MODELLING SOIL ORGANIC CARBON IN A SECOND-GROWTH COASTAL FOREST NEAR ELLSWORTH CREEK WASHINGTON USING FIELD MEASURABLE VARIABLES AND RANDOM FOREST REGRESSION

by

Derek L. Thedell

A Thesis
Submitted in partial fulfillment
Of the requirements for the degree
Master of Environmental Studies
The Evergreen State College
August 2023

This Thesis for the Master of Environmental Studies Degree

by

Derek L. Thedell

has been approved for

The Evergreen State College

by

_____

Erin Martin, Ph.D.

Member of Faculty

_____

Date

ABSTRACT

Modelling Soil Organic Carbon in a Second-Growth Coastal Forest Near Ellsworth Creek Washington Using Field Measurable Variables and Random Forest Regression.

Derek L. Thedell

As a result of anthropogenic climate change, discussion of carbon fluxes is of particular focus in modern research. Preservation of carbon sinks provides an opportunity to slow the impacts of climate change and allow room for natural carbon cycles to begin counteracting its effects. Soil represents the largest terrestrial carbon sink, the most sensitive portion of which is Soil Organic Carbon (SOC). This study sought to develop a methodological framework to build a predictive model for SOC storage in the O-horizon of a 318 acre second-growth forested watershed basin in Washington State. Predictive SOC models have been widely utilized on large spatial scales, but are generally under researched on the small, single-forest level. To build this model, six predictor variables were selected based on their capability to be measured in the field without laboratory analysis: (1) overstory cover, (2) understory cover, (3) stand age, (4) elevation, (5) slope, and (6) aspect. 66 soil samples were analyzed for SOC across a single 318-acre watershed basin. SOC was calculated in units of MgC/ha. This SOC data was used to build a random forest (RF) regression model using provided data on the predictor variables described above. To interpret this model, a digital soil map for SOC was generated using the RF model and spatial datasets of the predictor variables. The RF model generated in this study was highly capable of accurately modelling input data but showed limitation at generating predictions with novel data. With our model, we confirmed previously published relationships between stand age and SOC in addition to highlighting the potential for surface soil erosion and local precipitation or wind events to affect SOC storage. Overall, this study shows the capability of machine learning algorithms, such as random forest, in building small-scale predictive models that reflect ecological relationships. This methodological framework establishes the building blocks for low-cost predictive SOC models to contribute to management decisions through the lens of carbon sink preservation.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This study would not have been made possible without the tireless work, review, and guidance of many individuals. I want to thank Dylan Fischer and Steven Quick for providing access to their previously collected data and sharing generously their field and sample collection process. Thank you, Michael Case, for your guidance on project development and field practices. