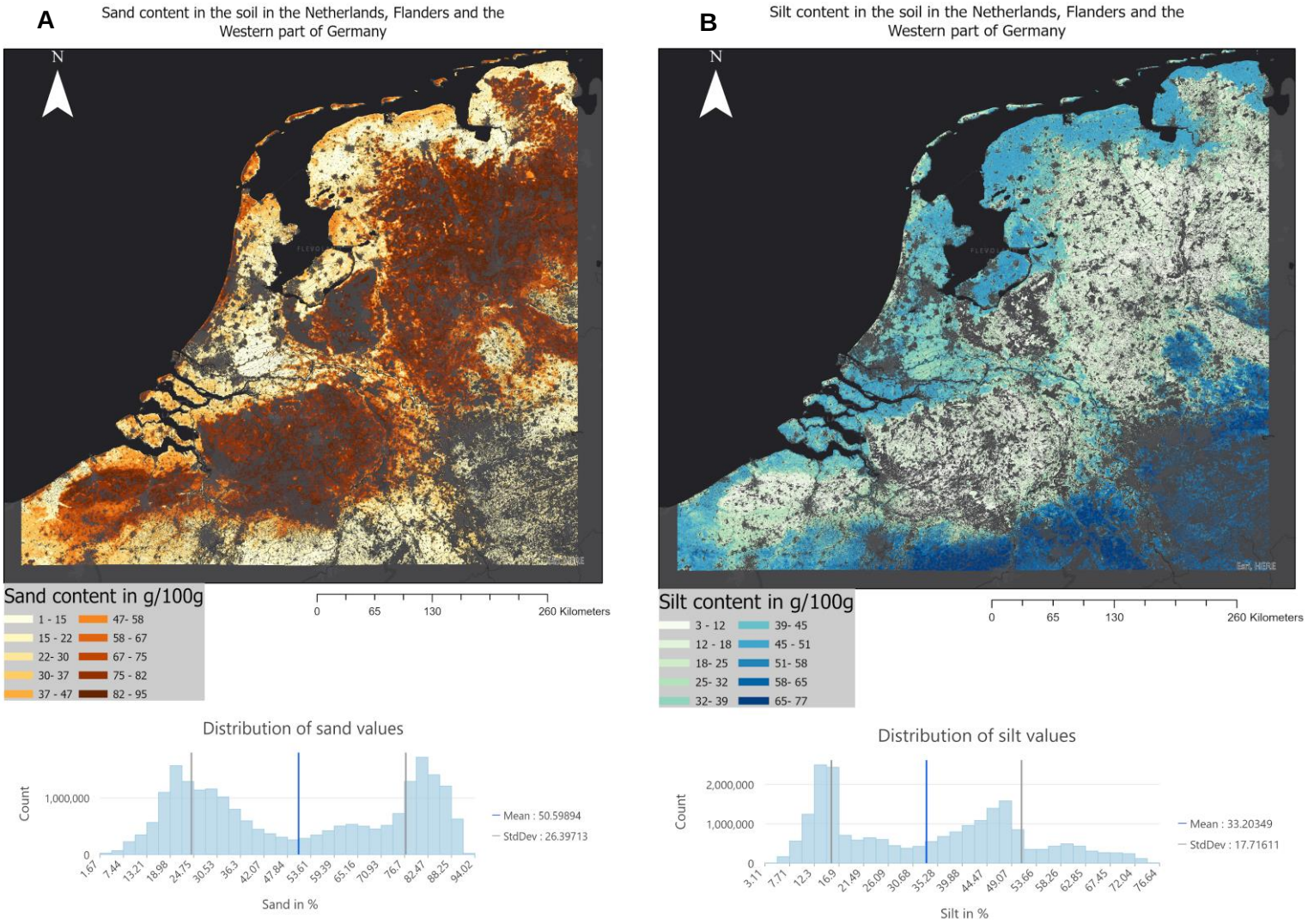
The Mean Error, measuring the average of the differences between the expected values and the observed ones without compensating for positive or negative differences, is relatively small for each soil property. The highest and therefore worst score is attained by organic content. This score can again be explained by the gigantic outliers in the organic content dataset. The fact that the values for the Mean Error are all relatively close to 0 means that the sum of errors that were too high equals the sum of errors that were too low.

4.4 Map Results

From the models used to predict the values of the soil properties sand, silt, clay, coarse fragments, organic content, and cation exchange capacity, six maps are generated. These maps visualise the soil properties' values on a scale ranging from their lowest to their highest value. The visualisation allows

for a study of patterns in the distribution of soil properties, without the ability to conclude whether the actual values for the soil properties are predicted accurately. Each map uses the same colour scale, but utilises a different range of values, meaning that the same colours do not infer similar values. The maps presented in this section are created using the ‘mean method’ as discussed in the methods. This was done as this method produced the most balanced statistical outcomes which are presented later in this section.

4.4.1 Sand, Silt and Clay maps



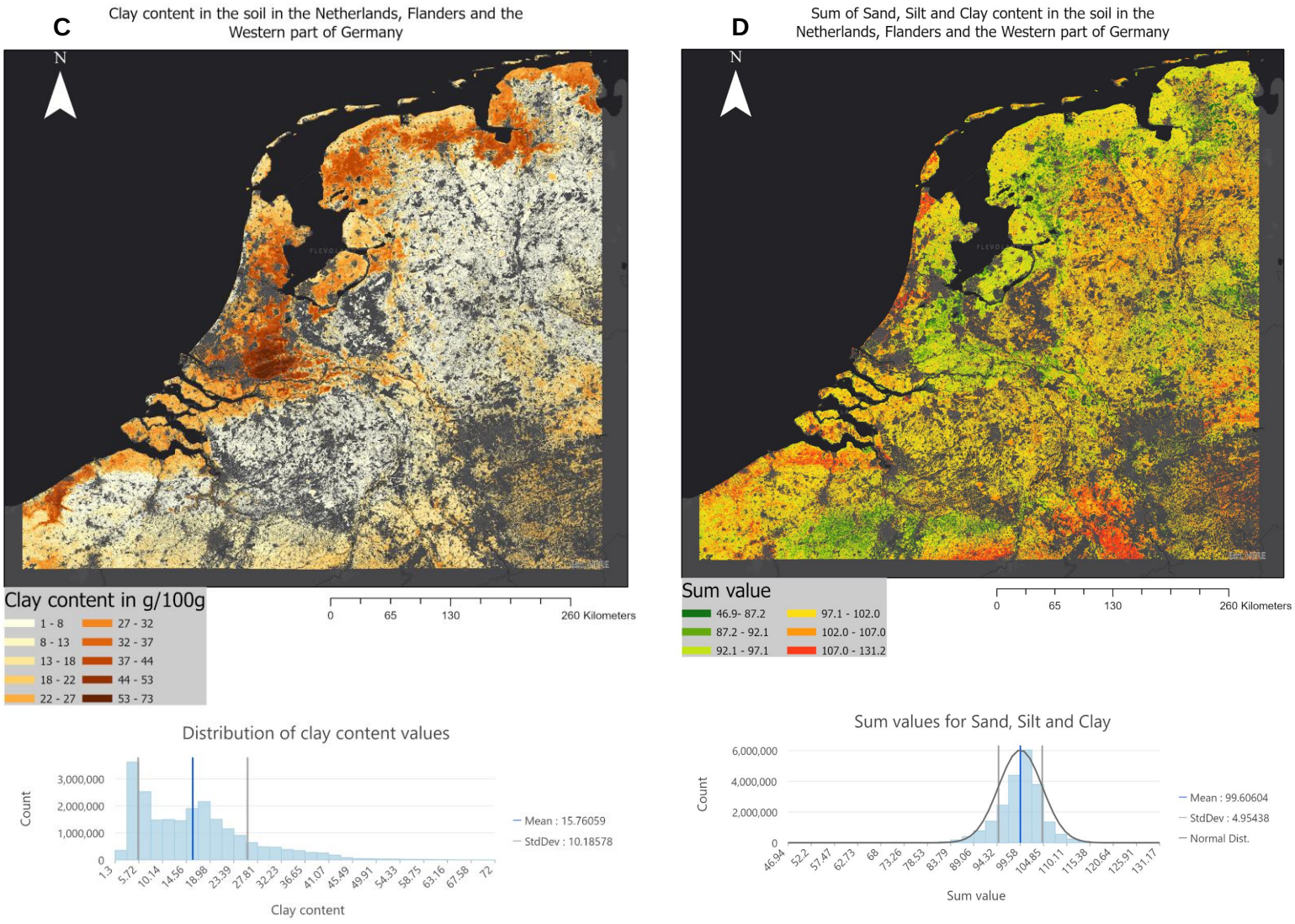


Figure 4.8: *visualisation of the spatial distribution of sand (a), silt (b) and clay (c). The sum of the three soil properties can be seen in (d) with red being too high and green being too low. Each image contains a histogram displaying the value distribution. Larger images can be found in appendix 1*

In figure 4.8, the maps depicting the spatial distribution of sand, silt and clay are displayed. The sand map (figure 4.8a) has high values in the coastal and dune areas, which is as expected since those areas almost completely consist of sand grains. Other areas with high concentrations of sand are located more inland. Soils in the province of Noord-Brabant, west Belgium, and the areas in the far-east of the Netherlands contain high proportions of sand. It should be noted that larger areas with high concentrations of sand are neither located close to rivers, nor located close to water bodies. When looking at the colour scale of the sand map, it is visible that the maximum value for this soil property is 92.7, meaning that dark red-ish areas have soils consisting of more than 90 percent sand. From the silt map in figure 4.8b, it is immediately discernible that only areas to the south east of the research area have high concentrations of silt, as they are displayed as dark red. In the lower elevation areas in the Netherlands, silt grains take up around 30-50 percent of the soil. The areas where silt is mostly absent are the same areas where concentrations of sand are high. The lowest values for silt content in the soil are achieved in Het Groene Hart (the green heart) which is likely due to the presence of peaty and

organic soils in this mostly nature-oriented region. Other relatively low-silt areas reside around and in between the large rivers in the Netherlands. The clay map in figure 4.8c displays high concentrations of clay content in Het Groene Hart, the province of Noord Holland and the continuous area ranging from Friesland into the Eems Harbour. Additionally, a relatively clayey area is present in the very-west part of Flanders. Clay content is also relatively high in some parts of the river area beneath the Veluwe. Another distinct area of high clay content is present just beneath Flevoland and north-west of Amersfoort. Similar to silt, the clay content in the soil is low in areas where sand content is high. In the map depicting the sum of all three presented soil properties (figure 4.8d) it is visible that there are some regions where the model predicts the soil properties too low, and there are areas where the sand, silt or clay content of the soil is overestimated. Ideally, the entirety of the research area would have the value '100'. The model, however, is not 100% accurate and therefore some areas have an overestimated or underestimated soil content value. The soil properties are predicted too high in areas that are red, which are the very north part of Noord-Holland, west from Antwerp, and the area east of Limburg. Areas in which the model estimated the soil properties to be too low are the area around Amsterdam, the southwest of Friesland, and the area around Groningen.

In the sand value histogram, it can be seen that the peaks in the graph reside around a sand content of 20 and 85, meaning that most soils have a sand content that is either relatively low, or very high. But given that no histogram bar passes the 2 million mark, it can also be concluded that sand content values are more evenly distributed when compared to silt and clay values. Moreover, the distribution of sand has the highest standard deviation out of these three soil property's distributions, meaning that the range in which sand content occurs is the largest amongst the three soil properties modelled above. From the histogram plotted using silt content values, it can be derived that most soils have a silt content lower than 50, and that a significant number of soils in the research area have even lower silt content values. The clay content histogram displays a distribution in which most observations have a clay content lower than 50 grams per 100 grams. Moreover, most clay content values reside between 3 and 30. There are however soils with a high clay content, visible in the clay content map. The histogram displaying the sum counts of the three soil properties displays a graph that looks normally distributed with a mean of 99.6 and a standard deviation of 4.95. This means that on average, the model predicts sand, silt and clay relatively well. According to the normal distribution, 68% of the sum values fall within the range of 94.7 and 104.5. The minimum and maximum, however, are both very low and very high respectively, meaning that there are some significant outliers still present in the model. Since they rarely occur it can be assumed that the model predicts the total sand, silt and clay content fairly accurately, especially given the fact that each soil property map is generated independently from each other map.

4.4.2 CEC, CF and OC maps

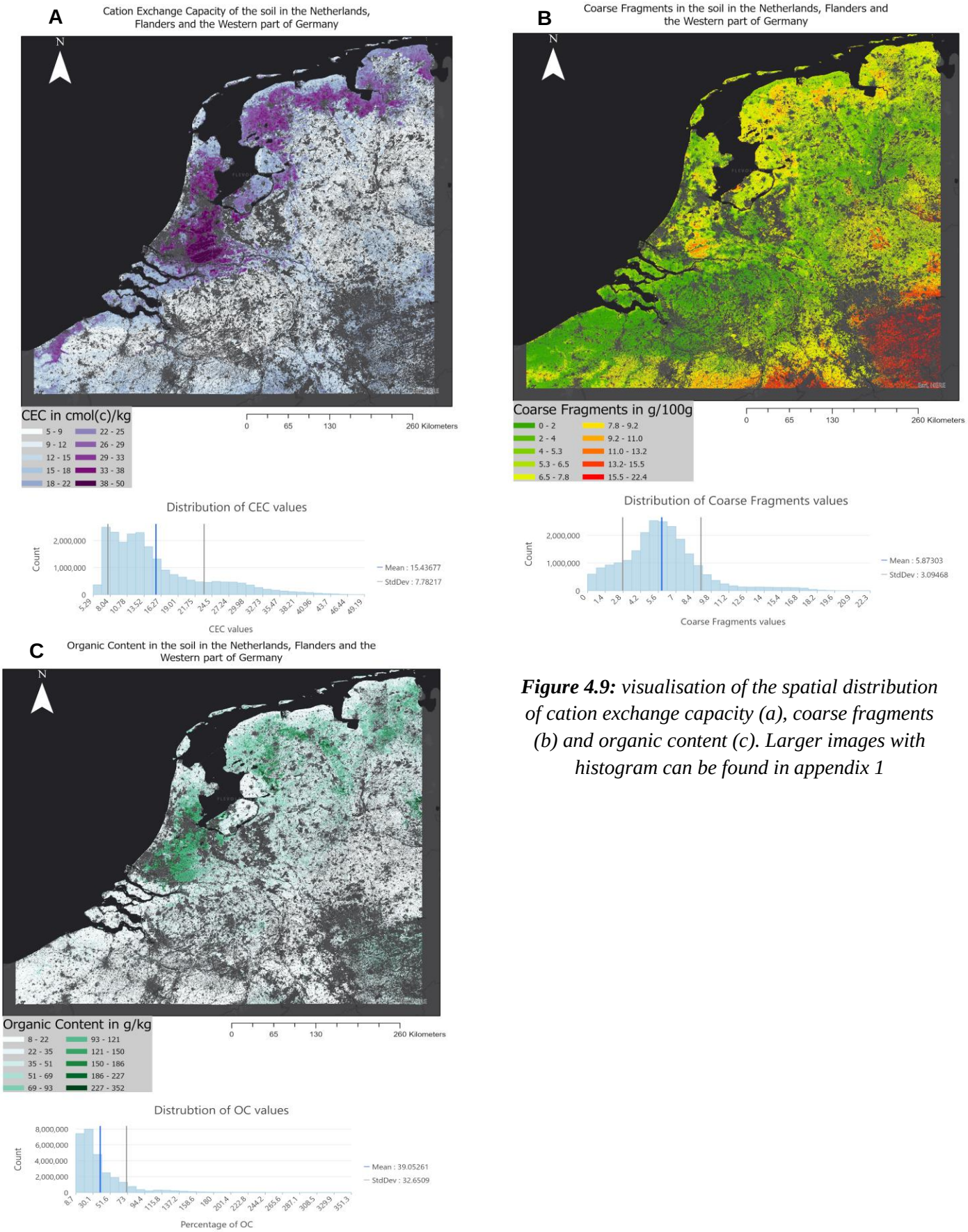


Figure 4.9: *visualisation of the spatial distribution of cation exchange capacity (a), coarse fragments (b) and organic content (c). Larger images with histogram can be found in appendix 1*

The cation exchange capacity map in figure 4.9a displays similar patterns to the clay content map. This similarity is however not coincidental since the cation exchange capacity is highly influenced by clay. The only notable difference between the two maps is the higher concentration of cation exchange capacity in the areas north and south-east of the Noordoostpolder. The map result for the coarse fragments in the soil (figure 4.8b) displays a clear distinction between high and low elevation areas. It must be noted that the coarse fragments values in this map range from 0 to 20, meaning that the area in the very south-east part of the research area has coarse fragments values of around 19 to 20. The value range is another topic of interest, as the predictions state that the coarse fragments values in the research area do not go over 21, which is relatively low considering a soil needs a coarse fragments value of 50 or higher to be labelled as gravel in the USCS classification. This means that while the coarse fragments values in the south-eastern part of the research area seem to be high, they are still relatively low on a global scale. Last is the map for organic content in figure 4.9c. At first glance, it is visible that the highest values for organic content are found in Het Groene Hart and the area around and north of Amsterdam. In addition, some areas in the north of the Netherlands have high concentrations of organic content as well, most notably the area north-east of the Noordoostpolder, and the Hondsrug, which is located south of Groningen. Organic content has high concentrations in certain spots, while having very low values in the other regions. This pattern is also visible in the map. Apart from the areas mentioned before, the organic content values are relatively low, meaning that organic content is either present in very high concentrations, or is almost not present at all.

From the histogram plotted beneath the cation exchange capacity map it is visible that most values reside on the low end of the scale, with most values falling within the range of 5 and 20. There is still a fair number of points higher than 20, as visible in the map of the cation exchange capacity. For the distribution of coarse fragments values, all values are relatively low. Again, this can be explained by the research area. The mean value is 4.8 with a standard deviation of 3, meaning that most soils have a relatively comparable coarse fragments content. The higher values for organic content that are visible in the histogram are mostly located in the southwest of the research area where the elevation is higher. The organic content value histogram shows that the distribution is very skewed. While most organic content values are below 50, there are some outliers that reach up to 344 g/kg. The high outliers also cause the largest standard deviation out of all soil properties.

4.5 Test Results

After the validation phase is completed and an average map results is produced from each k-fold-cross-validation, the test phase commences. This test phase incorporates the results produced in the earlier phases in combination with the testing data which was separated at the very beginning of the workflow. This data is therefore not used in any way to train or validate the model, thereby providing an independent source of testing data. Below, a table is presented containing the testing statistics for each soil characteristic. As stated in the methods, the test phase was performed on four different rasters. These rasters were constructed through the aforementioned methods: mean, median, standard deviation and restructured.

4.5.1 Test statistics per method

The tables below present the testing statistics generated from the four methods of creating a validated raster. In the tables, green coloured cells indicate the best statistic across the four methods used for

creating an output, while red coloured cells indicate the worst scoring statistic per soil characteristic. When multiple scores are both the best or worst scoring, then both are coloured green or red respectively. Across the four methods, the R-squared, RMSE and ME are quite similar to the validation statistics. The four methods are presented to assess what method of creating an output raster is the most accurate when compared to the testing data.

First is the method that calculates the mean from the 10 map outputs generated by the k-fold-cross-validation. From the table it is visible that for clay, the highest R-squared value has been reached, albeit shared with other methods. The same applies for the value of the mean error statistic of the clay content prediction, which is also reached by the median and standard deviation methods. While the lowest R-squared has been attained for sand using the mean method, it is only 0.001 lower than the highest score, making this an insignificant difference.

Table 4.3: *test statistics for mean method*

Soil Property	**R-Squared**	**RMSE**	**ME**
Sand	0.483	13.20	1.28
Silt	0.569	10.09	-0.78
Clay	0.685	4.13	-0.35
Coarse Fragments	0.173	3.90	-0.45
Organic Content	0.430	57.74	6.54
Cation exchange capacity	0.398	7.40	-1.21

The median method calculated the median cell values from the 10 validation rasters produced in the k-fold-cross validation. This method reaped the best R-squared statistic for organic content when compared to the other methods. It scored 0.013 higher than the second-best method, the standard deviation method, and scored 0.062 higher than the worst score, which was attained by the restructured method. For the same soil property, organic content, this method also scored the best Root Mean Squared Error but did not reap the best Mean Error for organic content. The high scores for organic content can likely be accounted to the method in which the validated map was produced. Since it is expected that the predictions for organic content vary greatly due to the wide range and skewed distribution of organic content values, the median method negates the issue where 1 of 10 values differs greatly from the other predicted values.

Table 4.4: *test statistics for median method*

Soil Property	**R-Squared**	**RMSE**	**ME**
Sand	0.484	13.20	1.29
Silt	0.567	10.13	-0.77
Clay	0.684	4.13	-0.35
Coarse Fragments	0.174	3.91	-0.46
Organic Content	0.445	57.31	6.78
Cation exchange capacity	0.399	7.41	-1.24

The standard deviation method eliminates outliers before calculating the mean for the remaining cell values. This method reaps the best statistics for the statistics describing cation exchange capacity. The R-squared, RMSE and ME are the highest using this method. The same applies to the clay content prediction, for which this method also scored the (shared) highest scores across the four methods. For

silt however, the standard deviation method scores both the worst R-squared and RMSE, while scoring the shared best mean error. The clay scores attained by the standard deviation method are also the shared highest across all used methods.

Table 4.5: *test statistics for standard deviation method*

Soil Property	R-Squared	RMSE	ME
Sand	0.484	13.17	1.25
Silt	0.567	10.14	-0.75
Clay	0.685	4.12	-0.35
Coarse Fragments	0.173	3.92	-0.45
Organic Content	0.432	57.81	6.80
Cation exchange capacity	0.400	7.38	-1.21

The restructured method does not calculate the cell values using a statistic, rather, it uses all of the training data to generate a map output. Using this method, the statistic scores are either the highest or the lowest scoring amongst methods. This is visible in the table, where all cells are either green or red. The most notable score in the table is the R-squared from the coarse fragments, which scores significantly higher than the other methods. Where the mean, median and standard deviation method scored 0.173 and 0.174, the reconstructed method generates an R-squared of 0.381, which is 0.208 higher. This means that using the reconstructed method, the created map output explains 20.8 percent more variation in coarse fragments. Contrastingly, the RMSE for coarse fragments is the worst-scoring statistic across the four methods. The sand content scores the highest statistics out of the four methods in the R-squared, RMSE and ME, albeit that the R-squared is equal to the median and standard deviation methods. Another high scoring statistic is the R-squared and RMSE for silt of which both are significantly better than the other methods. The R-squared is 0.012 higher and the RMSE is 0.16 lower and therefore better. Furthermore, the R-squared for the organic content is the lowest scoring out of all the four methods, being significantly lower than the best scoring (median method, 0.445) with a difference of 0.062. The RMSE for the organic content is also worse than the other methods with the difference being 1.83. All statistics for the cation exchange capacity are the worst scoring compared to the other methods.

Table 4.6: *test statistics for reconstructed method*

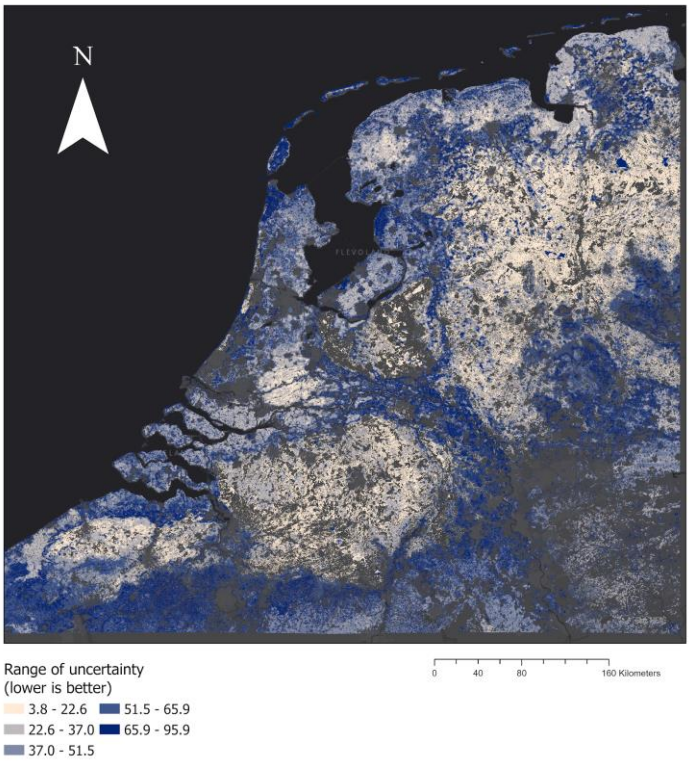
Soil Property	R-Squared	RMSE	ME
Sand	0.484	13.12	1.22
Silt	0.579	9.88	-0.90
Clay	0.685	4.16	-0.36
Coarse Fragments	0.381	4.27	-0.38
Organic Content	0.383	59.14	6.43
Cation exchange capacity	0.386	7.55	-1.44

4.6 Model Uncertainty

When predicting unknown variables such as soil properties, there is a degree of uncertainty present in this prediction. A feature of the ArcGIS forest-based regression tool is the option to create uncertainty rasters. One depicts the lower bound of the 90% confidence interval, and the other shows the upper

bound. Subtracting the lower from the upper bound therefore creates the range in which 90% of the trees in the Random Forest resulted in. If this range is large, the model is more uncertain than when this range is low. For each soil property, all the training data was used to generate such a map where 10% was selected for testing. The results are shown below. Layouts with histogram are shown in appendix 6. It should be noted that a high level of uncertainty does not mean that the final predicted outcome was wrong, it only means that the Random Forest model produced a wide range of values for a single pixel.

Uncertainty map for the prediction of Organic Content



Uncertainty map for the prediction of Silt Content

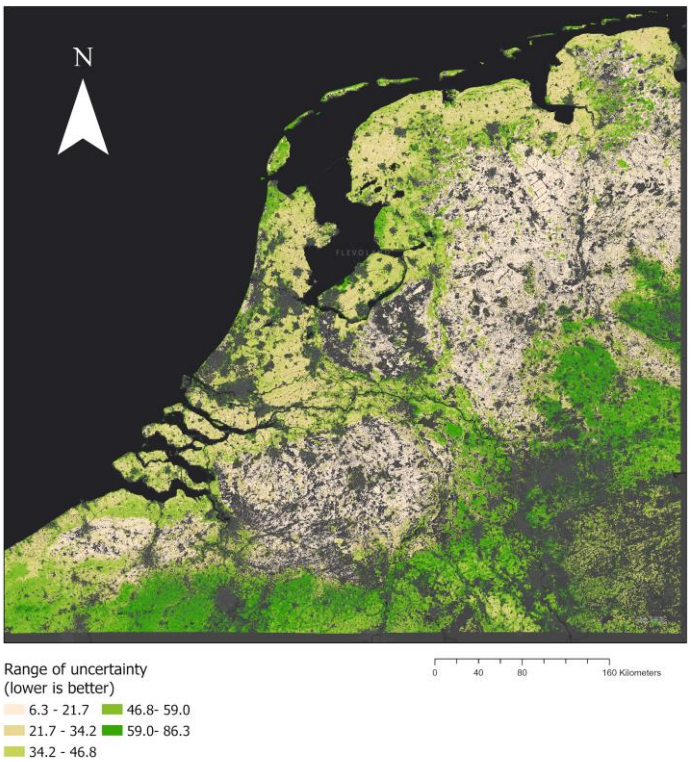


Figure 4.10: *uncertainty maps for each soil characteristic depicting the difference between the upper and lower bound of the 90% confidence interval where lower is better.*

From the maps it is visible that each of the soil property predictions has some geographical peaks in uncertainty. There are certain areas that score a low degree of uncertainty in each map. These areas are the same areas in which sand is highly present. For the soil properties: sand, silt and clay, the model has a high level of uncertainty in predicting the values in the Dutch river area. The soil property for which the 90% confidence interval is the highest is sand content, which can be related back to the high RMSE score for sand content prediction in the test statistics (table 4.4). The largest range of the 90% confidence interval is found in the Dutch river area, the south of Flanders, the Noordoostpolder and the northmost part of Noord-Holland. Areas where the predicted values for sand are high score low in model uncertainty. The opposite applies for the predicted silt content, where uncertainty is the highest in areas where the predicted values are also high. The uncertainty of the model for predicting clay content is the highest in Het Groene Hart where the predicted clay contents were also the highest in the research area. Cation exchange capacity and organic content show similar patterns in the model uncertainty. The highest uncertainty for both soil properties appears in Het Groene Hart, the Noordoostpolder and the Hondsrug. These are all regions containing large areas of peaty soils. Lastly, the uncertainty for coarse fragments is the highest in the southeast of the research area, again where the predictions for coarse fragments values is the highest.

4.7 USCS Classification

When all needed soil properties have been calculated, the final result can be classified according to the USCS classification. First, a general overview of the research area will be provided, and some resemblances and anomalies are outlined. Next, areas representing multiple different soil categories are chosen in order to display whether the algorithm has predicted every soil property accurately.

4.7.1 Research Area comparison

First, a global overview of the USCS classification is shown to indicate the general patterns in the predicted soil types. The generated result is compared to the reclassed national soil maps of the Netherlands and Flanders.

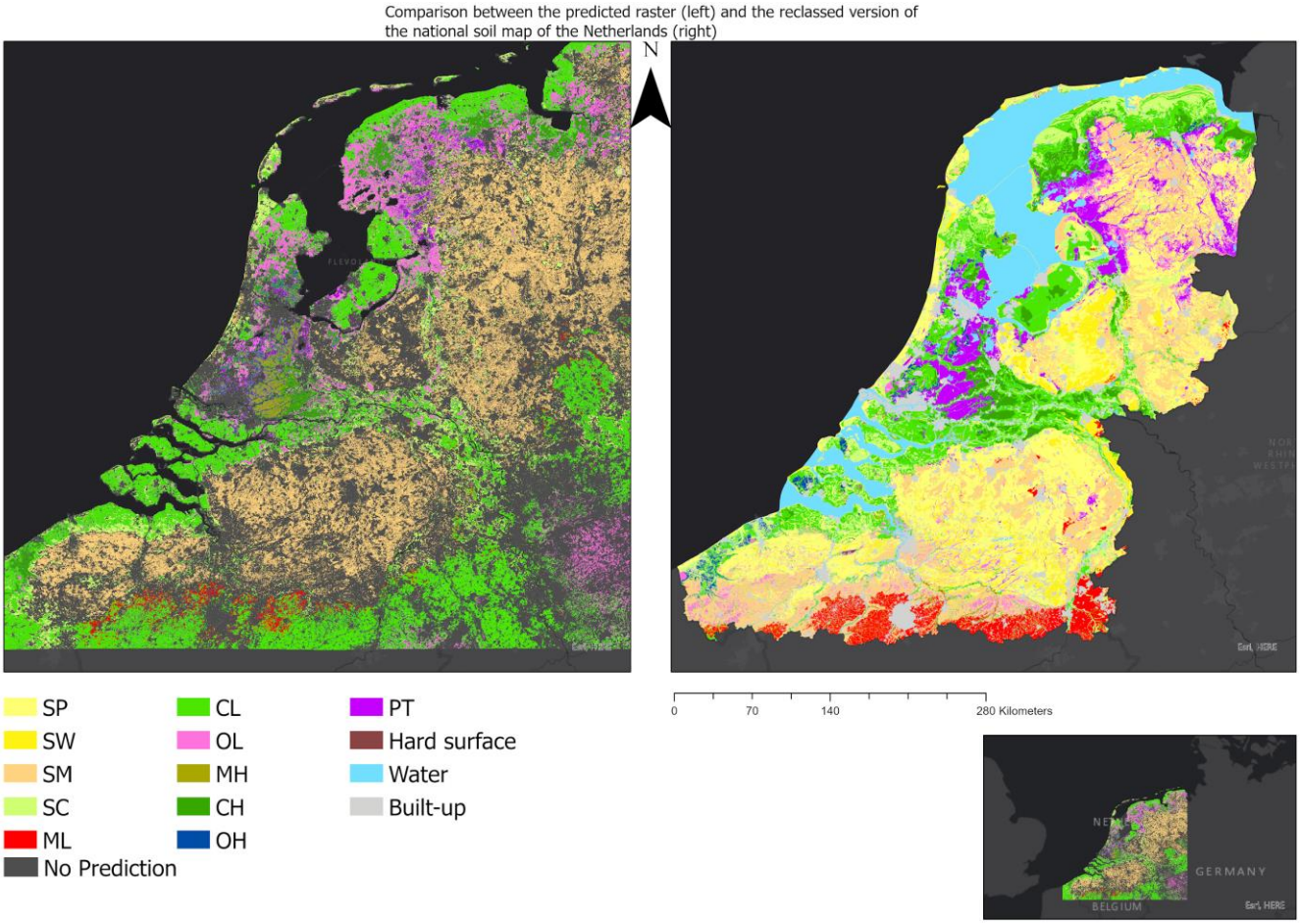


Figure 4.11: *comparison between the predicted soil types and the reclassed version of the national soil map of the Netherlands*

At first glance, the two maps share some resemblance when it comes to the regional patterns. For instance, there are great similarities in the sandy areas, which are yellow and beige in the maps. They appear most in the Dutch provinces of Overijssel, Drenthe and Noord-Brabant. Similarly, the predicted clayey area in the river area of the Netherlands is also quite similar to the national soil map. When it comes to peaty areas, some dissimilarities appear. In the predicted map, Het Groene Hart is shown as mostly MH and CH, meaning that the liquid limit is high, with both sandy and silty soils. In the reclassed version of the national soil map of the Netherlands, it is visible that the majority of Het Groene Hart consists of peaty soils and therefore differs from the predicted soil map. The areas around Het Groene Hart however are correctly predicted as peat. A similar anomaly occurs in the area around the Noordoostpolder, which consists of mostly peaty soils. However, the predicted map mainly predicts OL, meaning organic soils with a low plasticity. Another area that shows some anomalies is the southern part of Flanders and Limburg, where the national soil maps are coloured bright red, meaning that the region consists of silty soils with low plasticity. In the predicted soil map however, only a small part is actually coloured red, while the rest is defined as clayey soils with low plasticity. Overall, the predicted soil map presents the general patterns also visible in the reclassed national soil maps. When it comes to specific regions, some anomalies appear at a small scale. To further investigate these anomalies, two regions in the research area are selected for a zoomed-in assessment of the accuracy of the classification.

4.7.2 Hondsrug Comparison

The first of the selected areas is the Hondsrug in the province of Groningen. This area was selected as it houses many different soil types present in the USCS. The main point of interest is the Hondsrug itself, which consists mostly of peaty areas. Additionally, its sharp edge is relevant, since it indicates how well the Random Forest model has predicted the edges of this peaty area. The comparison is visible below.

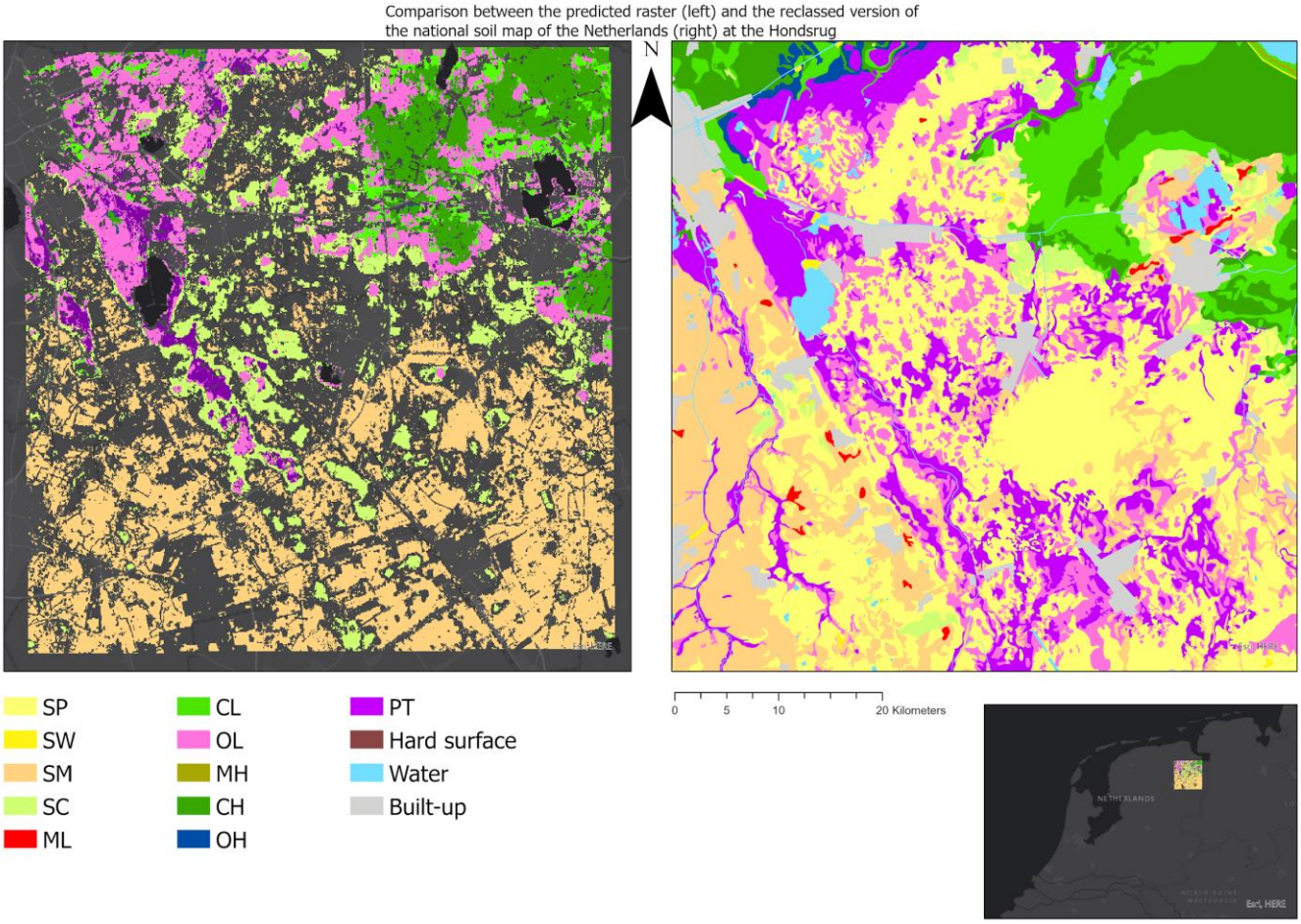


Figure 4.12: *comparison between the predicted soil types and the reclassed version of the national soil map of the Netherlands in the Hondsrug area*

At first glance, some general patterns in soil type distribution are visible in both the predicted map and the Dutch national soil map. In both maps, the sharp edge of the Hondsrug itself is visible. In the predicted map, this area is coloured purple and pink meaning that it predicted peaty soils and organic soils with low plasticity. While the OL soils do appear in this area, the Hondsrug consists of mostly peaty soils. The clayey area in the north-east of the selected region is identified fairly accurately, and its general outlines are also well defined. Still, the predicted map has an overrepresentation of CH soils, where CL should be present. Furthermore, the centre of the predicted map mostly displays clayey sands, where a patchwork of peaty soils and purely sandy soils should be present. The sand category is therefore predicted well, but the 'subcategory' is not predicted correctly. Similar to the entire research area, patterns of soil types are predicted fairly accurate in the Hondsrug area, while fine details are misinterpreted.

4.7.3 Eemvallei Comparison

The Eemvallei is chosen as an area for comparison as it also contains defined areas of peat and has a wide range of soil classes. Especially the peaty area in the south of this region is a point of interest, since it is relatively small and therefore might be difficult to detect.

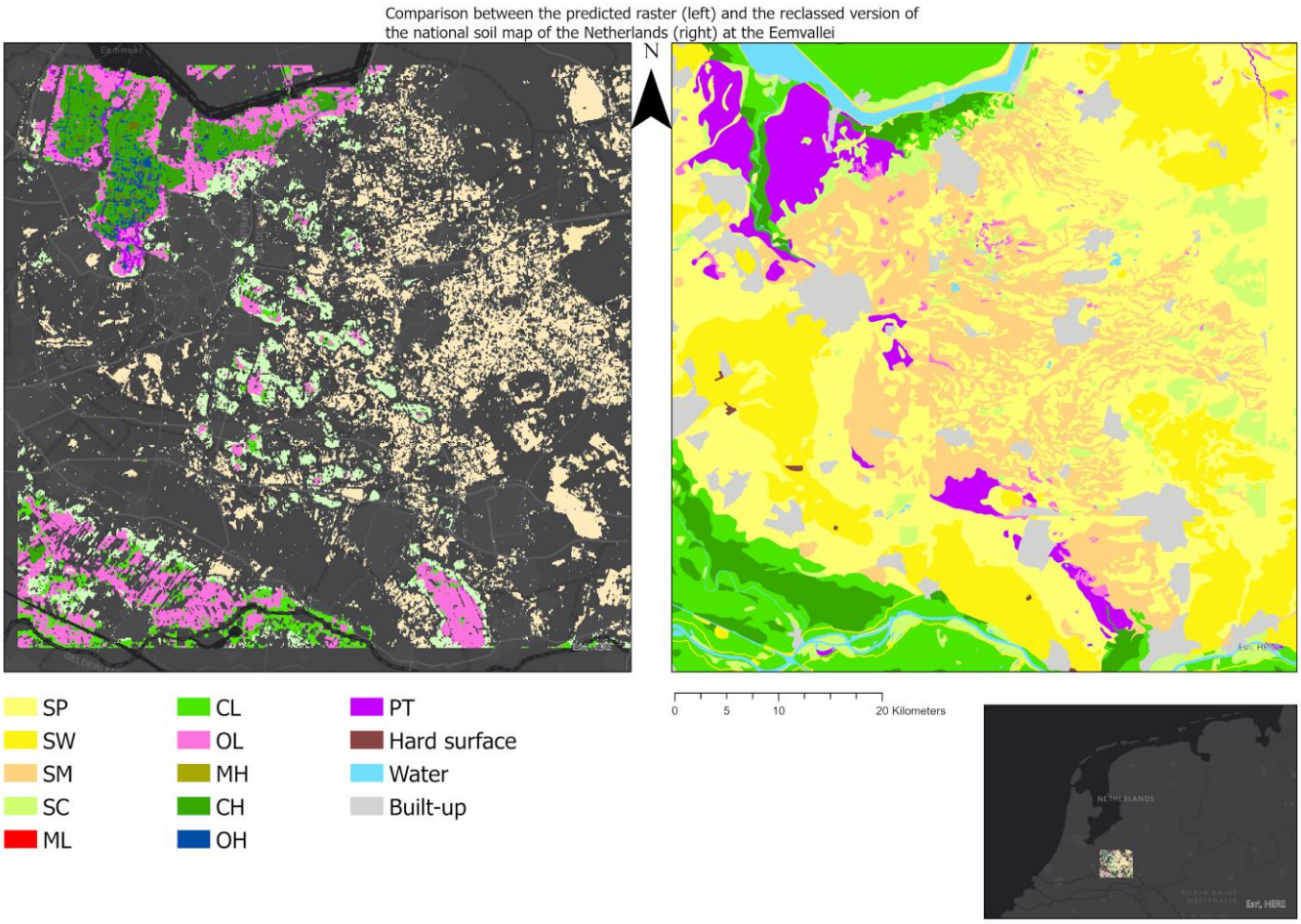


Figure 4.13: *comparison between the predicted soil types and the reclassed version of the national soil map of the Netherlands in the Eemvallei area*

From the figure above it is visible that the selected area is mostly a sandy area, with clayey and organic soils in both the far north and the far south of the area. The middle area is predicted somewhat accurately. The main soil group is predicted correctly, which is sand. Nevertheless, the subgroups are not entirely accurate. The subgroups present in the area are SP, SW and SM according to the national soil map. In the predicted map however, only SM is predicted, depicting a mix of sand and silt. In the northwest of the selected area, a peaty area is visible in the national soil map. In the predicted map, this same area is displayed as mostly clayey soils with high plasticity and organic soils with low plasticity, thereby diverging from the soil map's representation of that area. A similar pattern is visible in the south of the region, where clayey soils with both high and low plasticity should be present. Contrastingly, the predictions indicate mostly organic soils with low plasticity with small areas of clayey soils with low plasticity. Furthermore, the peaty area in the southeast of the region is identified as OL by the prediction map.

5 Discussion

In this discussion section, the research is reviewed by stating its key findings, highlighting their advantages followed by an evaluation of the sub questions posed in section 1. Next, this study is compared to previously conducted academic research to assess whether the results produced by this research are either similar or deviate from the results produced by academics. Even though some key findings were made in this study, its results are limited by some factors. These factors are explained and brought into perspective to the research performed. Next, specific findings that are deemed as surprising or inconclusive are discussed, and their assumed nature is explained. The discussion section ends with a list of possible topics that can be explored in future research. These are mainly topics which fell outside the initial definition of the scope of this research, or topics that were discovered during the execution of the study, but that were unattainable given the time and resources of this thesis.

5.1 Key Findings

In this thesis, some key findings were made that have relevance on both a practical and academic level. These findings are further elaborated on and their importance is discussed. One key finding was that an R-squared statistic of ~0.6 could be achieved for sand, silt and clay by only using open-source data such as the WOSIS and the LUCAS datasets and the open-source predictive variable used in the study. A test was performed to assess whether enough ground truth points were provided to the algorithm, of which the result was that there were in fact enough points, since the R-squared from using all the data was not significantly higher than the R-squared from using only 80 percent of the data (Byrd et al., 2012). This was proven true for each soil property. Subsequently, it can be stated that a similar approach can be used to map soil characteristics across Europe, which is ultimately the goal for the automated terrain analysis of the Dutch ministry of defence.

Similar to the relatively favourable R-squared statistics is the spatial resolution at which these predictions are made. While McBratney and Pringle (1999) state that autocorrelations in soil characteristics occur at 300 meters and less, the objective for this thesis was to create a raster containing the USCS classification of soils at a 30 meters spatial resolution composed of 6 predicted soil characteristic rasters in combination with the liquid limit and the plasticity index. The spatial resolution for this research was based on the required accuracy for the automated terrain analysis. The 30-meter pixel size is essential for the assessment of the smallest military unit: the platoon. This ambition for accuracy was maintained in the sense that no concessions were made regarding spatial resolution. Moreover, 30 meters is also the best possible accuracy when only using open-source data. While all rasters can be resampled into finer resolution maps, i.e., 10 meters, the choice was made to only perform the resampling on a select number of rasters to limit the loss of trustworthiness of information that occurs when a raster map is disaggregated. Even though some concessions were made in regard to the reliability of certain predictor maps, the accuracy that was attained in this study is remarkable given the fact that only open-source datasets were used.

The k-fold-cross validation used in this study was innovative to the point that it reaped a slightly lower R-squared than the built-in validation method in ArcGIS. This was likely due to the fact that the validation method in ArcGIS chooses its validation subset randomly from the ground truth dataset. While this random selection falls in line with the stochastic nature of the Random Forest algorithm, it is likely that the randomly selected subset does not include large outliers, which are essential for certain soil properties such as organic content. Only when a soil has a large concentration of organic content,

it can be classified as peat, which is a crucial soil class for the assessment of vehicle accessibility. The k-fold-cross validation ensures that 10 subsets are generated each containing a representative value range of data points that is geographically dispersed as well. Moreover, the fact that 10 cross-validations are performed, from which the average R-squared, RMSE, and ME are calculated, means that outliers in the statistics are nullified. From each cross validation, a map result is produced. These measures caused that the R-squared statistic was slightly lower than the one produced in the ArcGIS based tool but ensured consistent statistics when the validation was performed multiple times.

5.2 Research Questions

In section 2 of this thesis, several sub questions were introduced to divide the main research question up into manageable parts that can be answered individually. When they are combined, the main research question can be answered fully. This subsection will go over each of the sub questions and answer them to the fullest extent possible.

The sub question "*With what accuracy can the Random Forest algorithm predict distribution patterns of soil properties?"* can be answered indirectly by referring back to the map in figure 5.8. In the bottom-right figure, the summation of sand, silt and clay values is presented, where a total value of 100 is expected. While there are some outliers in this summation method, the overall accuracy in predicting the distribution of soil properties is relatively good. The mean of the sum sand, silt and clay is 99.6 with a standard deviation of 4.95, meaning that 68 percent of the values reside between 94.7 and 104.5.

The next sub question assesses the level of accuracy of the individual soil property predictions and is as follows: "*To what degree does the algorithm predict the soil property values correctly?"*. This question can be answered by the statistics generated by both the validation and the testing phase. The achieved R-squared statistics differ among the individual soil statistics with clay content having the best scoring statistic, followed by sand and silt. To answer the question: the prediction of soil property values is relatively good, but leaves space for improvement, since the RMSE statistics for the prediction of each soil property

As stated earlier, a key finding in this study was the linear regression formula to calculate the Atterberg limits using values of clay content and cation exchange capacity. This finding relates to the sub question: "*To what extent can the soil property maps be combined to calculate the liquid limit and plasticity index?"*. The regression analysis resulted in an R-squared statistic of 0.842 for the calculation of the liquid limit and an R-squared of 0.895 for the prediction of the plasticity index. The constructed formulas thereby exceeded the accuracy of prior research. For this reason, it can be stated that the calculation of the Atterberg limits is done with an exceptional degree of accuracy and significance.

Before any predictions of soil properties were made, a selecting process was conducted to assess what explanatory variables reaped the highest importance in predicting each soil characteristic. Using the results from this analysis in combination with the graphs created in section 5.3, the sub question "*What predictors are most influential in predicting soil properties?"* can be answered. It is evident that hydrological soil properties are most influential in predicting the six soil properties used in this research. Overall, this category of predictors reaped over 50 percent of the total importance in the prediction of all soil properties combined. This finding, that hydrological properties were most influential in predicting soil properties, fully applies to the prediction of sand, silt, and clay content. When looking

at the soil properties individually, coarse fragments, cation exchange capacity and organic content show diverging patterns. The most important predictors for coarse fragments in the soil is the digital elevation model, its derived properties, and several bands from the Sentinel 2 satellite platform. For the prediction of both cation exchange capacity and organic content, the nitrogen content of soils proved most influential. Furthermore, the soil organic carbon stock was important in explaining the variation of organic content of soils.

The role that the input data plays in the model is assessed by the sub question: "*In what way does the used data influence the model's outcome?"*. This question can be answered for both the dependent variables and the explanatory variables. For the dependent variables, two separate sources of soil samples are used to train, validate, and test the model. This reduces the dominance of one dataset over the other and therefore hopefully reduces the chance of a sampling bias. Both datasets originate from respectable and verified sources and are therefore expected to have accurate measurements. When reviewing the explanatory variables, there is dominance of some data source over the others. The two most influential datasets are the SoilGrids dataset and EU Hydro dataset. These two datasets supplied the predictors that had the most combined importance in predicting the six soil properties. While these sources explained the most variation in the six soil properties, the other data sources such as the Sentinel 2 platform and the Digital Elevation Model also explained a fair amount of variation. Were the EU Hydro and SoilGrids rasters left out, the result would likely be significantly different.

The sub question "*What level of uncertainty is present when using the Random Forest algorithm to predict soil properties?"* can be answered by referring back to the validation statistics, test statistics and the uncertainty maps. The RMSE differed across the six soil properties but indicated an overall slight level of uncertainty in the model's predictions. The highest and therefore worst scoring RMSE was observed with sand, which is also reflected in the map depicting model uncertainty for sand. Clay content and coarse fragments prediction resulted in the lowest and therefore best scoring RMSE statistic. While this observation is confirmed by the uncertainty map for coarse fragments prediction having only a small area with relatively high uncertainty, the clay content uncertainty reflects a different scene, as there are significantly large areas with high uncertainty. However, it is important to note that a high uncertainty does not mean that the prediction of the values is incorrect.

Using the results of this study in combination with the answers to the sub questions from this section, the main research question "*To what extent can soil be classified for the use of a terrain analysis through the Random Forest machine learning method using open-source data?*" can be answered. When classifying soil, one needs to be certain of the predicted soil values and one needs to have all properties needed to perform the classification. In this study, a fairly good regression statistic was achieved for most soil characteristics, and using these predicted values, two additional soil properties were calculated. These were then combined into the USCS with varying results. While exact soil types were not located where they were expected to be, the general pattern in the final soil map showed resemblance to the real-world situation.

5.3 Other research

In both the introduction and the theoretical framework, research was cited that also performed similar studies to the one performed in this thesis. To assess whether the conducted research holds up to the

other studies cited earlier, or that the results exceed the other literature, a comparison is made between a selection of academic sources and the findings of this study.

Da Silva Chagas et al. (2016) found that the Random Forest algorithm produced a better result than multiple linear regression when predicting sand, silt, and clay. They reached an R-squared of 0.63 and 0.56 for sand and clay respectively while using the Random Forest algorithm, meaning that the percentage of variation explained was slightly lower than the one achieved in this study. This slightly lower outcome can be accounted to the fact that no water related soil properties were used in the study performed by Da Silva Chagas et al. (2016). These soil properties were proven to be the most important in predicting sand, silt and clay content, and would likely have improved the percentage of variation explained in the mentioned research. Additionally, the inclusion of elevation data increased the accuracy of the predictions made in this study, and likely would have a similar effect on studies done elsewhere in the world. When looking at soil characteristics individually, other studies (Poppiel et al., 2019) also revealed that clay reaps the highest correlation statistics when compared to the other soil characteristics (Nussbaum et al., 2018). From the results section it is visible that this research also presented clay content as the soil property with the highest R-squared statistic, confirming the findings from previously conducted research.

In many aspects, the results and methods of this thesis relate to those of Gambill et al. (2016) in which the USCS soil classification was predicted using the Random Forest algorithm. Additionally, many of the explanatory variables used acted as predictors in this study as well, most notably the water-related soil properties. Both in this study and the one performed by Gambill et al. (2016), the water-related properties yielded a high importance rating within the random forest algorithm. It differed from this research since the prediction using the Random Forest algorithm was used on areas where soil properties such as sand, silt, clay, and organic matter were already known and recorded. While the goal of Gambill et al. (2016) was to only perform a classification of soil of which most relevant soil properties are already known, this research can act as an expansion on it, as the USCS classification can now be performed on soils of which the required soil properties for the USCS are not known. This means that the area on which this method can be performed is limited to the coverage of fewer soil property maps. Furthermore, the Atterberg limits were not used in Gambill et al. (2016), rather, they used the organic content soil property as a substitute in the Random Forest algorithm. They therefore used an indirect measure for the two highly influential soil properties in the USCS: the liquid limit and plasticity index. Using the linear regression analysis performed in this research, a more solid and accurate measure of the liquid limit and plasticity index could have been used.

The linear regression analysis that was performed to generate a formula to calculate the liquid limit and plasticity index was derived from previous research but expanded and improved on it. De La Rosa (1979), Mbagwa and Abeh (1998), and Seybold et al. (2008) each performed their take on a regression analysis to calculate the liquid limit and plasticity index. Seybold et al. (2008) used around 10.000 points to perform the regression analysis while the others used merely a fraction of this number. For the liquid limit regression, Seybold et al., (2008) found that clay content and CEC explained 81 percent of the variation in liquid limit. The regression analysis performed in this research found an R-squared of 0.842 by using a sample size of more than five times larger than Seybold et al. (2008). The R-squared statistic indicates that 84 percent of the variation in liquid limit was explained by clay content and CEC. Additionally, for the plasticity index Seybold et al. (2008) found that 71 percent of the variation in the liquid limit was explained by clay and CEC, while the method and data used in this research found that 89.5 percent of the variation was explained by the same soil properties. While high and significant correlations were measured, the sample size can be increased still by adding more datasets from the

SSURGO database, since only one county per state was selected. In addition to the improvement on sample size and percentage of variation explained, the used method also allows the calculation of the Atterberg limits to be performed in an easier way and in all areas containing data on clay content and cation exchange capacity. Stanchi et al. (2015) used soil groups from the World Reference Base and their horizon types to predict the liquid limit and plasticity index of soils in the north of Italy. These two soil properties are more difficult to derive than clay content and cation exchange capacity. This means that the method performed in this research provides a more accessible and general way of calculating the Atterberg limits in more places in the world.

The spatial resolution used in this study is comparable to some other articles that also perform a prediction of soil characteristics using machine learning methods. Wang et al. (2020) predict soil salinisation in China at a 30-meter scale using a combination of Landsat 8 bands, Sentinel 1, digital elevation models and Landsat derived indices. Using these predictors, they achieve an R-squared statistic of 0.75 for their predictions, being slightly higher than the accuracy attained in this study. However, since soil salinity is an entirely different soil characteristic than the ones predicted in this study, no fair comparison can be made between the accuracy statistics. Similarly, Wiesmeier et al. (2011) used the Random Forest algorithm to predict soil organic matter stocks in a semi-arid region in China. The produced predicted map consisted of pixels with a size of 90 by 90 meters, with an explained variation of 42 to 65 percent. When comparing the results of this thesis to the study performed by Wiesmeier et al. (2011), this study explains a similar amount of variation in the dependent variable but at a scale that is 9 times finer than the scale used by Wiesmeier et al. (2011) (30m vs 90m). Again, the dependent variables differ between both studies which makes the comparison the two degrees of explained variation an arbitrary one, but it does provide an insight into the difference in spatial resolution.

In regard to variable importance, this study showed both similarities and divergence from previously conducted research. Zhou et al. (2020) indicate that elevation and topography are key influences in the formation of soil, which corresponds with both the results of their study, and the results from this study. The digital elevation model is therefore relatively influential, with the DEM and its derived properties accounting for around one fifth of the total importance in predicting all soil characteristics. In the prediction of coarse fragments in the soil, the DEM was even the most important, accounting for one fourth of the total importance. It is also worth noting that the DEM was at least somewhat important in the prediction of all other soil properties as well. Furthermore, Gholizadeh et al. (2018) and Vaudour et al. (2019) state that indices derived from multispectral data also reap notable importance in the prediction of some soil properties, with the highest importance in organic content. According to Zhou et al. (2020), this is to be expected since the organic content in the soil is highly influenced by the vegetation present, which the Sentinel 2 indices are all a measure of. The results concerning variable importance differ from other studies when it comes to Sentinel 1 SAR importance. While Yang and Guo (2019) use Sentinel 1 as a primary predictor for deriving soil properties in coastal wetlands, this study only gains one percent overall importance for Sentinel 1 imagery. In contrast to Sentinel 1 imagery, data on hydrological soil properties reaped the highest importance by far. This significant importance was also observed by Gambill et al. (2016) in attempting to crosswalk between the USDA classification and the USCS classification. In the study, they used available water storage, which is similar to the available water capacity used in this research.

5.4 Research Limitations

While this thesis provided an insight into the modelling of soil properties and the workings of the Random Forest algorithm, and produced strong and promising results, there are still aspects to the research that pose as limitations to certain parts of the study. While these limitations do not impair the validity and results of the research, they are things that need to be considered when placing this study in the broader scope of digital soil mapping.

The first consideration to be made is the source of the datasets depicting water wilting point, soil organic carbon stock and nitrogen. While this dataset does come out as important in the statistics produced by the Forest-based regression tool in ArcGIS, there are some connotations to this dataset. The most notable of these is the fact that they are derived from the SoilGrids database. Even though this database provides rasters containing fairly accurate information on soil properties, it cannot be considered as ground truth data, since the mentioned rasters are generated through the use of an algorithm. This makes the values present in the rasters more of an assumption, rather than a fact. Moreover, the raster for soil organic carbon stock was created using values from the organic content variable combined with the bulk density. These datasets are both present in the SoilGrids database. The caveat resides in the fact that for instance organic content is predicted using a variable that is constructed through the use of organic content, albeit from another source. On top of this, both organic content maps, the one created in this thesis and the one constructed by SoilGrids, are trained using the same data, being the WOSIS soil samples. Nonetheless, the Random Forest algorithm still produced a fairly accurate result for organic content, with the soil organic carbon stock only having 9 percent importance in creating the trees.

As mentioned in the research design, two rasters predicting the permanent wilting point are added originating from the EU Hydro dataset, and the SoilGrids platform. From the results it is visible that the WWP raster from SoilGrids outperformed the raster from EU Hydro when it came to variable importance in predicting the soil properties. While it could be possible that the predictor originating from SoilGrids simply outperforms the other predictor, another possibility exists. In this study, the WOSIS and LUCAS ground-truth datasets were merged to form a single, large dataset containing soil sample points. For most soil properties, the WOSIS database provided the majority of the training points and therefore has the most influence in training the Random Forest algorithm. The caveat here is that the WWP from SoilGrids is also constructed using the WOSIS soil samples, meaning that it is likely that the WWP is constructed using the values from the WOSIS samples.

In the process of collecting the explanatory variables needed for the Random Forest algorithm, the decision was made to also include rasters that originate from the EU Hydro and Soil grids datasets. Since interpolation assumes the fact of spatial autocorrelation, it forms a threat to the model's accuracy when the spatial autocorrelation cannot be one hundred percent confirmed. As stated in the methods section, the choice was made to interpolate these rasters to prevent them from causing a pixelated result. The interpolation was also performed since it was expected that the predictor variables containing information on water and soil organic carbon stock do have some degree of spatial autocorrelation. Nevertheless, the rasters from EU Hydro and Soil grids are intended to have a spatial resolution of 250 meters for a reason, likely because this was the finest resolution at which the two organisations could produce reliable predictions. It is therefore a point of discussion whether the rasters should be interpolated or not, especially because the values in these rasters are already predictions themselves. Relating to the matter of spatial resolution versus accuracy and reliability, neglecting the military relevance of a 30-meter pixel size, is the question whether this study should have used a coarser

resolution and if this coarser resolution would have improved the overall accuracy of the model. On the one hand, this question could be answered with yes, since a coarser resolution would allow for less interpolation or no interpolation whatsoever, meaning that the data with a 250-meter pixel size is not edited and no assumptions are made regarding the spatial autocorrelation of its soil properties. On the other hand, performing the interpolation could have been the best possible option for the predictive rasters. While both organisations that published the mentioned datasets used the 250-meter spatial resolution assumed this to be the resolution with the highest reliability, it can still be the case that interpolation detracts only a fraction off of this reliability, improving the spatial resolution of the rasters. Additionally, most other datasets used in the Random Forest algorithm are in fact already at a spatial resolution of 30 by 30 meters, meaning their original resolution would need to change, which would likely lead to a significant amount of information loss.

The calculations for LL and PI are completely dependent on the SSURGO regression analysis, CEC, and clay values, meaning that this data is derived from elsewhere on the globe than the research area for this thesis. While this is not a direct shortcoming, it is something to consider when applying elsewhere on the globe, albeit that ground truth data on the liquid limit and plasticity index is almost exclusively available in the United States, making it difficult to perform a more geographically representative regression analysis. Moreover, the USCS classification largely relies on the liquid limit soil property, meaning that errors in the calculation of this soil property can have a detrimental effect on the final classification map. Another thing to note, while not directly related to this study, is the sharp threshold of 50 when examining the liquid limit in order to determine whether a fine-grained soil has a high or low liquid limit.

Additionally, the settings for the classification script can be tweaked or adjusted to meet the requirements by the one that will be using the USCS map. The exact values for the peat class are not defined in the classification table and were therefore estimated. Giving these a lower threshold might result in better classification of peat areas. However, in doing so, areas that should be another soil type can be wrongfully classified as peaty soil. The official USCS contains ‘inconclusive’ classes where a combination of two soil types is provided as a class. While this is scientifically correct, in practice this can be troublesome as a soil is essentially classed as two soil types.

5.5 Surprising or inconclusive results

An inconclusive result was the low R-squared value for the coarse fragments prediction. The correlation statistic indicated that the used predictor values explained a poor amount of the variation in coarse fragments values. This result is seen as inconclusive since it is unclear whether the low amount of variation explained can be accounted to the lack of proper predictive rasters, or to the research area. The latter is the most likely, given the research area used in this study contained extremely small amounts of coarse fragments in the soil, providing the Random Forest algorithm with a small range of values to create trees with. To solve this, an option would be to perform the same method on an area that has a greater, more representative range of coarse fragments values. This way, the Random Forest algorithm can be trained with a wider range of values and is therefore likely more able to explain a higher percentage of variation in coarse fragments.

The inconclusive result mentioned above is related to a concept of the ‘black box’. The black box is a machine learning term that refers to the unknown processes happening inside the algorithm. Due to this

lack of insight into the core functions of the algorithm, the exact cause for inaccuracies in the final result cannot be traced to their source. Therefore, some results that appear to be anomalous are not easily solved by looking at where the anomaly took place, since no insight is possible within the Random Forest model used.

5.6 Future research

This thesis provided a broad insight into the workings of the Random Forest regression algorithm when applied to the prediction of several soil characteristics with the ultimate goal to classify the predictions into the USCS classification. This study paved way into uncharted territories of digital soil mapping. It used limited resources to provide a relatively accurate prediction of soils in areas where no ground truth data is available. During the execution of the research, the aim of the final product was to be used in a military terrain analysis, however, it can also be widely applied in a multitude of instances, for example agriculture, urban planning, and infrastructural engineering. These fields of study also require information on soils, while this type of data is not readily available in most cases. During the process of conducting the research and processing the results, several questions were exposed for which not enough time or resources were available to answer adequately. Moreover, there are also some matters discussed in this study that could be elaborated on further in future research.

Currently, the prediction of soil properties is only performed on a single, limited research area. While this research area contains many different soil types of the USCS classification, it still does not cover the entire spectrum of soil types. For instance, the number of coarse fragments in the soil does not reach a number higher than 21 in the current research area, while other parts of Europe have coarse fragments values twice as high. For this reason, the model should be trained, validated, and tested in other regions of Europe as well to gain an insight into how the predictors perform in areas with different ranges of values for each soil characteristic. Since radar and hyperspectral reflectance differ across latitudes, it could be useful to perform large scale predictions using a rectangular research area that is both wide and short. By using a scope of this size and measure, the advantages of a large research area are attained, while limiting the possible negative effects that occur when using a research area that spans over a large latitude. Another solution could be to use the climate system designed by Köppen. The main differentiators between climate classes in this system are precipitation and temperature. All datasets concerning hydrology, vegetation, moisture, and temperature are affected by these two parameters. Especially when average reflectance values from multiple years are collected it can be useful to differentiate between regions that have different amounts and periods of precipitation.

