Master’s thesis
Geography
Geoinformatics

PREDICTING SOIL ORGANIC CARBON AND NITROGEN CONTENT USING AIRBORNE LASER SCANNING IN THE TAITA HILLS, KENYA

Jesse Hietanen

2016

Supervisors:
Janne Heiskanen, Petri Pellikka

UNIVERSITY OF HELSINKI
DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY
DIVISION OF BIOGEOSCIENCES

PO Box 64 (Gustaf Hällströmin katu 2)
FIN-00014 University of Helsinki

Tiedekunta/Osasto – Fakultet/Sektion – Faculty Faculty of Science	Laitos – Institution – Department Department of Geosciences and Geography

Tekijä – Författare – Author
Jesse Olavi Hietanen

Työn nimi – Arbetets titel – Title
PREDICTING SOIL ORGANIC CARBON AND NITROGEN CONTENT USING AIRBORNE LASER SCANNING IN THE TAITA HILLS, KENYA

Oppiaine – Läroämne – Subject
Geography (Geoinformatics)

Työn laji – Arbetets art – Level Master's thesis	Aika – Datum – Month and year November 2016	Sivumäärä – Sidoantal – Number of pages 92

Tiivistelmä – Referat – Abstract

Reducing greenhouse gas emissions and increasing carbon sequestration is critical for climate change mitigation. With the emergence of carbon markets and the development of compensatory mechanisms such as Reducing Emissions from Deforestation and Degradation in Developing Countries (REDD+), there is much interest in measurement and monitoring of soil organic carbon (SOC). Detailed information on the distribution of SOC and other soil attributes, such as nitrogen (N), across the landscape is necessary in order to locate areas where carbon stocks can be increased and loss of soil carbon slowed down. SOC has large spatial variability, which often demands intensive sampling in the field. Airborne laser scanning (ALS) provides very accurate information about the topography and vegetation of the measured area, and hence, possible means for improving soil properties maps.

In this thesis, the aim was to study the feasibility of ALS and free of cost ancillary data for predicting SOC and N in a tropical study area. The study area is located in the Taita Hills, in South-Eastern Kenya, and has highly fluctuating topography ranging between 930–2187 m. Land cover in the Taita Hills is very heterogeneous and consists of forest, woodlands, agroforestry and croplands.

The field data consisted of SOC and N measurements for 150 sample plots (0.1 ha). The soil samples along with several other soil and vegetation attributes were collected in 2013. ALS (Optech ALTM 3100, mean return density 11.4 m^{-1}) data was acquired in February 2013. ALS data was pre-processed by classifying ground, low- and high vegetation, buildings and power wires. ALS point cloud was used to calculate two types of predictors for SOC and N: 1) topographical variables based on the high resolution digital terrain model (DTM) and 2) ALS metrics describing the vertical distribution and cover of vegetation. The ancillary datasets included spectral predictors based on Landsat 7 ETM+ time series and soil grids for Africa at 250 m resolution. In total, over 500 potential predictors were calculated for the modelling.

Random Forest model was constructed from the selected variables and model performance was analysed by comparing the predicted values to the field measurements. The best model for SOC had pseudo R^2 of 0.66 and relative root mean square error (RRMSE) of 30.98 %. Best model for N had pseudo R^2 of 0.43 and RRMSE of 32.14 %. Usage of Landsat time series as ancillary dataset improved the modelling results slightly. For SOC, the most important variables were tangential curvature, maximum intensity and Landsat band 2 (green). Finally, the best model was applied for mapping SOC and N in the study area. The results of this study are in line with other remote sensing studies modelling soil properties in Africa. The soil properties in the study area do not correlate strongly with present vegetation and topography leading to intermediate modelling results.

Avainsanat – Nyckelord – Keywords
ALS; Soil organic carbon; Remote sensing; GIS; Random Forest

Säilytyspaikka – Förvaringställe – Where deposited
Kumpula Campus Library, Helsinki, Finland

Muita tietoja – Övriga uppgifter – Additional information

Tiedekunta/Osasto – Fakultet/Sektion – Faculty	Laitos – Institution – Department
Matemaattis-luonnontieteellinen tiedekunta	Geotieteiden ja maantieteen laitos

Tekijä – Författare – Author
Jesse Olavi Hietanen

Työn nimi – Arbetets titel – Title
PREDICTING SOIL ORGANIC CARBON AND NITROGEN CONTENT USING AIRBORNE LASER SCANNING IN THE TAITA HILLS, KENYA

Oppiaine – Läroämne – Subject
Maantiede (Geoinformatiikka)

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Pro gradu -tutkielma	Marraskuu 2016	92

Tiivistelmä – Referat – Abstract

Kasvihuonepäästöjen vähentäminen ilmakehästä on kriittistä ilmastonmuutoksen hillitsemisen kannalta. Hiilimarkkinoiden ja erilaisten korvausmekanismien kehittyminen on lisännyt kiinnostusta maaperässä olevan orgaanisen hiilen mittaamiseen ja monitoroimiseen. Yksityiskohtainen tieto maaperän ominaisuuksista, kuten orgaanisen hiilen ja typen alueellisesta jakaumasta, voi auttaa löytämään alueita, joissa hiilen osuutta voidaan kasvattaa tai sen vähentymistä voidaan hidastaa. Maaperän hiilen vaihtelevasta spatiaalisesta jakaumasta johtuen kalliita kenttämittauksia tarvitaan runsaasti. Lentolaserkeilaus tarjoaa tarkkaa tietoa kuvatun alueen topografiasta ja kasvillisuudesta, mikä voisi olla hyödyllistä maaperän karttojen laadun parantamisessa.

Tämän tutkimuksen tavoitteena oli selvittää, miten lentolaserkeilaus ja vapaasti saatavilla oleva lisäaineisto soveltuvat maaperän orgaanisen hiilen ja typen pitoisuuksien ennustamiseen. Tutkimusalue on Taitavuorilla Kaakkois-Keniassa, jossa topografia on hyvin vaihtelevaa, korkeuden vaihdellessa 930 ja 2187 metrin välillä. Taitavuorten maanpeite on hyvin heterogeenistä ja koostuu metsistä, metsämaasta, peltometsäviljelysmaista ja viljelysmaista.

Tutkimuksessa käytetty kenttäaineisto koostuu 150:sta maaperän hiili- ja typpimittauksista 0.1 hehtaarin kokoisilta koealoilta. Maaperän mittaukset suoritettiin vuonna 2013. Lentolaserkeilausaineisto (Optech ALTM 3100) kuvattiin helmikuussa 2013. Kuvattu lentolaserkeilausaineisto esikäsiteltiin luokittelemalla maaperä, matala ja korkea kasvillisuus, rakennukset ja voimalinjat. Lentolaserkeilausaineistoa käytettiin kahden tyyppisten muuttujien laskennassa: 1) topografiamuuttujat, jotka laskettiin erittäin korkearesoluutioisesta korkeusmallista ja 2) kasvillisuuden vertikaalisesta rakenteesta ja peitosta kertoviin muuttujiin. Lisäaineistona analyysissä oli mukana spektraalista tietoa sisältävä Landsat ETM+ aikasarja, sekä maaperäruudukot Afrikasta 250 m:n spatiaalisella resoluutiolla. Yhteensä noin 500 muuttujaa laskettiin mallinnusta varten.

Random Forest -malli rakennettiin valituista muuttujista ja mallien suorituskykyä arvioitiin vertaamalla ennustettuja arvoja havaittuihin arvoihin. Parhaan maaperän hiilimallin valeselitysaste oli 0.66 ja suhteellinen keskivirhe 30.98 %. Parhaan typpimallin valeselitysaste oli 0.43 ja suhteellinen keskivirhe 32.14 %. Tärkeimmät muuttujat maaperän hiilen ennustamiseen olivat tangentiaalinen kaarevuus (tangential curvature), maksimi-intensiteetti (maximum intensity) ja Landsatin kanava 2 (vihreä aallonpituus). Landsat aineiston käyttö avustavana aineistona johti pieniin parannuksiin mallinnuksessa. Lopulta maaperän hiili- ja typpikartat ennustettiin käyttämällä parhaita löydettyjä malleja. Tämän tutkimuksen tulokset ovat linjassa muiden kaukokartoitusta hyödyntävien maaperän ominaisuuksia tutkivien tutkimuksien kanssa. Maaperän ominaisuudet eivät korreloineet voimakkaasti kasvillisuuden ja topografian kanssa, mikä johti keskinkertaisiin tuloksiin.

Avainsanat – Nyckelord – Keywords
Lentolaserkeilaus; Maaperän orgaaninen hiili; kaukokartoitus; paikkatieto; Random Forest

Säilytyspaikka – Förvaringställe – Where deposited
Kumpulan kampuskirjasto, Helsinki

Muita tietoja – Övriga uppgifter – Additional information

Contents

LIST OF ABBREVIATIONS:

ALS	Airborne laser scanning
C	Carbon
CO_2	Carbon dioxide
CHM	Canopy height model
DEM	Digital elevation model
DTM	Digital terrain model
DSM	Digital surface model
GIS	Geographic information systems
GNSS	Global navigation satellite system
GPS	Global positioning system
IMU	Inertial measurement
INS	Inertial navigation system
LDSF	Land degradation surveillance framework
LiDAR	Light detection and ranging
N	Nitrogen
RMSE	Root mean square error
RRMSE	Relative root mean square error
RF	Random Forest
RS	Remote Sensing
SOC	Soil organic carbon
SOM	Soil organic matter
VSURF	Variable selection using Random Forests

1. Introduction

Soil is the thinnest and outmost layer of the Earth's surface, of which all living ecosystems and terrestrial organisms are solely dependent on (Breemen & Buurman 2002; Brearley & Thomas 2015). Soil is a complex, dynamic and living ecosystem, formed under influence of microscopic and larger organism performing vital functions with combination of water, gases and parent material such as sediments and solid rock (Bot & Benites 2005; Breemen & Buurman 2002; Singh et al. 2011). Fertility of soils is crucial for all living organisms through, for example, food, fibre, animal feed and timber production (Brearley & Thomas 2015). Healthy soil is critical for plant productivity, promoting plant, animal and human well-being (Singh et al. 2011). Water and air quality, diversity of soil organisms and animals are also highly dependent on soil health and quality (Singh et al. 2011).

In global scale, soils are depleting with accelerated speed as improper and abusive management, land clearing, erosion, salinization, acidification, desertification, pollution and appropriation of land for other use are destroying soils all around the world (Breemen & Buurman 2002). Over the next decades, climate change and soil erosion can lead to severe challenges in global food security (Lal 2010; Amundson et al. 2015). According to Rozanov et al. (1993) during the past 10000 years more productive soils have been lost than there is currently being farmed, while UNEP (1986) calculated that up to 2 billion ha of fertile land has been irreversibly degraded since 1000 AD.

Climate change is a long term change in temperature, precipitation, wind and in weather conditions. Climate change is caused by increase of greenhouse gasses, such as carbon dioxide (CO_2), methane and nitrous oxides. Atmospheric concentration of CO_2 has increased by over 30% since 1750 (Lal 2004b; Carbon dioxide concentration 2016). Soils are an important part of climate change mitigation, as they contain remarkable amount of carbon (C) (Jobbágy & Jackson 2010). Up to the depth of one meter of soil, ~1500 petagram (Pg) of organic C has been estimated to be found globally (Post et al. 1982; Scharlemann et al. 2014). However, there is large variation on global estimates (Scharlemann et al. 2014). According to Scharlemann et al. (2014) considerable uncertainty and debate remains about carbon emissions and storage in terrestrial ecosystems. Lal (2003) estimated the soil erosion-induced CO_2 emission of 0.8-1.2 Pg C per year globally and Pan et al (2011) estimated soil C declination of 7.7 % between 1990 and 2007 in tropical regions. However, soils have high potential for sequestrating C. According to the recent estimates, the potential of soils to sequestrate C is around 30-60 Pg C in 25-50 years (Lal 2004b).

In the global scale, most of the soil organic carbon (SOC) is stored in northern regions (Scharlemann et al. 2014). SOC levels are mostly related to climatic factors; in the cold and wet climates, rates of photosynthesis exceed decomposition resulting in high levels of SOC (Ontl & Schulte 2012). Low primary production in the arid regions leads to low levels of SOC, while intermediate levels are usually found in the tropics (Ontl & Schulte 2012). In the regional to local scales, other factors such as topography and vegetation have a large effect on the SOC levels (Ontl & Schulte 2012). Except in Europe, USA and China, most of the current SOC maps are based on coarse resolution FAO soil maps from the 1970s, leading to uncertainty and inconsistency in the maps (Scharlemann et al. 2014). According to Lal (2003) most of the currently available statistics on the extent and severity of soil erosion are subjective, obsolete, crude and unreliable.

The scientific and political initiatives for reducing C emissions and enhancing C sequestration promotes the development of accurate, cost-effective and repeatable methods for soil monitoring and modelling. Mechanisms such as Reducing emissions from deforestation and forest degradations (REDD and REDD+) are aiming at creating value for C storage (UN-REDD 2011). These mechanisms are promoting sustainable forest and soil practices in local and landscape level, while leading to enhancements even in larger scales. Mapping the soil properties is currently mostly based on in situ measurements, which are costly and have limited spatial coverage (Mulder et al. 2011; Vågen & Winowiecki 2013).

Remote sensing (RS) and geographical information system (GIS) based methods have been used for modelling and monitoring environmental phenomena and variables at various scales. Open data and development of RS sensors has multiplied the possibilities for all types of modelling works. In the past years, growing number of research has been made on vegetation and C modelling using optical sensors, such as multispectral and hyperspectral images or active sensors, such as airborne laser scanning (ALS) and radars (Brewer et al. 2011). Utilizing GIS and RS for soil modelling has been done for couple of the past decades (Florinsky 2012a; Mulder et al. 2011). Adewopo et al. (2014) ranked soil information systems (SIS) as one of the top priority research questions for soil science in the 21st century. SIS is broadly defined as the combination of soil science with GIS (Adewopo et al. 2014).

Accurate and high quality soil maps are needed to identify the most suitable locations for soil development and protection, and to allocate sparse resources and funds optimally. Landscape level maps of SOC content and other soil properties, such as nitrogen (N) content, provide scientific basis for structuring agricultural development plans and prioritizing assets for protecting or restoring the soils (Mayes et al. 2014). Soil properties modelling and mapping in heterogeneous tropical regions, in landscape or local scale are relatively little studied field. Modelling heterogeneous and complex landscapes is difficult and complexity of the landscapes will increase in future due to fragmentation of forests and land covers (Vågen & Winowiecki 2013). High accuracy and resolution of ALS data could provide valuable proxy information for the soil properties modelling, thus leading to more accurate and higher resolution soil properties maps. However, modelling soil properties using ALS data is relatively little studied field (Kristensen 2015).

The main objective of this thesis was to the study feasibility of ALS data for predicting SOC and N content across a tropical forest-agricultural landscape mosaic. The study area was located in the Taita Hills, South-Eastern Kenya. Random Forest (RF) algorithm was used for modelling and predicting SOC and N content using ALS data and Landsat time-series and coarse scale soil grids as ancillary datasets. Most of the modelling and analysis work was performed using open source-tools by creating automated Python and R scripts. More specifically, this study aimed to:

1) Test the performance of RF regression to model SOC and N content using ALS data
2) Analyse the importance of ancillary datasets to soil SOC and N content modelling
3) Find predictor variables that best explain variation in landscape-level SOC and N content
4) Analyse the effect of spatial resolution of variables and feature extraction on modelling performance
5) Produce SOC and N maps for the Taita Hills study area

2. Background

2.1 Soil organic carbon and nitrogen

2.1.1 Chemical background

Carbon (C) is a chemical element, located sixth in the atomic table, with atomic weight of 12. It is the fourth most common element in the universe, compared by mass. The extra-ordinary ability of C, polymer-forming with the diversity of organic compounds, makes the C chemical basis for all known life (Hartemink & McSweeney 2014). Soil C is crucial for sustaining life on Earth; it is basis of life, energy, fibre, food and shelter (Brearley & Thomas 2015; Hartemink & McSweeney 2014).

N is a chemical element located seventh in the atomic table, with atomic weight of 14. N is highly versatile, and it can transform forms easily, usually being available to plants as ammonium or nitrate (Hall 2008; Lamb et al. 2014). N is one of the most important nutrients of soils affecting vegetation growth and crop production (Miransari et al 2012). Too high levels of N should not be supplied to soils, as over fertilization can lead to environmental issues and unnecessary costs (Ju et al. 2009).

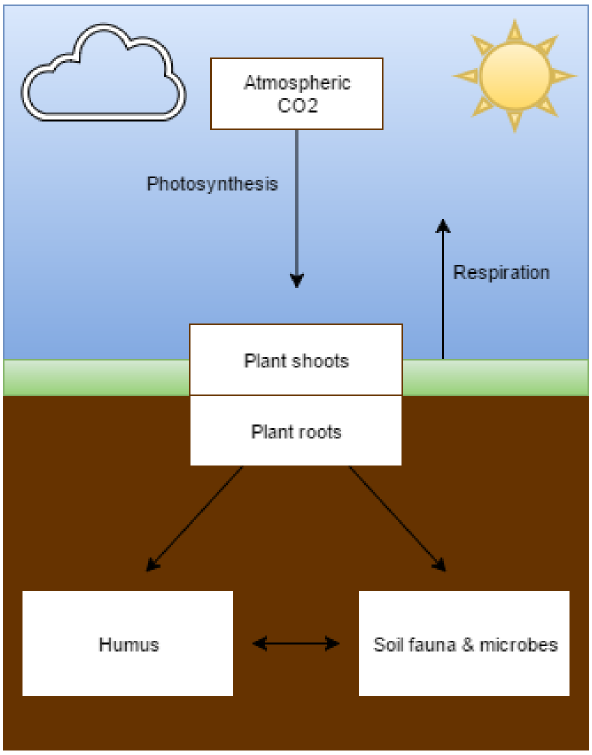


Figure 1. Photosynthesis, decomposition and respiration control the soil carbon balance (modified from Ontl & Schulte 2012)

Atmospheric CO_2 is converted to carbohydrates by the process of photosynthesis in vegetation (Figure 1). When the plants create litter or die, animals and other soil biota use the C compounds of the biomass (Hall 2008). Finally these C compounds are made available in the soil as humus, compost and different chemicals (Hall 2008: 22-23). When microbes decompose the biomass in soils, most of the CO_2 is released back to atmosphere (Ontl & Schulte 2012).

2.1.2 Global soil carbon pool

Global soil C pool is approximately 2500 Gt, making it larger than atmospheric (760 Gt) and biotic pool (560 Gt) together (Lal 2004a). Global soil C pool consists of 1550 Gt of SOC and 950 Gt of soil inorganic C (Post et al. 1982; Lal 2004a; Scharlemann et al. 2014). In the global scale, SOC has high spatial variability (Figure 2). In regional, landscape and local scales, other factors such as topography and vegetation can have large effect on the SOC levels (Florinsky 2012b; Ontl & Schulte 2012). On regional level, SOC concentrations have also high spatial variability, as for example seen in SOC map of Africa (Figure 3). The global SOC pool can be seen as dynamic equilibrium, consisting of inputs and losses to the SOC pools (Figure 4) (Lal 2004a).

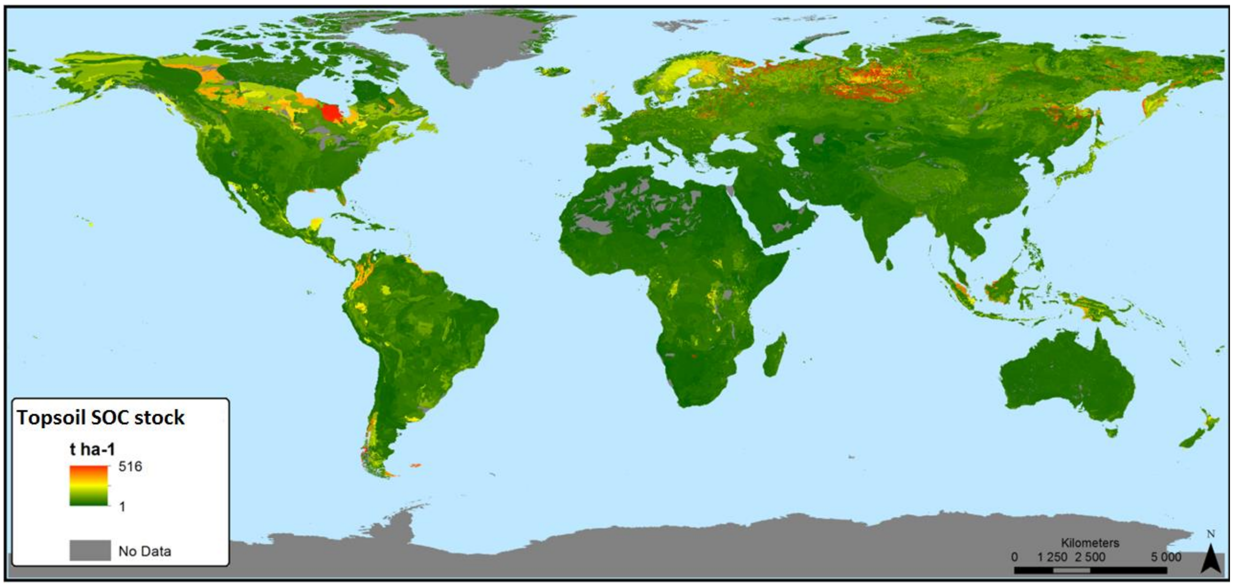


Figure 2. Global topsoil soil organic carbon (SOC) stocks have high spatial variability. Largest stocks are found in northern regions (data from Hiederer & Köchy 2011).

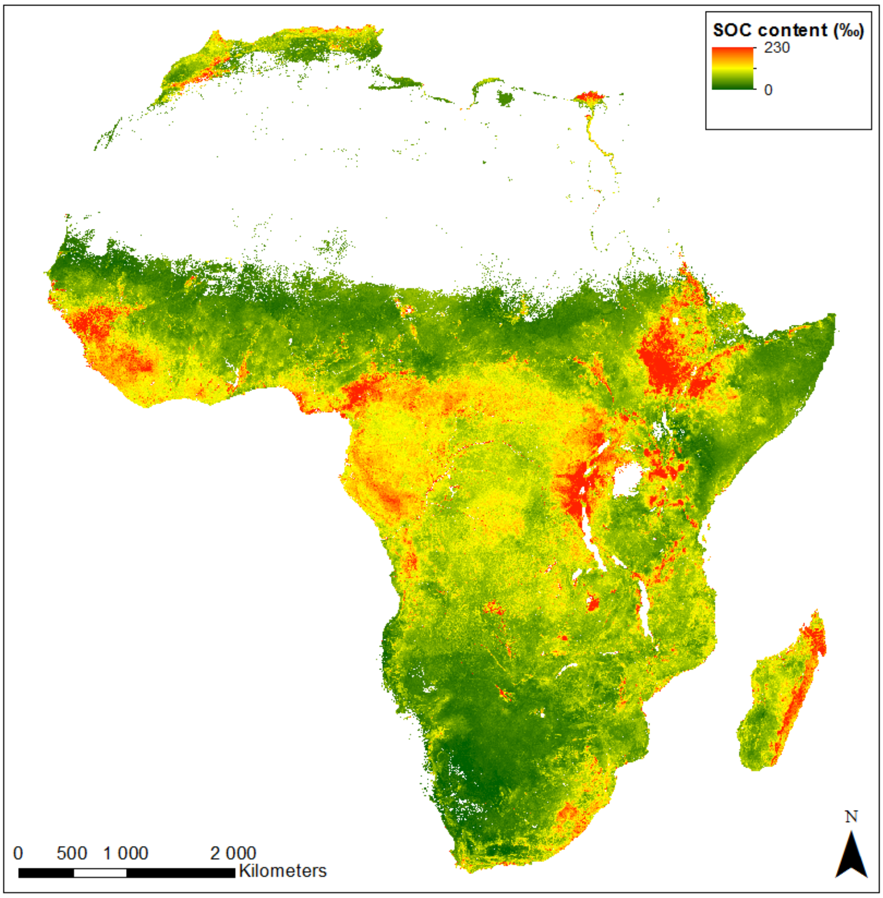


Figure 3. Topsoil soil organic carbon (SOC) content (‰) in Africa (data from Hengl et al. 2015)

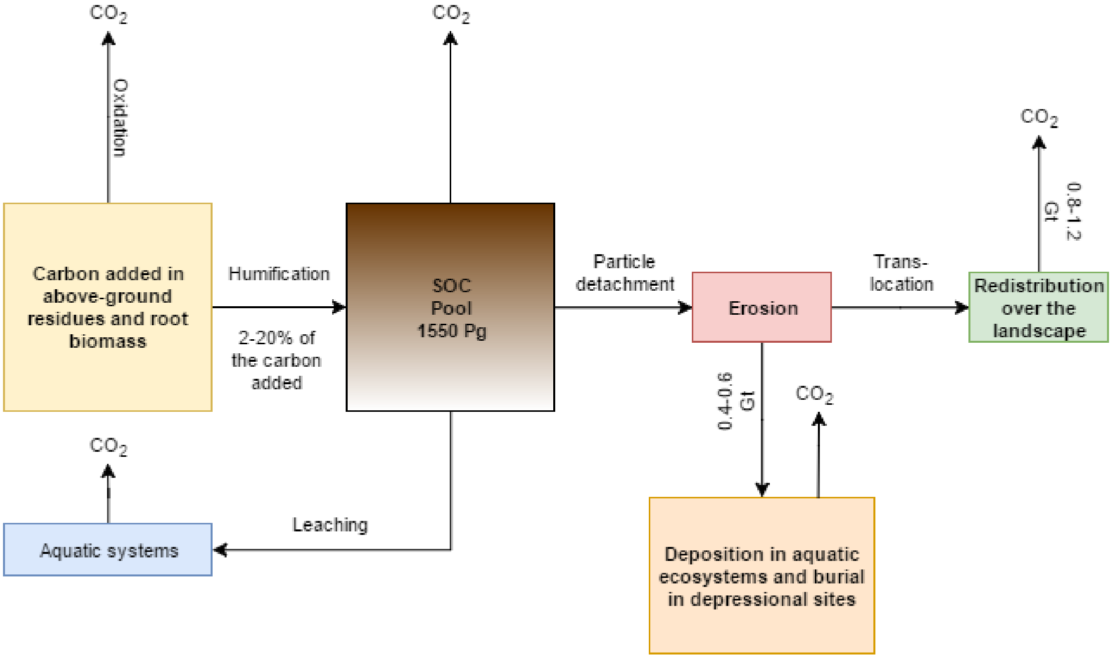

CO_2
Oxidation
CO_2
CO_2
0.8-1.2 Gt
Carbon added in above-ground residues and root biomass
Humification
2-20% of the carbon added
SOC Pool 1550 Pg
Particle detachment
Erosion
Trans-location
Redistribution over the landscape
0.4-0.6 Gt
CO_2
CO_2
Aquatic systems
Leaching
Deposition in aquatic ecosystems and burial in depressional sites

Figure 4. Global soil organic carbon dynamics are affected by several processes (modified from Lal 2004a).

2.1.3 Soil carbon sequestration

Soil carbon sequestration is a process where atmospheric CO_2 is removed from atmospheric circulation and stored in the soil C storages (Figure 4) (Lal 2004a). This process is mainly handled by vegetation through photosynthesis and decomposition (Hall 2008). Low levels of C sequestration can also happen in arid and semi-arid climates with minimal vegetation, through inorganic C formation, when CO_2 from air forms into secondary carbonates (Ontl & Schulte 2012).

Since the industrial revolutions, natural ecosystems have been converted to agricultural areas, resulting in vast depletion of SOC levels (Ontl & Schulte 2012). Lal (2004a) estimated losses of 60-90 Pg of C from soils to atmosphere. Soil C depletion has been caused mainly by reductions in the amount of plant root and residues returned to soil, increased decomposition from soil tillage and increased soil erosion (Lemus & Lal 2005).

Adoption of restorative land management practices (Table 1) can reduce the release of carbon dioxide to atmosphere and increase the soil C stocks (Lal 2004b). Potential soil C storage is affected by multiple factors, such as climatic controls, historic land use patterns, current land management strategies and topographic heterogeneity (Ontl & Schulte 2012). According to Lal (2004b) most of the depleted SOC stocks can be restored by converting only the marginal lands.

Table 1. Possible management practices for increasing soil organic carbon levels (modified from Ontl & Schulte 2012)

Management practice	Effect
Reduced tillage or no tillage	Reduced C loss
Erosion control	Reduced C loss
Addition of organic components (compost, crop residues)	Increased C input
Use of cover crops	Reduced C loss, increased C input

Continued increase in atmospheric CO_2 levels and rising of global temperatures has variety of consequences on soil C levels, which are still partly unknown, and uncertain (Ontl & Schulte 2012). Drake et al. (1997) found rising CO_2 levels to increase resource use efficiency in vegetation, leading to increased C consumption trough vegetation, thus producing more biomass. However, some studies also implicate that C loss might increase due to increased plant respiration from greater root biomass (Hungate et a. 1997). Increasing temperature could also lead to accelerated decomposition of soil organic matter (Pataki et al. 2003).

The resolution of soil maps from larger areas has been typically very coarse, making the analysis of local scale soil resources difficult (Scharlemann 2014). Especially in the tropics, the information about small scale changes in soil C stocks is needed to allocate the sparse funds to protect and develop the most important areas (Vågen & Winowiecki 2013). Current SOC models and databases are not able to predict and monitor the SOC levels and possible fate of SOC due to land-use change in scales relevant to management in developing countries (Vågen & Winowiecki 2013). The land cover change and land degradation are challenging problems especially in the developing countries. High resolution maps of soil C would provide possibilities to protect the most important soil C stocks.

Exact information about soil properties is necessary for environmental policy-making, resource management, targeting the management practices and monitoring the changes in environment (Mulder et al. 2011). Finding the areas with high SOC stocks combined with areas with high risk for soil erosion would be important when trying to find the best places to allocate funds to protect SOC stocks. As Vågen & Winowiecki (2013) explained, managing the soil erosion is one of the key strategies for reducing SOC losses. A better understanding of SOC stocks and flows is crucial for climate change mitigation and C management (Scharlemann 2014). Potential of the C sequestration can be determined when both historic SOC stocks under natural vegetation and current state are well understood and mapped (Vågen & Winowiecki 2013; Scharlemann 2014).

2.1.4 Environmental variables affecting soil properties

On global scale SOC levels are mostly related to climatic factors, such as temperature and rainfall (Ontl & Schulte 2013; Scharlemann 2014). However, on a local scale, environmental variables are somewhat different than the variables explaining global SOC levels. There is high variation in the important environmental variables, between different studies, scales, study areas and soil types (Powers & Schlesinger 2002). Thompson & Kolka (2005) concluded that relationships between soil properties and environmental variables are unique to soil property and environment.

Land use and land cover are one of the most important properties explaining the soil properties such as C and N (Islam & Weil 2000; Guo & Gifford 2002; Scharlemann 2014).

Emissions from land use and land cover change (LULCC) are the second largest anthropogenic source of C into the atmosphere (Scharlemann 2014). Studies are indicating SOC losses of 25-50 %, when conversion of vegetation to cropland (Scharlemann 2014). The losses are smaller from conversions of vegetation to pasture (Scharlemann 2014). Lal (2004b) estimated losses of 75 % or more in tropical soils, when converting from natural vegetation to cultivation. Similar results have been found for N (Islam & Weil 2000). However, Guo & Gifford (2002) found high variability on the loss and gain estimates regarding land cover changes on different studies. SOC and N levels and changes in both are also highly dependent on the soil types, as for example Mayes et al. (2014) found soil type being good explainer for SOC and N stock variability in landscape-scale. Type of soil can also have effect on the SOC losses on different land types. Scharlemann (2014) reported 20-40% losses on mineral soils when converting from forest to cropland, and even bigger losses on organic soils.

The soil properties levels are also influenced by local controls of the ecosystem processes (Ontl & Schulte 2012). Topographic based processes such as water infiltration, soil erosion and deposition of sediment and soil temperature can vary vastly in spatially heterogeneous landscapes, affecting the C and N input and loss rates in the soils (Seibert et al. 2007; Florinsky 2012b; Ontl & Schulte 2012). Florinsky et al. (2002) also noted the importance of temporal variability in soil-topogaphy relations, which can essentially influence the soil properties.

In general, most of the literature points that topography is one of the most important variables explaining spatial distributions of soil properties (Florinsky 2012b). Soil moisture, N and C are examples of soil properties that vary highly due to topographical conditions (Tsui et al. 2004; Wang et al. 2010; Florinsky 2012b; Ontl & Schulte 2012; Seid et al. 2013; Adhikari et al. 2014). According to Florinsky (2012b) topography influences the soil properties by spatially differentiating the temperature regimes of slopes and by the accumulation and movement of water or soil materials. Position and gradient of the soils in the landscape effect on levels of soil properties, such as SOC (Ontl & Schulte 2012). According to Tsui et al. (2013) movement of water and soil materials, leaching and degree of soil development are controlled by aspect and slope. Topography also affects the acquired solar radiation, thus affecting the soil properties (Wilson & Gallant 2000).

Accumulation of soil properties is usually found on bottom of the hills or areas with small slopes (Bot & Benite 2005). Topography influences the water balance, by controlling runoff and generating areas with more favourable water moisture levels, leading to differences in soil properties levels (Schwanghart & Jarmer 2011; Florinsky 2012b; Tsui et al. 2004). When evaporation and runoff area increases, the soil moisture reduces (Florinsky 2012b) and soil materials detached from rills erosion are transported by surface runoff (Schwanghart & Jarmer 2011). Good relationships between soil properties and topographic variables such as elevation, slope, curvature and several more complex indexes such as topographic wetness index (TWI) are reported by several studies (Florinsky 2012b; Schanghart & Jarmer 2011; Mulder et al 2011; Tsui et al. 2004; Adhikari et al. 2014). However, the importance of each topographic variable varies heavily between the studies.

Vegetation types have also important effect on soil properties. In general, highest SOC and N levels are found in natural or indigenous vegetation and lowest levels in areas with agriculture or areas without vegetation (e.g. Post et al. 1982; Guo & Gifford 2002; Lal 2004a; Wasige et al. 2014). Seid et al. (2013) studied how vegetation type affects the soil properties, and concluded that variability between vegetation classes can be high, but in general vegetation composition is useful when modelling soil properties. Maraseni & Pandey (2014) analysed SOC levels in five different forest types in Nepal and concluded that denser the canopy cover on forests, the higher amounts of SOC can be found. They also found levels of SOC to be related to presence of mixed species, N fixing trees and with the age of trees. Jobbágy & Jackson (2002) found a clear connection between biomes and SOC. Tsui et al. (2004) found that SOC concentrations in different slope levels are mainly explained by different litter decomposition conditions, which are explained by vegetation structure. Vegetation types are also important part of the land-cover change.

Soil type and depth has also effect on the explaining variables. Wang et al. (2010) found that effect of land use in SOC only significant in surface soils, and deeper soils were not that much affected by the land use and land cover changes. Topographic controls of soil properties also vary with depth, and the best results and correlations are usually achieved with the upper soil layers (Florinsky et al. 2002).

2.2 Remote sensing and airborne laser scanning

2.2.1 Principles

RS refers to collecting information without being in direct contact with the target. RS can be done in different scales and levels, thus sensors can be used for example on satellites, mounted on aircraft or unmanned aerial vehicles, cars, or in situ devices (NOAA 2015). There are two types of RS sensors: 1) passive and 2) active sensors (Figure 5). Passive sensor records the signal emitted or reflected by the surface, common source of energy being the sun. In active RS, the sensor sends the pulse/signal and records the signal it receives (Jensen 2000; NOAA 2015).

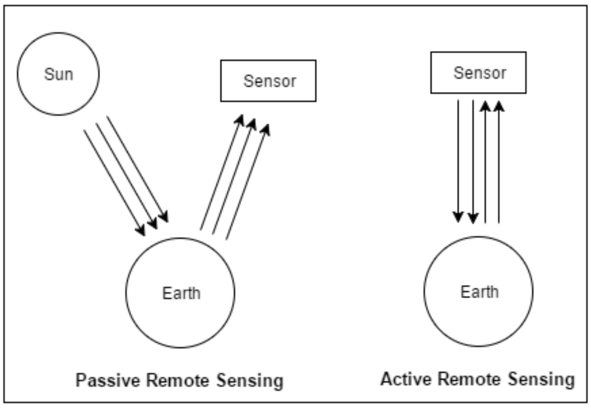


Figure 5. Passive and active remote sensing sensors.

Passive RS sensors are dependent on sun light, limiting the possible data collection conditions. Data collection can take place only during the time when the sun is illuminating the Earth, with exception for passive sensors measuring thermal- and microwave radiation (Jensen 2000; Natural Resources Canada 2015).

Active RS sensors are not dependent of sun light, thus data collection can be done during night. Active sensors can also be used for measuring wavelengths that are not sufficiently provided by the sun. Also better control about the illumination conditions of the target can be achieved (Jensen 2000; Natural Resources Canada 2015). ALS and synthetic aperture radar (SAR) are examples of active RS techniques.

2.2.2 Airborne laser scanning

2.2.3.1 History and terminology

During the last two decades, ALS has been stabilized as one of the most important methods for acquiring spatial information from the ground and vegetation (Petrie & Toth 2008). Increasing number of companies are designing and manufacturing laser scanning instruments, while private companies and public institutions use ALS for commercial and non-profit dedications. Instruments are complemented by a large selection of open source, free of cost and commercial software, intended for management, processing and visualization of the ALS data.

Well-developed instruments and software provide substantial means to use the data. High-quality 3D-points clouds can be used for example in forestry, urban planning, environmental modelling and many other purposes (Petrie & Toth 2008). ALS data can be also used for example erosion monitoring, vegetation modelling, topographic modelling, corridor mapping (e.g. roads, railway tracks), flood mapping, building extraction for 3D models, microclimate models and snow and ice-cover measurements (Wehr & Lohr 1999; Petrie & Toth 2008). Full potential of ALS has not been fully taken advantage of yet, as more understanding and information is needed. Finland has been one of the most successful countries using ALS in operational use, especially in forestry (Holopainen et al. 2013).

Most commonly used terms are the ALS and light detection and ranging (LiDAR), which both are based on the same basic principles. ALS includes always the positioning of the emitted pulse and determination of direction of the emitted pulse, as the LiDAR can also work without these. LiDAR can include methods such as profiling measurements (2D) and laser scanning (Petrie & Toth 2008). Profiling laser scanners detect only straight line below the flight line, as the laser scanning can detect wider angles (Holopainen et al. 2013). ALS data collection happens always from an airborne platform.

2.2.3.2 Principle of laser scanning

Laser scanning devices designed for measuring 3D surface, can be divided to two main categories by the method used: 1) light transit time estimation and 2) triangulation (Petrie & Toth 2008; Beraldin et al. 2010). In a certain medium, light waves always travels with a known velocity (Beraldin et al. 2010). Measurement of the distance is always based on the exact measurement of time. Some kind of laser-based ranging instrument, which can measure distance with high accuracy, is used as a basis for all laser ranging, profiling and scanning instruments (Petrie & Toth 2008; Beraldin et al. 2010).

Light transit time estimation based sensors calculate the distance by measuring the time delay for light to travel, from a source to a target and back to sensor (Figure 6). In continuous wave lasers, distance measurement can be indirectly calculated via phase measurement (Beraldin et al. 2010).

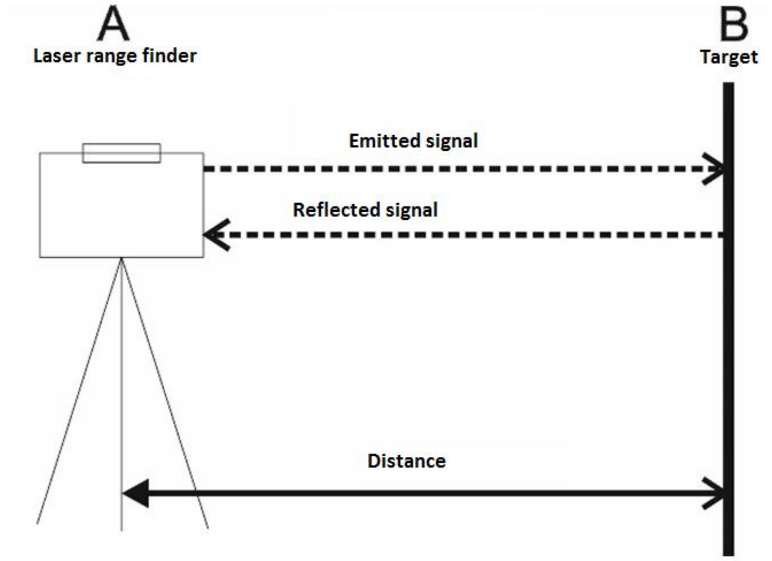


Figure 6. Basic principle of laser range finder. Laser range finder records the time between emitted and reflected signal and calculates the distance (modified from Petrie & Toth 2008).

To calculate the range or distance, following simplified formula can be applied:

$$p = \frac{c}{n}\frac{t}{2}$$

where c is speed of light in vacuum (299 792 485 m/s), t is round trip time from source to target and back, n is correction factor equal to refractive index, depending on air temperature, pressure and humidity (~1.00025) and p is the range between sensor and target (Beraldin et al. 2010).

Most of the ALS systems are able to capture multiple return echoes per one sent pulse, in case of certain site characteristics, such as vegetation canopies (Beraldin et al. 2010). The higher the sensor flies, the larger the area of one pulse is on the ground (footprint). Typically footprint of a laser beam is 0.3-3.8 meters, leading to sensor receiving couple of echoes back from the target from different heights (Figure 7) (Pyysalo 2000). Multiple echoes per pulse can be only detected if the distance between two echoes is larger than the half of the pulse width (Beraldin et al. 2010). For example, with pulse width of 5 ns, separate objects can be detected if their distance is larger than 0.75 m.

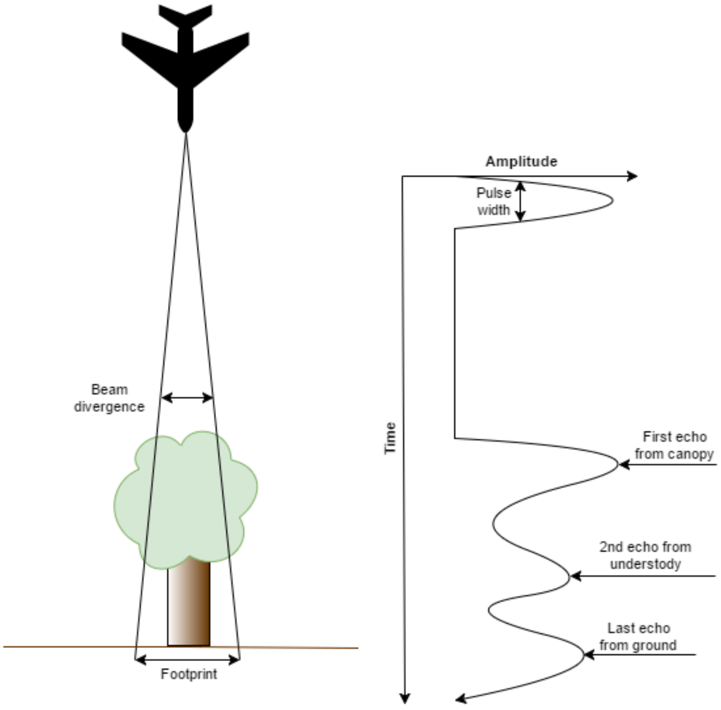


Figure 7. Most of the airborne laser scanning sensors can record multiple return echoes per one pulse.

2.2.3.3 Airborne laser scanning systems

To acquire densely sampled 3D measurement clouds, with a single laser source, a moving mechanism is needed to move the laser beam over the target surface. Moving the laser beam enables wider data collection and construction of the 3D surface (Beraldin et al. 2010). Typically these mechanisms are for example moving mirrors or prisms (Holopainen et al. 2013). Couple of different techniques are used in ALS sensors, such as Z-scanning, line based scanning and conical scanning. Different techniques have their pros and cons, though difference is usually seen as different ground patterns (Figure 8). Leica and Optech sensors are vibrating mirror based Z-scanning devices (Holopainen et al. 2013).

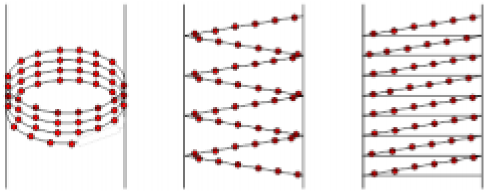


Figure 8. Different ground patterns of airborne laser scanning sensors (modified from Holopainen et al. 2013)

ALS system is usually made from two main components: the laser scanning system and Global Navigation Satellite System (GNSS) with Inertial Measurement Unit (IMU) unit. Laser scanning system calculates the distance to the ground while GNSS and IMU measure the exact position and orientation of the airplane (Beraldin et al. 2010; Holopainen et al. 2013). A control and data recording unit is also used for time synchronising, storing scanner, IMU and GNSS data and controlling the whole system (Beraldin et al. 2010). Typical construction of ALS system in airplane can be seen in Figure 9.

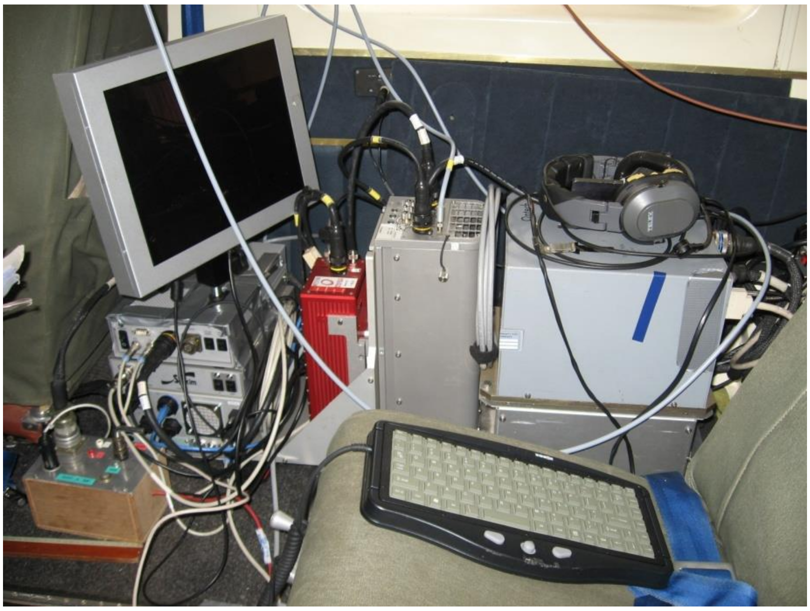


Figure 9. Optech ALTM 3100 ALS Sensor, AISA Eagle hyperspectral sensor and IMU system (photo by Tuure Takala)

2.2.3.4 Processing airborne laser scanning data to coordinate reference system

Raw data produced by the sensors on board, is usually not usable straight away. Depending on the purpose, several pre-processing steps are necessary. Calculating the complex and exact 3D point cloud is a combination of raw processing power and exact information produced by the sensors. Positional information collected by the GNSS is differentially corrected by using usually a minimum of one ground control devices (Lichti & Skaloud 2010).

During the data capture, laser scanner collects measurements in its own coordinate reference system while GNSS and IMU sensors collect the exact location and orientation data. By utilizing this information, the exact location of the laser scanning device in real coordinate reference system can be calculated up to accuracy of centimetres (Wehr 2008). Accurate measurements are combined using the exact timestamp produced by different sensors (Hovi 2013).

Direction vector of each laser pulse can be exported to world geodetic system 84 (WGS84) system by using coordinate system transformation matrices (Hovi 2013). This process is called registration (Lichti & Skaloud 2010) Example of a simplified transformation (Laser scanner, GNSS and IMU on same location and scanning angle directly below) can be calculated with the following equation:

$$G = r + s$$

where G is vector from centre of earth to the laser echo, r is vector from centre of earth to beginning of the laser beam and s is vector from begging of laser beam to the laser echo (Hovi 2013).

To be more accurate,

$$G_{wgs84} = r_{wgs84} + [\,]WGS84_H * [\,]IMU_H * [\,]IMU_L * s_{l,}$$

where [] are conversion matrices between coordinate systems and s_l is directional vector in laser scanner coordinate system (Hovi 2013). In reality, one should also take in account the scanning angle and distance between different sensors (Hovi 2013).

After the registration, different levels of calibrations are done to increase the accuracy of measurements (Lichti & Skaloud 2010). After the calibrations, strip adjustment is done to find and fix geometrical errors between flight lines (Lichti & Skaloud 2010). These errors are usually systematic, and known shapes such as buildings can be used to locate them (Hovi 2013).

2.2.3.5 Properties of airborne laser scanning data

Data collected by the laser scanner is usually called point cloud (Figure 10), as it is three dimensional and it contains vast amount of collected points. As the X and Y coordinate of each point cannot be stored as a regular grid, the exact location of X and Y is stored with the Z –value (Graham 2008). Performing basic operations, such as interpolation or searching for specific points in point clouds are more challenging and slower than for example in grid based datasets (Vosseman & Klein 2010). Though, there are benefits for storing the data as point cloud, as more complex shapes are retained, which can beneficial for different types of analyses and calculations (Vosselman & Klein 2010). Accuracy and level of detail usually depends on the used sensor and the flight height.

In addition to X, Y and Z measurements, several other parameters can measured and stored. Most of the modern sensors can record intensity, usually referred as intensity of the backscattered light or return strength of the pulse (Wehr & Lohr 1999). Intensity data has been mainly used on visualization and target classification purposes (Wehr & Lohr 1999; Kaasalainen et al. 2009). ALS intensity is however problematic, as each sensor provider has their own implementation of calculating it and calibrating intensity requires usually reference data (Kaasalainen et al. 2009; Korpela et al. 2010).

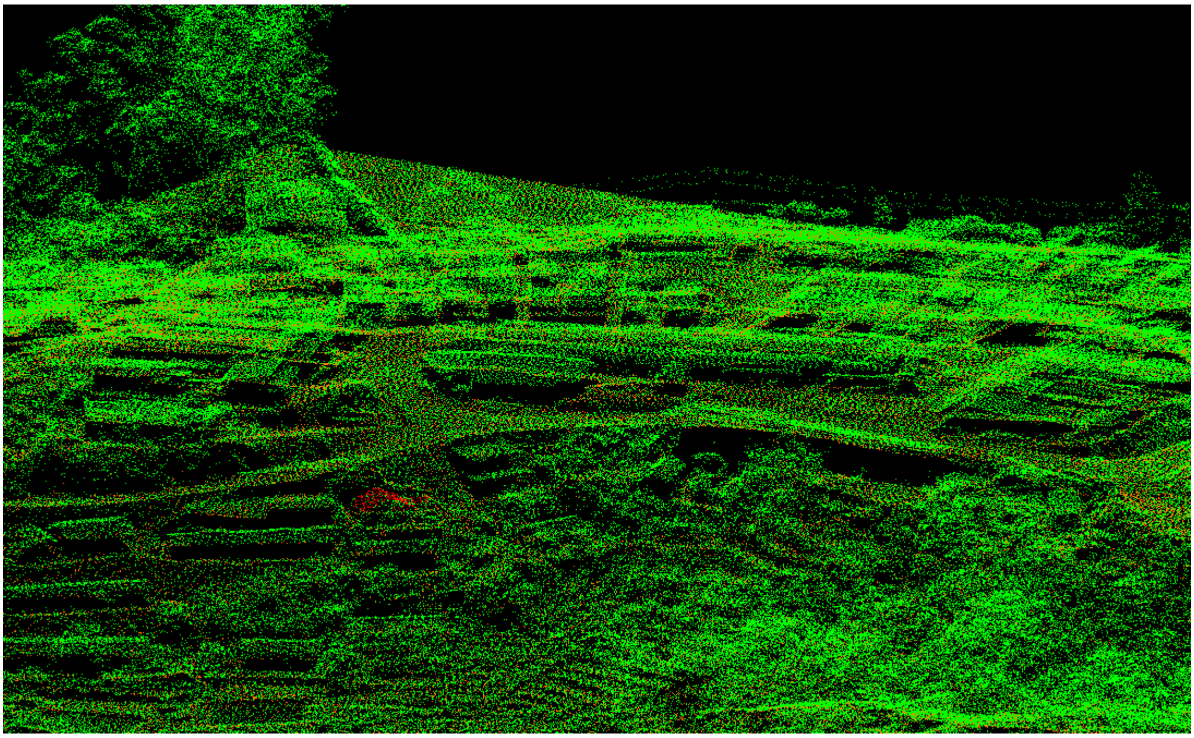


Figure 10. ALS data visualized as a point cloud.

There are several data formats than can be used to store the ALS data. During the past couple of years, extensive standardization work has been done and LAS-standard seems to be widely accepted through the community (Graham 2008). However, still many of the commercial software are implementing their own formats, to increase speed or reduce disk space usage, such as Fast Binary of TerraSolid and LasZip of LasTools. Most of the commercial and open source GIS and RS software can read the LAS format (Graham 2008).

2.2.3.6 Classification of airborne laser scanning data

After the pre-processing and calibration steps, ALS data can be classified. ALS point classification refers to assigning classification code for every point. For LAS formats, classification codes are standardized by the American Society for Photogrammetry and Remote Sensing (ASPRS).

Classification of the points usually starts with classifying the ground points, which is the bare minimum needed for separating other points from ground. Several automatic algorithms have been invented for ground point filtering (Briese 2010). After ground points have been filtered, other classes can be identified. Manmade objects such as buildings and wires can be identified by using algorithms based on geometric shapes or intensity values (Brenner 2010). Vegetation can be classified based on the return echo type (e.g. first of many echoes) or by for example classifying all points above the ground as vegetation (Maas 2010). Classification of different objects can be sometimes improved by using ancillary datasets, such as multispectral RS data (Brenner 2010). Level of classification needed is usually dependent on the purpose of the output dataset. Different use cases provide different needs for the level of classification.

2.2.3.7 Digital terrain model, digital surface model and canopy height model

Digital Terrain Model (DTM), Digital Surface Model (DSM) and Canopy Height Model (CHM) are concepts for digital representations of earth surface. Digital Elevation Model (DEM) is usually seen as umbrella term to include all representation of ground, however these terms can be context and country specific (Oksanen 2006). DTM usually refers to a digital grid based model having elevation above sea level as the value for bare ground (Figure 11). DSM (Figure 12) is similar to DTM in other aspects, but it includes the vegetation and possibly manmade objects, such as buildings (Briese 2010). CHM refers to a model, which values are elevation above the ground, instead of elevation above sea level, as DTM and DSM have. CHM is also sometimes referred as normalized digital surface model (nDSM) (Briese 2010). These models can be derived from different data sources, such as ALS, terrestrial laser scanning, radars and tachymetric measurements. This study concentrates only on models generated from ALS data.

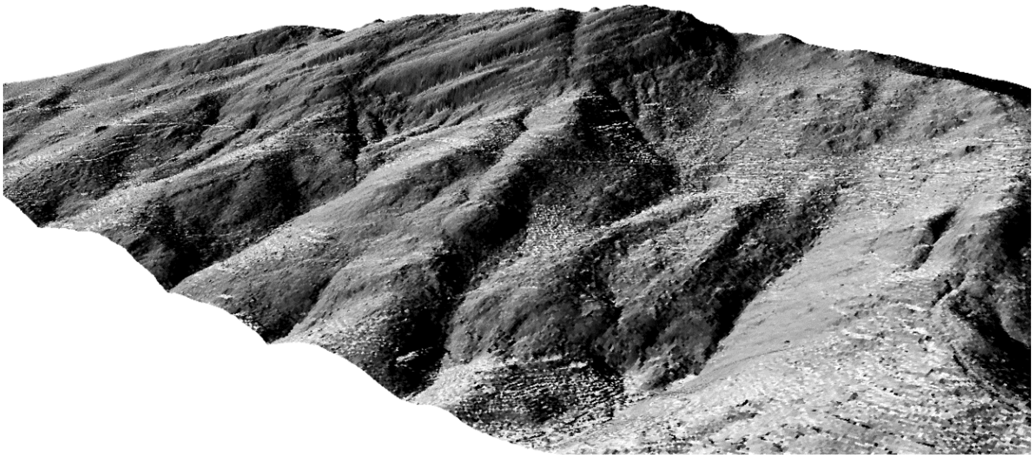


Figure 11. 3D view of hill shaded DTM showing only bare ground in the Taita Hills.

For generating accurate and high quality DTM from ALS data, non-ground points must be carefully filtered out. Typical approach is to use only last echo points or only echoes, but more sophisticated approaches have been developed, such as segment based filtering, surface based filtering, progressive densification and morphological filtering (Briese 2010). In general, most automated filters work well on areas with low complexity, but areas with more complex terrain and vegetation tend to be challenging for automated tools (Briese 2010).

After the ground point classification, classified points are used to generate the DTM, which can be a full 3D presentation, or more typical 2.5D raster representation (Briese 2010). Raster models are usually generated using different interpolation methods or using triangulation (Axelsson 2000; Briese 2010)

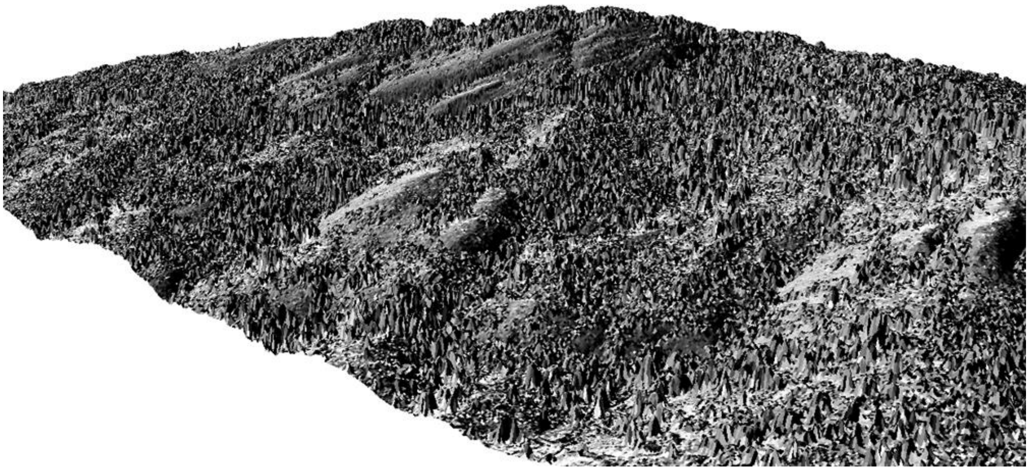


Figure 12. 3D view of hill shaded DSM showing bare ground with vegetation in the Taita Hills.

DSM models are usually generated similarly, but only based on all first echoes (Briese 2010). CHM (Figure 13) can be calculated directly from height-normalized ALS points or by subtracting DTM from DSM (Isenburg 2014).

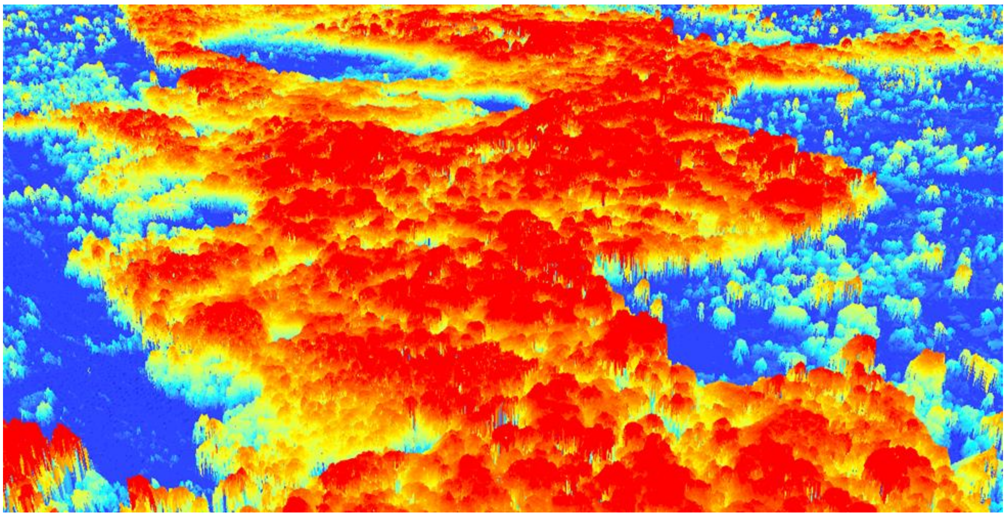


Figure 13. 3D view of CHM showing normalized height (height from ground) for non-ground objects, such as vegetation.

2.2.3.8 Airborne laser scanning based variables for environmental modelling

ALS data can be used to generate vast amount of variables for environmental modelling, including prediction vegetation and soil attributes. Variables calculated from ALS can be roughly divided to two groups, topographic variables generated from ALS based DTM models and vegetation variables generated from the point cloud.

Topographic variables are usually generated from the ALS based DTM. Calculation of topographic variables from DTMs is usually very well integrated into all GIS tools. Basic variables such as slope, aspect and curvatures can be calculated with almost any tool, and more rare variables such as topographic wetness index, topographic position index can be calculated with several open-source and commercial tools. The accuracy and quality of ALS based topographic variables are usually superior compared to variables calculated from other DTM sources (Briese 2010).

Derived vegetation parameters can be used for example measuring canopy cover, tree height, crown diameter, tree density, biomass estimations and determination of forest borders (Wehr & Lohr 1999; Maas 2010). DSM and CHM are also representations of vegetation, as they contain the height of the object from ground. Several tools and software are available for vegetation variable extraction for ALS data. Tools such as FUSION (McGaughey 2016), LasTools (Isenburg 2016), GRASS (GRASS Development Team 2016) and ArcGIS (ESRI 2014) can be used for extraction of vegetation variables. Most of the tools are also capable for deriving ALS intensity variables.

2.3 Remote sensing of soil properties and digital soil mapping

As explained in previous sections, soil properties are in global scale, highly related to climate and rainfall in global scale. In more regional scale, factors like topography and vegetation cause the variation in the soil properties levels. RS and GIS provides methods and data sources for modelling most these variables as proxies of soil properties (McBratney et al. 2003).

The most important aspect of RS and spatial modelling is the capability to reduce the need of soil sampling in the field and improve the accuracies of current databases and maps. RS and GIS has already been recognized as potentially cost-efficient technology in digital soil mapping; however it is not yet routinely used (Mulder et al. 2011). According to several scientists, the knowledge on how to use RS in soil mapping is still incomplete and lots of uncertainties remain (Mulder et al. 2011).