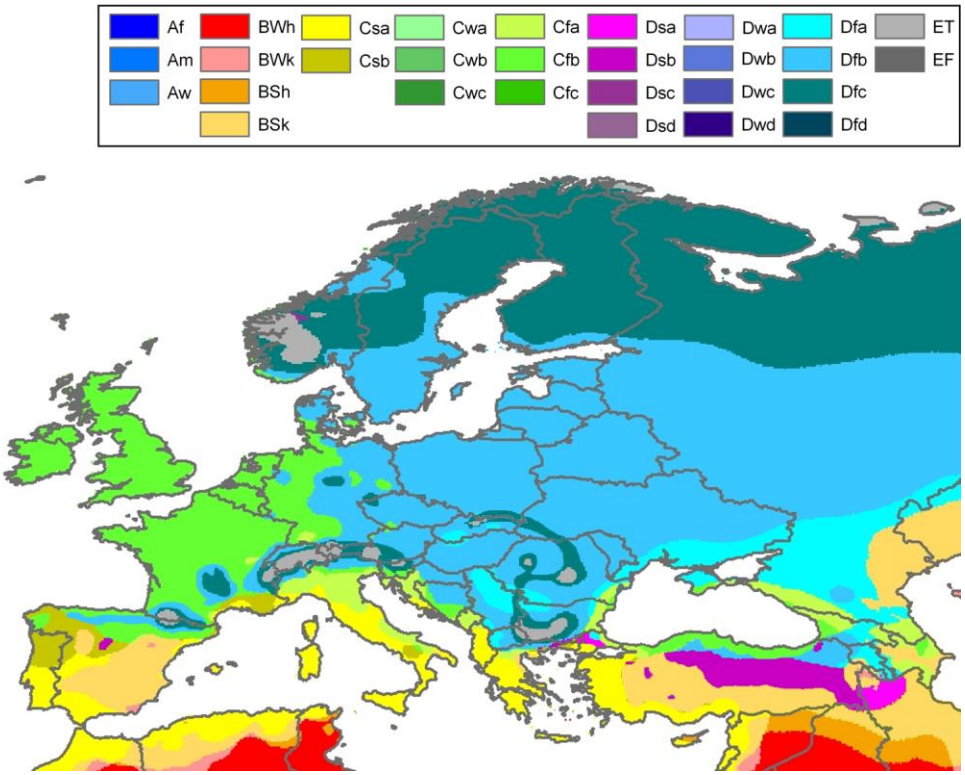


Figure 5.1: *map of Europe according to the Köppen climate system (from Peel et al., 2007)*

While the Random Forest regression algorithm in this thesis resulted in fairly accurate predictions of soil characteristics at a 30-meter spatial resolution, there might be other methods that allow for a more precise estimate of the contents and properties of a soil. Chang and Islam (2000) suggest using an Artificial Neural Network (ANN) to predict physical soil properties at a fine resolution using passive microwave remote sensing and moisture properties. They conclude that an Artificial Neural Network operates exceptionally well when used in tandem with spatio-temporal data. Another deep learning technique that is applied to predict soil characteristics is the Convolutional Neural Network. Ng et al. (2019) describe its capability to handle raw spectral data and its efficiency with high dimensional data. The Convolutional Neural Network approach might prove beneficial when attempting to predict larger areas across Europe, or when more explanatory variables are added. When a larger area is used as a research area, more data is fed into the model, and the Convolutional Neural Network deep learning method performs better when more data is used as input. This phenomenon is described by Ng et al. (2020) and illustrated in the graph below. This graph also depicts that given a lower number of samples, the machine learning (i.e. Random Forest) perform better than the deep learning (i.e. convolutional neural networks) techniques. This is why for this study, a machine learning technique was expected to perform better.

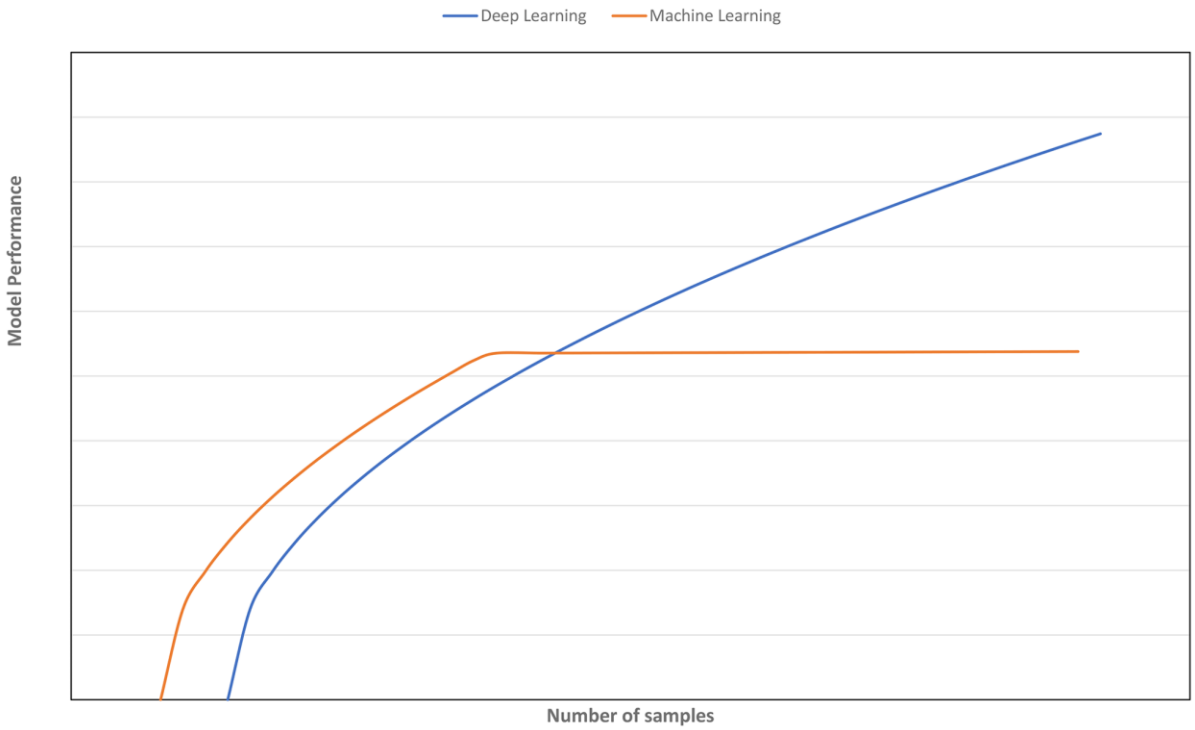


Figure 5.2: *comparison of machine learning and deep learning in their performance when fed with an imaginative number of samples (from Ng et al., 2020)*

In this study, no distinction was made between well graded and poorly graded soils, even though the well graded and poorly graded sands are two separate soil classes in the USCS. To assess whether a sandy soil is either well or poorly graded, one must calculate the uniformity coefficient. This calculates the ratio of the finest 60% size of the finest 10% grain size, and thereby determines whether a sandy soil is a mix of different grain sizes, or that it consists of equally sized grains. The former corresponds to the well graded sands, and the latter to poorly graded sands. Determining the exact uniformity coefficient can only be done accurately through laboratory tests, making it difficult to predict using remotely sensed imagery and derived properties. When a method could be developed that can accurately predict the uniformity coefficient, each class in the USCS can be determined.

During the examination of the results, it became evident that the range of predicted values is limited by the range of the observed values. The model does not expect any higher values than the ones observed in the soil samples. To remedy this, a qualitative investigation can be performed to search spots with the highest value for a soil property in the given research area. When these locations are identified, a self-performed soil sample can be taken to include the highest observed soil property value in the dataset. This allows the machine learning model to predict soil characteristic values across the entire existing range of values.

In this study, a soil map is created for an automated military terrain analysis. It is classified according to the USCS using predicted individual soil properties. Through this map, it is possible for military engineers to assess what vehicles can cross certain terrain according to their technical properties, which were tested through experimentation. A possibility for future research would be to, instead of using soil samples, use classified field tests. These field tests would be points where military vehicles are tested, and their accessibility is assessed. This could result in different accessibility classes, for example: accessible, difficult terrain, or not accessible. Using field tests with a wide geographical dispersion, a Random Forest classification can be performed where each cell in a raster would result in either 'accessible', ''difficult terrain', or 'inaccessible'. This could relieve the issues that were encountered in the process of turning soil samples into a soil map according to the USCS using Random Forest regression. The primary task for this research would be testing the vehicles at a multitude of points to gather enough samples to run the classification reliably. The result of such a Random Forest classification would look something like figure 5.3, where for example blue indicates accessible terrain, purple indicates difficult terrain, and yellow indicates inaccessible terrain.

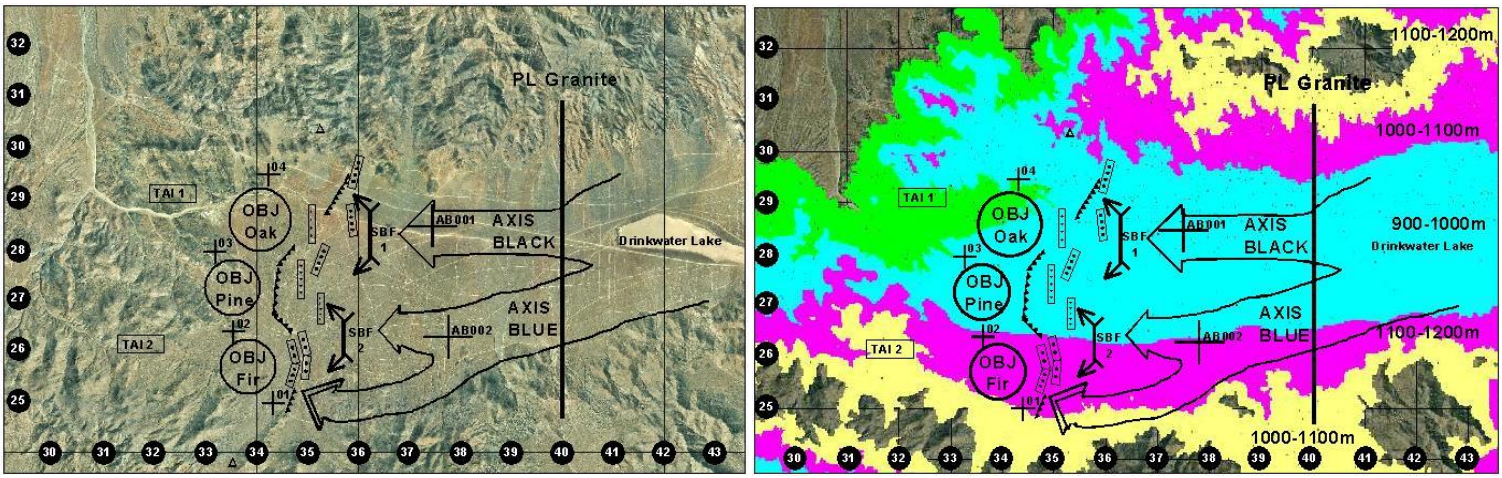


Figure 5.3: *an example of a terrain accessibility map (from GlobalSecurity, 2021)*

In this study, the variables that had the most importance in predicting were derived properties from two independent data suppliers. These variables were related to hydrology and chemical properties of the soil, which means that they are not related to reflectance from satellites. Additionally, the digital elevation model was also important in predicting soil properties. Given this information, it can be stated that the same Random Forest regression can be applied to regions where tree cover is present, as the most influential predictors are not obstructed by tree cover. Moreover, the generated soil map can be used in combination with tree cover and tree density maps to gain further insight into where certain vehicles can pass successfully. This way, the accessibility map can be extended to all types of soft terrain, instead of only bare soil and grass.

6 Conclusion

This thesis explored the potential of the Random Forest algorithm to predict soil properties at a 30-meter spatial resolution which are classified into the USCS soil classification system. This result can be applied in an automated terrain analysis to accelerate and refine decisions on military movements. The results of this research indicate that the Random Forest algorithm is a suitable machine learning method to assess the properties of a soil, given the right amount of training points and provided a varied and finetuned supply of explanatory variables. For each soil property the results were different. Not all explanatory variables reaped similar importance values for each soil property, and there were predictors that only were important for predicting a single property. While the distributions of the soil property values are considered to be very accurate, their specific predicted values tend to differ slightly from the observed values. Even though the accuracy is not one hundred percent, the goal of achieving a fair accuracy by only using open-source datasets has been achieved through the Random Forest algorithm. Moreover, from the findings presented in this research, a map displaying the soils according to the USCS classification can be constructed. This is done through a classification script where the classification is executed utilising the rules from the USCS. For this classification, more soil properties were required in addition to the six that were generated using the Random Forest algorithm. These were the liquid limit and the plasticity index. Since no soil samples in the research area contained data on these soil properties, a regression analysis was performed using data points using the USA-based SSURGO dataset. From this, a highly correlating, statistically significant regression formula was constructed using clay content and cation exchange capacity as variables. Combining the set of soil properties, a classification can be performed.

This study exemplified the strength of machine learning in predicting soil properties and made way for research following up on this topic. Digital soil mapping is expected to become increasingly more accurate over the coming years with the development of newer and more intelligent algorithms. Remote sensing will play a defining role in this advancement by delivering accurate and reliable imagery. The future direction for this field of study firstly revolves around exploring the possibilities of other methods that use artificial intelligence. Deep learning offers improved accuracy when more samples are used in the algorithm and is therefore potentially a good method to use when analysing larger areas than the one used in this study. Another key point to investigate is the effect of the research area on the results of the analysis. When performing the investigation, the difference in reflectance and hydrological properties should be taken into account. Therefore, an analysis of a larger research area is proposed either over a longer longitude, or according to the climate system of Köppen. Furthermore, additional soil properties are required to predict the full range of classes present in the USCS. Last, another method for determining accessible terrain can be performed by using the Random Forest classifier on sample points containing data on vehicle accessibility.

7. References

Abdel-Rahman, E. M., van den Berg, M., Way, M. J., & Ahmed, F. B. (2009, July). Hand-held spectrometry for estimating thrips (Fulmekiola serrata) incidence in sugarcane. In *2009 IEEE International Geoscience and Remote Sensing Symposium* (Vol. 4, pp. IV-268). IEEE.

Abdel-Rahman, E. M., Way, M., Ahmed, F., Ismail, R., & Adam, E. (2013). Estimation of thrips (Fulmekiola serrata Kobus) density in sugarcane using leaf-level hyperspectral data. *South African Journal of Plant and Soil*, *30*(2), 91-96.

Aber, J. D., & Melillo, J. M. (1980). Litter decomposition: measuring relative contributions of organic matter and nitrogen to forest soils. *Canadian Journal of Botany*, *58*(4), 416-421.

Agus, F., Hairiah, K., & Mulyani, A. (2010). *Measuring carbon stock in peat soils: practical guidelines*. World Agroforestry Centre.

Atterberg, A. (1911). Die plastizitat der Tone. *Intern mitt. boden.*, 4-37.

Ballabio, C., Panagos, P., & Monatanarella, L. (2016). Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma*, *261*, 110-123

Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, *13*(2), 111-122.

Batjes, N. H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., & de Jesus, J. M. (2017). WoSIS: providing standardised soil profile data for the world. *Earth System Science Data*, *9*(1), 1.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24-31.

Berg, B., & Staaf, H. (1981). Leaching, accumulation and release of nitrogen in decomposing forest litter. *Ecological Bulletins*, 163-178.

Beucher, A., Møller, A. B., & Greve, M. H. (2019). Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark. *Geoderma*, *352*, 351-359.

Blake, G. R., & Hartge, K. H. (1986). Bulk density. *Methods of soil analysis: Part 1 Physical and mineralogical methods*, *5*, 363-375.

Bookdown (2020). Regression and Classification Trees. *Machine Learning for Biostatistics*. [online] available at: https://bookdown.org/tpinto_home/Beyond-Additivity/regression-and-classification-trees.html

Bourdev, L., & Malik, J. (2009, September). Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 1365-1372). IEEE.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Briggs, L. J., & McLane, J. W. (1910). Moisture equivalent determinations and their application. *Agronomy journal*, *2*(1), 138-147.

Brouwer, C. (1985). Irrigation Water Management: Training Manual No. 1 - Introduction to Irrigation. *FAO land and water development division*

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards Jr, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, *239*, 68-83.

Byrd, R. H., Chin, G. M., Nocedal, J., & Wu, Y. (2012). Sample size selection in optimization methods for machine learning. *Mathematical programming*, *134*(1), 127-155.

Casagrande, A. (1948). Classification and identification of soils. *Transactions, Asce*, *113*, 901-991.

Cassel, D. K., & Nielsen, D. R. (1986). Field capacity and available water capacity. *Methods of Soil Analysis: Part 1 Physical and Mineralogical Methods*, *5*, 901-926.

Chaudhari, P. R., Ahire, D. V., Ahire, V. D., Chkravarty, M., & Maity, S. (2013). Soil bulk density as related to soil texture, organic matter content and available total nutrients of Coimbatore soil. *International Journal of Scientific and Research Publications*, *3*(2), 1-8.

Clark, L. A., & Pregibon, D. (1992). Tree-based Models. Ch. 8 in Statistical Models in S, eds. JM Chambers and T. Hastie. Pacific Grove, California: Wadsworth & Brooks.

Colman, E. A. (1947). A laboratory procedure for determining the field capacity of soils. *Soil Science*, *63*(4), 277-284.

Conrad, O., & Olaya, V. (2012). SAGA-GIS module library documentation (v2. 2.3). *Module Valley Depth. Available online: http://www. saga-gis. org/saga_tool_doc/2.2*, *3*.

Cootes, T. F., Ionita, M. C., Lindner, C., & Sauer, P. (2012, October). Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision* (pp. 278-291). Springer, Berlin, Heidelberg.

Criminisi, A., Shotton, J., Robertson, D., & Konukoglu, E. (2010, September). Regression forests for efficient anatomy detection and localization in CT studies. In *International MICCAI Workshop on Medical Computer Vision* (pp. 106-117). Springer, Berlin, Heidelberg.

Criminisi, A., Shotton, J., & Konukoglu, E. (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, *5*(6), 12.

Crist, E. P., & Cicone, R. C. (1984). A physically-based transformation of Thematic Mapper data---The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote sensing*, (3), 256-263.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

da Silva Chagas, C., de Carvalho Junior, W., Bhering, S. B., & Calderano Filho, B. (2016). Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena*, *139*, 232-240.

Chagas, C. D. S., Carvalho Júnior, W. D., Pinheiro, H. S. K., Xavier, P. A. M., Bhering, S. B., Pereira, N. R., & Calderano Filho, B. (2018). Mapping soil cation exchange capacity in a semiarid region through predictive models and covariates from remote sensing data. *Revista Brasileira de Ciência do Solo*, *42*.

De la Rosa, D. (1979). Relation of several pedological characteristics to engineering qualities of soil. *Journal of Soil Science*, *30*(4), 793-799.

Dixon, J. B., Weed, S. B., & Parpitt, R. L. (1990). Minerals in soil environments. *Soil Science*, *150*(2), 562.

FAO. (2019). Measuring and modelling soil carbon stocks and stock changes in livestock production systems: Guidelines for assessment (Version 1). *Livestock environmental assessment and performance (LEAP) partnership*, 170.

Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., & Koch, B. (2014). Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, *154*, 102-114.

Franklin, S. E., Lavigne, M. B., Wulder, M. A., & McCaffrey, T. M. (2002). Large-area forest structure change detection: An example. *Canadian Journal of Remote Sensing*, *28*(4), 588-592.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367-378.

Gambill, D. R., Wall, W. A., Fulton, A. J., & Howard, H. R. (2016). Predicting USCS soil classification from soil property variables using Random Forest. *Journal of Terramechanics*, *65*, 85-92.

García-Gaines, R. A., & Frankenstein, S. (2015). USCS and the USDA soil classification system: Development of a mapping scheme.

Gardner, W. R. (1971). Laboratory measurement of available soil water. *Soil Science Society of America Journal*, *35*(5), NP-NP.

Gholizadeh, A., Žižala, D., Saberioon, M., & 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.

Ghorbani, M. A., Shamshirband, S., Haghi, D. Z., Azani, A., Bonakdari, H., & Ebtehaj, I. (2017). Application of firefly algorithm-based support vector machines for prediction of field capacity and permanent wilting point. *Soil and Tillage Research*, *172*, 32-38.

Girshick, R., Shotton, J., Kohli, P., Criminisi, A., & Fitzgibbon, A. (2011, November). Efficient regression of general-activity human poses from depth images. In *2011 International Conference on Computer Vision* (pp. 415-422). IEEE.

GlobalSecurity. (2021) Terrain Analysis Considerations. *US Army Engineer School* [online] available at: https://www.globalsecurity.org/military/library/report/call/call_01-19_ch3.htm

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma*, *146*(1-2), 102-113.

Grunwald, S., Thompson, J. A., & Boettinger, J. L. (2011). Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Science Society of America Journal*, *75*(4), 1201-1213.

Gülser, C., & Candemir, F. (2014). Using soil moisture constants and physical properties to predict saturated hydraulic conductivity. *Eurasian Journal of Soil Science*, *3*(1), 77.

Gupta, S., & Larson, W. E. (1979). Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. *Water resources research*, *15*(6), 1633-1635.

Håkansson, I. (1990). A method for characterizing the state of compactness of the plough layer. *Soil and tillage research*, *16*(1-2), 105-120.

Håkansson, I., & Lipiec, J. (2000). A review of the usefulness of relative bulk density values in studies of soil structure and compaction. *Soil and Tillage Research*, *53*(2), 71-85.

Hazelton, P., & Murphy, B. (2016). *Interpreting soil test results: What do all the numbers mean?*. CSIRO publishing.

Hemond, H. F. (1983). The nitrogen budget of Thoreau's bog. *Ecology*, *64*(1), 99-109.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., ... & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, *12*(2), e0169748.

Hengl, T. (2018). Soil texture classes (USDA system) for 6 soil depths (0, 10, 30, 60, 100 and 200 cm) at 250 m (Version v0.2) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.2525817

Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional-scale: a random forest approach. *Geoderma*, *214*, 141-154.

Hong, Y., Yu, L., Chen, Y., Liu, Y., Liu, Y., Liu, Y., & Cheng, H. (2018). Prediction of soil organic matter by VIS–NIR spectroscopy using normalized soil moisture index as a proxy of soil moisture. *Remote Sensing*, *10*(1), 28.

Hounkpatin, O. K., de Hipt, F. O., Bossa, A. Y., Welp, G., & Amelung, W. (2018). Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso). *Catena*, *166*, 298-309.

Hudson, B. D. (1994). Soil organic matter and available water capacity. *Journal of soil and water conservation*, *49*(2), 189-194.