Traditionally RS imagery have been classified to land cover or soil cover maps and used as background information when planning the sampling of different soil landscape-units (Mulder et al. 2011). The rapid development of RS sensors has enhanced possibilities for soil properties mapping and creating more reliable global and regional soil databases (Mulder et al. 2011). Digital soil mapping is strongly based on the availability of covariates explaining the spatial patterns of soil properties (Callant & Austin 2015). Topographic variables can be derived from DTM models, while high-resolution RS data provides spectral information about soil and vegetation (Seid et al. 2013).

One of the most interesting approaches in soil properties mapping is the usage of ALS. ALS provides accurate information about the vegetation and terrain structure (Li e al. 2016). ALS has the ability to produce accurate measurements of vegetation structure such as volumetric forest properties and species composition, which all are linked to soil C stocks (Kristensen et al. 2015). DTMs calculated from ALS data, which are used to calculated topographic variables, are in general thought to be very accurate (Briese 2010). Laser scanning sensors have developed rapidly in the past years, increasing the pulse density and accuracy and reducing the total costs. However, the use of ALS is relatively understudied field in soil properties mapping (Kristensen et al. 2015). ALS could provide information and proxies that the passive RS sensors cannot, however combining ALS data with optical RS data could be beneficial approach.

The problem with direct RS of soil properties is that the passive sensors cannot directly access the bare ground or soil. However, in areas where with little or no vegetation (e.g. agricultural lands) good results have been achieved using multispectral and hyperspectral RS data (McBratney et al. 2003; Mulder at al. 2011). In densely vegetated areas, soil analysis has been typically relied on indirect proxies of soil properties, such as topographic variables and vegetation, which can be derived indirectly or directly from RS data (Daughtry et al. 2012; Mulder at al. 2011). However, mixed results have been achieved using indirect proxies for soil properties derivation (Mulder et al. 2011). Recent development of active and passive RS sensors and opening of data has created possibilities to map many of the important biophysical variables of vegetation and soil with reasonable costs and high enough accuracy. Especially the development of spaceborne sensors should improve the retrievals of soil based information at larger scales, with good price quality efficiency (Mulder et al. 2011).

Multispectral and hyperspectral data has been widely used in soil properties modelling in several studies. Gomez et al. (2008) used hyperspectral Hyperion-1 data with very good results in Australian agricultural areas, promoting the potential of RS for soil mapping. Vågen & Winowiecki (2013) explored approaches using moderate to high resolution satellite imagery in soil C assessing in three East African countries. They achieved 0.67 and 0.65 coefficient of determination (R^2) values using Landsat ETM+ ground reflectance images. Mirzaee et al. (2016) estimated soil organic matter (SOM) using geostatistical methods, such as kriging, using Landsat 7 ETM+ data. Based on the results, authors concluded that ancillary datasets such as Landsat 7 ETM+ are very important for improving the SOM estimates.

In addition to pure RS data, DTMs based on radars or ALS are very important for soil properties modelling. DTMs can be used to calculate the topographic variables such as slope, aspect and curvatures and more complex indices such as topographic wetness index or topographic position index (Florinsky et al. 2002; McBratney et al. 2003; Mulder et al. 2011). As an example, Seid et al. 2013 analysed soil properties using RS and soil-landscape modelling techniques in Ethiopian watershed. According to the authors, the use of accurate DTM and RS data with minimum field data provides alternative source to capture the spatially continuous soil attributes.

There are several methodological approaches to model the soil properties. Different modelling techniques such as kriging, regression or decisions trees have been used with successful results (Mulder et al. 2011). Bui et al. (2006) used data mining technologies for predicting pH, organic C, total N and total phosphorus with good results. Piecewise linear tree models were built to choose variables from the large amount of input data, including climate variables, DEM based variables, Landsat bands, land use and lithology maps. Yang et al. (2016) used RF with environmental covariates to model the soil depth functions and SOC stocks with relatively good results.

In general, there seems to be no clear consensus on best available method for soil properties mapping using RS and GIS-tools. Modelling methods varied between studies, and most of them seemed to perform relatively well. Important variables for SOC and N modelling differed between study areas and used models, indicating high dependency on the local environment for soil modelling. However, combining several datasets has been done successfully on several studies.

3. Study area

The Taita Hills (03 º 25’ S, 38 º 20’ E) are located in South-Eastern Kenya, near Tanzanian border (Figure 14). The Taita Hills are part of coastal province and the Taita Taveta district. The district consists of two topographically diverse areas, lowlands of the Tsavo plains and the mountainous Taita Hills. The hills are in the middle of Tsavo national park and Tsavo plains (Geography of Taita 2006). The Taita Hills cover approximately an area of 1000 km^2 (Pellikka et al. 2013). The study area of this thesis was 10 km × 10 km in size and located in the higher parts of the hills (Figure 14).

Capital of the district is Wundanyi, a small agricultural and trading centre in the region. Population of the Taita Hills has doubled within 30 years, increasing the pressure on land and environment (Geography of Taita 2006). Main livelihood of the local people (78 %) is intensive agriculture. Unemployment remains a critical issue as about 44 % of the total labour force remains unemployed (Geography of Taita 2006).

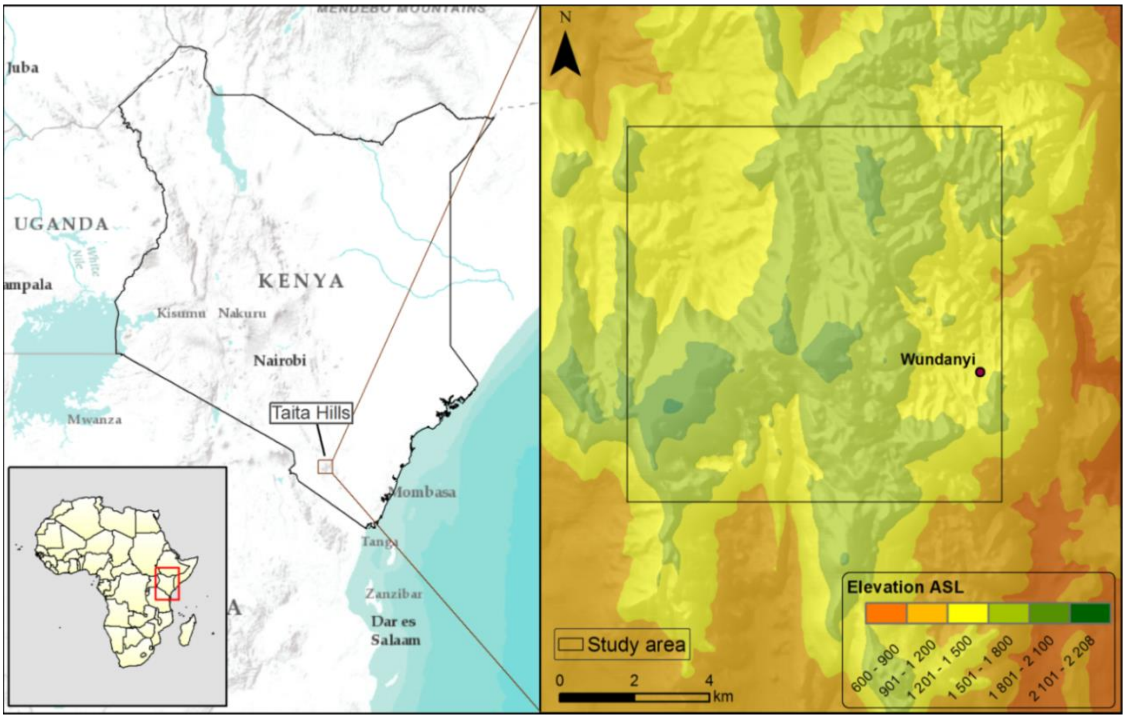


Figure 14. Study area of this thesis is located in the Taita Hills, Kenya.

Closeness to the equator can be seen on climate, as rainy seasons take place from March to June and October to December. The variability in the amount of rain can be high depending on the year. The mean annual rainfall varies from 500 mm in the lowlands to 1500 mm in higher elevations (Geography of Taita 2006). Closeness of Indian Ocean brings orographic rains to the Hills and cloud and moist precipitation occurs throughout the whole year (Pellikka et al. 2013).

The Taita Hills are part of the Eastern Arc Mountains, which are considered one of the most important biodiversity hotspots in the World (Myers et al. 2000). Only few indigenous mountain rain forest fragments are left in the hills. With the latest estimates, only 1% of the original forested areas remain preserved (Pellikka et al. 2009). Most of the forest loss can be explained with human population growth and conversion of land for agriculture (Pellikka et al. 2009). These rain forest fragments have variety of threatened and endemic fauna and flora, which cannot be found elsewhere in the World (Geography of Taita 2006. Current threats to the forests and biodiversity are harvesting of fuel wood, grazing of cattle in the forests and invasive exotic tree species (Thijs et al. 2014).

The Taita Hills are characterized by highly fluctuating topography, ranging between 500 m – 2200 m above sea level (Pellikka et al. 2013). The highest peak (2208 m) in the Taita Hills is

Vuria (Figure 15). The hills are part of Precambrian mountains of Eastern Arc chain (Geography of Taita 2006).

Figure 15. View from Vuria mountain (photo by Jesse Hietanen).

Geology of the Taita Hills consists of mainly undifferentiated basement rocks and the main soil types in the Taita Hills are humic cambisols, acrisols, ferrasols and rankers (Jaetzold et al. 2010). However, soil types reported in Soil atlas of Africa (Jones et al. 2013) are slightly different. Figure 16 shows the soil types of the study area and the Taita Hills based on the Soil atlas of Africa data (Jones et al. 2013). Fertility of the soil in the Taita Hills varies between the soil type (Jaetzold et al. 2010).

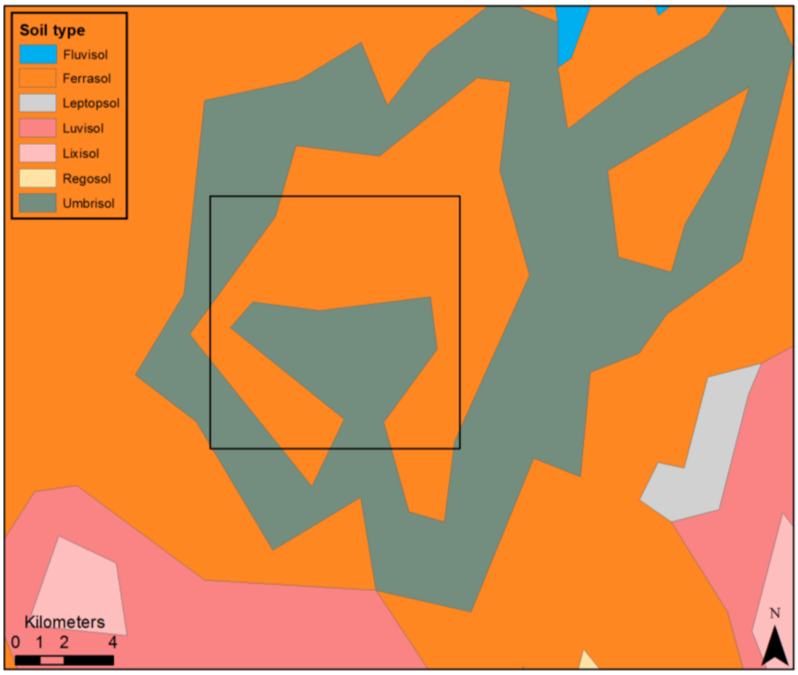


Figure 16. Soil types of the Taita Hills (data from Jones et al. 2013)

Soils of the the Taita Hills are also vulnerable for environmental change. The risk of soil erosion is high due to poor agricultural management, erodible soils, rough topography and land cover change (Erdogan et al. 2011; Pellikka et al. 2013). Erdogan et al. (2011) estimated that soil-loss potential increased from 7 % to 12 % between 1987 and 2003.

Very little information about soil properties of the study area is available. Several continental datasets have been calculated, which include SOC, however, the spatial resolution and level of detail tends to be coarse. Figure 17 shows the topsoil SOC stock map of Kenya and Taita Hills, based on the Soil atlas of Africa (Jones et al. 2013) dataset, with spatial resolution of 1 km. Figure 18 shows the topsoil SOC content map of the Taita Hills, based on African soil grids (Hengl et al. 2015), with spatial resolution of 250 m. Spatial resolution of the African soil grids is clearly superior compared to the Soil atlas of Africa, however still being relatively coarse.

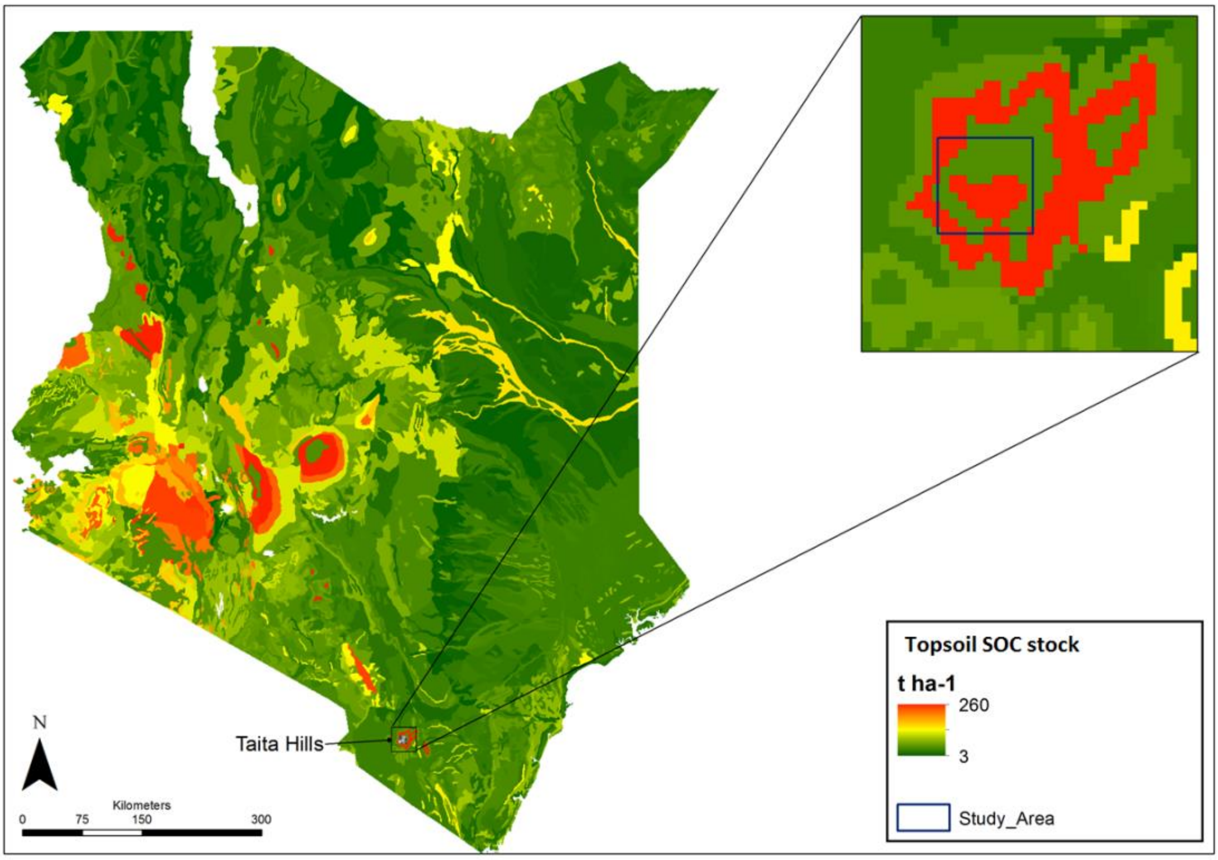


Figure 17. Topsoil soil organic carbon (SOC) stock map of Kenya and the Taita Hills (data from Jones et al. 2013).

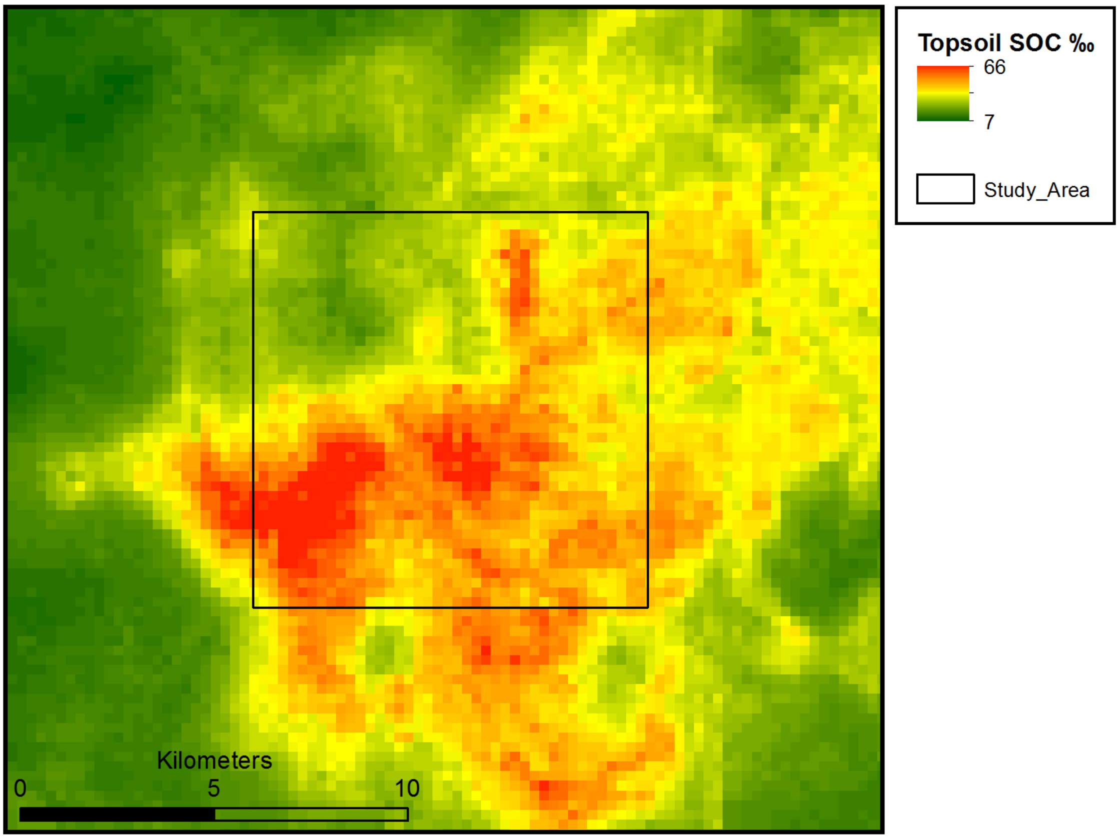


Figure 18. Topsoil soil organic carbon (SOC) content map of the Taita Hills (data from Hengl et al. 2015)

Omoro et al. (2013) studied the SOC densities in differerent forest types of the Taita Hills and found out that SOC densities were generally lower in the plantations than in the indigenous forests. Authors concluded that this could be due to different litter conditions, or due to insufficient time for SOC levels to recover in plantations.

4. Material

4.1 Data sources

Datasets used in this study are listed in Table 2. Field measurements were used to derive response variables for in the modelling. All other datasets were used as a source of the predictor variables. ALS data was the primary source for independent variables. Landsat time series and African soil grids were used as ancillary datasets due to possible complementary information on ALS data. Ancillary datasets are available free of cost for the whole of Africa. DTM, DSM, CHM and ALS datasets were also used for visualization purposes.

Table 2. Datasets used in this study.

Dataset	Date of Collection	Coverage	Source
Field measurement	2013	Study area	This study
Airborne Laser Scanning	2013	Study area	This study
Landsat time series	2012 - 2013	South Eastern Kenya *	Adhikari et al. 2016
African Soil Grids	2008 - 2014	Africa	Hengl et al. 2015

* L7 Path 167, Row 62

4.2 Field measurements

Field data was collected between January and February 2013 (locations in Figure 19). Hierarchical field survey and sampling strategy was based on land degradation surveillance framework (LDSF) guidelines (Vågen et al. 2015). Study area (10 km × 10 km), was divided to 16 tiles (2.5 km × 2.5 km). For each of these tiles, a random centroid was generated for the clusters of study plots. Each cluster consists of 10 plots. Centre location for each plot was randomized, falling within a 564 m radius from the cluster centroid. Each plot was 0.1 ha in size, and consisted of 4 subplots (0.01 ha in size) (Figure 20).

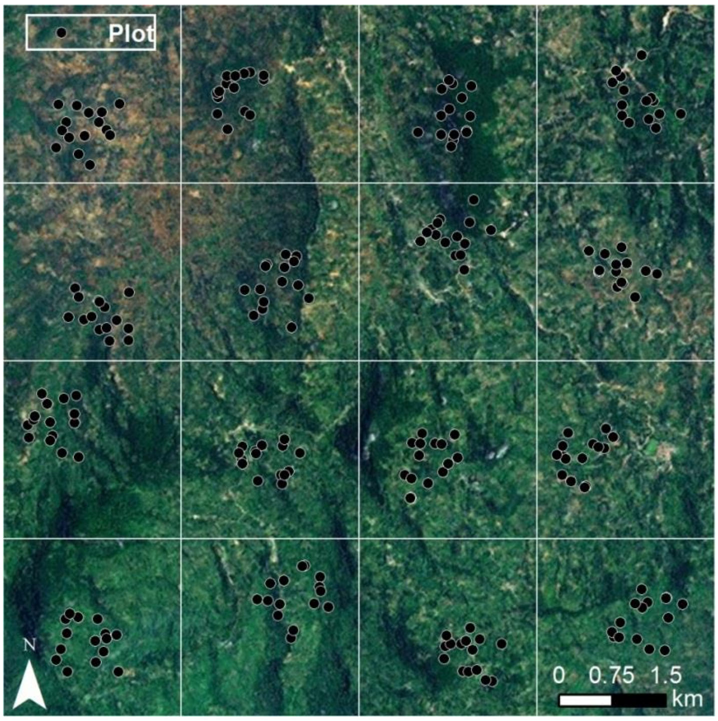


Figure 19. Location of the field plots in the study area.

Navigation to the study plot and centre position measurement in the field was done by using a consumer-grade GPS receiver (Trimble Juno 3B). Positional accuracy reported by the GPS device was in average 2.3 m, ranging between 1.6 m and 8.7 m. However, positional accuracy was not recorded for all study plots.

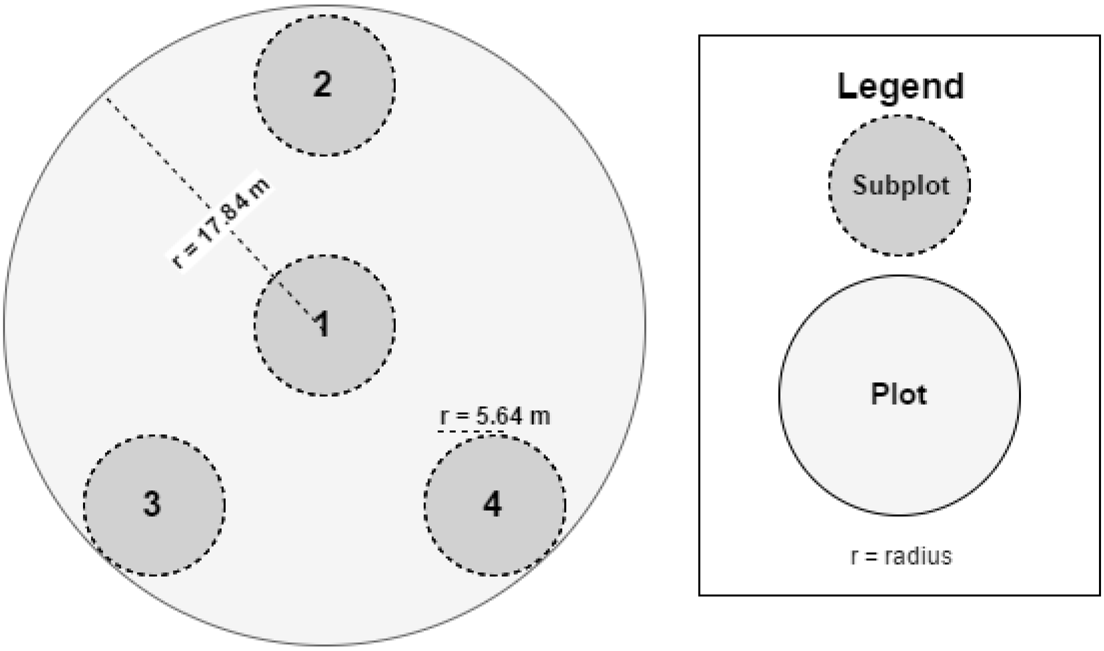


Figure 20. Study plot and subplots (modified from Vågen et al. 2015).

Soil samples were collected at the centre of each subplot, topsoil samples from 0-20 cm depth and subsoil samples from 20-50 cm depth using an auger. Soil samples from each subplot were pooled thoroughly into one sample, representing the soil characteristics of one plot (Vågen et al. 2015). Soil samples were delivered to World Agroforestry Organization headquarters (Nairobi, Kenya) for the soil properties analysis. Before the analysis, soil samples were dried and sieved through 2 mm sieve. SOC and N concentrations were calculated for each plot, by utilizing thermal oxidation method. In this study, only topsoil samples were used. Due to issues in field data collection or in analysis, topsoil samples were collected only from 150 unique study plots (Table 3). Samples with errors in positioning or in data collection were discarded from the analysis.

Table 3. Summary statistics for topsoil samples.

Statistics	SOC (%)	N (%)
Mean	2.093	0.170
Min	0.749	0.043
Max	6.759	0.479
Standard Deviation	1.111	0.076

4.3 Airborne laser scanning

ALS data was acquired by German operator TopScan GmBH in January 2013 using Optech ALTM 3100 sensor. Mean point density for the LiDAR data was 9.6 points/m^2. First, last and maximum of two intermediate pulses were recorded, including intensity. The dataset was pre-processed by the operator and delivered in LAS1.2 format. Detailed parameters and summary of the ALS data are given in Table 4.

Table 4. ALS survey and sensor specifications.

Parameter	Value
Date of acquisition	4–5 February, 2013
Sensor	Optech ALTM 3100
Flying height (m AGL)	213–1168 (760)
Range (m)	216–1170 (764)
Flying speed (knots)	116–126
Pulse rate (kHz)	100
Scan rate (Hz)	36
Scan angle (degrees)	±16
Pulse density (pulses m−2)	9.6
Return density (returns m−2)	11.4
Maximum number of returns per pulse	4
Beam divergence at 1/e2 (mrad)	0.3
Footprint diameter (cm)	6–35 (23)

4.4 Ancillary data

4.4.1 Landsat time series

Landsat time series dataset used in this study was generated by Adhikari et al. (2016). Adhikari et al. (2016) used 17 Landsat 7 Enhanced Thematic Mapper Plus (ETM+) images, from between June 2012 and October 2013. Time period was defined to start eight months before and to stop eight months after ALS data collection (January 2013).

Adhikari et al. (2016) calculated topographic correction to all images using Shuttle Radar Topography Mission (SRTM) DEM. Five different vegetation indices were calculated: 1) normalized difference vegetation index (NDVI), 2) reduced simple ratio (RSR), 3) brightness, 4) greenness and 5) wetness. Seasonal features were computed for all layers based on the statistical distribution of annual vegetation index values (Adhikari et al. 2016).

In this study, all seasonal features of vegetation indices and separate Landsat bands were used in the modelling process. However, only 50 % percentile of the seasonal features was included in this study. Due to issues in SRTM data, there were gaps in the produced Landsat images.

4.4.2 African soil grids

African soil grid dataset used in this study was provided by the Hengl et al. (2015). Hengl et al. (2015) generated soil properties maps (e.g. organic carbon, pH, sand and silt content) in spatial resolution of 250 m for whole of Africa. Datasets used for creating these maps were from two different databases: 1) the Africa Soil Profiles (legacy) and 2) Africa Soil Information Service (AfSIS). Authors used RF for predicting the values based on the input soil databases, MODIS, SRTM DEM based topographic variables, GlobeLand30 land covers and 1 km soil grids. In total over 28000 sampling locations were used (Hengl et al. 2015).

African soil grid data for this study was downloaded from ICRIS web mapping server (www.soilgrids.com), using web coverage service standard. After the data was downloaded, top- and subsoil layers were combined to match the top- and subsoil definitions used in this study.

5. Methods

5.1 Overview

The main steps of the methodology are summarized in Figure 21. First, ALS data was pre-processed, which included manual quality inspection, improvements to ground classification, vegetation, wire and building classification and removal of faulty points. Pre-processed ALS data was used to calculate DTM, DSM and CHM with spatial resolution of 1 m. These models were resampled to 5 m, 10 m, 25 m, 50 m and 100 m spatial resolutions. Resampled models were used to calculate several topographic variables. Pre-processing of the field data included conversion to shapefiles and buffering of the point locations with radii of 17.84 m (0.1 ha), 25.23 m (0.2 ha), 35.68 m (0.4 ha), 50.46 m (0.8 ha) and 71.37 m (1.6 ha).

Feature extraction (Zonal Statistics) was performed for the calculated topographic variables, Landsat time series and African soil grids, using the five different shapefiles with different study plot sizes. ALS based vegetation and intensity variables were calculated for the same locations using the same study plot sizes. Vegetation and intensity variables were joined to the shapefiles by study plot ids and sizes. Vegetation and intensity variables calculated with

0.1 ha study plot size were joined to the shapefiles, which had the study plot size of 0.1 ha. This was repeated to all five study plot sizes.

All calculated variables were inserted to VSURF variable selection tool, and most important variables were selected. RF modelling was performed using all datasets with five different study plot sizes. After the best model was identified, SOC and N maps were predicted for the whole study area. Different combinations of datasets (ALS, ALS + Landsat, ALS + African soil grids, Landsat + African soil grids) were also tested in order to analyse how different datasets contribute to modelling. Only the best performing study plot size found in the first modelling was used for this analysis.