Huete, A. (1988). A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment. *Remote sensing of environment*, *25*, 295-309.

Ismail, R., & Mutanga, O. (2010). A comparison of regression tree ensembles: Predicting Sirex noctilio induced water stress in Pinus patula forests of KwaZulu-Natal, South Africa. *International Journal of Applied Earth Observation and Geoinformation*, *12*, S45-S51.

ISRIC. (2020) WoSIS database [Online] Available at: www.isric.org

Kauth, R. J., & Thomas, G. S. (1976, January). The tasselled cap--a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat. In *LARS symposia* (p. 159).

Koch, D. J., Ayers, P. D., Howard, H. R., & Siebert, G. (2012). *Vehicle Dynamics Monitoring and Tracking System (VDMTS): Monitoring Mission Impacts in Support of Installation Land Management* (No. ERDC/CERL-TR-12-11).

Lehner, K., & Hartmann, D. (2007). Using knowledge based fuzzy logic components in the design of underground engineering structures. *Proc., 1st EURO: TUN (Computational Methods in Tunneling)*.

Logsdon, S. (2019). Should Upper Limit of Available Water be Based on Field Capacity?. *Agrosystems, Geosciences & Environment*, *2*(1), 1-6.

Maindonald, J., & Braun, J. (2006). *Data analysis and graphics using R: an example-based approach* (Vol. 10). Cambridge University Press.

Malinowski, R., Lewinski, S., Rybicki, M., Jenerowicz, M., Gromny, E., Krupinski, M., Wojtkowski, C., Krupinski, M., Günther, S., Krätzschmar, E. (2019) S2GLC Final Report. *Scientific Exploitation of Operational Missions project*.

Martín, J. R., Álvaro-Fuentes, J., Gonzalo, J., Gil, C., Ramos-Miras, J. J., Corbí, J. G., & Boluda, R. (2016). Assessment of the soil organic carbon stock in Spain. *Geoderma*, *264*, 117-125.

Mbagwa, J. S. C., & Abeh, O. G. (1998). Prediction of engineering properties of tropical soils using intrinsic pedological parameters. *Soil Science*, *163*(2), 93-102.

McBratney, A. Á., & Pringle, M. J. (1999). Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. *Precision Agriculture*, *1*(2), 125-152.

McBride, R. A. (2002). 2.9 Atterberg Limits. *Methods of Soil Analysis: Part 4 Physical Methods*, *5*, 389-398.

Melillo, J. M., Aber, J. D., Linkins, A. E., Ricca, A., Fry, B., & Nadelhoffer, K. J. (1989). Carbon and nitrogen dynamics along the decay continuum: plant litter to soil organic matter. *Plant and soil*, *115*(2), 189-198.

Millard, K., & Richardson, M. (2015). On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping. *Remote sensing*, *7*(7), 8489-8515.

Mohanty, M., Sinha, N. K., Painuli, D. K., Bandyopadhyay, K. K., Hati, K. M., Reddy, K. S., & Chaudhary, R. S. (2015). Modelling soil water contents at field capacity and permanent wilting point using artificial neural network for Indian soils. *National Academy Science Letters*, *38*(5), 373-377.

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological processes*, *5*(1), 3-30.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., & McBratney, A. B. (2019). Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*, *352*, 251-267.

Ng, W., Minasny, B., Mendes, W. D. S., & Demattê, J. A. M. (2020). The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *Soil*, *6*(2), 565-578.

Odeha, I. O. A., McBratney, A. B., & Chittleborough, D. J. (1994). Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, *63*(3-4), 197-214.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & 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., Borrelli, P., Meusburger, K. (2015) A New European Slope Length and Steepness Factor (LS-Factor) for Modeling Soil Erosion by Water. Geosciences, (5): 117-126.

Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. (2007). Random forest models to predict aqueous solubility. *Journal of chemical information and modeling*, *47*(1), 150-158.
Pluske, W., Murphy, D., Sheppard, J. (2021). Total organic carbon factsheet. *soil quality australia* [online] available at: http://www.soilquality.org.au/factsheets/organic-carbon last accessed 03/01/2021

Peel, M. C. and Finlayson, B. L. and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* 11: 1633-1644

Poppiel, R. R., Lacerda, M. P., Safanelli, J. L., Rizzo, R., Oliveira, M. P., Novais, J. J., & Demattê, J. A. (2019). Mapping at 30 m Resolution of Soil Attributes at Multiple Depths in Midwest Brazil. *Remote Sensing*, *11*(24), 2905.

Post, W. M., Emanuel, W. R., Zinke, P. J., & Stangenberger, A. G. (1982). Soil carbon pools and world life zones. *Nature*, *298*(5870), 156-159.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181-199.

Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning* (pp. 307-323). Springer, Boston, MA.

Rab, M. A., Chandra, S., Fisher, P. D., Robinson, N. J., Kitching, M., Aumann, C. D., & Imhof, M. (2011). Modelling and prediction of soil water contents at field capacity and permanent wilting point of dryland cropping soils. *Soil Research*, *49*(5), 389-407.

Rad, M. R. P., Toomanian, N., Khormali, F., Brungard, C. W., Komaki, C. B., & Bogaert, P. (2014). Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. *Geoderma*, *232*, 97-106.

Raynolds, M. K., & Walker, D. A. (2016). Increased wetness confounds Landsat-derived NDVI trends in the central Alaska North Slope region, 1985–2011. *Environmental Research Letters*, *11*(8), 085004.

Rezaei, S. A., & Gilkes, R. J. (2005). The effects of landscape attributes and plant community on soil chemical properties in rangelands. *Geoderma*, *125*(1-2), 167-176.

Richards, L. A., & Weaver, L. R. (1943). Fifteen-atmosphere percentage as related to the permanent wilting percentage. *Soil Science*, *56*(5), 331-340.

Sarmadian, F., & Taghizadeh Mehrjardi, R. (2008). Modeling of some soil properties using artificial neural network and multivariate regression in Gorgan Province, North of Iran. *Global Journal of Environmental Research*, *2*(1), 30-35.

Santanello Jr, J. A., Peters-Lidard, C. D., Garcia, M. E., Mocko, D. M., Tischler, M. A., Moran, M. S., & Thoma, D. P. (2007). Using remotely-sensed estimates of soil moisture to infer soil texture and hydraulic properties across a semi-arid watershed. *Remote Sensing of Environment*, *110*(1), 79-97.

Scholl, P., Leitner, D., Kammerer, G., Loiskandl, W., Kaul, H. P., & Bodner, G. (2014). Root induced changes of effective 1D hydraulic properties in a soil column. *Plant and soil*, *381*(1-2), 193-213.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

Seybold, C. A., Elrashidi, M. A., & Engel, R. J. (2008). Linear regression models to estimate soil liquid limit and plasticity index from basic soil properties. *Soil science*, *173*(1), 25-34.

Singh, S. K., Singh, A. K., Sharma, B. K., & Tarafdar, J. C. (2007). Carbon stock and organic carbon dynamics in soils of Rajasthan, India. *Journal of Arid Environments*, *68*(3), 408-421.

Slatyer, R. O., & Markus, D. K. (1968). Plant-water relationships. *Soil Science*, *106*(6), 478.

Soilmapper (2020) Soil observations and variables. *Predictive soil mapping with R.* [online available at https://soilmapper.org/soil-variables-chapter.html]

Soil Survey Division Staff, (1993). Soil Survey Manual. Soil ConservationService, *U.S. Department of Agriculture Handbook 18 (Chapter 3)*.

Stanchi, S. D., D'Amico, M., Zanini, E., & Freppaz, M. (2015). Liquid and Plastic limits of mountain soils as a function of the soil and horizon type. *Catena*, *135*, 114-121.

Strobl, C., & Zeileis, A. (2008). Danger: High power!–exploring the statistical properties of a test for random forest variable importance.

Sumner, M. E., & Miller, W. P. (1996). Cation exchange capacity and exchange coefficients. *Methods of Soil Analysis: Part 3 Chemical Methods*, *5*, 1201-1229.

Tesfa, T. K., Tarboton, D. G., Chandler, D. G., & McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. *Water Resources Research*, *45*(10).

Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the RPART routines* (Vol. 61, p. 452). Mayo Foundation: Technical report.

Thompson, J. A., Bell, J. C., & Butler, C. A. (2001). Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, *100*(1-2), 67-89.

Tolk, J. A. (2003). Soils, permanent wilting points. *Encyclopedia of water science*, *120010337*, 92.

Tóth, B., Weynants, M., Pásztor, L., & Hengl, T. (2017). 3D soil hydraulic database of Europe at 250 m resolution. *Hydrological Processes*, *31*(14), 2662-2666.

Twarakavi, N. K., Sakai, M., & Šimůnek, J. (2009). An objective analysis of the dynamic nature of field capacity. *Water Resources Research*, *45*(10).

US Air Force Engineering Support Agency/Civil Engineering Squad (AFCESA/CES) (1997), Criteria and Guidance for C-17 Contingency and Training on Semi-Prepared Airfields, Engineering Technical Letter, 97–99.

VandenBygaart, A. J., & Angers, D. A. (2006). Towards accurate measurements of soil organic carbon stock change in agroecosystems. *Canadian Journal of Soil Science*, *86*(3), 465-471.

Vaudour, E., Gomez, C., Loiseau, T., Baghdadi, N., Loubet, B., Arrouays, D., ... & Lagacherie, P. (2019). The impact of acquisition date on the prediction performance of topsoil organic carbon from Sentinel-2 for croplands. *Remote Sensing*, *11*(18), 2143.

Vaysse, K., & Lagacherie, P. (2015). Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). *Geoderma Regional*, *4*, 20-30.

Veihmeyer, F. J., & Hendrickson, A. H. (1931). The moisture equivalent as a measure of the field capacity of soils. *Soil Science*, *32*(3), 181-194.

Verbyla, D. L. (1987). Classification trees: a new discrimination tool. *Canadian Journal of Forest Research*, *17*(9), 1150-1152.

Vereecken, H., Maes, J., Feyen, J., & Darius, P. (1989). Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. *Soil science*, *148*(6), 389-403.

Wang, N., Xue, J., Peng, J., Biswas, A., He, Y., & Shi, Z. (2020). Integrating Remote Sensing and Landscape Characteristics to Estimate Soil Salinity Using Machine Learning Methods: A Case Study from Southern Xinjiang, China. *Remote Sensing*, *12*(24), 4118.

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and soil*, *340*(1-2), 7-24.

Wilcox, J. C., & Spilsbury, R. H. (1941). Soil Moisture Studies: II. Some Relationships Between Moisture Measurements and Mechanical Analysis. *Scientific Agriculture*, *21*(8), 459-472.

Yang, R. M., & Guo, W. W. (2019). Modelling of soil organic carbon and bulk density in invaded coastal wetlands using Sentinel-1 imagery. *International Journal of Applied Earth Observation and Geoinformation*, *82*, 101906.

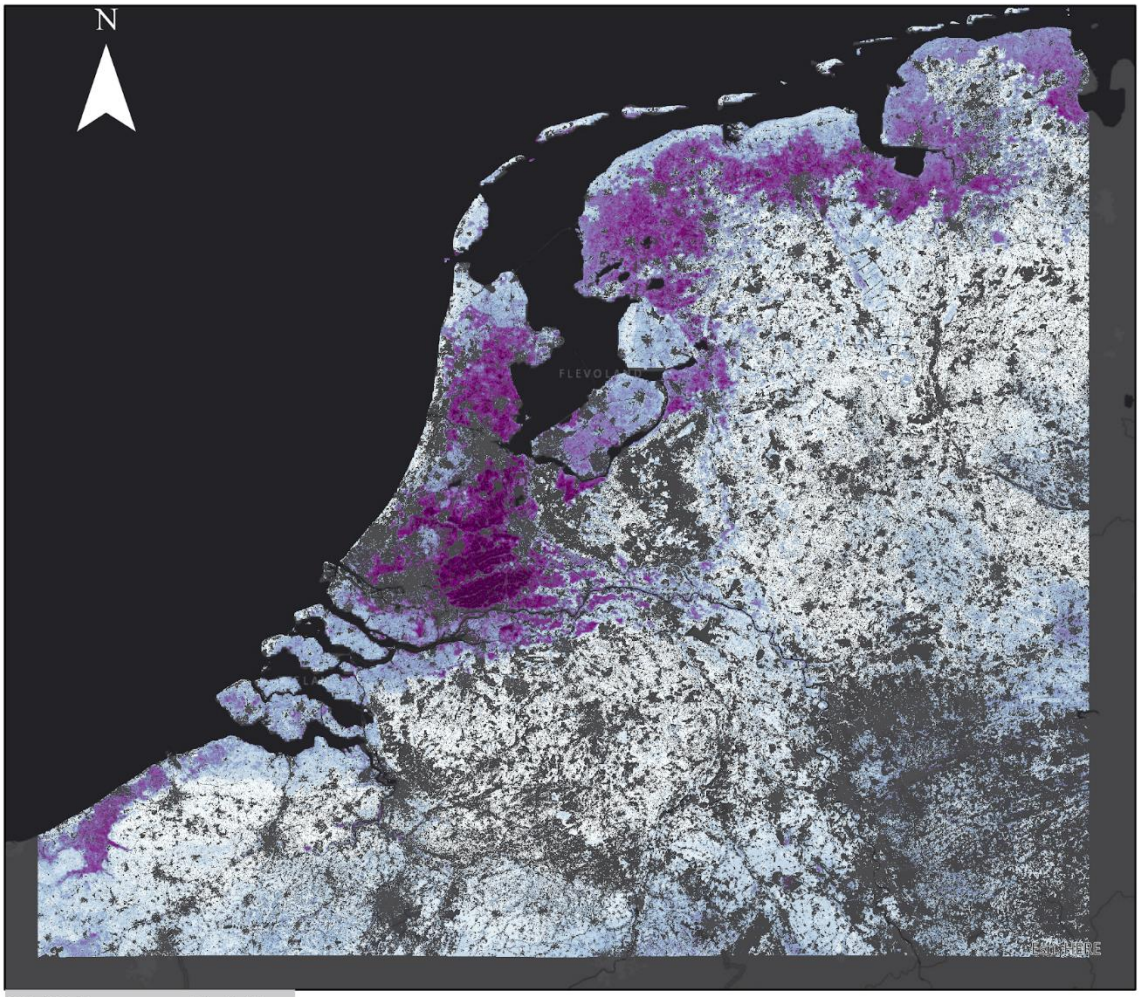
Young, A. (1973). Soil survey procedures in land development planning. *Geographical Journal*, 53-64.

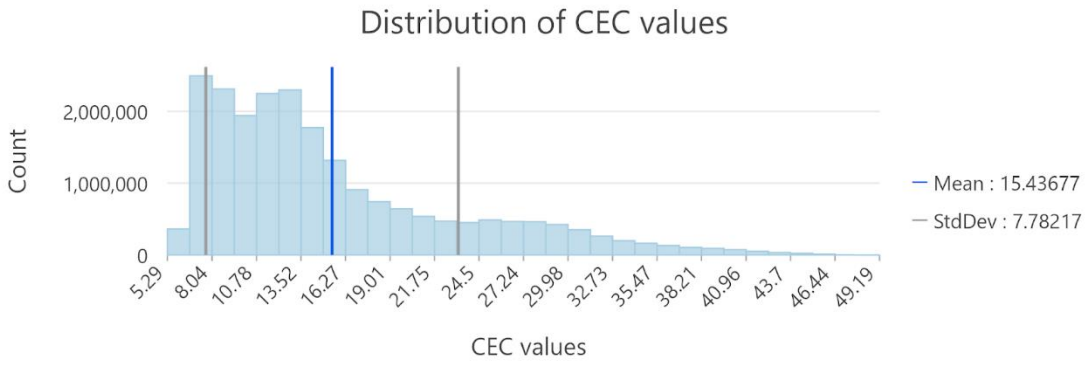
Zhang, X., Schaaf, C. B., Friedl, M. A., Strahler, A. H., Gao, F., & Hodges, J. C. (2002, June). MODIS tasseled cap transformation and its utility. In *IEEE International Geoscience and Remote Sensing Symposium* (Vol. 2, pp. 1063-1065). IEEE.

Zhu, Z., Bergamaschi, B., Bernknopf, R., Clow, D., Dye, D., Faulkner, S., ... & Liu, S. (2010). A method for assessing carbon stocks, carbon sequestration, and greenhouse-gas fluxes in ecosystems of the United States under present conditions and future scenarios. *Scientific Investigations Report 2010-5233. Reston, VA: US Geological Survey. 188 p.*

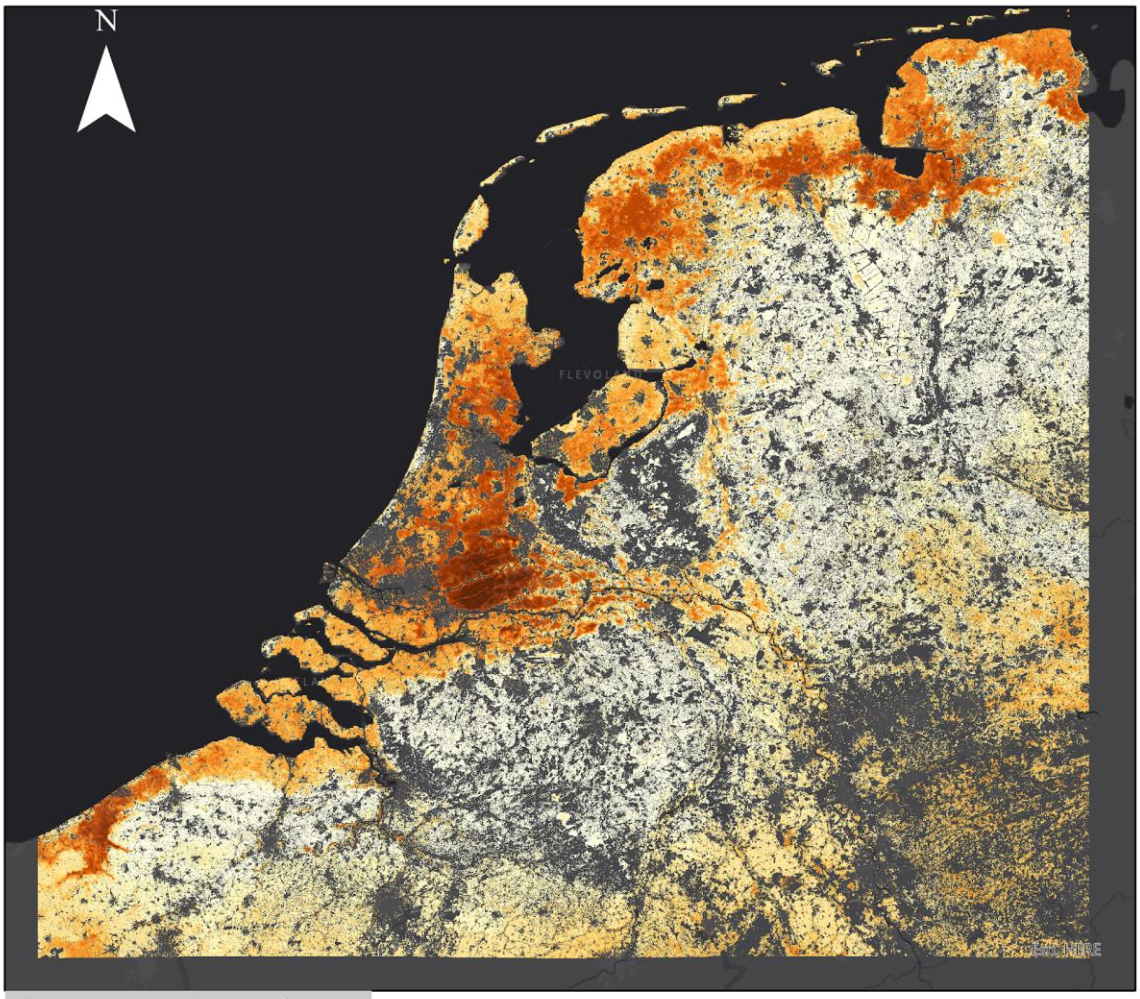
Appendix 1: visual map results

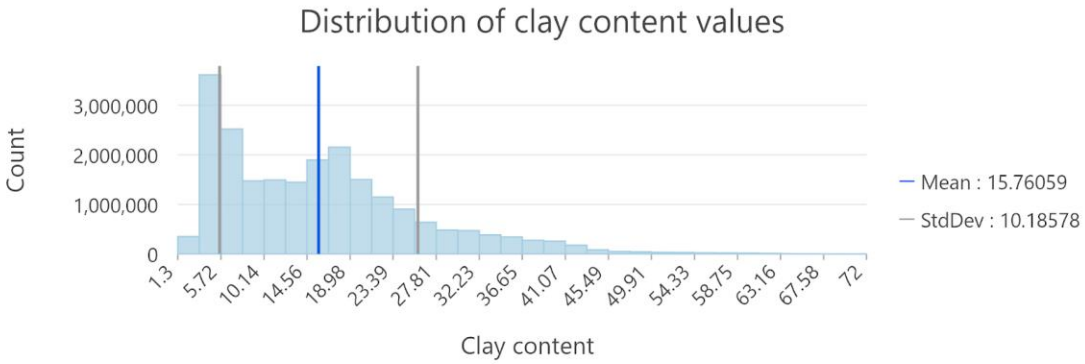
Cation Exchange Capacity of the soil in the Netherlands,
Flanders and the Western part of Germany