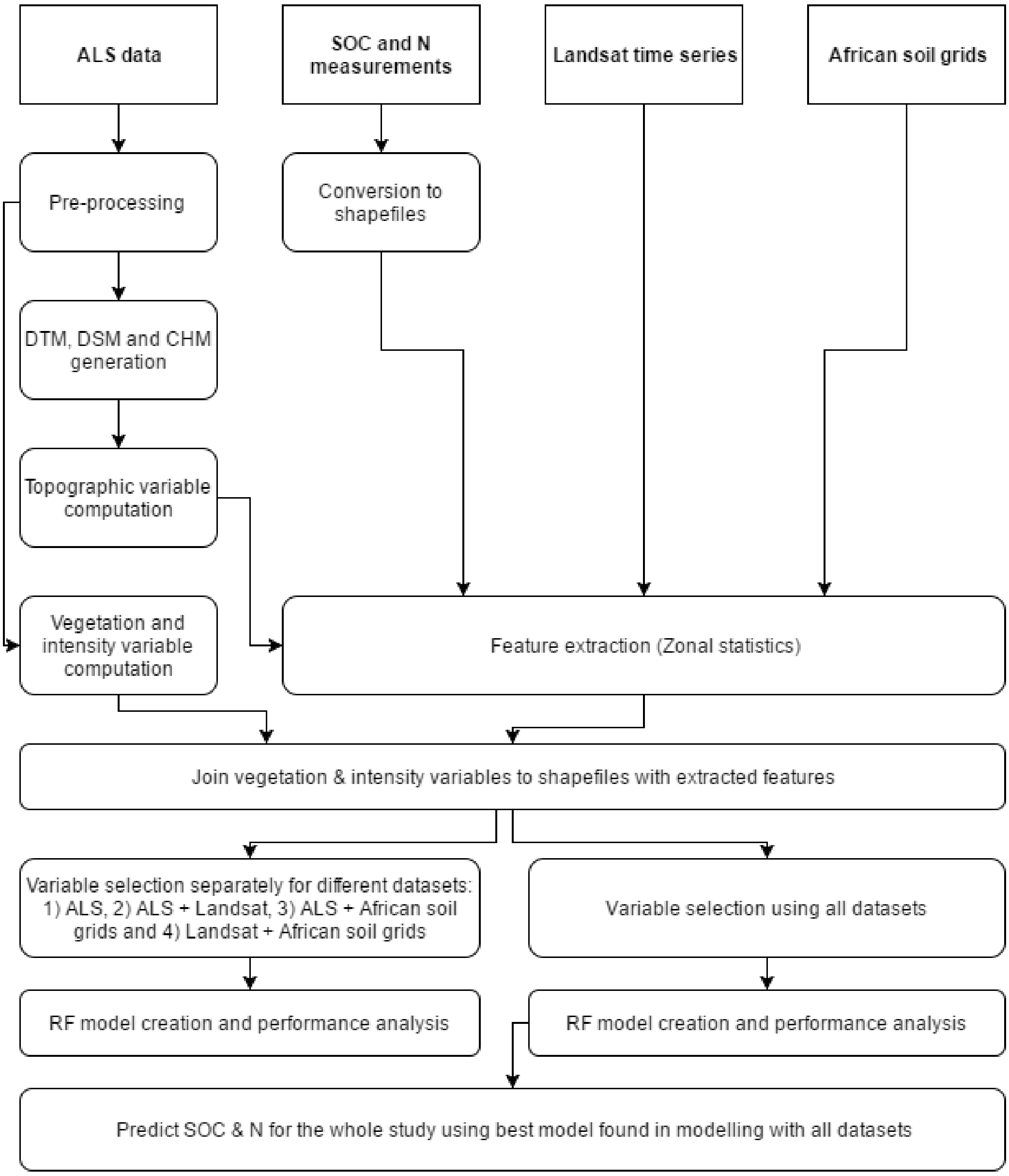

ALS data
SOC and N measurements
Landsat time series
African soil grids
Pre-processing
Conversion to shapefiles
DTM, DSM and CHM generation
Topographic variable computation
Vegetation and intensity variable computation
Feature extraction (Zonal statistics)
Join vegetation & intensity variables to shapefiles with extracted features
Variable selection separately for different datasets: 1) ALS, 2) ALS + Landsat, 3) ALS + African soil grids and 4) Landsat + African soil grids
Variable selection using all datasets
RF model creation and performance analysis
RF model creation and performance analysis
Predict SOC & N for the whole study using best model found in modelling with all datasets

Figure 21. Overview of the methodology and materials used in this study.

5.2 Pre-processing airborne laser scanning data

5.2.1 Quality checking

To make the data pre-processing and manual quality inspection more systematic and repeatable, the study area was divided to 50 m × 50 m tiles. Each tile had unique identifier, and all verification and improvements were done to each tile, one at a time. Results were written down to Excel for each tile, and necessary steps to fix issues or improve quality were completed. Figure 22 shows the ALS pre-processing flow, including the four main steps: 1) ground classification improvement, 2) low and high vegetation classification, 3) building classification and 4) wire classification. Temporary versions of the DTM, DSM and CHM were calculated after each step to help identify issues in the ALS pre-processing steps. If any issues were found, it was fixed and inspected again.

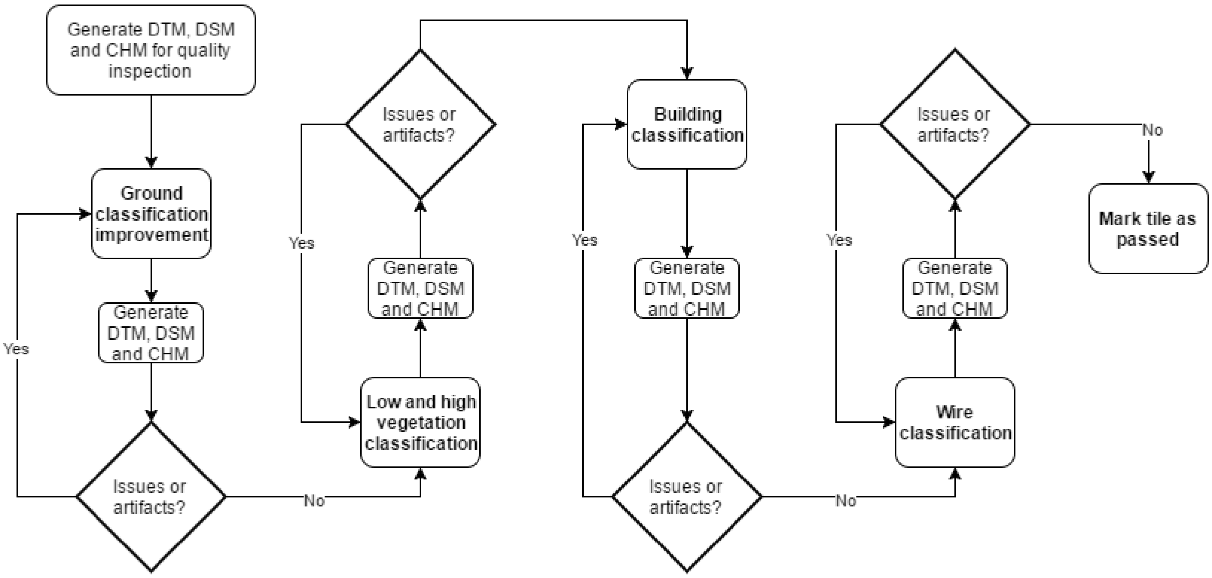

Generate DTM, DSM and CHM for quality inspection
Ground classification improvement
Generate DTM, DSM and CHM
Issues or artifacts?
Yes
No
Low and high vegetation classification
Generate DTM, DSM and CHM
Issues or artifacts?
Yes
Building classification
Generate DTM, DSM and CHM
Issues or artifacts?
Yes
No
Wire classification
Generate DTM, DSM and CHM
Issues or artifacts?
Yes
No
Mark tile as passed

Figure 22. Pre-processing steps for ALS data.

Classification schema used in this study did not follow the LAS 1.2 standard exactly as several classes had different class code. Classification used in this study, and the LAS 1.2 standard are given in Table 5.

Table 5. ALS data classification schema (LAS 1.2 standard based on ASPRS Board 2008).

Class	LAS Standard	Classification in this study
0	Created, never classified	-
1	Unclassified	Unclassified
2	Ground	Ground
3	Low Vegetation	Low Vegetation
4	Medium Vegetation	-
5	High Vegetation	High Vegetation
6	Building	Buildings
7	Low Point (noise)	-
8	Model Key-Point	-
9	Water	Error
10	Reserved for ASPRS Definition	Power Lines & Towers
11	Reserved for ASPRS Definition	-

5.2.1.1 Improved ground classification

Improvements to the classification were needed to ensure data quality for different analyses and DTM, DSM and CHM creation. As topography of the Taita Hills is highly variable, automatic algorithms had difficulties filtering the ground points.

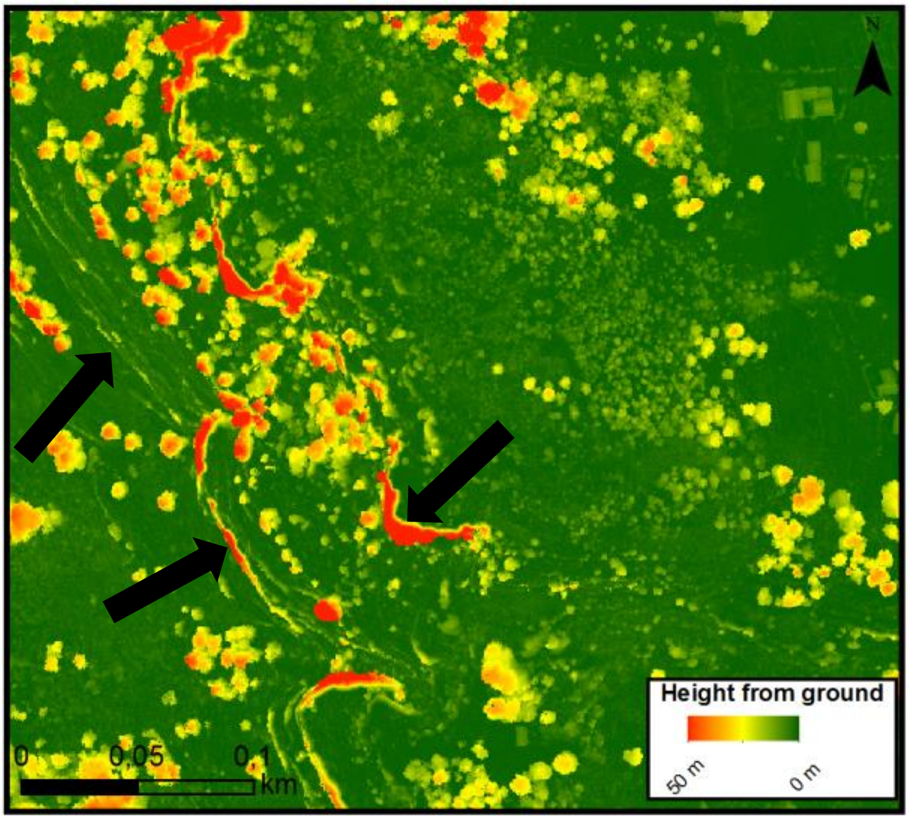


Figure 23. Ground point classification errors shown in CHM. Areas with steep cliffs are classified as vegetation, thus seen as long shaped geometries following the cliffs.

Initial automatic ground point classification had vast amount of classification errors. Classification errors were found especially in steep cliffs or ridges, where vegetation was classified to ground points (Figure 23). Areas with faulty classifications were identified from the DTM, DSM, CHM or ALS point cloud, and manually reclassified to ground point, error or as unclassified. Identifying and re-classifying faulty points was an iterative process, where temporary DTM, DSM and CHM models were calculated for this purpose only (Figure 24). Reclassification of the ALS points was made in TerraScan software. Example of cleaned ground can be seen in Figure 25.

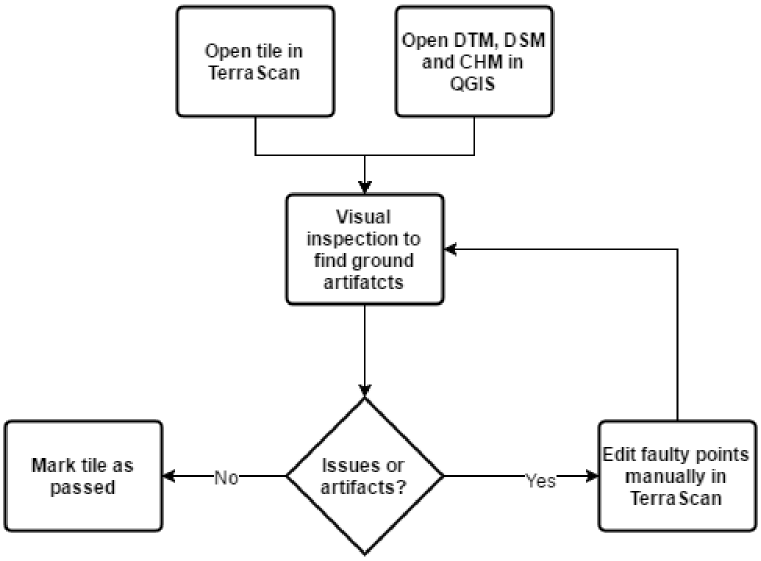

Open tile in TerraScan
Open DTM, DSM and CHM in QGIS
Visual inspection to find ground artifatcts
Issues or artifacts?
No
Mark tile as passed
Yes
Edit faulty points manually in TerraScan

Figure 24. ALS ground point improvement process

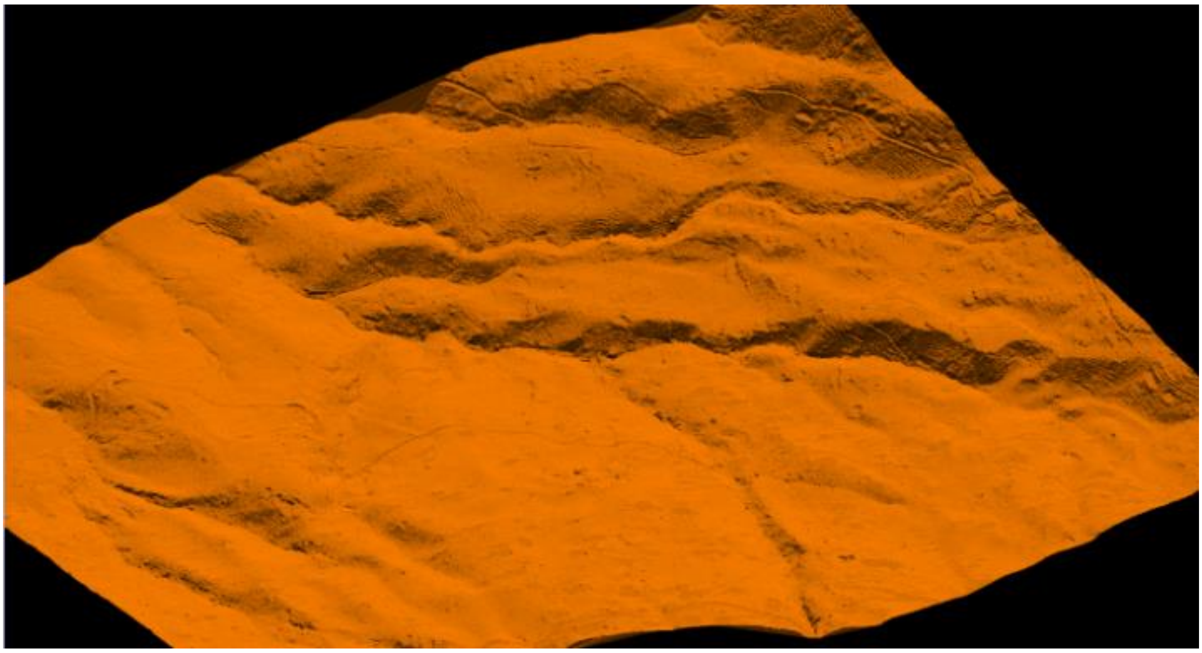


Figure 25. Triangulated 3D view of the cleaned ground points.

5.2.1.2 Vegetation classification

Vegetation points were classified from the point cloud by using algorithm based on height from ground using TerraScan. Points from 0 m to 1.5 m were classified as low vegetation, while everything above 1.5 m was classified as high vegetation. Separation between low and high vegetation was needed for the building and wire classification. To remove faulty points, all points above 50 m from the ground were removed.

5.2.1.3 Building and wire classification

Classifying buildings were done by using automatic algorithm in TerraScan. Results from the algorithm were not good and manual verification and editing was necessary. Due to irregularity of building shapes and sizes, a lot of classification errors were found from results of the automatic classification. All tiles were inspected and missing buildings were identified from DSM, CHM (Figure 29) and visualized point cloud.

Due to difficult topography and small power lines, no automatic algorithm could detect power wires or towers in the study area. Therefore, power lines were manually edited in TerraScan by using CHM as background for identification (Figure 26). Results of classification can be seen in Figure 27 as triangulated visualization.

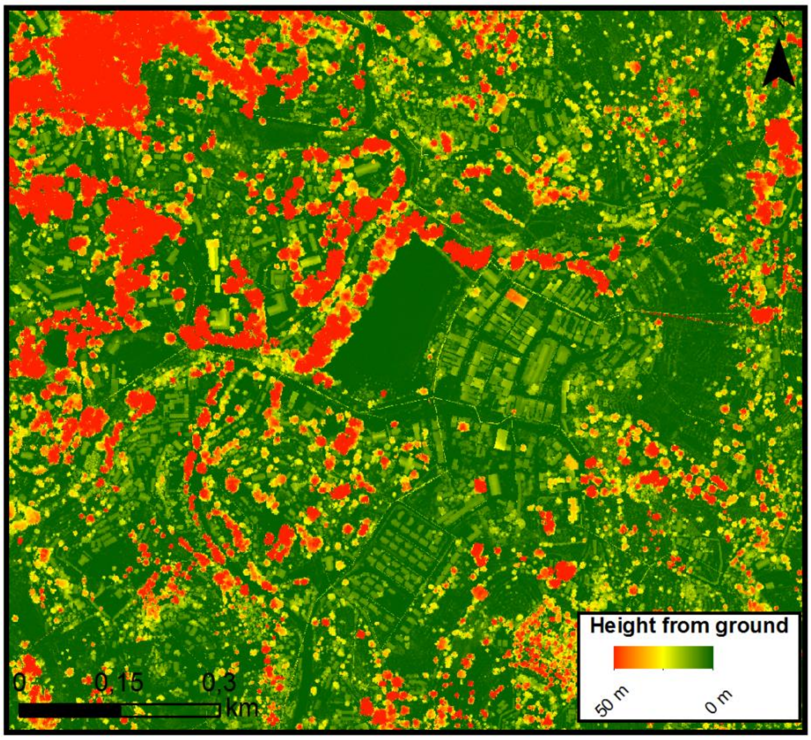


Figure 26. Buildings, wires and vegetation visible in CHM. Different objects can be identified based on their geometrical shapes and heights.

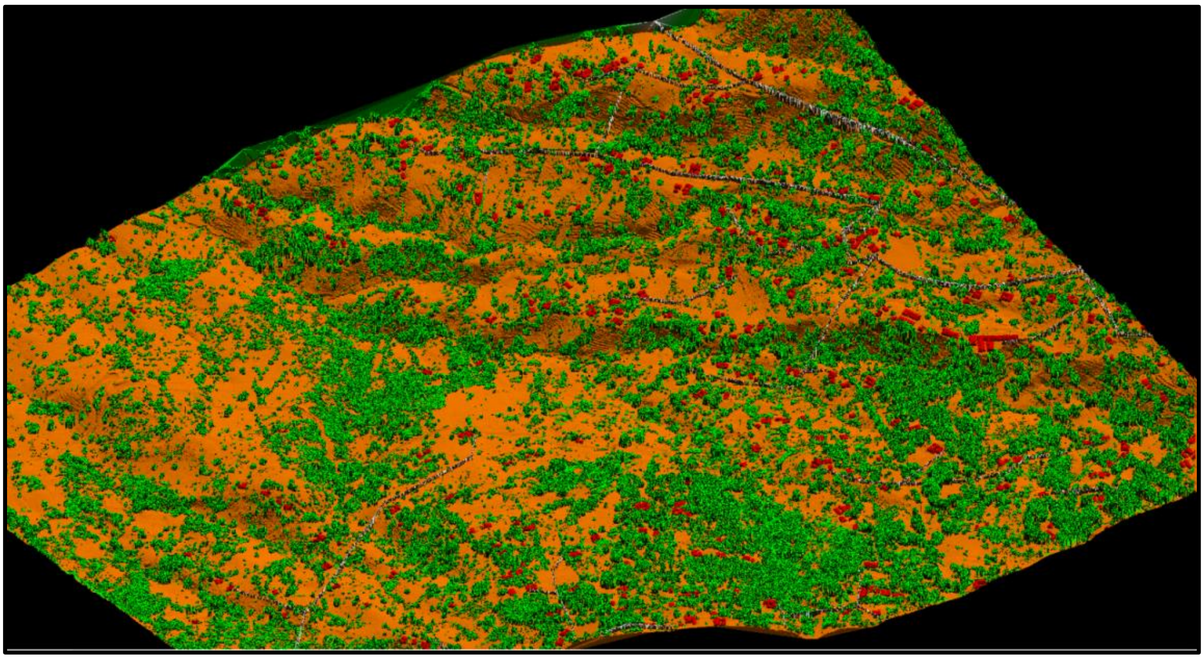


Figure 27. Triangulated 3D view of classified ALS point cloud (red = buildings, white = power wires, orange = ground, green = vegetation).

5.2.2 Digital terrain model, digital surface model and canopy height model

DTM and DSM were calculated from the cleaned and pre-processed ALS data. DTM was calculated by using adaptive TIN-surfaces (Axelsson 2000; Soininen 2016) algorithm implemented in TerraScan. Spatial resolution for DTM was 1 m. Only points classified as ground were used for DTM calculation. DSM was calculated from the ALS data by using first only or first of many echoes. Highest Hit Z algorithm in TerraScan (Soininen 2016) was used to generate the grid. CHM was calculated by subtracting DTM from DSM in QGIS. All models were clipped to the study area extent by using a shapefile in QGIS.

5.3 Variable computation and extraction

The main steps of the variable computation and extraction are summarized in Figure 28. In total, over 500 potential predictors were calculated for the modelling. This included 11 variables from Landsat time series, four variables from African soil grids, 14 topographic variables and 77 vegetation and intensity variables. In addition, Landsat time series variables were extracted (Zonal Statistics) using mean and range statistics, African soil grids and topographic variables were extracted using max, mean, min, range and standard deviation. Topographic variables also were computed using six different spatial resolutions: 1 m, 5 m, 10 m, 25 m, 50 m and 100 m. Feature extraction and vegetation and intensity variable computation was performed using the five different study plot sizes. All computed predictor variables are listed in Appendix 1, without the different feature extraction methods or spatial resolutions.

All variables were created or calculated using R or Python by utilizing several open source libraries and tools. QGIS Processing framework was used to calculate the topographic variables, utilizing System for Automated Geoscientific Analyses (SAGA) (Conrad et al. 2015), Geospatial Data Abstraction Library (GDAL) (GDAL Development Team 2016) and Geographic Resources Analysis Support System (GRASS) (GRASS Development Team 2016) libraries through the Python application programming interface (API). Vegetation and intensity variables were calculated using FUSION software (McGaughey 2016). Feature extraction was performed using Zonal Statistics method in QGIS. All the variables were combined to a CSV file which was then used in the variable selection. This procedure was repeated five times using the different study plot sizes for feature extraction, vegetation and intensity variable computation.

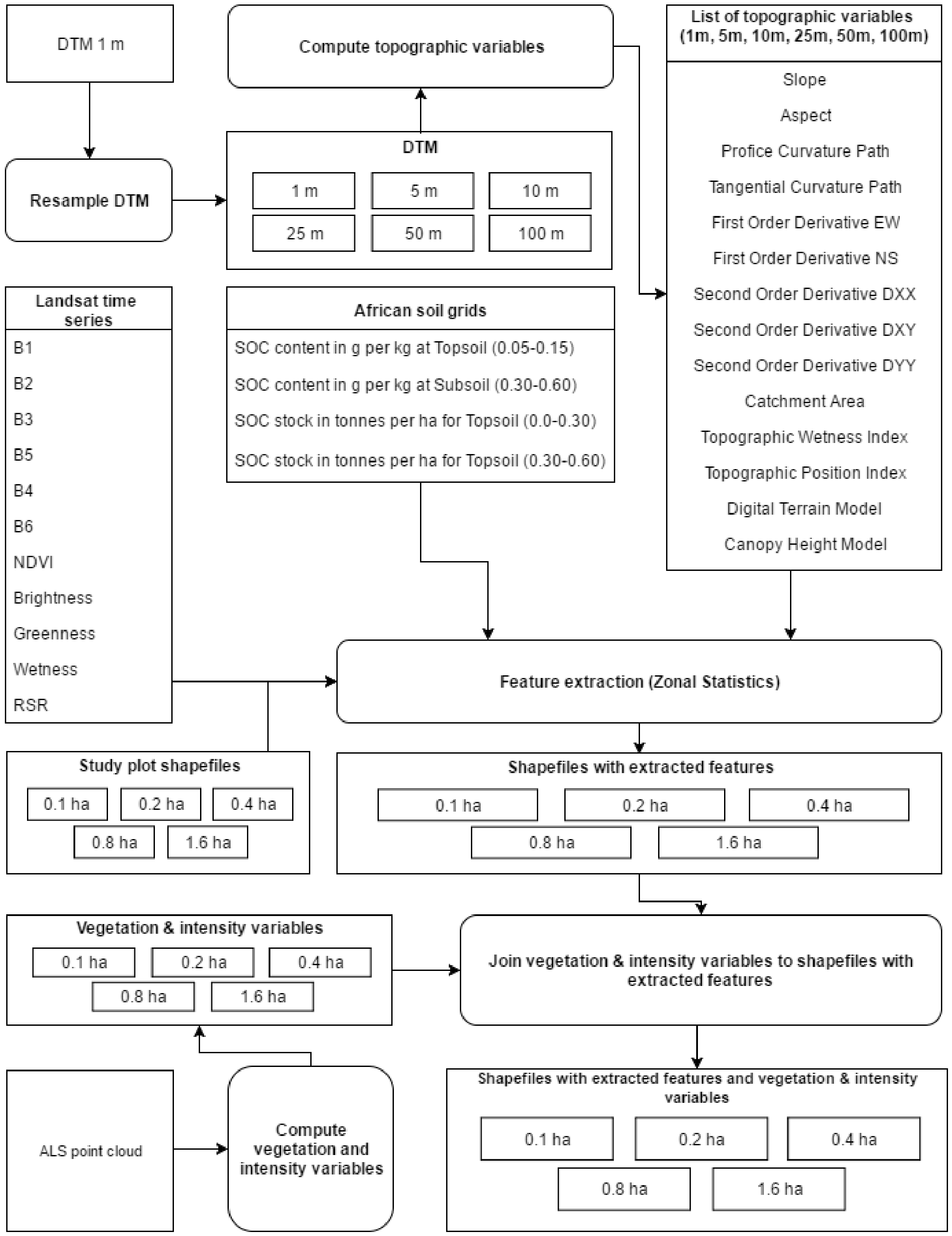

DTM 1 m
Compute topographic variables
List of topographic variables (1m, 5m, 10m, 25m, 50m, 100m)
Slope
Aspect
Profice Curvature Path
Tangential Curvature Path
First Order Derivative EW
First Order Derivative NS
Second Order Derivative DXX
Second Order Derivative DXY
Second Order Derivative DYY
Catchment Area
Topographic Wetness Index
Topographic Position Index
Digital Terrain Model
Canopy Height Model
Resample DTM
DTM
1 m
5 m
10 m
25 m
50 m
100 m
Landsat time series
B1
B2
B3
B5
B4
B6
NDVI
Brightness
Greenness
Wetness
RSR
African soil grids
SOC content in g per kg at Topsoil (0.05-0.15)
SOC content in g per kg at Subsoil (0.30-0.60)
SOC stock in tonnes per ha for Topsoil (0.0-0.30)
SOC stock in tonnes per ha for Topsoil (0.30-0.60)
Feature extraction (Zonal Statistics)
Study plot shapefiles
0.1 ha
0.2 ha
0.4 ha
0.8 ha
1.6 ha
Shapefiles with extracted features
0.1 ha
0.2 ha
0.4 ha
0.8 ha
1.6 ha
Vegetation & intensity variables
0.1 ha
0.2 ha
0.4 ha
0.8 ha
1.6 ha
Join vegetation & intensity variables to shapefiles with extracted features
ALS point cloud
Compute vegetation and intensity variables
Shapefiles with extracted features and vegetation & intensity variables
0.1 ha
0.2 ha
0.4 ha
0.8 ha
1.6 ha

Figure 28. Overview of the variable computation and feature extraction process. Digital terrain model (DTM) was resampled to six different spatial resolutions and topographic variables were computed from the outputs. Statistics (max, mean, min, range and standard deviation) for topographic variables, Landsat time series and African soil grids were extracted for different study plot extents. Vegetation and intensity variables were computed from the airborne laser scanning (ALS) data with the same study plot sizes and joined with the extracted features resulting in shapefiles with all calculated variables for each plot size.

5.3.1 Vegetation and intensity variables

ALS based vegetation and intensity variables were calculated using the ALS data processing software called FUSION (MgGaughey 2016). FUSION has a command line interface (CLI), which was utilized to calculate the variables. To make the processing more repeatable and less error prone, an automatic python script was developed to construct all necessary calls to FUSION CLI. Script generates the commands to call dynamically, based on the user defined input parameters. These generated calls were then transferred to CLI using Python subprocess-module.

As FUSION do not support traditional raster formats, a DTM was calculated in FUSION using PLANS DTM format. Next, study plots were clipped using ClipData tool (McGaughey 2016). Study plot locations are based on user configurable text file, which contains study plot identifier and coordinates in user defined CRS. Clipped files were normalized using the Plans DTM file created in the first step. This procedure was repeated five times for each study plot size.

After study plot files were clipped, the vegetation and intensity variables were calculated using CloudMetrics tool in FUSION (McGaughey 2016). Variables were calculated three times for each study plot, using different minimum height break values (0 m, 2 m and 4 m). As the minimum height break only affected certain cover estimates, output CSVs for each study plot were combined and all identical variables were deleted.

5.3.2 Topographic variables

DTM calculated in 1 m resolution was resampled to 5 m, 10 m, 25 m, 50 m and 100 m spatial resolutions using automatic script implemented in R. RGDAL (Bivand et al. 2016) library was used for resampling. Bilinear interpolation was used as the resampling method. Topographic variables (Table 6) were calculated from the DTMs, using QGIS Processing-framework with GDAL, GRASS and SAGA libraries. Topographic variables were calculated for each input DTM resolution. Whole process was automated with a Python script. The effect of spatial resolution is demonstrated in Figure 29.

Table 6. Computed topographic variables.

Variable	Tool
Slope	GRASS
Aspect	GRASS
Profile Curvature	GRASS
Tangential Curvature	GRASS
First Order Derivative (E-W slope)	GRASS
First Order Derivative NS (N-S slope)	GRASS
Second Order Partial Derivative (DXX)	GRASS
Second Order Partial Derivative (DYY)	GRASS
Second Order Partial Derivative (DXY)	GRASS
Catchment Area	SAGA
Topographic Position Index	SAGA
Topographic Wetness Index	SAGA
Elevation	TerraScan/GDAL*
Elevation from ground	QGIS

*Resampled with GDAL

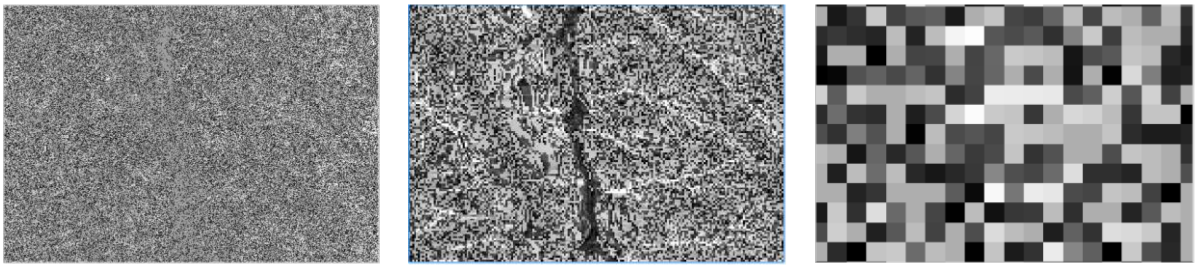


Figure 29. Topographic wetness index (TWI) with different spatial resolutions (1 m, 10 m and 100 m).

5.4 Variable selection

Due to the high number of predictors in this study, an automatic variable selection method was needed. RF based R package called variable selection using Random Forests (VSURF) (Genuer et al. 2016) was used to select the most important predictors. VSURF utilizes a three step variable selection procedure, which is designed for high dimensional data, where count of predictors exceeds the count of observations (Genuer et al. 2016). VSURF was run separately for SOC and N, and for each study plot size. An output CSV was generated containing the selected variables for each run.

5.5 Random Forest modelling and accuracy assessment

RF (Breiman 2001) is a popular ensemble machine learning algorithm used in several scientific applications. RF is based on multiple decision trees. When a new subsequent tree is built, a technique called bagging is used. In bagging, the RF algorithm does not look at the previous trees, which reduces the risk for overfitting the model. In inclusion with bagging, a random sample of predictors is taken before splitting the node to new trees, yielding to improved error rates (Breiman 2001). The final predictions are derived by taking the average of each individual tree (Breiman 2001). Benefits of RF are that it has no requirements for probability distribution of the target variable, leading to improved fitting for non-linear relationships (Hengl et al. 2015). RF has been also proven to perform well on complex patterns.

Selected variables by the VSURF were used for RF regression and final modelling results were computed. The number of RF trees was set to 1000. Due to the slight variation between each run, RF was run 50 times for each variable and resolution. Accuracy statistics were calculated as average of these 50 runs. Accuracy statistics were computed by using leave-one-out cross validation (Packalén et al. 2012).

The accuracy statistics computed were root mean square error (RMSE), relative root mean square error (RRMSE, %, RMSE divided by the mean), bias (mean residual error), relative bias (%) and pseudo coefficient of determination (R^2), which was computed based on Pearson correlation coefficient between the predicted and observed values. Variable importance was analysed by comparing increase in mean square error (%IncMSE).

5.6 Soil organic carbon and nitrogen content maps

SOC and N content maps were generated only for the best model found in the previous step. Variables selected by VSURF were computed for the whole study area. In the case of topographic variables, they were already computed for the whole area. ALS vegetation and intensity metrics were calculated for the whole study area using GridMetrics tool in FUSION (McGaughey 2016).

Vegetation and intensity metrics created in GridMetrics were converted to shapefile. Feature extraction was performed for the selected topographic, Landsat time-series and/or African soil grids. Shapefiles with selected variables were imported to R and RF model created in previous step was used to predict SOC and N values for each pixel in the study area. Predicted values were imported back to QGIS and converted to raster for visualization.

6. Results

6.1 Modelling results based on airborne laser scanning and ancillary data

6.1.1 Model performance

The modelling results varied considerably among the response variables and resolutions (Table 7). R^2 varied between 0.47 and 0.66 for SOC and between 0.36 and 0.44 for N (Figure 31). RRMSE varied between 31 and 40 % for SOC and between 32 and 34 % for N (Figure 30).

Table 7. Accuracy statistics for RF models of soil organic carbon (SOC) and nitrogen (N) for various plot sizes.

Response	Plot size (ha)	RMSE	RRMSE	Bias	RBias	Pseudo R^2
SOC (%)						
	0.1	0.76	36.94	0.031	1.49	0.47
	0.2	0.74	36.05	0.014	0.67	0.50
	0.4	0.76	37.31	**0.002**	**0.10**	0.46
	0.8	**0.63**	**30.98**	-0.027	-1.31	**0.66**
	1.6	0.73	35.92	-0.008	-0.40	0.50
N (%)						
	0.1	0.054	32.39	0.0006	0.39	0.41
	0.2	**0.053**	**31.80**	-0.0005	-0.33	**0.44**
	0.4	0.054	32.43	-0.0023	-1.37	0.41
	0.8	**0.053**	32.14	**-0.0005**	**-0.29**	0.43
	1.6	0.056	33.88	-0.0003	-0.18	0.36

For SOC, plot size of 0.8 ha performed relatively well, RRMSE being 31 % and pseudo R^2 0.66. For N, all the models performed similarly and none of the models was clearly the best. In general, SOC results were slightly better than N results.

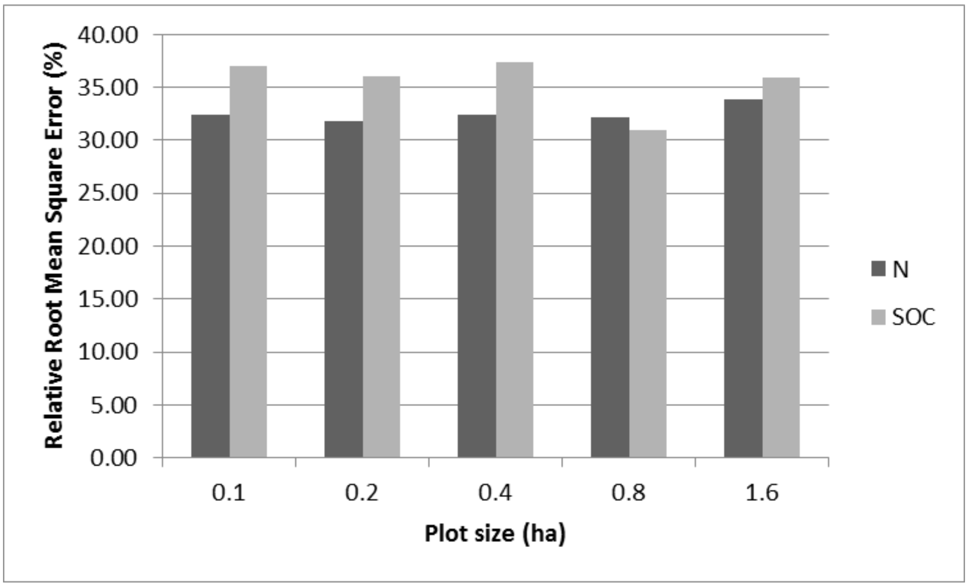


Figure 30. Relative root mean square error (%) values of soil organic carbon (SOC) and nitrogen (N) models for various plot sizes.

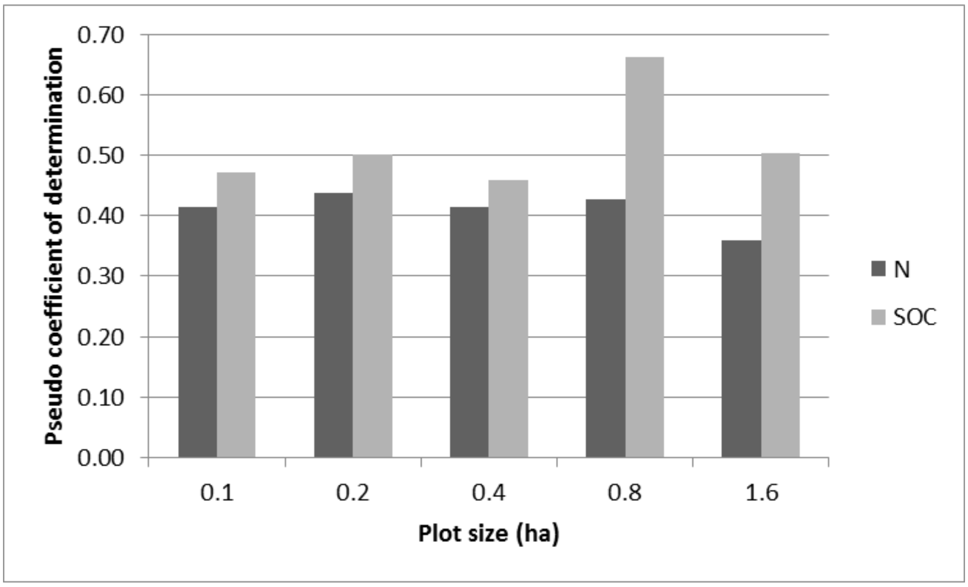


Figure 31. Pseudo coefficient of determination (R^2) values of soil organic carbon (SOC) and nitrogen (N) models for various plot sizes.

Figure 32 shows the relationships between observed and predicted SOC values. Most models were rather similar but model with plot size of 0.8 ha provided clearly the best model fit as indicated by the accuracy statistics (Table 7). Most of the observed and predicted SOC values are on the lower end of the scale. In general, all the models seem to under-predict SOC as the highest predicted value in the 0.8 ha model was 4.6 and the highest observed value was 6.5.

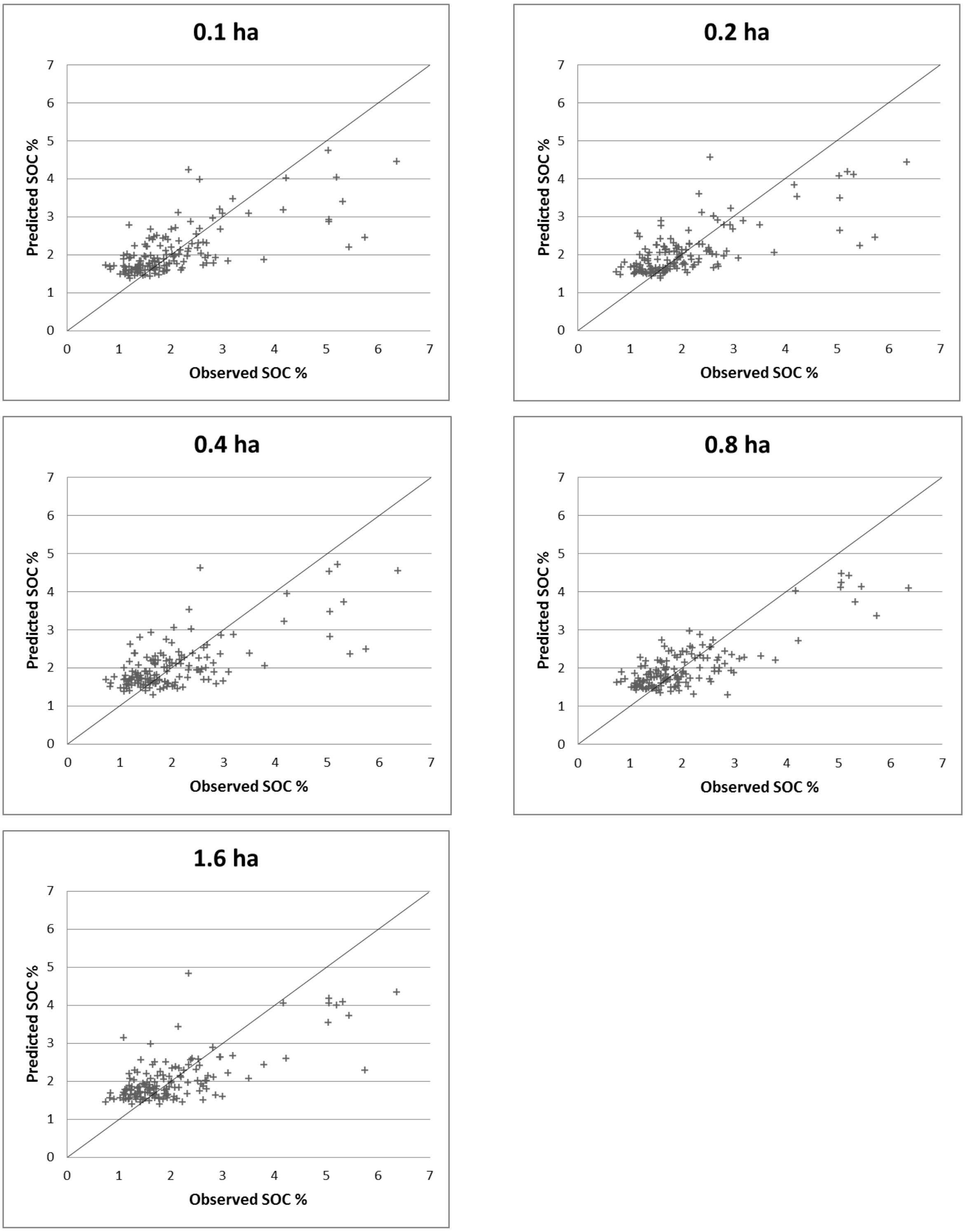


Figure 32. Relationships between observed and predicted soil organic carbon (SOC) values.

Figure 33 shows the relationships between observed and predicted N values. All the models are relatively scattered, and no clear differences can be distinguished between the models. Most of the observed and predicted N values are relatively small. In general, all the models seem to under-predict N as the highest predicted value in the 0.8 ha model was 0.32 and the highest observed N was 0.49.

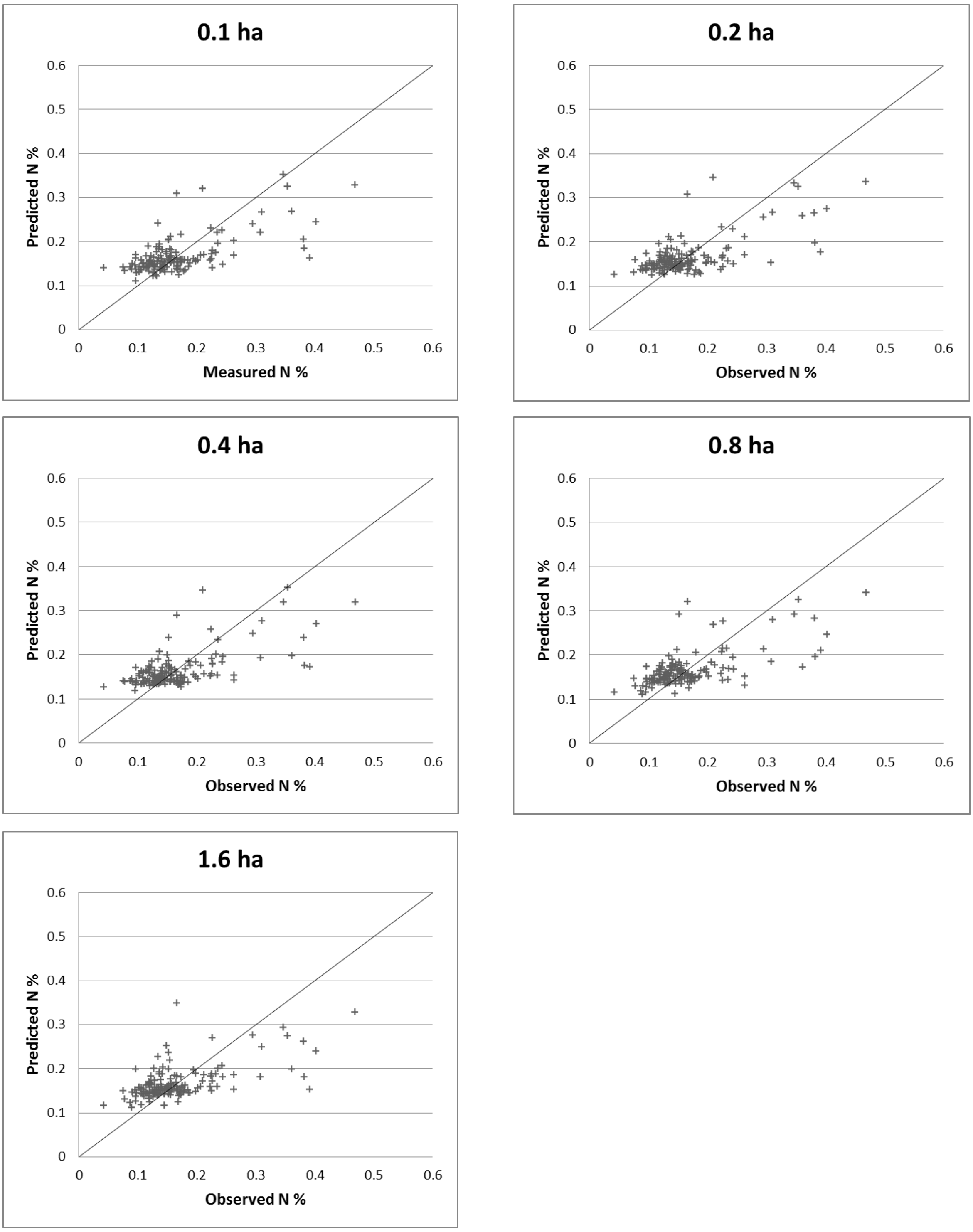


Figure 33. Relationships between observed and predicted nitrogen (N) values.

6.1.2 Variable selection

Results of the variable selection are shown in Table 8. The variables are coded using the coding schema defined in Appendix 1. Number of variables selected varied highly between the plot size and response variable. On average, SOC had lower number of predictors (9.6) than N (14.4). The lowest number of predictors was 3 for SOC and 10 for N. The highest number of predictors was 16 for SOC and 17 for N. SOC modelled with 0.8 ha plot size had only three predictors: Landsat band 2 (Mean as feature extraction statistics), tangential curvature (5 m spatial resolution, using range as feature extraction statistics) and maximum intensity.

Selected predictors varied highly between models, though couple of them appeared regularly in the models: tangential curvature, intensity based variables, DTM (elevation) and Landsat band 2 or 3. Models had also same variables with different spatial resolution. Due to the high number of selected predictors, only the best SOC model is described more closely. The results of N modelling were worse and the number of predictors very high, and hence closer analysis was not done.

As described above, SOC model with 0.8 ha plot size performed best. Figure 34 shows the relationships between the observed SOC and three selected predictors: maximum intensity, tangential curvature (5 m spatial resolution, range as feature extraction method) and Landsat band 2 (mean as feature extraction method). Reflectance values of the Landsat band 2 have been multiplied with 10000 (Adhikari et al. 2016).

Table 8. Results of the VSURF variable selection. Variables are coded using the coding schema defined in Appendix 1. Statistical method used in the feature extraction (for topography, African soil grid and Landsat variables) with the original spatial resolution (for topography variables) are defined inside the parenthesis. For certain canopy cover estimates minimum height break (in meters) is defined inside the parenthesis, with word above.

Response	Plot size (ha)	No X	X1	X2	X3	X4
SOC (%)						
	0.1	16	I.stddev	I.L2	I.AAD	TAC (Max, 5m)
	0.2	14	I.L2	I.variance	E.P50	PARA (Above 2)
	0.4	7	ARATFR (Above 2)	I.AAD	I.L2	E.MAD.med
	0.8	3	TAC (Range, 5m)	RB2 (Mean)	I.maximum	
	1.6	8	RB2 (Mean)	I.maximum	I.P99	SYY(Min, 10m)
N (%)						
	0.1	17	I.L2	I.variance	I.AAD	I.IQ
	0.2	10	I.variance	I.L2	I.AAD	E.P50
	0.4	15	I.AAD	I.L2	ARATFR (Above 2)	E.P50
	0.8	17	E.P50	E.MAD.med	PARA (Above 2)	E.MAD.mo
	1.6	13	RB2 (Mean)	I.maximum	RB1 (Mean)	Elev.MAD.mode

Response	Plot size (ha)	X5	X6	X7	X8	X9
SOC (%)						
	0.1	PFRA (Above 2)	RB3 (Mean)	I.IQ	SXY (Range, 5m)	I.P99
	0.2	I.maximum	SXY (Range, 5m)	ARATFR (Above 2)	TAC (Max, 5m)	RB1 (Mean)
	0.4	I.maximum	DTM (Max, 50m)	FNS (Range, 5m)		
	0.8					
	1.6	RB3 (Mean)	SXX (Min, 1m)	E.MAD.mo	SXX (Range, 1m)	
N (%)						
	0.1	PFRA (Above 2)	RB3 (Mean)	SXX (Mean, 1m)	ASP (Mean, 1m)	TAC (Mean, 1m)
	0.2	ASP (Mean, 1m)	ASP (Range, 25m)	TAC (Range, 5m)	DTM (Max, 1m)	TAC (Max, 5m)
	0.4	E.MAD.med	DTM (Max, 50m)	PFRA (Above 2)	ASP (Min, 5m)	TWI (Min, 25m)
	0.8	I.AAD	TAC (Range, 5m)	RB2 (Mean)	I.maximum	I.IQ
	1.6	DTM (Mean, 100m)	E.P05	DTM (Max, 100m)	DTM (Max, 1m)	RSR (Stddev)

Response	Plot size (ha)	X10	X11	X12	X13	X14
SOC (%)						
	0.1	SLO (Mean, 50m)	DTM (Max, 10m)	TAC (Mean, 1m)	DTM (Mean, 25m)	DTM (Min, 10m)
	0.2	ASP (Mean, 1m)	ASP (stdev, 25m)	RSR (Mean)	TAC (Min, 1m)	DTM (Max, 25m)
	0.4					
	0.8					
	1.6					
N (%)						
	0.1	TAC (Max, 5m)	DTM (Max, 10m)	SXY (Range, 5m)	DTM (Max, 25m)	I.maximum
	0.2	DTM (Min, 5m)				
	0.4	I.maximum	SLO (Min, 25m)	TAC (Max, 5m)	E.maximum	I.P99
	0.8	I.P99	RB1 (Mean)	RB3 (Mean)	DTM (Max, 5m)	DTM (Max, 25m)
	1.6	RSR (Mean)	ARATFR (Above 2)	SXX (Min, 1m)	SLO (Min, 100m)	

Response	Plot size (ha)	X15	X16	X17
SOC (%)				
	0.1	DTM (Mean, 10m)	SXY (Min, 50m)	
	0.2			
	0.4			
	0.8			
	1.6			
N (%)				
	0.1	TAC (Mean, 5m)	RWT (Stddev)	DTM (Min, 5m)
	0.2			
	0.4	TPI (Min, 5m)		
	0.8	DTM (Mean, 50m)	SLO (Min, 100m)	SLO (Min, 5m)
	1.6			

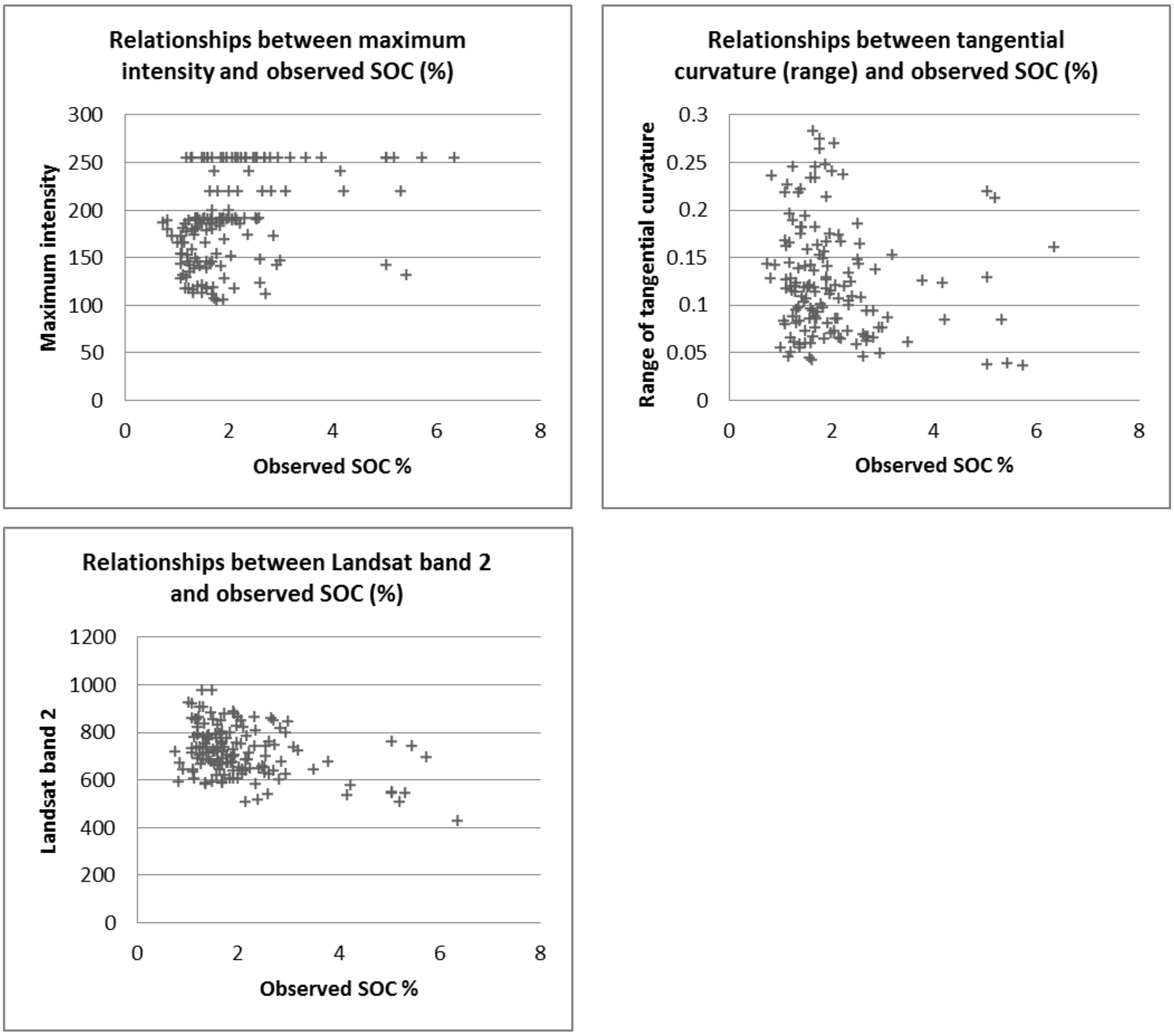


Figure 34. Relationships between selected predictors and observed soil organic carbon (SOC).

Variable importance was also analysed for the SOC (Figure 35) and N (Figure 36) models with 0.8 ha plot size. The higher the increase in mean square error (%IncMSE) value, the more important variable. For SOC, the most important predictor was Landsat band 2, with small difference to maximum intensity and tangential curvature. For N, the predictor importance was more varying among the selected variables. The most important predictors for N were the tangential curvature, intensity P99 and maximum intensity.

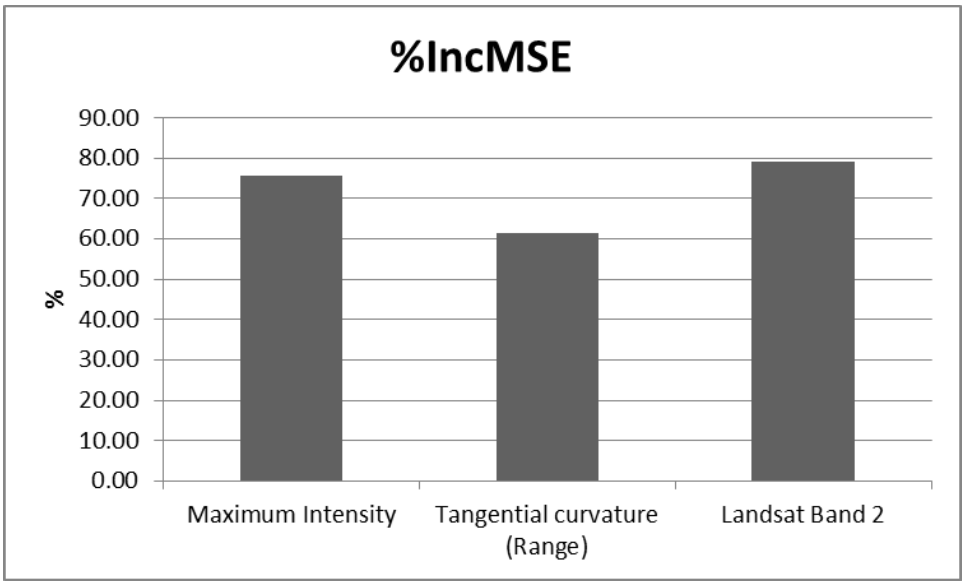


Figure 35. Increase in mean square error (%) for selected soil organic carbon variables.

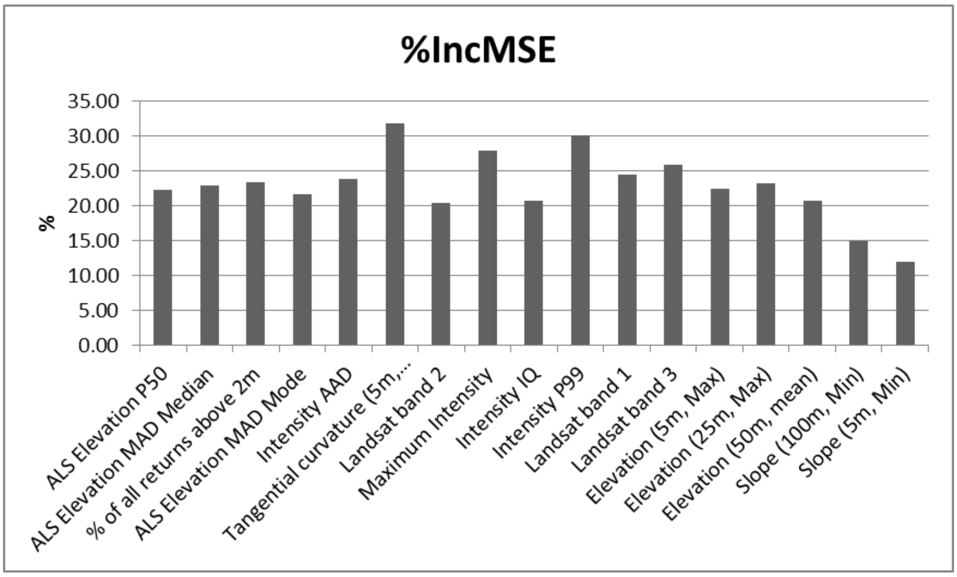


Figure 36. Increase in mean square error (%) for selected nitrogen variables.

6.1.3 Maps of soil organic carbon and nitrogen contents

RF model described in the previous step was used to predict SOC and N content for the whole study area. For both response variables, only the model with 0.8 ha plot size was used as it was clearly the best performing model for SOC and with N there were no clear differences between the models. Circular plot with 50.46 m radius (0.8 ha), equals to 89.45 m × 89.45 m in area, as square. However, for the map creation, the spatial resolution was rounded to 90 m × 90 m.

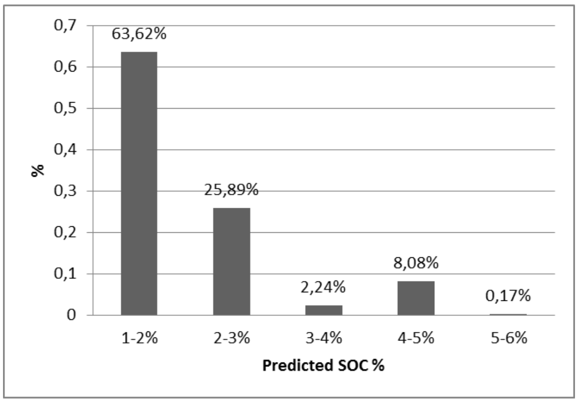
Due to the poor results and high number of predictors for N, no detailed analysis was done. Summary statistics of the predictor variables and predicted SOC values for the whole study area are given in Table 9. There were in total 12991 pixels in the study area after removal of artefacts from the Landsat images and corner pixels.

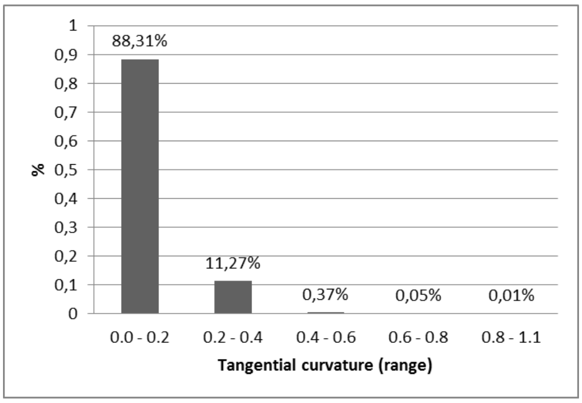
Table 9. Summary statistics of input data and predicted values for the whole study area.