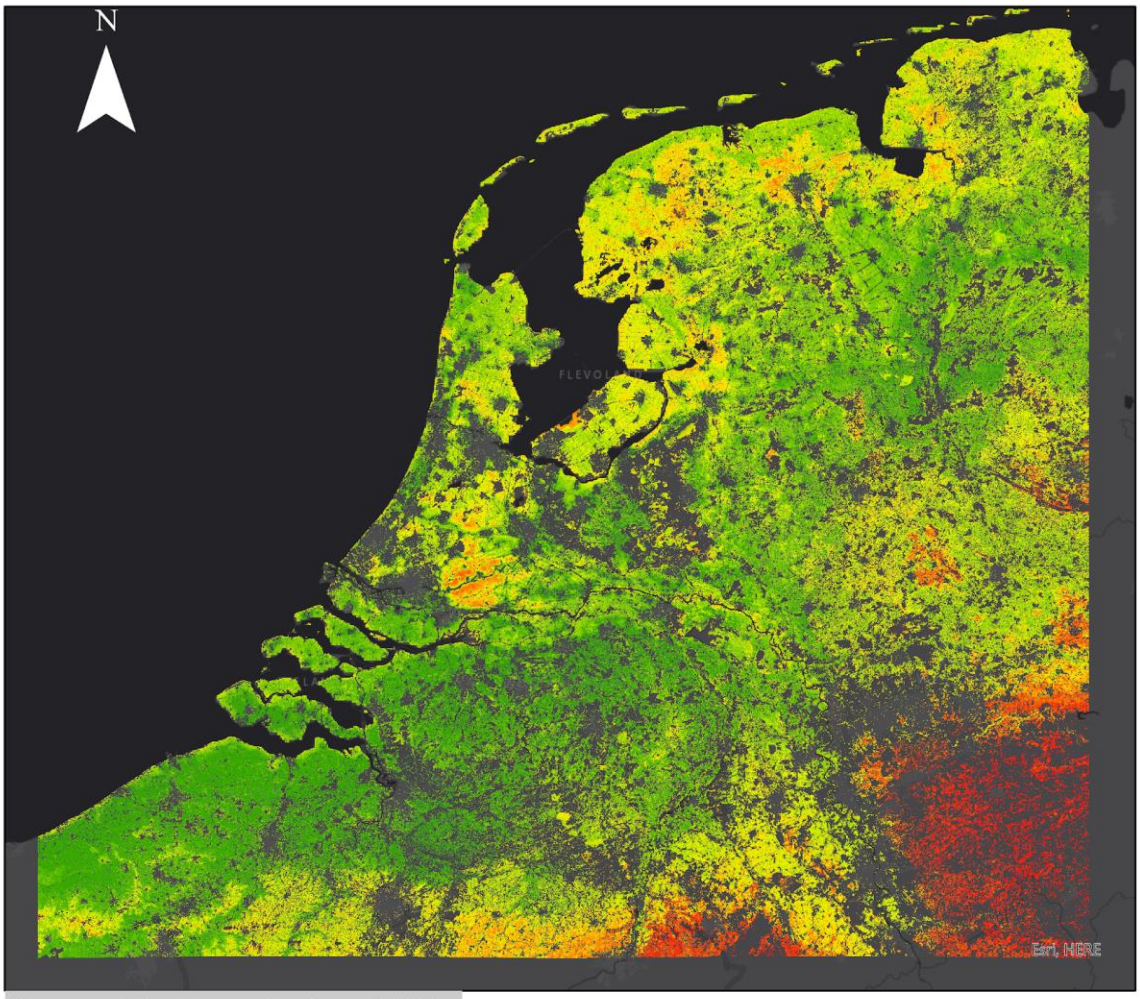


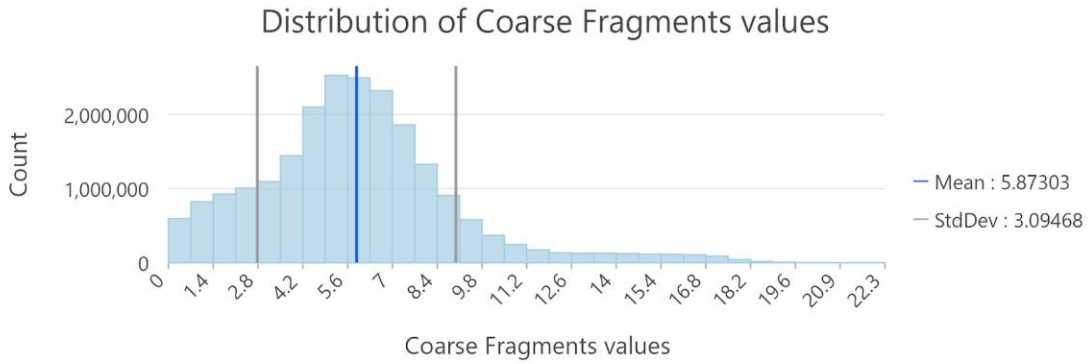
Clay content in the soil in the Netherlands, Flanders and the Western part of Germany



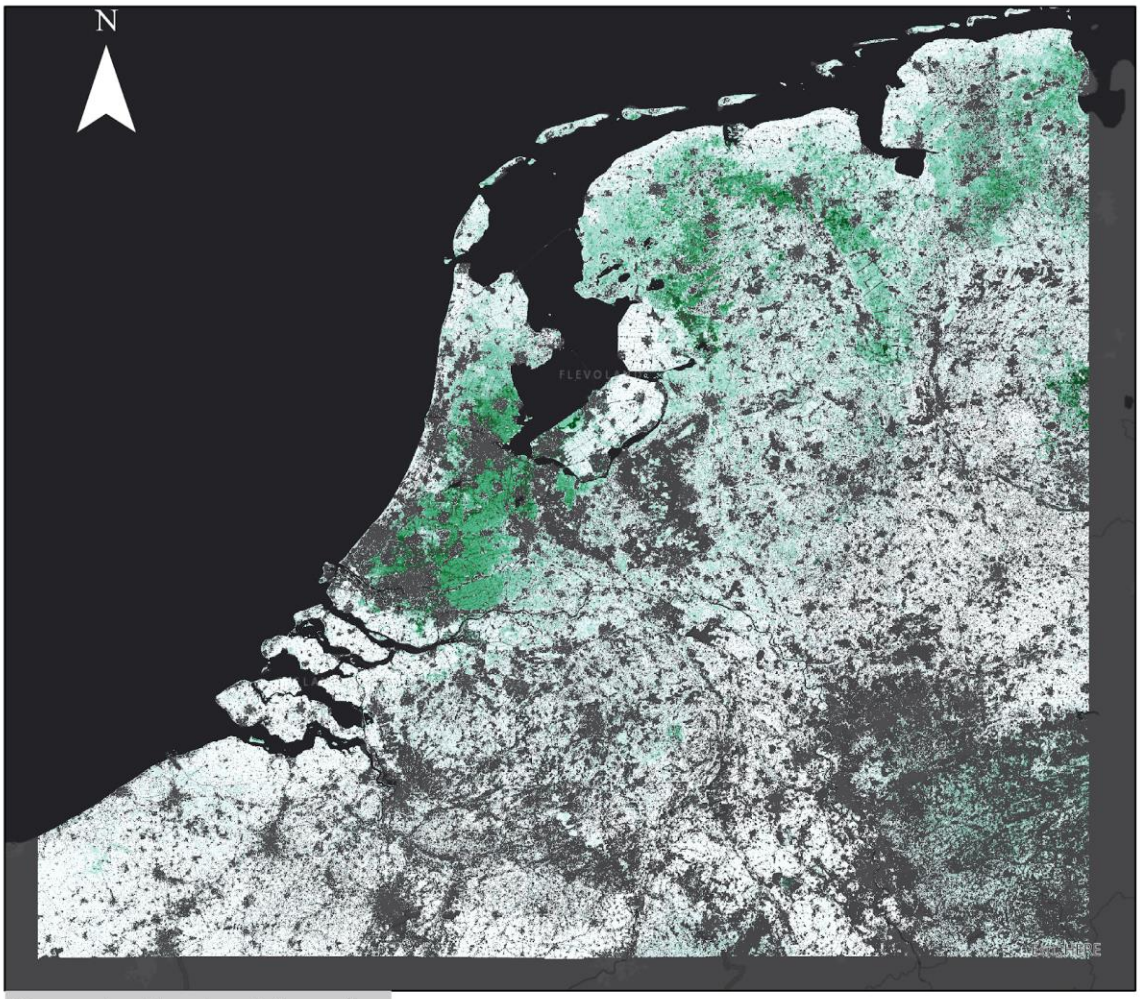


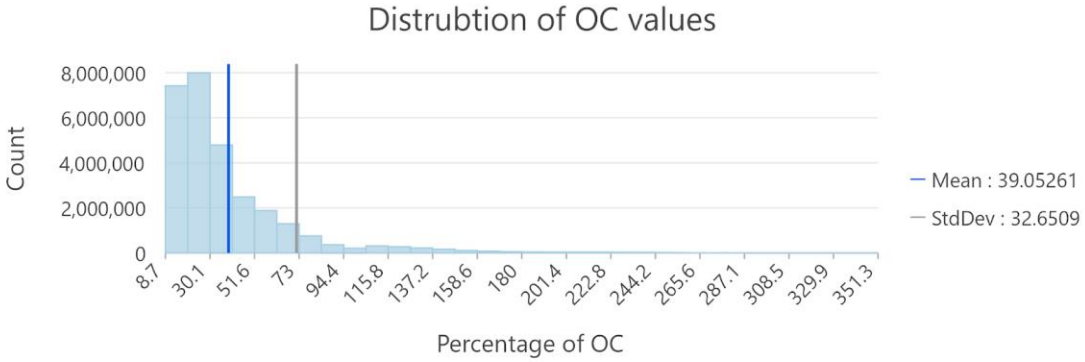
Coarse Fragments in the soil in the Netherlands, Flanders and the Western part of Germany



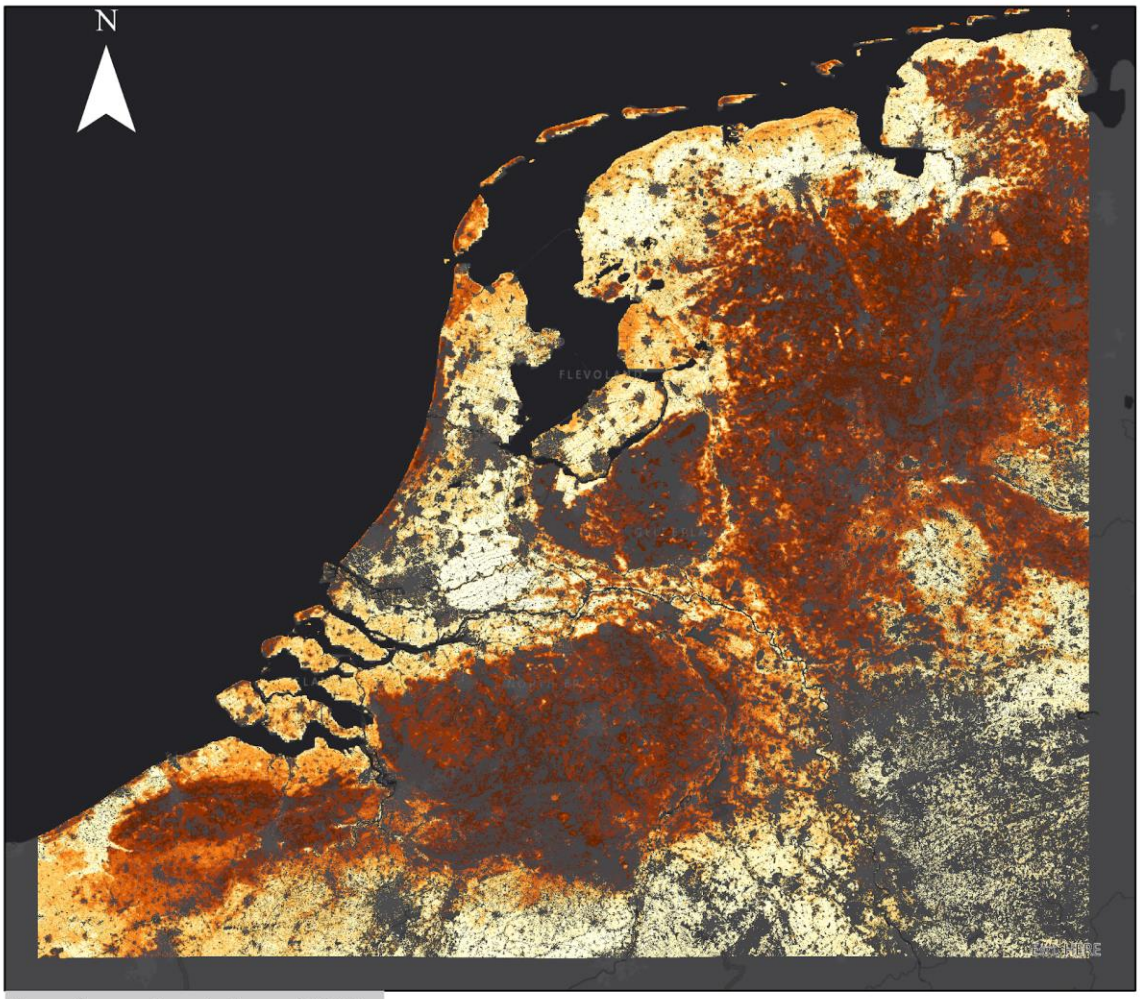


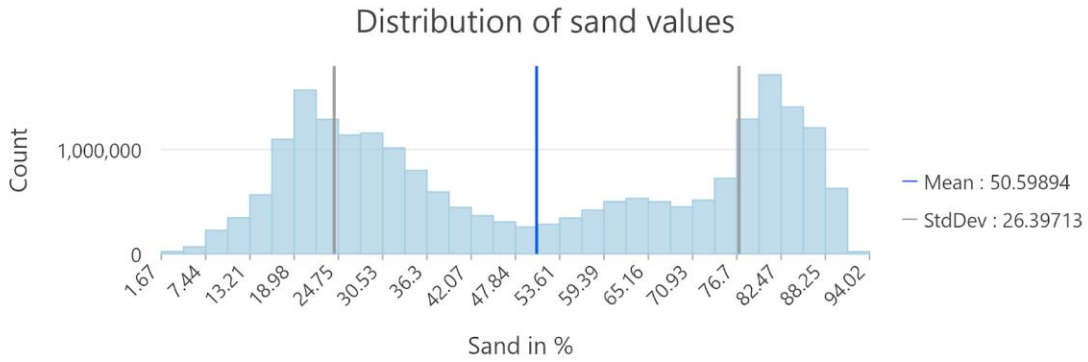
Organic Content in the soil in the Netherlands, Flanders and the Western part of Germany



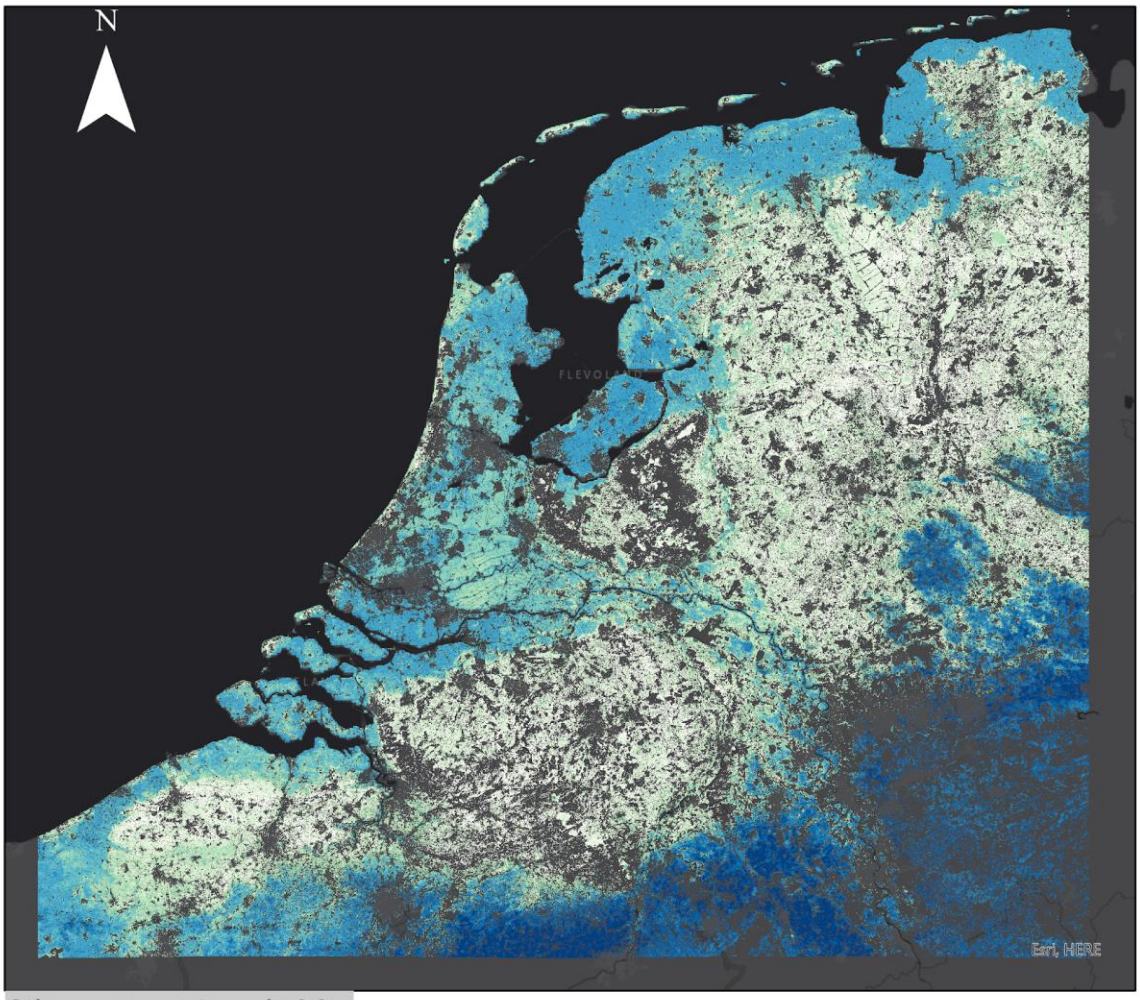


Sand content in the soil in the Netherlands, Flanders and the Western part of Germany





Silt content in the soil in the Netherlands, Flanders and the Western part of Germany

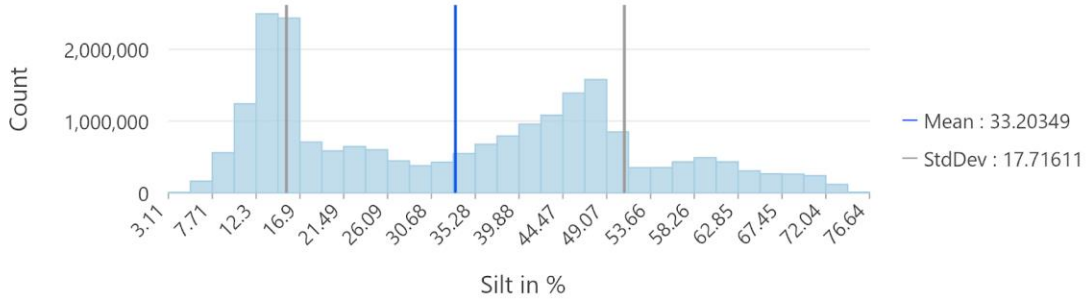


Silt content in g/100g

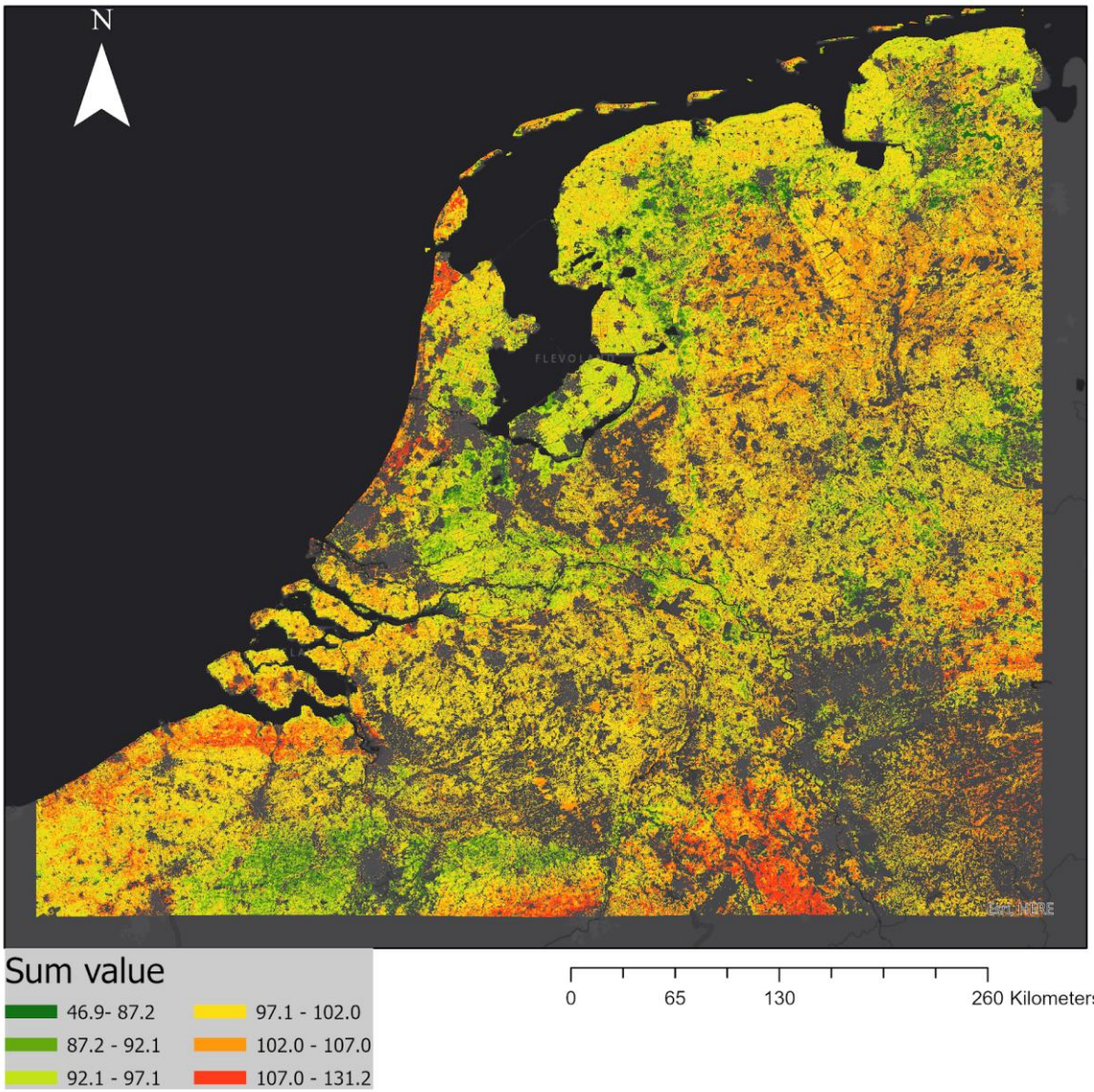
3 - 12	39- 45
12 - 18	45 - 51
18- 25	51- 58
25- 32	58- 65
32- 39	65- 77