Statistics	Maximum intensity	Tangential curvature	Landsat band 2*	Predicted SOC (%)
Minimum	68.00	0.016	348.42	1.23
Maximum	255.00	1.045	1706.01	5.04
Mean	186.38	0.127	708.92	2.11
Standard deviation	45.21	0.066	120.39	0.82

* Reflectance values can be optained by diving by 10000

Figure 37 shows the frequencies of the values of each variable. All the predictor variables had high values for certain range of data, and similar structure can be seen in the predicted SOC values. For tangential curvature, most of the values (88.31 %) are between 0 and 0.2, while values between 0.8 and 1.1 have only 0.01 %. Feature extraction method for tangential curvature was range, so values are the difference of minimum and maximum in the pixel. Similar structures can be seen also in Landsat band 2, where 76.66% of the values are between 400 and 800, while only 0.04% of the values were between 1200 and 1600. With maximum intensity, percentages were more widely dispersed, though values between 150 and 200 were dominant (44.39%).





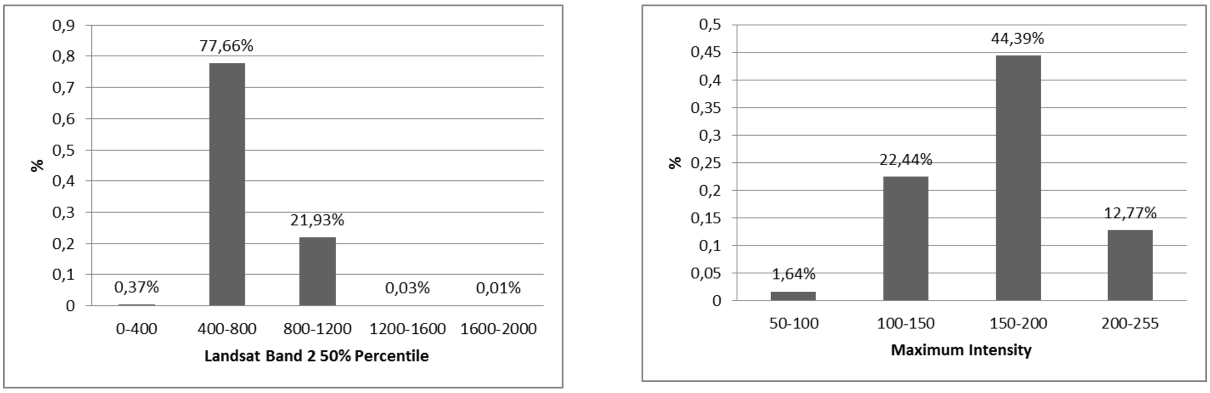


Figure 37. Relative frequencies (%) of the predicted SOC (%) values and predictor variables.

Predicted SOC content map can be seen in Figure 38 and N content map in Figure 39. There are gaps in the maps due to the missing data in the Landsat images. From both of the maps, areas with large and small SOC and N values can be identified. For both maps, the large values seem to be found in the forests and areas of dense vegetation, and small values are found near non-vegetated areas.

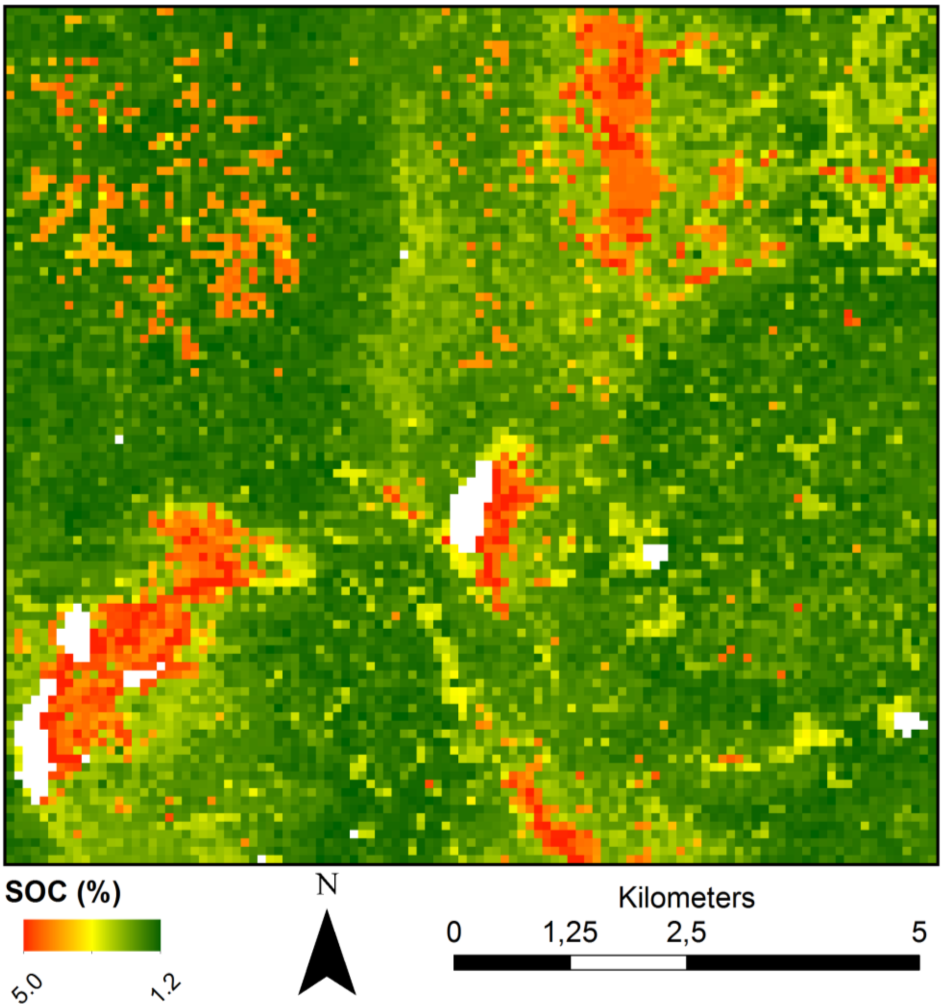


Figure 38. Predicted soil organic carbon (SOC) content (%) map for the study area.

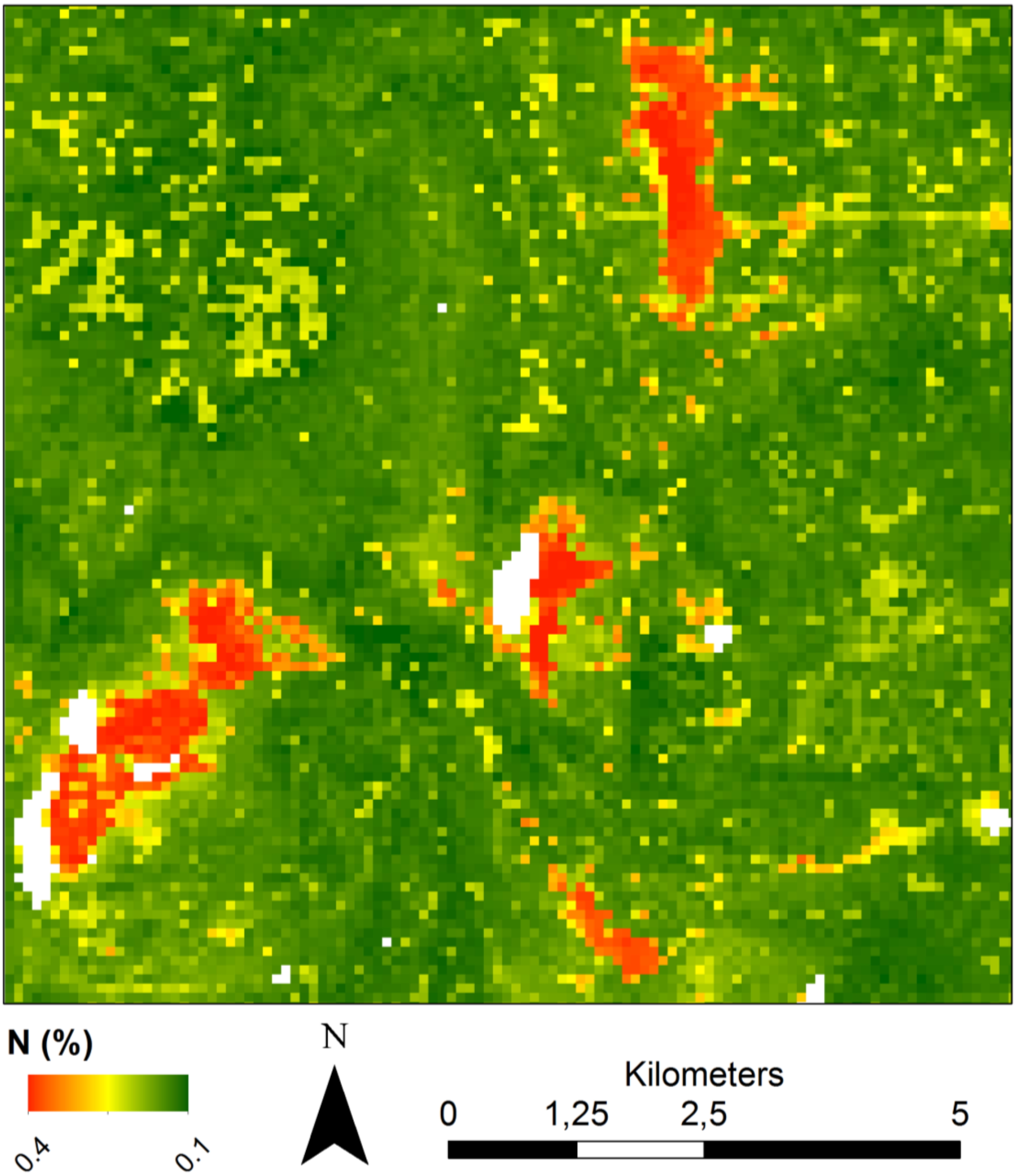


Figure 39. Predicted nitrogen (N) content (%) map for the study area.

Figure 40 shows the selected predictor variables for SOC model and Figure 41 the selected predictor variables for N model. Patterns in the SOC and N map clearly follow the patterns of the predictor variables. Low reflectance in Landsat band 2 and high value of maximum intensity seem to lead to high levels of SOC and N. Smaller range of tangential curvature leads to higher predicted values, as higher range leads to smaller values.

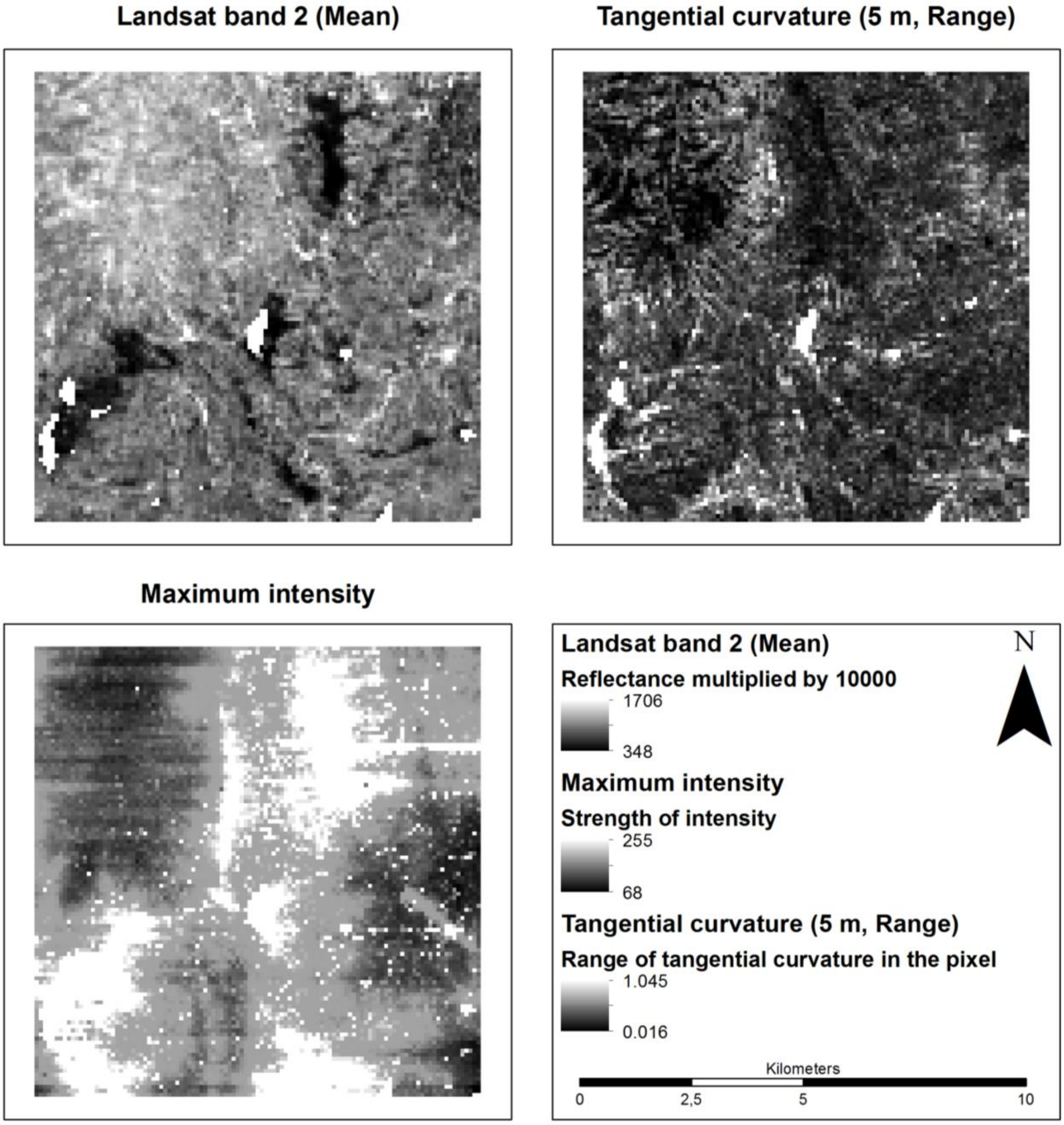

Landsat band 2 (Mean)
Tangential curvature (5 m, Range)
Maximum intensity
Landsat band 2 (Mean)
Reflectance multiplied by 10000
1706
348
Maximum intensity
Strength of intensity
255
68
Tangential curvature (5 m, Range)
Range of tangential curvature in the pixel
1.045
0.016
N
Kilometers
0
2,5
5
10

Figure 40. Maps of the selected variables for soil organic carbon model (0.8 ha). Statistical method used in the feature extraction (for topography and Landsat variables) with the original spatial resolution (for topography variables) are defined inside the parenthesis.

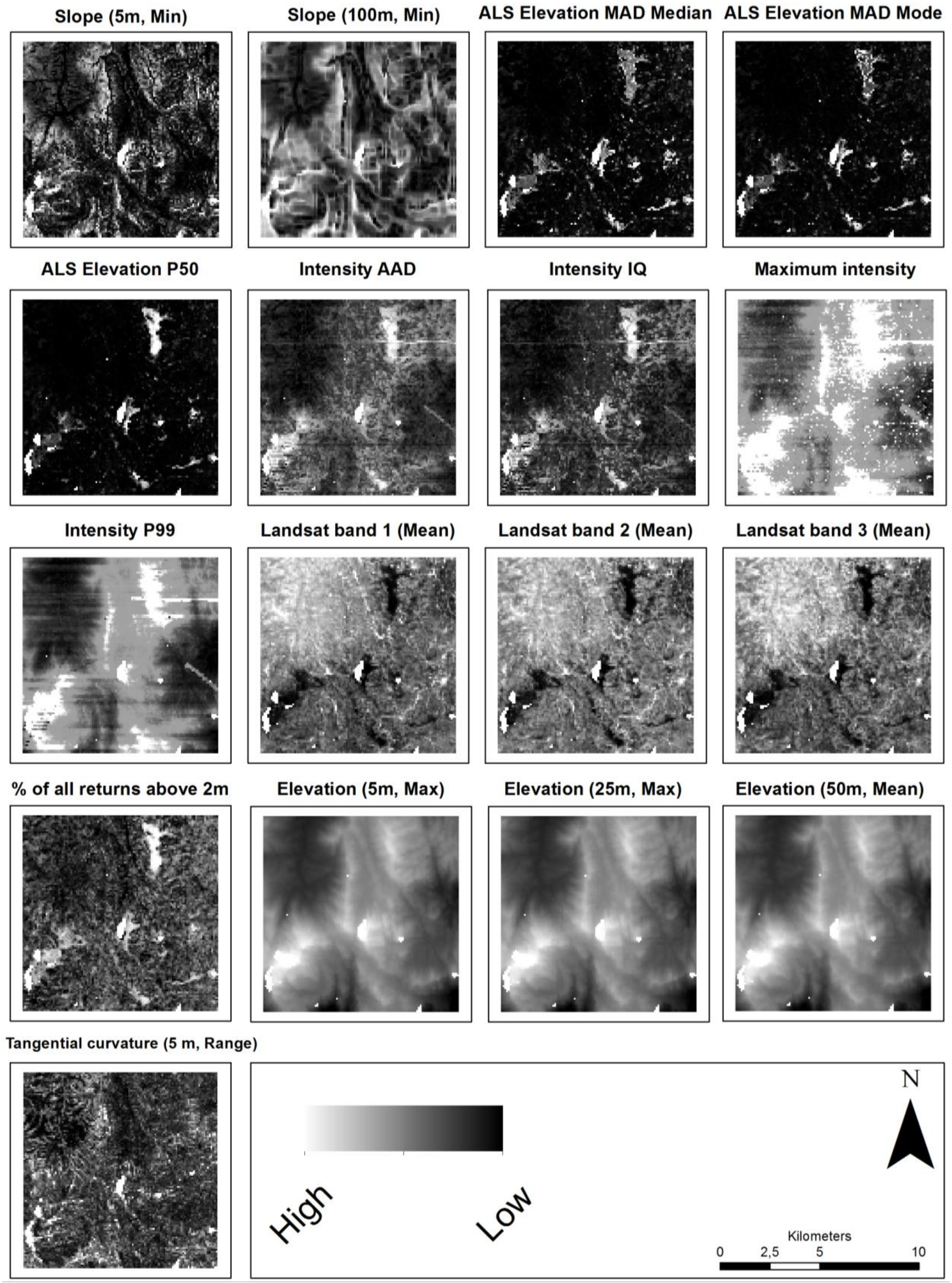


Figure 41. Maps of the selected variables for nitrogen model (0.8 ha). Statistical method used in the feature extraction (for topography and Landsat variables) with the original spatial resolution (for topography variables) are defined inside the parenthesis.

6.2 Modelling results based on different data combinations

6.2.1 Model performance

Soil properties were also modelled separately using different combinations of datasets. Results for SOC are shown in Table 10 and results for N in Table 11. Comparison of the RRMSE values is shown in Figure 42 and pseudo R^2 in Figure 43. All the models were calculated with plot size of 0.8 ha, which was found to perform best when using the all datasets.

Table 10. Statistical comparison of soil organic carbon models using different data combinations.

Dataset	RMSE	RRMSE	Bias	Rbias	Pseudo R^2
All	0.63	30.98	-0.027	-1.31	0.66
ALS	0.71	34.57	-0.018	-0.90	0.55
ALS + Landsat	0.63	30.82	-0.025	-1.24	0.66
ALS + African soil grids	0.72	35.05	-0.020	-0.97	0.53
Landsat + African soil grids	0.82	40.07	-0.019	-0.94	0.38

Table 11. Statistical comparison of nitrogen models using different data combinations.

Dataset	RMSE	RRMSE	Bias	Rbias	Pseudo R^2
All	0.053	32.14	-0.0005	-0.29	0.43
ALS	0.055	33.07	-0.0008	-0.49	0.39
ALS + Landsat	0.054	32.59	-0.0008	-0.49	0.41
ALS + African soil grids	0.055	32.84	-0.0013	-0.80	0.40
Landsat + African soil grids	0.061	36.62	-0.0013	-0.79	0.25

RRMSEs (Figure 42) of the SOC models varied between 30.89 % and 40.07 % and pseudo R^2 (Figure 43) values between 0.38 and 0.66. The best results were achieved with ALS + Landsat datasets and the worst when using Landsat and African soil grid data. Model with ALS data only performed intermediately, with RRMSE of 34.57 % and R^2 of 0.55.

RRMSEs (Figure 42) of the N models varied between 32.14 % and 36.62 % and pseudo R^2 (Figure 43) values between 0.25 and 0.43. The best results were achieved with ALS + Landsat datasets and the worst when using Landsat and African soil grid data. Model with only ALS data performed intermediately, with RRMSE of 33.07 % and R^2 of 0.39.

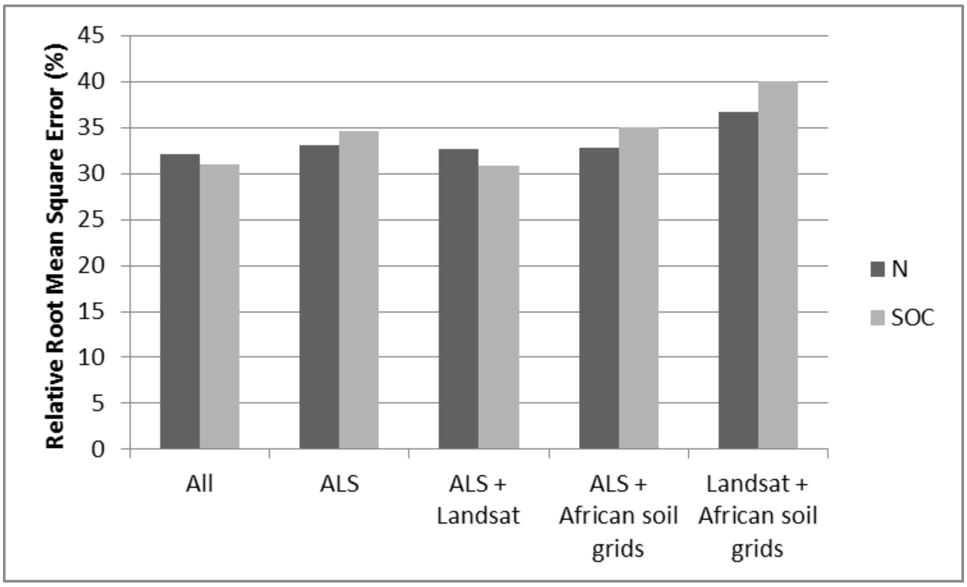


Figure 42. RRMSE (%) values of SOC and N models using different data combinations.

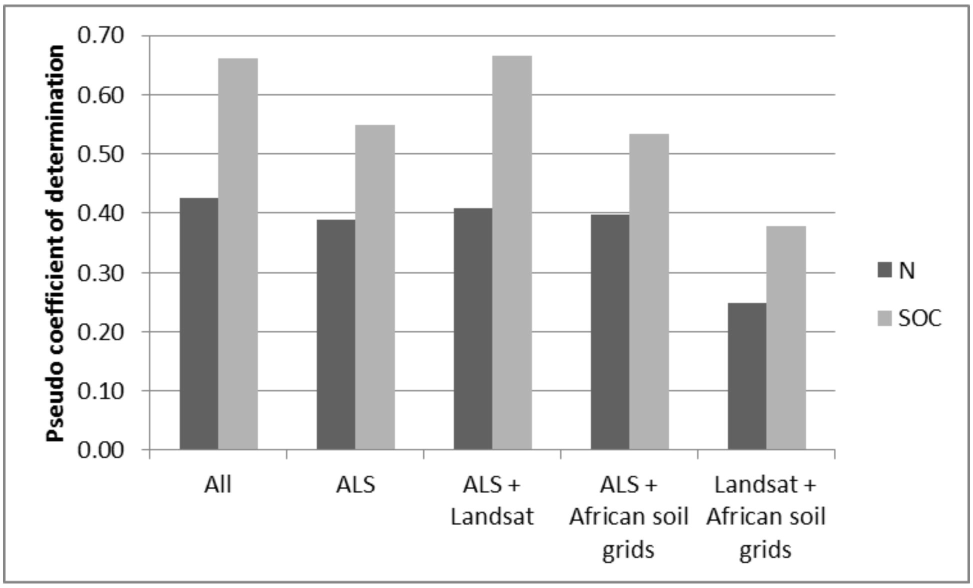


Figure 43. Pseudo R^2 values of SOC and N models using different data combinations.

Model with all datasets and model with ALS + Landsat performed equally well, with just slight difference in accuracy statistics due to random variations in the RF predictions. Thus, the African soil grids did not provide any value to the modelling and could be discarded. This was also indicated by the fact that those data were not included in the selected predictors in the earlier analyses. The model based on only ALS data performed slightly worse than the model with all data. Therefore, Landsat time series dataset seem to improve pseudo R^2 by 0.1 in the case of SOC, and by 0.02-0.04 in the case of N. RMSE improved by 0.08 for SOC, but remained same for N.

6.2.2 Variable selection

Selected variables for SOC models are shown in Table 12. For all SOC models with ALS data, maximum intensity and tangential curvature (with spatial resolution of 5m and range as feature extraction method) were selected. Model with ALS data only had also percentage of all returns above 2 meters and median of the absolute deviations from the overall median from elevation as selected variables. Model with Landsat and African soil grids had different Landsat bands (RB1, RB2 and RB3), topsoil soil organic carbon stock in tonnes per ha (TTH) and topsoil soil organic carbon content (fine earth fraction) in g per kg (TKG).

Table 12. Results of the VSURF variable selection for soil organic carbon models. Variables are coded using the coding schema defined in Appendix 1. Statistical method used in the feature extraction (for topography, African soil grid and Landsat variables) with the original spatial resolution (for topography variables) are defined inside the parenthesis. For certain canopy cover estimates minimum height break (in meters) is defined inside the parenthesis, with word above.

Dataset	No X	X1	X2	X3	X4	X5
All	3	TAC (Range, 5m)	RB2 (Mean)	I.maximum		
ALS	4	TAC (Range, 5m)	PARA (Above 2)	E.MAD.med	I.maximum	
ALS + Landsat	3	TAC (Range, 5m)	RB2 (Mean)	I.maximum		
ALS + African soil grids	4	TAC (Range, 5m)	E.P50	PARA (Above 2)	I.maximum	
Landsat + African soil grids	5	RB2 (Mean)	RB3 (Mean)	RB 1 (Mean)	TKG (Max)	TTH (Max)

Selected variables for N are shown in Table 13. N models had clearly more selected variables than the SOC models. Most of the N models had tangential curvature (TAC), 50% percentile of elevation (E.P50), percentage of all returns above a specified height (PARA), maximum intensity (I.maximum) and several others. Nitrogen models had also several elevation (DTM) variables, which were calculated using different feature extraction methods or had different spatial resolution. For model with Landsat and African soil grids, six variables were selected from the Landsat data, though reason for this is most likely that the African soil grids did not have any N data. As shown in chapter 6.2.1, performances of the N models were poor, which can be seen as large amount of predictor variables.

Table 13. Results of the VSURF variable selection for nitrogen models. Variables are coded using the coding schema defined in Appendix 1. Statistical method used in the feature extraction (for topography, African soil grid and Landsat variables) with the original spatial resolution (for topography variables) are defined inside the parenthesis. For certain canopy cover estimates minimum height break (in meters) is defined inside the parenthesis, with word above.

Dataset	No X	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
All	17	E.P50	E.MAD.med	PARA (Above 2)	E.MAD.mo	I.AAD	TAC (Range, 5m)	RB2 (Mean)	I.maximum	I.IQ	I.P99	RB1 (Mean)	RB3 (Mean)	DTM (Max, 5m)	DTM (Max, 25m)	DTM (Mean, 50m)	SLO (Min, 100m)	SLO (Min, 5m)
ALS	13	PARA (Above 2)	E.P50	E.MAD.med	E.MAD.mo	I.variance	TAC (Range, 5m)	I.maximum	I.P99	I.IQ	DTM (Mean, 100m)	DTM (Max, 25m)	DTM (Mean, 25m)	TWI (Min, 25m)				
ALS + Landsat	17	E.P50	E.MAD.med	PARA (Above 2)	E.MAD.mo	I.AAD	TAC (Range, 5m)	RB2 (Mean)	I.maximum	I.IQ	I.P99	RB1 (Mean)	RB3 (Mean)	DTM (Max, 5m)	DTM (Max, 25m)	DTM (Mean, 50m)	SLO (Min, 100m)	SLO (Min, 5m)
ALS + African soil grids	13	PARA (Above 2)	E.P50	E.MAD.med	E.MAD.mo	I.stddev	TAC (Range, 5m)	I.maximum	I.IQ	I.P99	DTM (Max, 25m)	DTM (Max, 10m)	DTM (Max, 5m)	SLO (Min, 100m)	DTM (Max, 5m)			
Landsat + African soil grids	6	RB2 (Mean)	RB1 (Mean)	RSR (Std dev)	RB3 (Mean)	RWT (Mean)												

7. Discussion

7.1 Modelling performance

When comparing SOC modelling results to similar studies, the results seem to be quite nicely in line. There are not many studies having similar data, study area and scale, but most of the other studies have something in common with this study. Rather similar study was conducted by Li et al. (2016) who used ALS based intensity and topography variables for modelling soil properties in Korean pine forests. Authors achieved similar results, as R^2 ranged between 0.46 and 0.66 for SOM (related to SOC), N, pH and soil depth. However, study area of Li et al. (2016) was pine forest, but study area in this study was very heterogeneous. Were et al.

(2015) achieved R^2 of 0.64 when modelling SOC content in eastern Kenya using topographic variables and Landsat imagery as predictor variables. Furthermore, Vågen & Winowiecki (2013) achieved R^2 of 0.65 using Landsat images for SOC modelling in four different study sites in eastern Africa.

Even higher results have been achieved in more homogenous landscapes. Thompson & Kolka (2005) explained 71 % of the variability of SOC in forested study area in USA. Vågen et al. (2013) achieved R^2 of 0.79 when using Landsat data for modelling SOC on four study areas in Ethiopia. However, Vågen et al. (2013) had training data from 38 sites, all around the Africa, including 3378 topsoil SOC samples, and this study had only 150 SOC samples from one site.

In contrast to SOC, the N modelling results are surprisingly poor when compared to other studies. Li et al. (2016) achieved R^2 of 0.6 for N and R^2 of 0.6 for SOC as explained previously. Other studies have usually found similar performance for both SOC and N. Also in this study, observed SOC and N contents are strongly correlated (Figure 44), and hence more similar performance could have been expected.