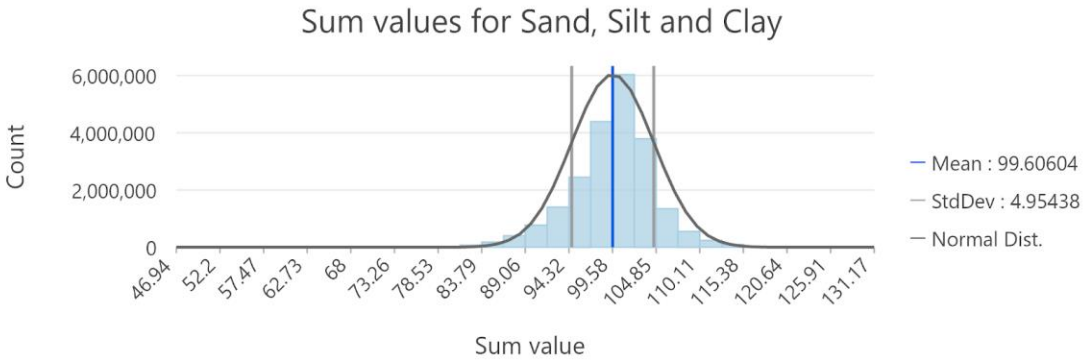
0 65 130 260 Kilometers

Distribution of silt values

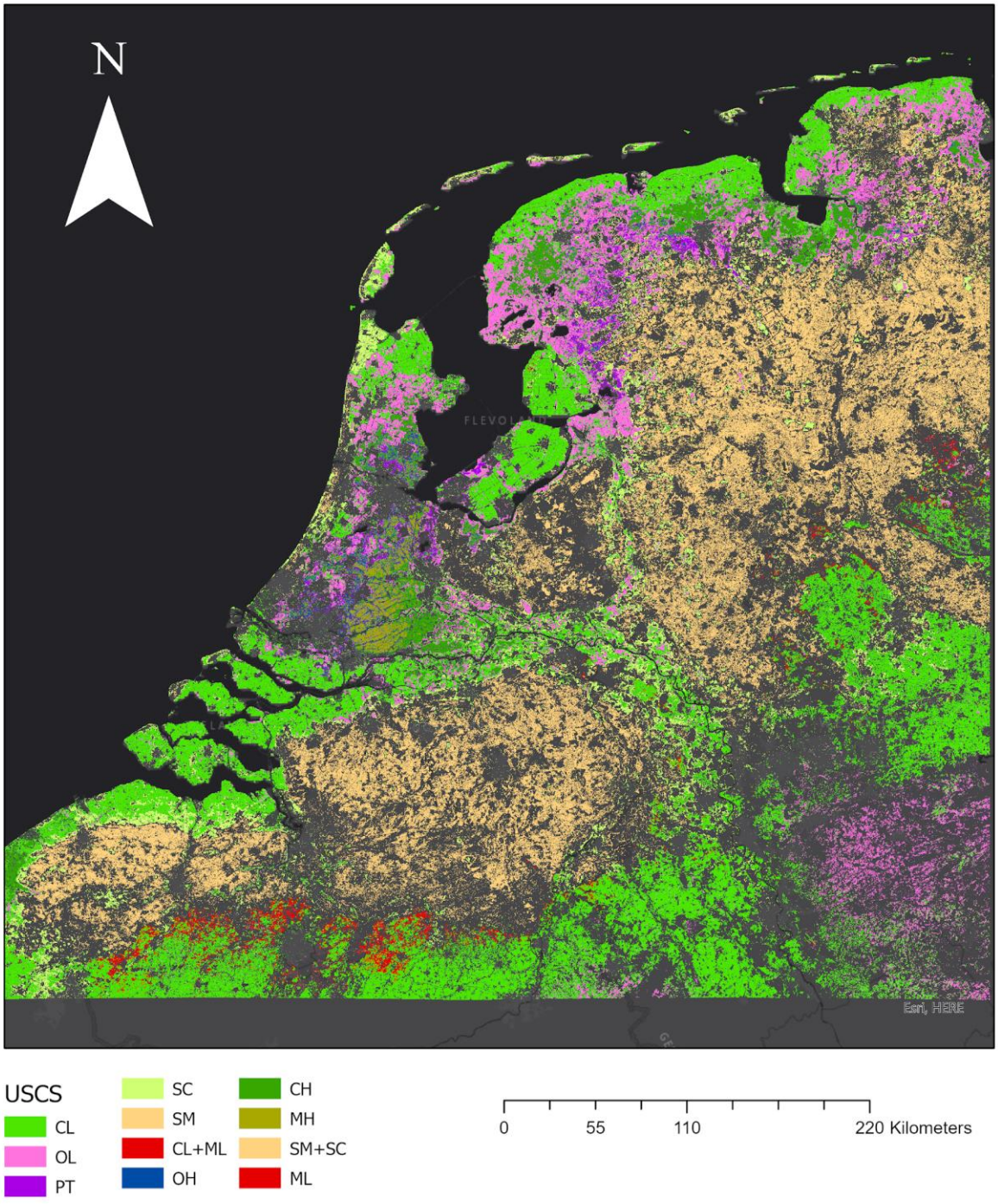


Sum of Sand, Silt and Clay content in the soil in the Netherlands, Flanders and the Western part of Germany





Appendix 2: Predicted USCS map



Appendix 3: validation statistics

Sand

OID_	R2	RMSE	ME
1	0.382536	11.56754	-2.90877
2	0.251239	14.1541	0.10464
3	0.783347	8.718687	1.411628
4	0.689471	16.11794	3.515201
5	0.682209	11.0232	2.186823
6	0.463737	15.29048	-0.02009
7	0.371278	11.51602	0.37522
8	0.574025	11.44826	-1.1984
9	0.502702	12.41695	0.303271
10	0.549211	12.05295	1.466833

Silt

OID_	R2	RMSE	ME
1	0.41397	12.91125	-0.55843
2	0.373196	8.737799	0.504787
3	0.321656	8.273326	-2.34643
4	0.544883	7.777219	-0.03951
5	0.491083	7.908244	-0.13913
6	0.301909	12.47397	-1.42255
7	0.617934	5.659647	-0.09313
8	0.466651	10.53275	0.330671
9	0.412803	12.92444	-0.85605
10	0.50515	9.799262	2.482383

Clay

OID_	R2	RMSE	ME
1	0.553443	4.216695	-0.54955
2	0.5994	4.957827	0.042787
3	0.670413	5.334832	0.174679
4	0.691866	4.345061	0.146909
5	0.661848	3.976653	0.065317
6	0.644158	4.885663	-0.23192
7	0.707844	3.111484	-0.08854
8	0.64729	4.35282	-0.19755
9	0.735294	4.527861	0.018146
10	0.612459	4.587067	-0.3677

Coarse

OID_	R2	RMSE	ME
1	0.327784	3.427412	-0.83629
2	0.24168	4.669921	-0.24707
3	0.161367	4.79318	0.075572
4	0.100971	3.940338	-0.53059
5	0.298569	2.32609	-0.85441
6	0.244114	4.962674	-0.38935
7	0.253823	2.615525	-0.73344
8	0.10644	3.477423	-1.00212
9	0.283675	5.093079	0.279469
10	0.198671	4.956453	-0.3761

OC

OID_	R2	RMSE	ME
1	0.626183	43.33958	2.863014

2	0.317036	78.40125	15.30676
3	0.698208	22.68093	0.048671
4	0.358225	53.94252	-5.68612
5	0.003689	7.221516	-3.61718
6	0.639527	23.56075	-8.38792
7	0.497782	19.42566	-4.99211
8	0.196094	35.0895	-6.85944
9	0.64799	59.86942	5.936936
10	0.232116	27.96243	-8.18239

CEC

OID_	R2	RMSE	ME
1	0.309494	5.253405	-0.45591
2	0.809465	5.057553	-1.26347
3	0.413203	7.102216	0.337787
4	0.82931	4.668059	-1.34783
5	0.688103	3.9661	-0.68056
6	0.525912	7.717776	1.192632
7	0.567106	7.681029	0.43598
8	0.351628	10.97353	0.797329
9	0.426723	7.85578	2.072408
10	0.654622	6.301507	0.215755

Appendix 4: classification script

Python script to classify points into the USCS classification using the field calculator in ArcGis

```
Reclass(!clay!, !cec!, !sand!, !silt!, !coarse!, !OC!, !LL!, !PI!)

def Reclass(clay, cec, sand, silt, coarse, OC, LL, PI):
      if (coarse > 50):
      return "gravel"
      if (coarse < 50):
      if (sand > 50):
            if (clay + silt) < 5:
                  return "SW or SP"
            if (clay + silt) > 12:
                  if ((PI < 4) or (PI < (0.73 * (LL-20)))):
                  return "SM"
                  if ((PI > 7) and (PI > (0.73 * (LL-20)))):
                  return "SC"
            else:
                  return "SM+SC"
      if (sand < 50):
            if ((LL >= 50) and (OC > 200)) or ((LL < 50) and (OC >
150)):
                  return "PT"
            if LL < 50:
                  if (OC > 40):
                  return "OL"
                  if ((PI < 4) or (PI < (0.73 * (LL-20)))):
                  return "ML"
                  if ((PI > 7) and (PI > (0.73 * (LL-20)))):
                  return "CL"
                  else:
                  return "CL+ML"
            if (LL >= 50):
                  if (OC > 150):
                  return "OH"
                  if ((PI < 4) or (PI < (0.73 * (LL-20)))):
                  return "MH"
                  if ((PI > 7) and (PI > (0.73 * (LL-20)))):
                  return "CH"
                  else:
                  return "wtf"
            else:
                  return "wtf"
      else:
      return "wtf"
```

Appendix 5: google earth engine script

Sentinel 2 median for multiple spring season

```
var ROI = /* color: #0b4a8b */ee.Geometry.Polygon([
      [-26.60888671875, 66.01801815922043],
      [-18.1494140625, 32.565333160841035],
      [15.954649288064843, 35.39010560236367],
      [44.80224609375, 36.756490329505176],
      [43.59375, 70.85188122123132],
      [11.88720703125, 73.23937702441908]
  ]);
      Map.centerObject(ROI,4);

var s2 = ee.ImageCollection('COPERNICUS/S2')
  .filterBounds(ROI)
  .filterMetadata('CLOUDY_PIXEL_PERCENTAGE','less_than',0.05);


var voorjaar2016 = s2.filterDate('2016-04-01','2016-05-31');
var voorjaar2017 = s2.filterDate('2017-04-01','2017-05-31');
var voorjaar2018 = s2.filterDate('2018-04-01','2017-05-31');
var voorjaar2019 = s2.filterDate('2019-04-01','2019-05-31');
var voorjaar2020 = s2.filterDate('2020-04-01','2020-05-31');

var median = s2.reduce(ee.Reducer.median());


var vis_param = {bands: ['B1_median', 'B2_median', 'B3_median'], gamma:
3.6, scale: 30};

Map.addLayer(median, vis_param);
Map.setCenter(5, 47, 4);

var polygon001 = table;
Export.image.toDrive({
  image: median,
  description: 'test1_S2',
  scale: 30,
  region: polygon001,
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});

var polygon001 = table2;
Export.image.toDrive({
  image: median,
  description: 'test2_S2',
  scale: 30,
  region: polygon001,
```

```
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});

var polygon001 = table3;
Export.image.toDrive({
  image: median,
  description: 'test3_S2',
  scale: 30,
  region: polygon001,
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});

var polygon001 = table4;
Export.image.toDrive({
  image: median,
  description: 'test4_S2',
  scale: 30,
  region: polygon001,
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});
```

Sentinel 1 median for multiple spring seasons

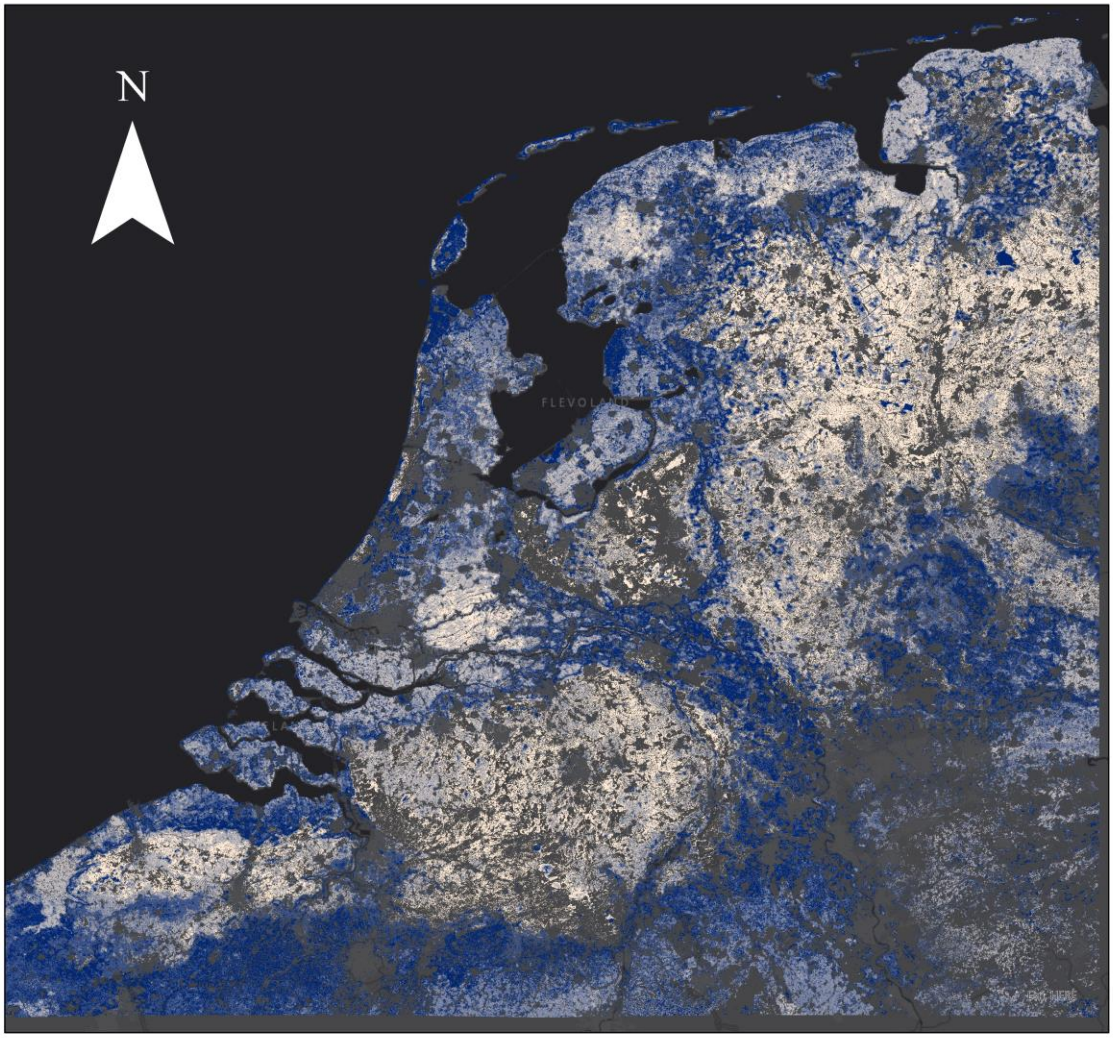
```
var imgVV = ee.ImageCollection('COPERNICUS/S1_GRD')
      .filter(ee.Filter.listContains('transmitterReceiverPolarisation',
'VV'))
      .filter(ee.Filter.eq('instrumentMode', 'IW'))
      .filter(ee.Filter.eq('orbitProperties_pass', 'DESCENDING'))
      .select('VV')
      .map(function(image) {
            var edge = image.lt(-30.0);
            var maskedImage = image.mask().and(edge.not());
            return image.updateMask(maskedImage);
      });


var desc = imgVV.filter(ee.Filter.eq('orbitProperties_pass',
'DESCENDING'));

var voorjaar2016 = ee.Filter.date('2016-04-01','2016-05-31');
var voorjaar2017 = ee.Filter.date('2017-04-01','2017-05-31');
```

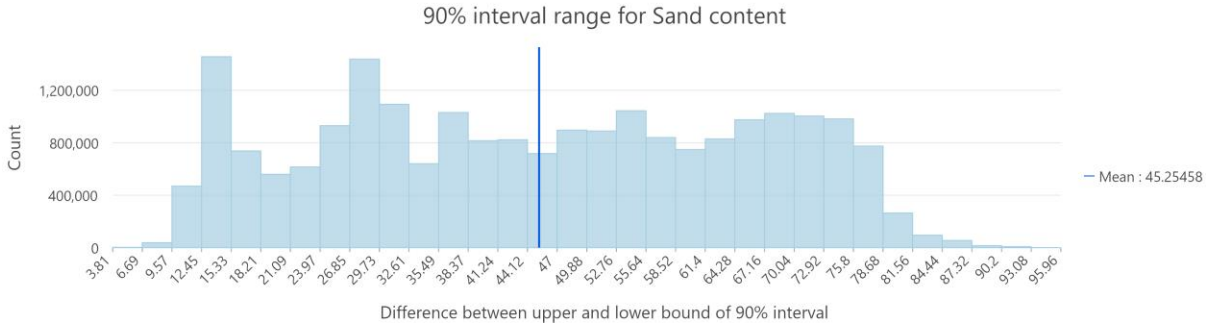
```
var voorjaar2018 = ee.Filter.date('2018-04-01','2017-05-31');
var voorjaar2019 = ee.Filter.date('2019-04-01','2019-05-31');
var voorjaar2020 = ee.Filter.date('2020-04-01','2020-05-31');

var spring = ee.Filter.or(voorjaar2016, voorjaar2017, voorjaar2018,
voorjaar2019, voorjaar2020);


var descView = ee.Image.cat(
      desc.filter(spring).median());

Map.setCenter(5, 47, 4);
Map.addLayer(descView, {min: -30, max: 30}, 'Multi-T Mean ASC', true);

var polygon001 = table;
Export.image.toDrive({
  image: descView,
  description: 'test1_S1',
  scale: 30,
  region: polygon001,
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});

var polygon001 = table2;
Export.image.toDrive({
  image: descView,
  description: 'test2_S1',
  scale: 30,
  region: polygon001,
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});

var polygon001 = table3;
Export.image.toDrive({
  image: descView,
  description: 'test3_S1',
  scale: 30,
  region: polygon001,
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});

var polygon001 = table4;
Export.image.toDrive({
  image: descView,
```

```
  description: 'test4_S1',
  scale: 30,
  region: polygon001,
  fileFormat: 'GeoTIFF',
  formatOptions: {
      cloudOptimized: true
  }
});
```

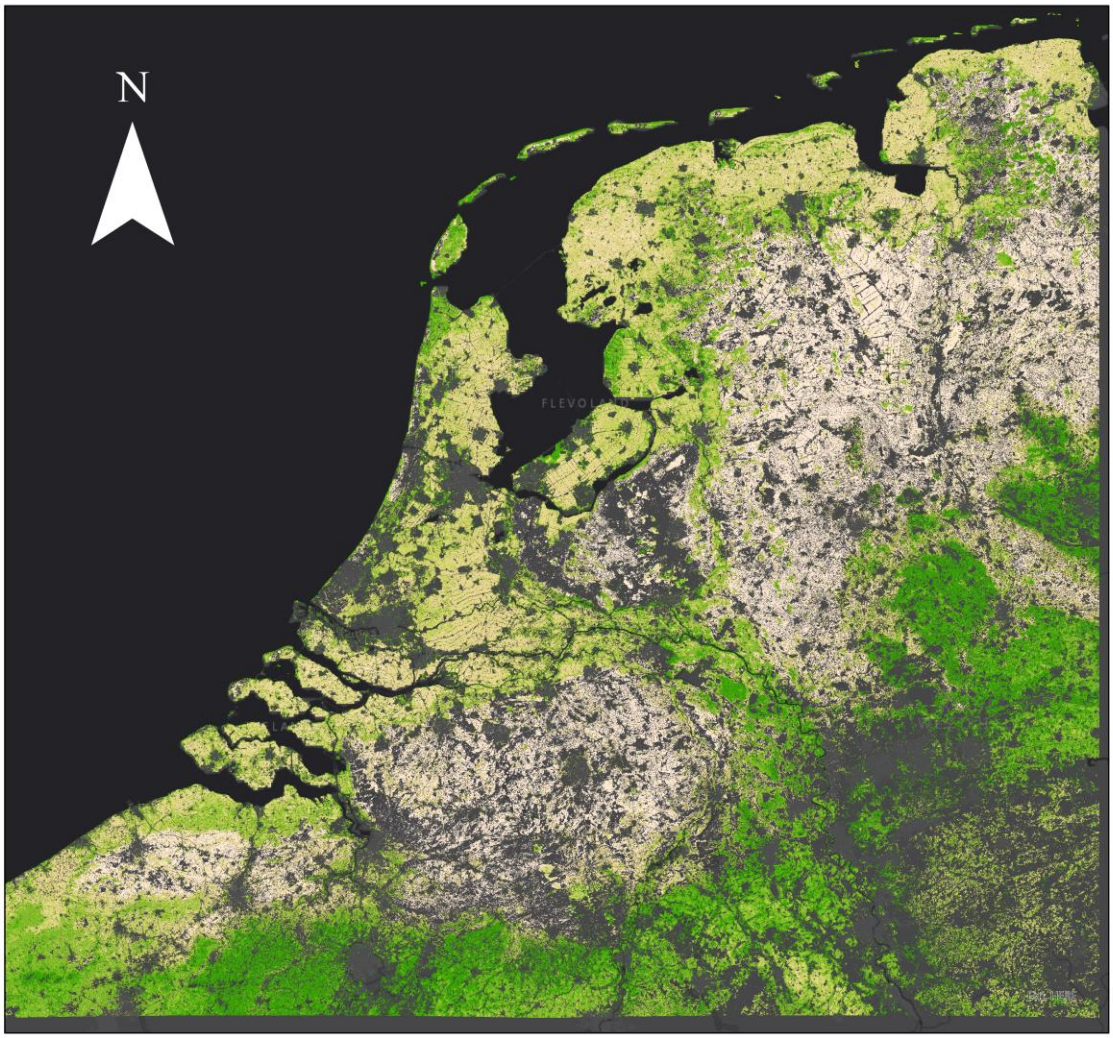
Appendix 6: Uncertainty Maps

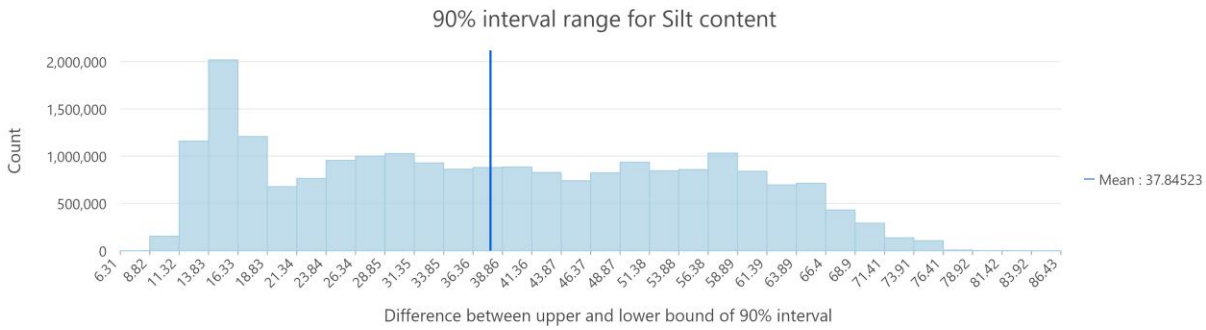
Uncertainty map for the prediction of Sand Content



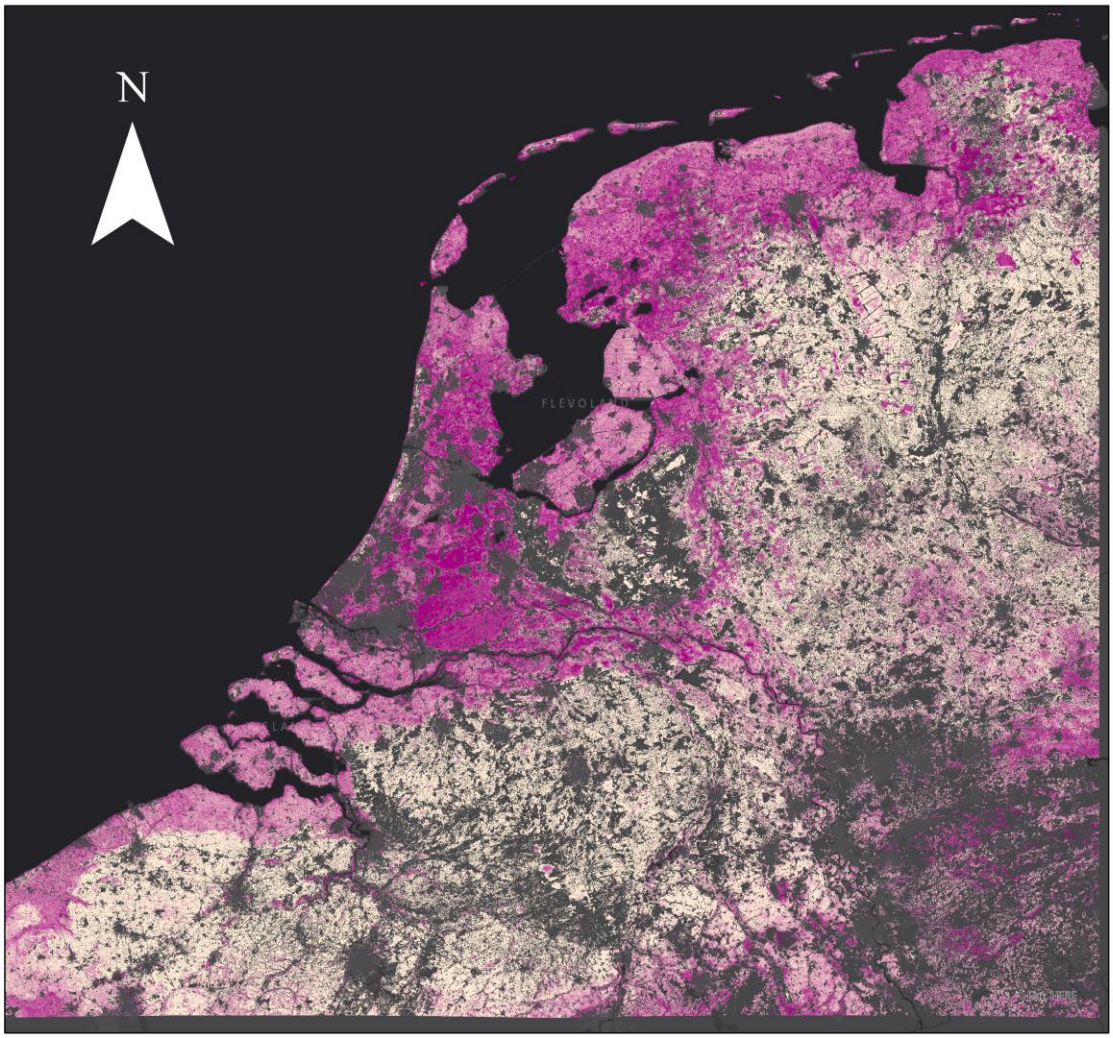


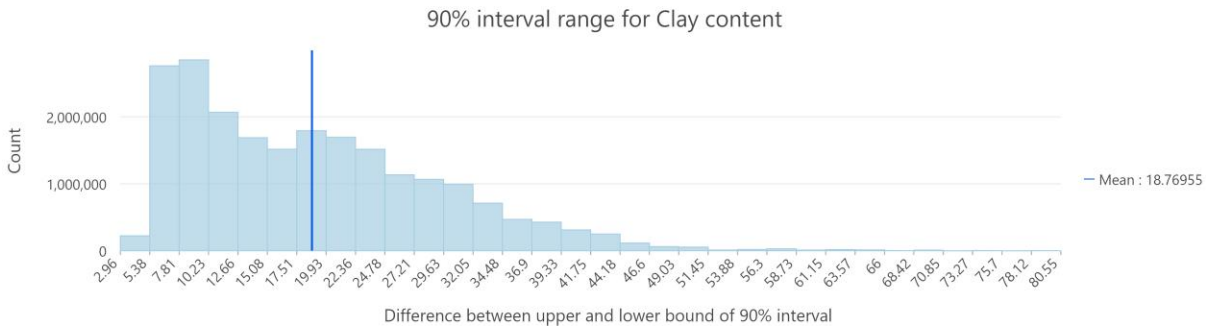
Uncertainty map for the prediction of Silt Content



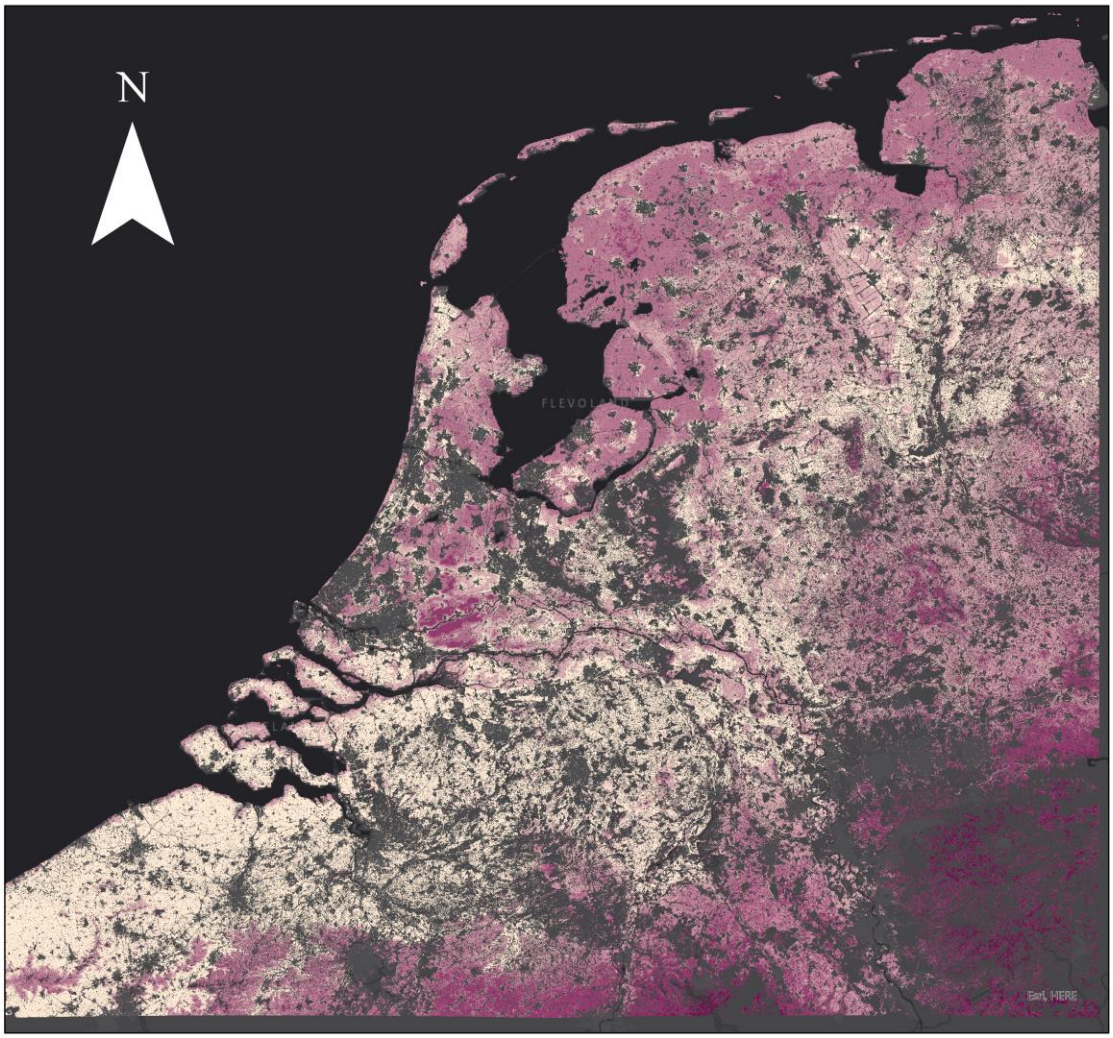


Uncertainty map for the prediction of Clay Content





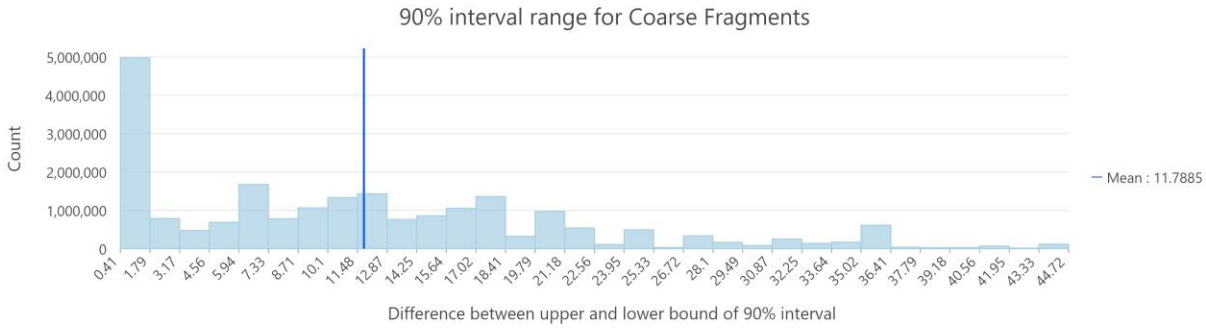
Uncertainty map for the prediction of Coarse Fragments



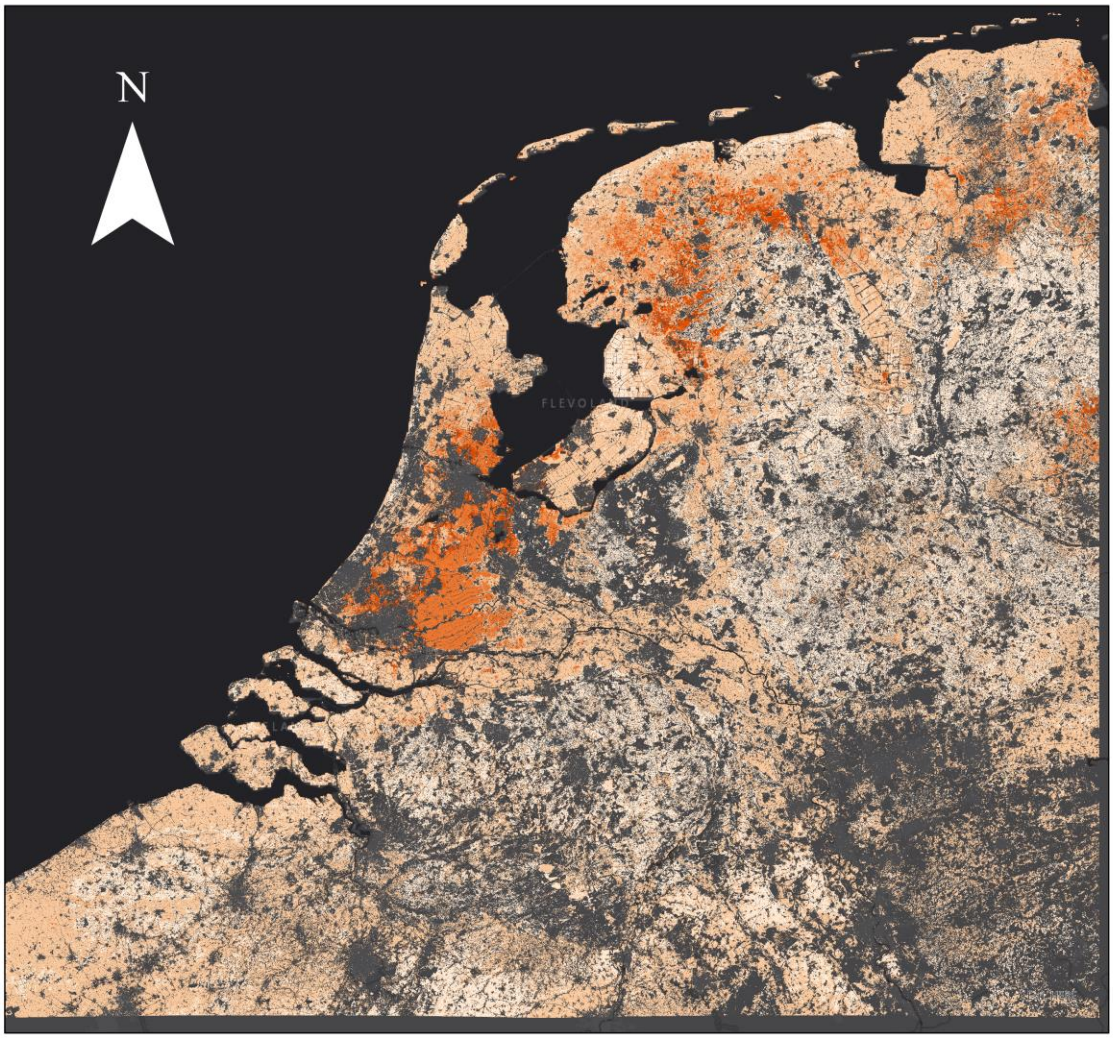
0 40 80 160 Kilometers

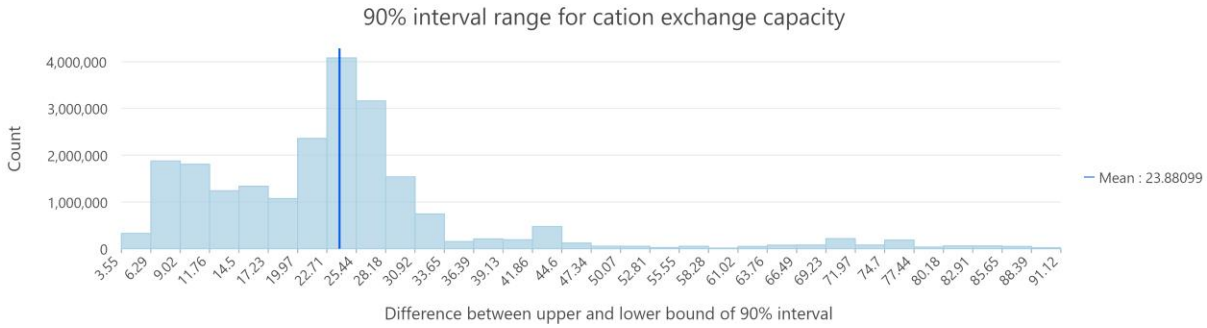
Range of uncertainty
(lower is better)

- 0.4 - 5.4
- 5.4 - 14.1
- 14.1 - 25.4
- 25.4 - 36.8
- 36.8- 44.7

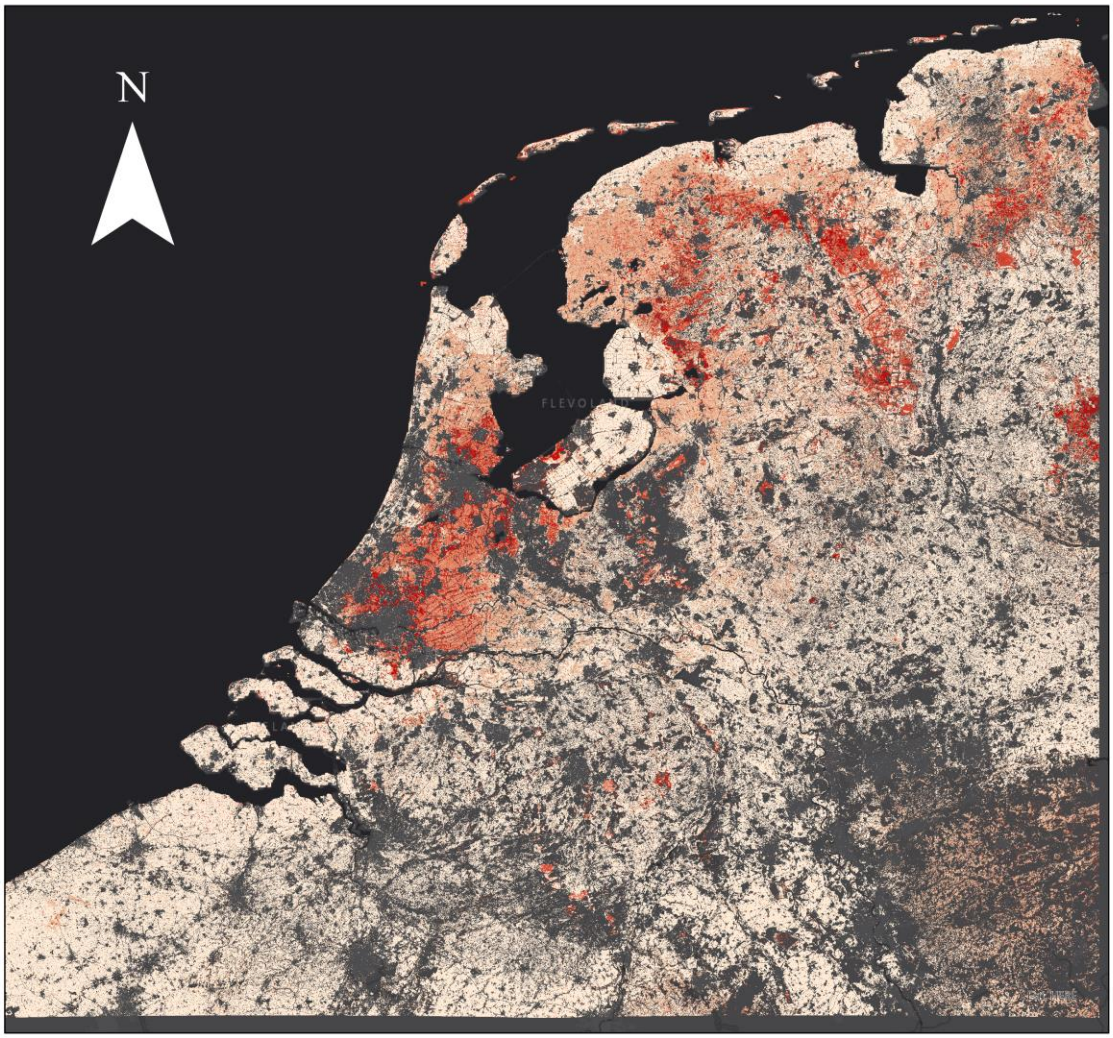


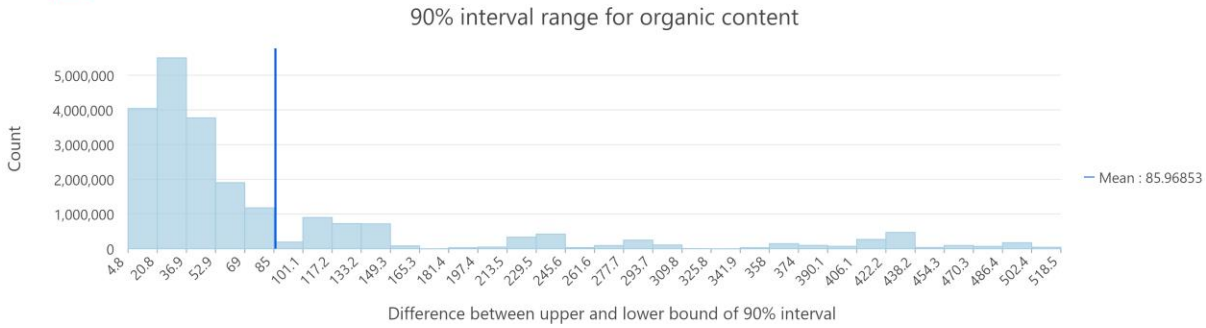
Uncertainty map for the prediction of Cation Exchange Capacity





Uncertainty map for the prediction of Organic Content





Appendix 7: S2GLC used land covers

Mineral extraction sites 7	Mineral extraction sites 7
Sport and leisure facilities 11	Sport and leisure facilities 11
Non-irrigated arable land 12	Non-irrigated arable land 12
Pastures 18	Pastures 18
Complex cultivation patterns 20	Complex cultivation patterns 20
Land principally occupied by agriculture, with significant areas of natural vegetation 21	Land principally occupied by agriculture, with significant areas of natural vegetation 21
Natural grasslands 26	Natural grasslands 26
Moors and Meathland 27	Moors and Meathland 27