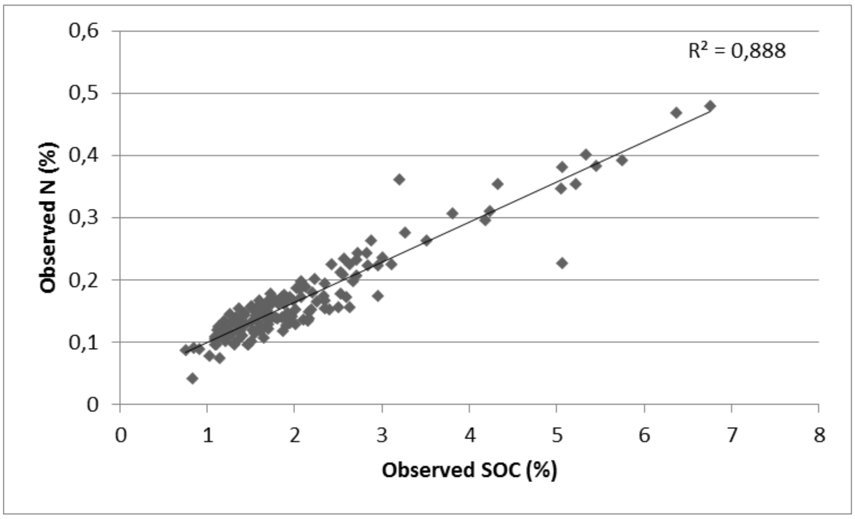


Figure 44. Relationship between observed soil organic carbon (SOC) and observed nitrogen (N).

One possible limitation of the model performance is relatively low number of field measurements, especially inside forests and other areas with dense vegetation. Vågen & Winowiecki (2013) had also relatively low number of forested plots and achieved similar results (R^2 0.67). They had four study sites, using same LDSF sampling methodology. Authors expect the results to improve if more study sites were available. Low number of forested areas is also seen in the concentration of SOC and N values to the lower end of the values, as only nine study plots had SOC content higher than 4 % (Figure 32 and Figure 33). The SOC and N predictions are also under-predicted, leading to under-estimations on SOC and N levels on the study area.

When analysing the first objective of this study, the results are looking somewhat promising. ALS seems viable option for modelling SOC in heterogeneous landscape when combined with ancillary datasets, but does it bring enough value and is it cost-efficient enough for operational use. Furthermore, is ALS cost-efficient enough for monitoring purposes? Cost of the ALS data collection remains relatively high though data collection is getting cheaper all the time as sensors are improving. Also unmanned aerial vehicle based ALS sensors have been released and even spaceborne laser scanning sensors have been planned.

7.2 Importance of the ancillary datasets

Performances of the different combinations of the datasets were tested to analyse the importance of each ancillary dataset on soil properties modelling (Table 10 and Table 11). Most important finding was the improvement of the results when using ALS with Landsat data compared to model with ALS only. Model with ALS only performed slightly worse than the models with ALS and Landsat data. Statistically, using Landsat data improved the RMSE to 0.63 from 0.71 for SOC and for N to 0.053 from 0.055. Pseudo R^2 improved from 0.55 to 0.66 for SOC and from 0.39 to 0.43 for N.

Two most performing models for both SOC and N were the model with all datasets and model with ALS and Landsat. These two models were performing almost identically, with very small variations in the statistics, due to differences in RF runs. Exactly same variables were also selected for the same models. Not a single variable from African soil grids were used in these two models.

Clearly the worst model was the one without ALS data (R^2 of 0.38 for SOC and 0.25 for N). Even that African soil grids were not tested alone, they do not seem to bring any value to the modelling as none of the models with ALS data had variables from the soil grids. Coarse resolution (250 m) could explain the low performance of African soil grids for the SOC and N modelling in such as heterogeneous study area. Maps of the African soil grids were modelled for whole Africa (Hengl et al. 2015), thus inaccuracies are possible in heterogeneous regions.

However, the lack of predictive power indicates that the SOC map from African soil grids (Hengl et al. (2015) is not accurate estimator for the SOC patterns of the Taita Hills.

According to the results, ALS is promising data source for soil properties modelling but model performance can be improved by adding optical reflectance data to the models. Optical RS data has been also important in other studies. Hengl et al. (2015) used MODIS data in their African soil grid modelling. Vågen & Winowiecki (2013) achieved good results with Landsat data only (R^2 0.67) and Vågen et al. (2013) even better results (R^2 0.79).

Heterogeneity of the topography and vegetation could be the reason for low performance for models without ALS. ALS brings the high resolution accuracy on topographic variables and information about the vegetation and its structure (Kristensen et al. 2015). It is also somewhat surprising how good results Vågen & Winowiecki (2013) and Vågen et al. (2013) achieved only with Landsat data and without topographical variables. This could be however explained by the simpler landscapes or greater variation in SOC content (data was collected from several sites).

7.3 Most important variables

7.2.1 Selected variables

Third objective of this study was to find the best predictor variables explaining SOC and N content in a heterogeneous landscape. Relatively good results were achieved using the SOC model with plot size of 0.8 ha. For N, no clear conclusions could be drawn because of the weak performance of the models. In this section, the selected variables for SOC model with 0.8 ha study plot size are discussed and compared to literature.

For SOC, three good predictor variables were identified, and those can be easily computed from the ALS and Landsat data. The importance of the selected variables was close to each other, Landsat band 2 being slightly the most important one. However, it is good to keep in mind that SOC content and stocks are dependent on the site-specific conditions as well as on the land cover, and current and historical land management practices (Thompson & Kolka 2005; Bou et al. 2010).

7.2.2 Range of tangential curvature

Curvature is an important DTM based surface property used in several applications, such as geomorphology and hydrology (Schmidt et al 2003). A surface has different curvatures in different directions (Neteler & Mitasova 2008). The curvature in a direction perpendicular to the gradient is called tangential curvature, measured in the normal plane. Tangential curvature reflects the change in aspect angle and influences the divergence and convergence of water flow (Evans & Cox 1999; Fraisse et al. 2001; Neteler & Mitasova 2008).

The tangential curvature is expressed as 1/m and value of 0.05 corresponds to radius of curvature of 20 m (Shapiro & Waupotitsch 2015). Convex form areas have positive and concave form areas have negative values (Shapiro & Waupotitsch 2015). Convex (ridges) forms of tangential curvature exhibits converging flow, while concave (valleys) forms exhibit diverging flows (Mitasova & Hofierka 1993).

In this study, most models selected range of tangential curvature as one of the predicting variables. The highest SOC and N seem to concentrate in areas with low or intermediate values of range of tangential curvature, and the lowest values are near values with high values of range of tangential curvature. In this case, the range of tangential curvature explains the complexity of topography. As the feature extraction was performed with 0.8 ha plot size and spatial resolution of tangential curvature was 5 m, dozens of pixels fitted inside the polygon. The higher the range of tangential curvature, the more complex study plot it was (Figure 45).

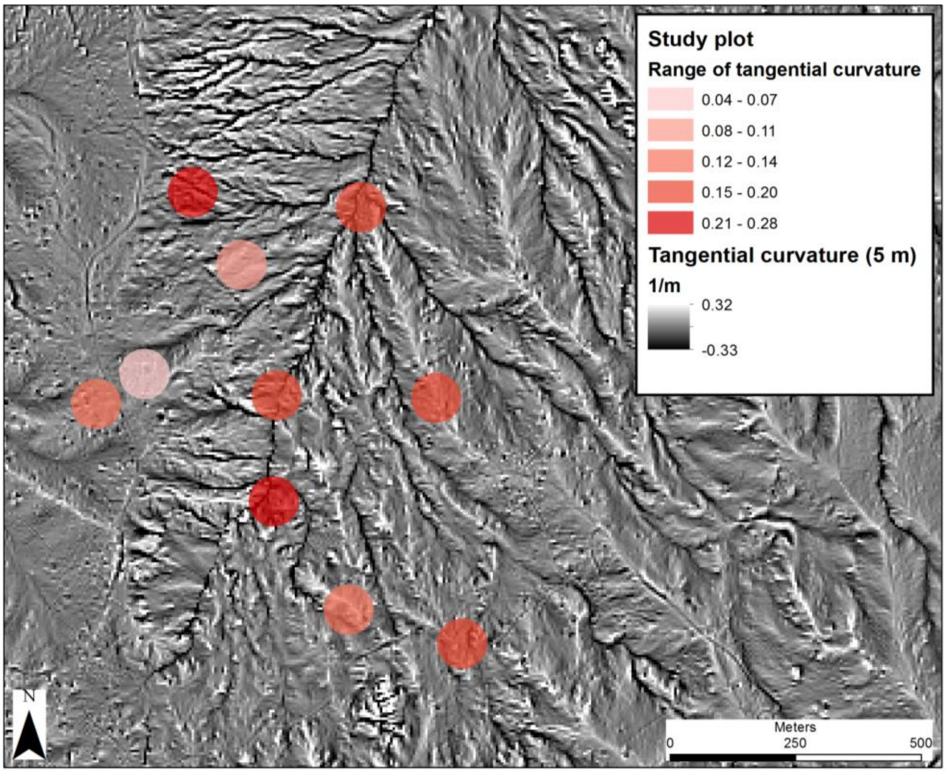


Figure 45. Tangential curvature (5m) and study plots (0.8 ha) of cluster 2. Extracted values of the tangential curvature are shown for the study plots (using range as statistical method).

Tangential curvature has been used in several soil properties modelling studies, and it has proved to be good predictor variable for soil properties in some of the studies. Timlin et al. (2003) used DTM based terrain indices to model soil water holding capacity (WHC), and tangential curvature and slope proved to be important variables. Bou et al. (2010) found tangential curvature to be good predictor of SOC and the most important terrain based variable. On the other hand, Thompson & Kolka (2005) did not find significant correlation (R^2 0.109) between SOC and tangential curvature in a study area in Kentucky, USA.

When comparing to literature, tangential curvature is potentially important variable when studying SOC and in general soil properties. Range of the tangential curvature is also very interesting as no other study seem to have used it. However, the range could make more sense than the actual values of the tangential curvature as range depicts the complexity of the landscape.

7.2.3 Maximum intensity

ALS intensity has not been studied much for soil mapping. Intensity has been used for modelling vegetation and soil moisture with intermediate results (Garroway et al. 2011). Another study was done in China using intensity as predictor for SOC and other topsoil properties (Li et al. 2016). Results of Li et al. (2016) indicate that ALS intensity could be effective for estimating topsoil properties in forests with complicated topography and dense canopy cover. The ALS sensors operate in the near-infrared region, and the relationships between near-infrared wavelengths and soil properties has been identified previously (e.g. Ge et al. 2011). According to Li et al. (2016), ALS intensity could provide information about the relative proportions of organic compounds in soils, which could explain its power on explaining some of the soil properties.

In this study, the intensity variables contained both ground and vegetation points, thus not having only values from the ground or soil. Laser scanning sensors operate on near infrared spectral region, in which vegetation reflects strongly. The importance of the maximum intensity could be also related to the high reflectance of vegetation. When comparing maximum intensity and Landsat band 2 (Figure 40), similar spatial patterns related to structure of the vegetation can be found. However, also areas with high intensity that are not visible in the Landsat image can be identified.

Using intensity is not always straightforward. Due to changes in flying height and scanning angle, intensity data should be calibrated and normalized before using it for analyses (Korpela et al. 2010; Li et al. 2016). Even simple range-based intensity normalization has been shown to improve classification results and compensate for losses in target-to-sensor path (Korpela et al. 2010). Issues with the intensity data were also noticed in this study, as clear artefacts can be seen in the intensity images (Figure 46) and predicted SOC and N content maps. Stripes are going across the data on two locations. This could be explained by issues in the sensor, flight lines or in some of the pre-processing steps. More close investigations should be done. Even with the issues in intensity data, it proved to be important variable in this study. Results might improve if the proper calibration steps were done for the data. The higher the intensity values, more SOC and N was predicted in the maps.

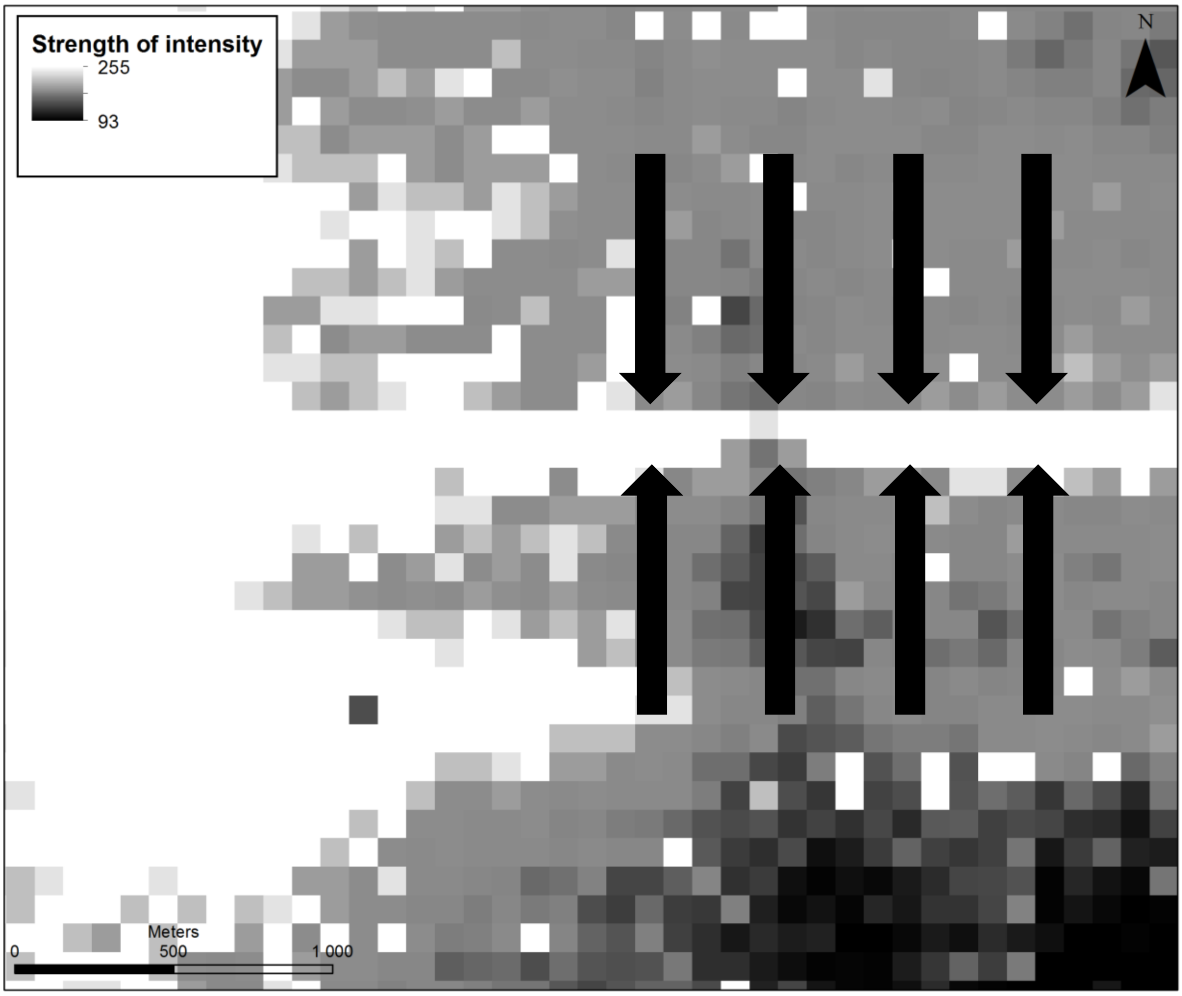


Figure 46. Striped artefact in the maximum intensity data.

7.2.4 Landsat band 2 (reflectance in green spectral region)

Landsat ETM+ band 2 was the third variable selected for the SOC model. Landsat band 2 corresponds to green spectral region (0.52-0.62 μm). In the green wavelength, bare ground reflects more than vegetation, thus forests and dense vegetation are seen darker in the images (Jensen 2000). Landsat data has been widely used for soil properties modelling with good results (Vågen & Winowiecki 2013; Vågen et al. 2013; Mulder et al. 2011). Mirzaee et al. (2016) found that Landsat ETM+ imagery accounted from 22.3 % to 47.9 % of the SOM variation, concluding that RS data can improve the soil predictions. According to Minasny et al. (2008), Landsat is the most often used remotely sensed imagery in soil properties modelling. Vågen & Winowiecki (2013) used all bands of Landsat 7 ETM+ data with good results. However, authors did not explain the relative importance of the separate bands.

It is little surprising that the band 2 was selected for most of the models. The modelling also included every other Landsat bands and several vegetation indices. Most likely the function of green band in the SOC model is to measure the type and amount of vegetation. However, some of the importance of the green band could be also explained by relationships between soil properties. Schwanghart & Jarmer (2011) found that the highest correlation between the organic C and multispectral imagery were found between 0.5 and 0.7 μm. They also found that correlations decreased with increase in wavelength.

7.4 Effect of plot size and spatial resolution of topographic variables

Usually, soil modelling has been done using only one plot size or one spatial resolution for the variables, typically being limited by the data used for modelling. This study had different approach as several study plot sizes were tested and variables were calculated with several spatial resolutions. The best model for estimating SOC was the model with 0.8 ha plot size. The results of the modelling were weaker with larger and smaller study plot sizes. Reasons for this could be several; data quality, properties of the data and properties of the soil dynamics in the Taita Hills.

The positioning data of the field plots used in the modelling raised also some concerns. The study plot coordinates were written on paper sheets and there is possibility of human errors. When analysing the data, several writing mistakes were identified and those plots were excluded from the analysis. However, a possibility for some inaccuracies remains. The Taita Hills is a difficult area for GPS devices due to the variable topography and dense vegetation, which could cause extra inaccuracy to the location measurements. Thus larger study plot size could take these inaccuracies into account in comparison to smaller plots. The relatively large study plot size also relate to the scale of soil and vegetation spatial variations in the Taita Hills.

Florinsky (2012c) noted the choice of spatial resolution as one of the main problems for soil studies using topographical data. Wrong choice of spatial resolution can lead to incorrect results or artefacts (Florinsky 2012c). Topographic variables of this study were calculated using six different spatial resolutions. Most often selected spatial resolution for topographic variables was 5 m. The effect of spatial resolution and neighbourhood size on terrain attributes regarding soil mapping has been studied and the general conclusion is that terrain variables calculated from too high resolution DTM are full of noise and errors, as too coarse resolution DTMs lose the exact information (Timlin et al. 2003; Budiman & Bishop 2005; Roecker & Thompson 2010). No clear conclusions can be drawn on the best spatial resolution or neighbourhood size as it is usually related to the study area. A good guide for selecting proper resolution and neighbourhood size would be to understand the landforms within the study area (Roecker & Thompson 2010). Bou et al. (2010) concluded that several studies have demonstrated that relationships between soil properties and terrain attributes could be unique to each environment and soil property. Florinsky (2012c) proposed calculating topographic variables with several spatial resolutions and identified the most suitable for the specific study area. Similar approach was used in this study, with relatively good results. Spatial resolution of 5 m for topographic variables sounds reasonable for the Taita Hills, while not being full of noise, but still being accurate enough.

7.5 Feasibility of the mapping results

SOC and N content maps were produced using 90 m × 90 m spatial resolution (0.8 ha). The spatial resolution of 90 m should be accurate enough for almost all applications, and compared to most of the global or continental datasets with 250 m - 1 km resolutions, the maps are very detailed. Areas with high and low SOC and N content can be easily identified from the maps, though as the accuracy statistics suggest, the accuracy may not be very high. Especially the high values are missing from the predicted data, due to under-prediction in the modelling, which might limit the usability of the maps.

The highest SOC levels were found in the indigenous forests of the Taita Hills, which is in line with the results of Vågen et al. (2016). Vågen et al. (2016) found the highest levels of

SOC in tropical forest systems, including montane forests in East Africa. Lower levels of SOC were found in the areas without vegetation or complex topography. Spatial patterns for both SOC and N were similar.

When comparing the SOC maps of this study (Figure 38) and African soil grids (Figure 18), similar spatial patterns can be found, however the map produced in this study seems to reveal more details. Though, the SOC and N maps of this study are somewhat noisy, and some level of smoothing could improve the visual appearance. Even with the listed faults, the SOC and N maps could be valuable for certain applications. Eswaran et al. (1995) noted that the spatial distribution of SOC and other soil properties is complicated and accepting errors in the geographical variation of SOC could be necessary when creating soil maps.

7.6 Methodological considerations

7.6.1 Suitability of Random Forest for soil properties modelling

RF is a popular machine learning tool used in several scientific applications (Wiesmeier 2011). However, RF has not been yet widely used on soil properties modelling (Wiesmeier 2011; Vågen et al. 2016) but it has high potential to be powerful tool for digital soil mapping (Vågen et al. 2016).

RF was used in this study for modelling the SOC and N, producing reasonable results. When comparing to other studies, results of this study seem to be well in line with them. Wiesmeier et al. (2011) used RF to predict soil properties with good results, R^2 being 0.74 for SOC and 0.78 for N in a semi-arid steppe ecosystem, concluding that RF is promising framework for the spatial prediction of soil properties. Stum et al. (2010) used RF with Landsat and DTM based data to predict individual soil classes. Their conclusion was that RF provided an effective and objective method for their purposes. RF has been also used successfully by Vågen & Winowiecki (2013), Vågen et al. (2013), Vågen et al. (2016) and Hengl et al. (2015) for soil properties modelling. Bou et al. (2010) concluded that if the goal of an analysis is to predict something, then decision-tree based modelling is recommended approach.

However, Were et al. (2015) compared different machine learning algorithms on soil properties modelling, and concluded that RF had highest tendency for overestimation and lowest R^2. Hengl et al. (2015) also noted the slowness of RF modelling, when working with large number of observations and predictors, which was also seen in this study. Also the RF model created in this study has not been validated using other data. RF are also slowly getting more integrated into the field of GIS, which will most likely increase their usage on different purposes. GRASS for example has implementation of RF, which can be used easily with the raster data (Pawley 2016).

7.6.2 Automated variable selection

Automated variable selection methods are gaining popularity and they have been successfully used in vast amount of studies. Selecting variables in regression and classification studies is an important challenge (Hastie et al. 2009). Removing irrelevant variables, selecting all important variables or determining sufficient subset is beneficial for statistical analysis and prediction (Genuer et al. 2010; Genuer et al. 2016). Selecting sufficient variables aids with diagnosis, interpretation and speeds up the data processing (Genuer et al. 2010; Genuer et al. 2016). As Minasny et al. (2008) discussed, the variable selection is an acute problem for soil scientists as the number of possible predictor variables is growing.

One of the goals of this thesis was to identify the most important predictor variables explaining SOC and N content in the study area. The vast number of variables was calculated from ALS and ancillary datasets, almost everything that could be related to soils and found in open source GIS tools. Calculated variables were inserted to automatic variable selection tool. An alternative approach would have been to carefully examine the literature and previous studies to find the most suitable variables explaining this phenomena and use only those for modelling. The latter approach would have been somewhat limiting and time consuming as very little information about important variables or parameters could be found from literature for such as study area. Furthermore, the variables found were not always very consistent between different studies and study areas. Also the uniqueness of the study area added its own challenge on finding information about the phenomena.

It is also important to understand the dangers of using automated variable selection methods. Minasny et al. (2008) discussed the problems of variable selection regarding soil mapping. Authors noted that the likelihood of finding important variables for the models increases when number of variables grows, however this could lead to only good performance on the specific dataset. However, authors also promoted the power of computers in variable selection as it can lead to identification of previously unknown relationships.

The automatic variable selection worked relatively well in this study, and the selected variables made mostly sense. The best SOC model had only three predictor variables, as most of the other models had several selected predicting variables. The models with the highest number of variables generally also performed poorly when comparing to the models with fewer variables. The selected variables in the poorly performing models did not make sense every time. For example SOC model with 0.1 ha study plot size had four elevation variables, only difference being spatial resolution and feature extraction method. The variables were most likely highly inter-correlated.

7.6.3 Scripted workflow

Most of the work was done using open source tools and scripts to make the analysis more repeatable, faster, easier and accurate. Open source tools used in this study are freely available for everyone to use. The used open source tools lacked the easiness and visual niceness of commercial tools, but took it back in the flexibleness and modifiability. The scripts written for this study can be found in GitHub: https://github.com/jehie/soil-modelling . The scripts could be used to understand how the study was implemented, even reworked and extended for other research purposes. In addition to the scripts used in this study, also ancillary datasets used in this study are freely available or based on open access data.

Scripting and automation saved dozen hours of work in this study, if compared to doing work by hand. Adding, removing and modifying variables or changing parameters was fast and easy, though because of the large datasets, some operations took several hours of computing time. Computer time is however cheap compared to human time.

Automated workflows and analyses are also gaining popularity in the soil sciences and in the field of GIS. Soil information systems have been ranked as one of the top priority research question for soil science in the 21st century (Adewopo et al. 2014). Automation is important part of the digital soil mapping and soil information systems.

An example of automation in soil science and in GIS is a project called SoilGrids, which is a system for automated soil mapping. SoilGrids contains collection of soil properties and soil class maps of the world at 1 km and 250 m spatial resolution. These maps are automatically produced using RF models (Hengl et al. 2014; Hengl et al. 2016). With the automation, whenever someone updates the data or modifies some input parameter, a new map can be generated with minimal human work. SoilGrids data is available freely to everyone, and can be downloaded from a web mapping server using web coverage service (WCS) standard. The African soil grid data used in this study was downloaded from this service.

7.7 Future research suggestions

Several improvement ideas or future research suggestions regarding this study were identified during the work. Selected variables for the best SOC model were somewhat surprising, and not among the usually reported variables in the literature. Intensity was identified as promising variable explaining the SOC and N content in the soils. Intensity has not been widely used in soil properties, and further analysis should be done. More accurate pre-processing steps should also be done when using intensity data from ALS (Korpela et al. 2010; Kaasalainen et al. 2009). Range of tangential curvature is not also widely used variable in soil properties modelling, however it makes sense. More analysis should be done.

In several articles, land use and land cover was found to be the most important variable explaining the soil properties (Wiesmeier et al. 2011; Scharlemann et al. 2014; Guo et al. 2002). Using land use and land cover data as part of the modelling could improve the results. Land use data created by Heikinheimo (2015) could provide valuable information regarding the land use in the study area, which also includes information about historical land use in the area. SOC levels are highly related to changes in land cover, and resources can deplete rapidly. However, accumulation of the SOC levels can take up to decades to revert back to natural state (Lal 2004a). Thus, understanding historical land cover changes could provide valuable information for the modelling. For example, even dense vegetation measured by the RS sensor could have been agricultural field for the past decades, thus usually having low levels of SOC. On the other hand, the ALS based variables about vegetation structure, height and density were not very important for the modelling. However, this could be explained partly by the continuously changing land cover in the study area. Also splitting the study area into several classes, based on the land use could be useful. Thompson & Kolka (2005) concluded that creating a single model for all soil types in an area is highly unlikely to be successful.

Imaging Spectroscopy (IS) could provide valuable information on the SOC and N content. Using IS for soil properties modelling have been studied by several studies and generally good results has been achieved (Wulf et al. 2015). Spectral indices from IS have been also powerful (Bartholomeus et al. 2008). Though, a general issue remains vegetative areas, where optical sensors have no direct access to bare soil (Mulder et al. 2011; Wulf et al. 2015). Though, combining ALS data with IS could be very powerful. Identifying different vegetation types from the IS data could be valuable for the SOC and N modelling. For example, Omoro et al. (2013) found that SOC densities in the Taita Hills were generally lower in the plantations than in the indigenous forests.

This study focused on analysing the suitability of ALS for soil properties modelling. Another potential aspect of the ALS in soil mapping could be identifying potential locations for soil restoration and protection. Locations which have the highest SOC and N content are not necessarily the same locations where most of the C sequestration could happen. Studies have shown that different land cover and vegetation types have different potential for soil C sequestration. As Paustian et al. (2016) stated, there are two alternative mitigation approaches, avoiding conversion and degradation of native ecosystems or restoring degraded ecosystems to forests or grasslands. ALS could most likely be effective also on the latter.

The size of plot size should be investigated more close. Same study plot size was used for calculating the ALS vegetation and intensity metrics and for the feature extraction from raster based variables. In other words, area of the feature extraction was similar to the area where the ALS metrics were calculated. This was done to limit the number of predictor variables to reasonable limits. It is possible that ALS vegetation and intensity metrics would explain SOC and N content better in smaller resolutions than topographic variables. An improved approach could be to create models with mixed study plot sizes and spatial resolutions for ALS and topographic variables.

8. Conclusions

The use of ALS and free of cost ancillary datasets was studied for predicting SOC and N content in a heterogeneous landscape in the Taita Hills, Kenya. The field data of this study consisted of 150 topsoil measurements. Topographic and vegetation variables were calculated from the pre-processed ALS dataset, while Landsat time series and African soil grids were used as ancillary datasets, to provide valuable extra information. Several RF models were created for predicting the SOC and N content, and performance of the models was compared and evaluated trough statistical analysis.

Relatively good results were achieved for SOC, while N models performed poorly. Combining ALS with Landsat data resulted in approximately 10 % better results than modeling with only ALS data. The modelling performance of SOC content was in line previous studies from similar environment, but the performance of N models was weaker.

Important predictor variables explaining SOC and N were selected and analyzed. SOC and N maps were produced for the Taita Hills at 90 m spatial resolution. Spatial patterns of SOC and N content can be identified from the generated maps but also inaccuracies were identified. There are clear limitations and challenges to applying ALS for SOC and N mapping, but the automated modelling approach presented in this study could be further developed and additional datasets should be tried in order to improve the model.

9. Acknowledgements

I would like to thank the Ministry of Foreign Affairs of Finland for supporting this thesis through the Building Building Biocarbon and Rural Development in West Africa (BIODEV) project conducted at the Department of Geoscience and Geography, University of Helsinki. I would also like to thanks Ermias Betemariam and World Agroforestry Centre for collecting the field data and providing the SOC and N measurements. A travel grant for visiting the Taita Hills for field work was provided by the Department of Geoscience and Geography. Special thanks for Dr. Janne Heiskanen for excellent guidance and encouragement through the whole process of doing this thesis. I’m also grateful for Prof. Petri Pellikka for the opportunity of being part of his research team and for the opportunity to visit the Taita Hills. Big thanks also for Vuokko Heikinheimo and Elisa Schäfer for field data collection and general support.

10. References

Adewopo, J.B., C. VanZomeren, R.K. Bhomia, M. Almaraz, A.R. Bacon, E. Eggleston, J.D. Judy, R.W. Lewis, M. Lusk, B. Miller, C. Moorberg, E.H. Snyder & M. Tiedeman (2014). Top-ranked priority research questions for soil science in the 21st century. *Soil Science Society of America Journal,* 78: 2, 337-343.

Adhikari, H., J. Heiskanen, E.E. Maeda & P.K.E. Pellikka (2016). The effect of topographic normalization on fractional tree cover mapping in tropical mountains: An assessment based on seasonal Landsat time series. *International Journal of Applied Earth Observation and Geoinformation,* 52, 20-31.

Adhikari, K., Alfred E. Hartemink, Mogens H. Greve, Mette B. Greve, Rania Bou Kheir & Budiman Minasny (2014). Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark. *PLoS ONE, 9*: 8, 0105519.

Amundson, R., A.A. Berhe, J.W. Hopmans, C. Olson, A.E. Sztein & D.L. Sparks (2015). Soil science. Soil and human security in the 21st century. *Science,* 348: 6235, 647-653.

ASPRS Board (2008). *LAS Specification - Version 1.2,* American Society of Photogrammetry and Remote Sensing.

Axelsson, P. (2000). DEM generation from laser scanner data using adaptive TIN models. *International Archives of Photogrammetry and Remote Sensing,* XXXIII: Part B4, 110-117.

Bartholomeus, H.M., M.E. Schaepman, L. Kooistra, A. Stevens, W.B. Hoogmoed & O.S.P. Spaargaren (2008). Spectral reflectance based indices for soil organic carbon quantification. *Geoderma,* 145: 1–2, 28-36.

Beraldin, J., F. Blais & U. Lohr (2010). Laser Scanning Technology. *In* Vosselman, G. & H.G Maas (eds.). *Airborne and terrestrial laser scanning*, 1-44. Whittles Publishing, Dunbeath.

Bivand, R., B. Rowlingson, E. Pebesma, M. Sumner, R. Hijmans & E. Rouault. (2016). Bindings for the Geospatial Data Abstraction Library.

Bot, A. & J. Benites (2005). The importance of soil organic matter: key to drought-resistant soil and sustained food production, 78 p. Food and Agriculture Organization of the United Nations, Rome.

Bou, R., M.B. Greve, M.H. Greve, P.K. Bøcher, R. Larsen & K. McCloy (2010). Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. *Journal of Environmental Management,* 91: 5, 1150-1160.

Brearley, F.Q. & A.D. Thomas (2015). Land-use change impacts on soil processes in tropical and savannah ecosystems: an introduction. *In* Brearley, F.Q & A.D Thomas (eds.). *Land-use change impacts on soil processes: tropical and savannah ecosystems*, 1-7. CABI, Wallingford.

Breemen, N. v. & P. Buurman (2002). *Soil formation,* 2nd edn. 415 p. Kluwer Academic Publishers, New York.

Breiman, L. (2001). Random Forests. *Machine Learning,* 45: 1, 5-32.

Brenner, C. (2010). Building extraction. *In* Vosselman, G. & H.G Maas (eds.). *Airborne and Terrestrial Laser Scanning*, 169-212. Whittles Publishing, Dunbeath.

Brewer, K., J. Monty, A. Johnson, D. Evans & H. Fisk (2011). *Forest carbon monitoring: A review of selected remote sensing and carbon measurement tools for REDD+,* 35 p. Department of Agriculture, Salt Lake City.

Briese, C. (2010). Extraction of Digital Terrain Models. *In* Vosselman, G. & H.G Maas (eds.). *Airborne and Terrestrial Laser Scanning*, 135-168. Whittles Publishing, Dunbeath.

Budiman, M. & T. Bishop (2005). Digital Soil-Terrain Modeling. *In* Grunwald, S. (eds.). *Environmental Soil-Landscape Modeling*, 185-213. CRC Press, Boca Raton.

Bui, E.N., B.L. Henderson & K. Viergever (2006). Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling,* 191: 3, 431-446.

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, J. Wehberg, V. Wichmann & J. Böhner (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.,* 8: 7, 1991-2007.

Daughtry, C.S.T., E.R. Hunt Jr., P.C. Beeson, S. Milak, M.W. Lang, G. Serbin, J.G. Alfieri, G.W. McCarty & A.M. Sadeghi (2012). Remote Sensing of Soil Carbon and Greenhouse Gas Dynamics across Agricultural Landscapes A2 - Liebig, Mark A. *In* Franzluebbers, A. J. & R.F. Follett (eds.). *Managing Agricultural Greenhouse Gases*, 385-408. Academic Press, San Diego.

Drake, B.G., M.A. Gonzàlez-Meler & S.P. Long (1997). More Efficient Plants: A Consequence of Rising Atmospheric CO2? *Annual Review of Plant Physiology and Plant Molecular Biology*, 48: 1, 609-639.

Erdogan, H.E., P.K.E. Pellikka & B. Clark (2011). Modelling the impact of land-cover change on potential soil loss in the Taita Hills, Kenya, between 1987 and 2003 using remote-sensing and geospatial data. *International Journal of Remote Sensing*, 32: 21, 5919-5945.

ESRI (2014). *ArcGIS 10.3: The Next Generation of GIS Is Here*. 10.11.2016. < https://blogs.esri.com/esri/arcgis/2014/12/10/arcgis-10-3-the-next-generation-of-gis-is-here/>

Eswaran, H., E. Van den Berg, P. Reich & J. Kimble (1995). Global Soil Carbon Resources. *In* Kimble, J., E. Levine & B. Stewart (eds.). *Soils and Global change*, 27-45. CRC Press, Boca Raton.

Evans, I.S. & N.J. Cox (1999). Relations between land surface properties: Altitude, slope and curvature. *In* Hergarten, S. & H.J. Neugebauer (eds.). *Process Modelling and Landform Evolution*, 13-45. Springer, Berlin.

Florinsky, I.V., R.G. Eilers, G.R. Manning & L.G. Fuller (2002). Prediction of soil properties by digital terrain modelling. *Environmental Modelling and Software*, 17: 3, 295-311.

Florinsky, I.V. (2012a). Digital Terrain Modeling: A Brief Historical Overview. *In* Florinsky, I.V (eds.). *Digital Terrain Analysis in Soil Science and Geology*, 1-4. Academic Press, Boston.

Florinsky, I.V. (2012b). Influence of Topography on Soil Properties. *In* Florinsky, I.V (eds.). *Digital Terrain Analysis in Soil Science and Geology*, 145-149. Academic Press, Boston.

Florinsky, I.V. (2012c). Adequate Resolution of Models. *In* Florinsky, I.V (eds.). *Digital Terrain Analysis in Soil Science and Geology*, 151-165. Academic Press, Boston.

Fraisse, C.W., K.A. Sudduth & N.R. Kitchen (2001). Delineation of Site-Specific Management Zones by Unsupervised Classification of Topographic Attributes and Soil Electrical Conductivity. *American Society of Agricultural Engineers,* 44: 1, 155-166.

Gallant, J.C. & J.M. Austin (2015). Derivation of terrain covariates for digital soil mapping in Australia. *Soil Research*, 53: 8, 895-906.

Garroway, K., C. Hopkinson & R. Jamieson (2011). Surface moisture and vegetation influences on lidar intensity data in an agricultural watershed. *Canadian Journal of Remote Sensing*, 37: 3, 275-284.

GDAL Development Team (2016). *GDAL - Geospatial Data Abstraction Library*.

Ge, Y., J. Thomasson & R. Sui (2011). Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science,* 5: 3, 229-238.

Geography of Taita (2006). University of Helsinki. 13.11.2016. <http://blogs.helsinki.fi/taita-research-station/geography-of-taita/>.

Genuer,R., J. Poggi & C. Tuleau-Malot. (2016). *VSURF: Variable Selection Using Random Forests*.

Genuer, R., J. Poggi & C. Tuleau-Malot (2010). Variable selection using random forests. *Pattern Recognition Letters,* 31: 14, 2225-2236.

Gomez, C., R.A. Viscarra Rossel & A.B. McBratney (2008). Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma,* 146: 3, 403-411.

Graham, L. (2008). Management of LiDAR data. *In* Petrie, G. & C.K. Toth (eds.). *Topographic Laser Ranging and Scanning*, 295-306. CRC Press, Boca Raton.

GRASS Development Team. (2016). *Geographic Resources Analysis Support System (GRASS GIS) Software*.

Guo, L.B. & R.M. Gifford (2002). Soil carbon stocks and land use change: a meta analysis. *Global Change Biology,* 8: 4, 345-360.

Hall, R. (2008). Soil essentials : managing your farm's primary asset, 182 p. Landlinks, Collingwood.

Hartemink, A.E. & K. McSweeney (2014). Foreword. *In* Hartemink, A. E. & K. McSweeney (eds.). *Soil carbon*, v. Springer, Cham.

Hastie, T., R. Tibshirani & J. Friedman (2009). Model Assessment and Selection. *In* Hastie, T., R. Tibshirani & J. Friedman (eds.). *The Elements of Statistical Learning,* 1-41. Springer, New York.

Heikinheimo, V. (2015). *Impact of land change on aboveground carbon stocks in the Taita Hills, Kenya*. 95 p. Master's thesis. Department of Geosciences and Geography, Faculty of science, University of Helsinki.

Hengl, T., G.B.M. Heuvelink, B. Kempen, J.G.B. Leenaars, M.G. Walsh, K.D. Shepherd, A. Sila, R.A. MacMillan, J. Mendes de Jesus, L. Tamene & J.E. Tondoh (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *PloS one,* 10: 6, e0125814.

Hengl, T., J.M.d. Jesus, R.A. MacMillan, N.H. Batjes, G.B.M. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J.G.B. Leenaars, M.G. Walsh & M.R. Gonzalez (2014). SoilGrids1km — Global Soil Information Based on Automated Mapping. *PLOS ONE,* 9: 8, e105992.

Hiederer, R. & M. Köchy (2011). *Global Soil Organic Carbon Estimates and the Harmonized World Soil Database*, Publications Office of the European Union, Luxembourg.

Holopainen, M., J. Hyyppä & M. Vastaranta (2013). *Laserkeilaus metsävarojen hallinnassa,* 75 p. Helsingin yliopiston metsätieteiden laitos, Helsinki.

Hovi, A. (2013). *Laser- ja kuva-aineistojen radiometria ja geometria - Aineistojen yhdistäminen.,* 35 p. Helsingin yliopiston metsätieteiden laitos, Helsinki.

Hungate, B. A., E.A. Holland, R.B. Jackson, F.S. Chapin, H.A. Mooney & C.B. Field (1997). The fate of carbon in grasslands under carbon dioxide enrichment. *Nature*, 388: 6642, 576-579.

Isenburg, M. (2016). *LAStools: converting, filtering, viewing, processing, and compressing LIDAR data in LAS format*.

Isenburg, M. (2014). *Rasterizing Perfect Canopy Height Models from LiDAR*. 4.11.2016. <https://rapidlasso.com/2014/11/04/rasterizing-perfect-canopy-height-models-from-lidar/>.

Islam, K.R. & R.R. Weil (2000). Land use effects on soil quality in a tropical forest ecosystem of Bangladesh. *Agriculture, Ecosystems & Environment,*79: 1, 9-16.

Jaetzold, R., H. Scmidth, B. Hornetz & C. Shisanya (2010). *Farm management handbook of Kenya, VOL. II.* 1-46 p. Ministry of Agriculture, Kenya, in Cooperation with the German Agency for International Cooperation (GIZ), Nairobi.

Jensen, J.R. (2000). *Remote sensing of the environment : an earth resource perspective,* 544 p. Prentice Hall, Upper Saddle River (NJ).

Jobbágy, E.G. & R.B. Jackson (2000). The Vertical Distribution of Soil Organic Carbon and Its Relation to Climate and Vegetation. *Ecological Applications,* 10: 2, 423-436.

Jones, A., H. Breuning-Madsen, M. Brossard, A. Dampha, J. Deckers, O. Dewitte, T. Gallali, S. Hallett, R. Jones, M. Kilasara, P. Le Roux, E. Micheli, L. Montanarella, O. Spaargaren, L. Thiombiano, E. Van Ranst, M. Yemefack, R. Zougmoré, A. Jones, L. Montanarella & R. Jones (2013). *Soil atlas of Africa,* 176 p. Publications Office of the European Union, Luxembourg.

Ju, X., G. Xing, X. Chen, S. Zhang, L. Zhang, X. Liu, Z. Cui, B. Yin, P. Christie, Z. Zhu & F. Zhang (2009). Reducing environmental risk by improving N management in intensive Chinese agricultural systems. *Proceedings of the National Academy of Sciences,* 106: 9, 3041-3046.

Kaasalainen, S., H. Hyyppa, A. Kukko, P. Litkey, E. Ahokas, J. Hyyppa, H. Lehner, A. Jaakkola, J. Suomalainen, A. Akujarvi, M. Kaasalainen & U. Pyysalo (2009). Radiometric

Calibration of LIDAR Intensity With Commercially Available Reference Targets. *IEEE Transactions on Geoscience and Remote Sensing*, 47: 2, 588-598.

Kahn, B. (2016). *The World Passes 400 PPM Threshold. Permanently*. 12.11.2016. <http://www.climatecentral.org/news/world-passes-400-ppm-threshold-permanently-20738>.

Korpela, I., H.O. Ørka, M. Maltamo, T. Tokola & J. Hyyppä (2010). Tree species classification using airborne LiDAR - effects of stand and tree parameters, downsizing of training set, intensity normalization, and sensor type. *Silva Fennica*, 44, 319-339.

Kristensen, T., E. Næsset, M. Ohlson, P.V. Bolstad & R. Kolka (2015). Mapping Above- and Below-Ground Carbon Pools in Boreal Forests: The Case for Airborne Lidar. *PLOS ONE*, 10: 10, e0138450.

Lal, R. (2004a). Soil carbon sequestration impacts on global climate change and food safety. *Science*, 304: 5677, 1623-1627.

Lal, R. (2004b). Soil carbon sequestration to mitigate climate change. *Geoderma*, 123: 1–2, 1-22.

Lal, R. (2003). Soil erosion and the global carbon budget. *Environment international*, 29: 4, 437-450.

Lal, R. (2010). Managing Soils and Ecosystems for Mitigating Anthropogenic Carbon Emissions and Advancing Global Food Security. *BioScience*, 60: 9, 708-721.

Lamb, J., F. Fernandez & D. Kaiser (2014). *Understanding nitrogen in soils*. 30.10.2016. <http://www.extension.umn.edu/agriculture/nutrient-management/nitrogen/understanding-nitrogen-in-soils/>.

Lemus, R. & R. Lal (2005). Bioenergy Crops and Carbon Sequestration. *Critical Reviews in Plant Sciences*, 24: 1, 1-21.

Li, C., Y. Xu, Z. Liu, S. Tao, F. Li & J. Fang (2016). Estimation of Forest Topsoil Properties Using Airborne LiDAR-Derived Intensity and Topographic Factors. *Remote Sensing*, 8: 7, 1-13.

Lichti, D. & J. Skaloud (2010). Registration and Calibration. *In* Vosselman, G. & H.G Maas (eds.). *Airborne and Terrestrial Laser Scanning*, 83-134. Whittles Publishing, Dunbeath.

Maas, H. (2010). Forestry Applications. *In* Vosselman, G. & H.G Maas (eds.). *Airborne and Terrestrial Laser Scanning*, 213-236. Whittles Publishing, Dunbeath.

Maraseni, T.N. & S.S. Pandey (2014). Can vegetation types work as an indicator of soil organic carbon? An insight from native vegetations in Nepal. *Ecological Indicators,* 46, 315-322.

Mayes, M., E. Marin-Spiotta, L. Szymanski, M. Akif Erdoğan, M. Ozdoğan & M. Clayton (2014). Soil type mediates effects of land use on soil carbon and nitrogen in the Konya Basin, Turkey. *Geoderma,* 232–234, 517-527.

McBratney, A.B., M.L. Mendonça Santos & B. Minasny (2003). On digital soil mapping. *Geoderma,* 117: 1–2, 3-52.

McGaughey, R.J. (2016). *FUSION/LDV: Software for LIDAR Data Analysis and Visualization*, United States Department of Agriculture.

Minasny, B., A.B. McBratney & R.M. Lark (2008). Digital Soil Mapping Technologies for Countries with Sparse Data Infrastructures. *In* Hartemink, A. E., A. McBratney and M.D.L Mendonça-Santos (eds.). *Digital Soil Mapping with Limited Data*, 15-30. Springer Netherlands.

Miransari, M., H. Omidi, N. Korde, M. Amini & S. Maleki (2012). Uptake of Nitrogen by Arbuscular Mycorrhizal Fungi. *In* Miransari, M (eds.). *Soil Nutrients*, 311-320. Nova Science Publishers, New York.

Mirzaee, S., S. Ghorbani-Dashtaki, J. Mohammadi, H. Asadi & F. Asadzadeh (2016). Spatial variability of soil organic matter using remote sensing data. *Catena,* 145118-127.

Mitasova, H. & J. Hofierka (1993). Interpolation by regularized spline with tension: II. Application to terrain modeling and surface geometry analysis. *Mathematical Geology,* 25: 6, 657-669.

Mulder, V.L., S. de Bruin, M.E. Schaepman & T.R. Mayr (2011). The use of remote sensing in soil and terrain mapping — A review. *Geoderma*, 162: 1–2, 1-19.

Myers, N., R.A. Mittermeier, C.G. Mittermeier, da Fonseca, Gustavo A B & J. Kent (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403: 6772, 853-858.

Natural Resources Canada (2015). *Passive vs. Active Sensing*. 26.10.2016. <http://www.nrcan.gc.ca/earth-sciences/geomatics/satellite-imagery-air-photos/satellite-imagery-products/educational-resources/14639>.

Neteler, M. & H. Mitasova (2008). Working with raster data. *In* Neteler, M. and Helena Mitasova (eds.). *Open Source GIS*, 83-168. Springer US, New York.

NOAA (2015). *What is remote sensing?*. 26.10.2016. <http://oceanservice.noaa.gov/facts/remotesensing.html>.

Oksanen, J. (2006). *Digital elevation model error in terrain analysis*, 51 p. Helsinki University Press, Helsinki.

Omoro, L. M. A., M. Starr & P.K.E. Pellikka (2013). Tree biomass and soil carbon stocks in indigenous forests in comparison to plantations of exotic species in the Taita Hills of Kenya. *Silva Fennica*, 47: 2, 1-18.

Ontl, T.A. & L.A. Schulte (2012). Soil Carbon Storage. *Nature Education Knowledge*, 3: 35. 26.10.2016. <http://www.nature.com/scitable/knowledge/library/soil-carbon-storage-84223790>.

Packalén, P., H. Temesgen & M. Maltamo (2012). Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. *Canadian Journal of Remote Sensing*, 38: 5, 557-569.

Pan, Y., R.A. Birdsey, J. Fang, R. Houghton, P.E. Kauppi, W.A. Kurz, O.L. Phillips, A. Shvidenko, S.L. Lewis, J.G. Canadell, P. Ciais, R.B. Jackson, S.W. Pacala, A. McGuire, S. Piao, A. Rautiainen, S. Sitch & D. Hayes (2011). A Large and Persistent Carbon Sink in the World's Forests. *Science*, 333: 6045, 988-993.

Pataki, D.E., D.S. Ellsworth, R.D. Evans, M. Gonzales-Meler, J. King, S.W. Leavitt, G. Lin, R. Matamala, E. Pendall, R. Siegwolf, C. Van Kessel & J.S. Ehleringer (2003). Tracing Changes in Ecosystem Function under Elevated Carbon Dioxide Conditions. *BioScience,* 53: 9, 805-818.

Paustian, K., J. Lehmann, S. Ogle, D. Reay, G.P. Robertson & P. Smith (2016). Climate-smart soils. *Nature,* 532: 7597, 49-57.

Pawley, S. (2016). *r.randomforest - Supervised classification and regression of GRASS rasters using the python scikit-learn package*. 30.10.2016. <https://grass.osgeo.org/grass70/manuals/addons/r.randomforest.html>.

Pellikka, P.K.E., B.J.F. Clark, G.A. Gosa, N. Himberg, P. Hurskainen, E. Maeda, J. Mwang'ombe, L.M.A. Omoro & M. SIljander (2013). Agricultural Expansion and Its Consequences in the Taita Hills, Kenya. *In* Paron, P., D.O. Olagi & C.T. Omuto (eds.). *Developments in Earth Surface Processes*, 165-179.

Pellikka, P.K.E., M. Lötjönen, M. Siljander & L. Lens (2009). Airborne remote sensing of spatiotemporal change (1955–2004) in indigenous and exotic forest cover in the Taita Hills, Kenya. *International Journal of Applied Earth Observations and Geoinformation,* 11: 4, 221-232.

Petrie, G. & C.K. Toth (2008). Introduction to Laser Ranging, Profiling, and Scanning. *In* Shan, J. & C.K. Toth (eds.). *Topographic Laser Ranging and Scanning,* 1–28. CRC Press, Boca Raton.

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

Powers, J.S. & W.H. Schlesinger (2002). Relationships among soil carbon distributions and biophysical factors at nested spatial scales in rain forests of northeastern Costa Rica. *Geoderma,* 109: 3–4, 165-190.

Pyysalo, U. (2000). A method to create a three dimensional forest model from laser scanner data. *Photogrammetric Journal of Finland,* 17: 1, 34-42.

Roecker, S.M. & J.A. Thompson (2010). Scale Effects on Terrain Attribute Calculation and Their Use as Environmental Covariates for Digital Soil Mapping. *In* Boettinger, D. J. L., D.W. Howell, A.C. Moore, A.E. Hartemink, & S. Kienast-Brown (eds.). *Digital Soil Mapping*, 55-66. Springer Netherlands.

Rozanov, B.G., V. Targulian & D.S. Orlov (1990). Soils. *In* Turner, B. L., W.C. Clark, R.W. Kates, J.F. Richards, J.T. Mathews and W.B. Meyer (eds.). *The Earth As Transformed by Human Action*, 203-214. Cambridge University Press With Clark University, Cambridge.

Scharlemann, J., E.V.J. Tanner, R. Hiederer & V. Kapos (2014). Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Management,* 5: 1, 81-91.

Schmidt, J., I.S. Evans & J. Brinkmann (2003). Comparison of polynomial models for land surface curvature calculation. *International Journal of Geographical Information Science,* 17: 8, 797-814.

Schwanghart, W. & T. Jarmer (2011). Linking spatial patterns of soil organic carbon to topography — A case study from south-eastern Spain. *Geomorphology,* 126: 1–2, 252-263.

Seibert, J., J. Stendahl & R. Sørensen (2007). Topographical influences on soil properties in boreal forests. *Geoderma,* 141: 1, 139-148.

Seid, N., B. Yitaferu, K. Kibret & F. Ziadat (2013). Soil-Landscape Modeling and Remote Sensing to Provide Spatial Representation of Soil Attributes for an Ethiopian Watershed. *Applied and Environmental Soil Science,* 20131-11.

Shapiro, M. & O. Waupotitsch (2015). *GRASS GIS Manual: r.slope.aspect*. 13.11.2016. <https://grass.osgeo.org/grass70/manuals/r.slope.aspect.html>.

Singh, B.P., A.L. Cowie & K.Y. Chan (2011). Preface. *In* Singh, B.P, A.L Cowie & K.Y. Chan. *Soil Health and Climate Change,* vii. Springer, Heidelberg.

Soininen, A. (2016). *TerraScan User's Guide,* 1-586 p. TerraSolid, Oy, Helsinki.

Stum, A.K., J.L. Boettinger, M.A. White & R.D. Ramsey (2010). Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah. *In* Boettinger, D.J.L., D.W. Howell, A.C.

Moore, A.E. Hartemink, & S. Kienast-Brown (eds.). *Digital Soil Mapping*, 179-190. Springer Netherlands, Amsterdam.

Thijs, K.W., I. Roelen & W.M. Musila (2014). Field Guide to the Woody Plants of Taita Hills, Kenya. *Journal of East African Natural History,* 102: 1-2, 1-272.

Thompson, J. A. & R.K. Kolka (2005). Soil Carbon Storage Estimation in a Forested Watershed using Quantitative Soil-Landscape Modeling. *Soil Science Society of America Journal*, 69: 4, 1086-1093.

Timlin, D.J., Y.A. Pachepsky & C.L. Walthall (2003). A Mix of Scales: Topographic Information, Point Samples and Yield Maps.&nbsp; *In* Pachepsky, Y., D. Radcliffe and M.H. Selim (eds.). *Scaling Methods In Soil Physics,* 227-240. CRC Press, Boca Raton.

Tsui, C., Z. Chen & C. Hsieh (2004). Relationships between soil properties and slope position in a lowland rain forest of southern Taiwan. *Geoderma,*123: 1–2, 131-142.

UNEP (1987). *Sands of Change: Why Land Becomes Desert and What Can Be Done about It. UNEP Environment Brief No. 2*. United Nations Environment Programme, Nairobi.

UN-REDD (2011). *The UN-REDD Programme Strategy 2011-2015.*

Vågen, T., L.A. Winowiecki, D.L. Tamene & J.E. Tondoh (2015). *The Land Degradation Surveillance Framework (LDSF) - Field Guide v4.1.* World Agroforestry Centre, Nairobi, Kenya.

Vågen, T., L.A. Winowiecki, A. Abegaz & K.M. Hadgu (2013). Landsat-based approaches for mapping of land degradation prevalence and soil functional properties in Ethiopia. *Remote Sensing of Environment,* 134, 266-275.

Vågen, T., L.A. Winowiecki, J.E. Tondoh, L.T. Desta & T. Gumbricht (2016). Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. *Geoderma,* 263, 216-225.

Vågen, T. & L.A. Winowiecki (2013). Mapping of soil organic carbon stocks for spatially explicit assessments of climate change mitigation potential. *Environmental Research Letters,* 8: 1, 1-9.

Vosselman, G. & R. Klein (2010). Visualisation and Structuring of Point Clouds. *In* Vosselman, G. & H.G. Maas (eds.). *Airborne and Terrestrial Laser Scanning,* 45-82. Whittles Publishing, Dunbeath.

Wang, Y., B. Fu, Y. Lü, C. Song & Y. Luan (2010). Local-scale spatial variability of soil organic carbon and its stock in the hilly area of the Loess Plateau, China. *Quaternary Research,* 73: 1, 70-76.

Wasige, J. E., T.A. Groen, B.M. Rwamukwaya, W. Tumwesigye, E.M.A. Smaling & V. Jetten (2014). Contemporary land use/land cover types determine soil organic carbon stocks in south-west Rwanda. *Nutrient Cycling in Agroecosystems,* 100: 1, 19-33.

Wehr, A. (2008). LiDAR Systems and Calibration. *In* Shan, J. and C.K. Toth (eds.). Topographic Laser Ranging and Scanning, 129-172. CRC Press, Boca Raton.

Wehr, A. & U. Lohr (1999). Airborne laser scanning—an introduction and overview. *ISPRS Journal of Photogrammetry and Remote Sensing,* 54: 2–3, 68-82.

Were, K., D.T. Bui, ØB. Dick & B.R. Singh (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators,* 52394-403.

Wiesmeier, M., F. Barthold, B. Blank & I. Kögel-Knabner (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.

Wilson, J. & J. Gallant (2000). Digital Terrain Analysis. *In* Wilson, J. and J. Gallant (eds.). *Terrain Analysis: Principles and Applications*, 1-29. Wiley, New York.

Wulf, H., V.L. Mulder, M. Schaepman, A. Keller & P.C. Jörg (2015). *Remote Sensing of Soils,* 71 p. National Point of Contact for Satellite Images 2 Infosol Unit Remote Sensing Laboratories, Zurich.

Yang, R., G. Zhang, F. Yang, J. Zhi, F. Yang, F. Liu, Y. Zhao & D. Li (2016). Precise estimation of soil organic carbon stocks in the northeast Tibetan Plateau. Scientific Reports, 621842.

11. Appendices

Appendix 1. All predictor variables used in this study.

Input Data	Abbrevation	Description
DTM		
	DTM	Elevation
	CHM	Elevation from ground
	ASP	Aspect
	PRC	Profile curvature
	TAC	Tangential curvature
	FEW	First order derivative (East-West)
	FNS	First order derivative (North-South)
	SXX	Second order derivative (DXX)
	SXY	Second order derivative (DXY)
	SYY	Second order derivative (DYY)
	CAA	Catchment area
	TWI	Topographic wetness index
	TPI	Topographic position index
African soil grids		
	TTH	Soil organic carbon stock in tonnes per ha (Topsoil)
	STH	Soil organic carbon stock in tonnes per ha (Subsoil)
	TKG	Soil organic carbon content (fine earth fraction) in g per kg (Topsoil)
	SKG	Soil organic carbon content (fine earth fraction) in g per kg (Subsoil)
Landsat time series		
	RB1	Band 1, 50 % percentile
	RB2	Band 2, 50 % percentile
	RB3	Band 3, 50 % percentile
	RB4	Band 4, 50 % percentile
	RB5	Band 5, 50 % percentile
	RB6	Band 6, 50 % percentile
	RBR	Vegetation index: Brightness, 50 % percentile
	RGR	Vegetation index: Greenness, 50 % percentile
	RND	Vegetation index: NDVI, 50 % percentile
	RSR	Vegetation index: RSR, 50 % percentile
	RWT	Vegetation index: Wetness, 50 % percentile
ALS		
	E.mode	Elevation mode
	E.stddev	Elevation standard deviation

E.variance	Elevation variance
E.CV	Elevation coefficient of variation
E.IQ	Elevation interquartile distance
E.skewness	Elevation skewness
E.kurtosis	Elevation kurtosis
E.AAD	Elevation average absolute deviation
E.MAD.med	Elevation median of the absolute deviations from the overall median
E.MAD.mo	Elevation median of the absolute deviations from the overall mode
E.L1	Elevation L-moment 1
E.L2	Elevation L-moment 2
E.L3	Elevation L-moment 3
E.L4	Elevation L-moment 4
E.L CV	Elevation L-moment coefficient of variation
E.L skewness	Elevation L-moment skewness
E.L kurtosis	Elevation L-moment kurtosis
E.P01	Elevation 1st percentile
E.P05	Elevation 5th percentile
E.P10	Elevation 10th percentile
E.P20	Elevation 20th percentile
E.P25	Elevation 25th percentile
E.P30	Elevation 30th percentile
E.P40	Elevation 40th percentile
E.P50	Elevation 50th percentile
E.P60	Elevation 60th percentile
E.P70	Elevation 70th percentile
E.P75	Elevation 75th percentile
E.P80	Elevation 80th percentile
E.P90	Elevation 90th percentile
E.P95	Elevation 95th percentile
E.P99	Elevation 99th percentile
CRR	Canopy relief ratio
E.SQRT.MEAN.SQ	Generalized means for 2nd power
E.CURT.MEAN.CUBE	Generalized means for 3rd power
I.minimum	Intensity minimum
I.maximum	Intensity maximum
I.mean	Intensity mean
I.mode	Intensity mode
I.stddev	Intensity standard deviation
I.variance	Intensity variation
I.CV	Intensity coefficient of variation
I.IQ	Intensity interquartile distance
I.skewness	Intensity skewness
I.kurtosis	Intensity kurtosis

I.AAD	Intensity average absolute deviation
I.L1	Intensity L-moment 1
I.L2	Intensity L-moment 2
I.L3	Intensity L-moment 3
I.L4	Intensity L-moment 4
I.L CV	Intensity L-moment covariance
I.L skewness	Intensity L-moment skewness
I.L kurtosis	Intensity L-moment kurtosis
I.P01	Intensity 1st percentile
I.P05	Intensity 5th percentile
I.P10	Intensity 10th percentile
I.P20	Intensity 20th percentile
I.P25	Intensity 25th percentile
I.P30	Intensity 30th percentile
I.P40	Intensity 40th percentile
I.P50	Intensity 50th percentile
I.P60	Intensity 60th percentile
I.P70	Intensity 70th percentile
I.P75	Intensity 75th percentile
I.P80	Intensity 80th percentile
I.P90	Intensity 90th percentile
I.P95	Intensity 95th percentile
I.P99	Intensity 99th percentile
PARA	Percentage of all returns above a specified height
PFRA	Percentage of first returns above a specified height (canopy cover estimate)
ARATFR	Number of returns above a specified height / total first returns * 100
PFRAME	Percentage of first returns above the mean height/elevation
PFRAMO	Percentage of first returns above the mode height/elevation
PARAME	Percentage of all returns above the mean height/elevation
PARAMO	Percentage of all returns above the mode height/elevation
ARAMETFR	Number of returns above the mean height / total first returns * 100
ARAMOTFR	Number of returns above the mode height / total first returns * 100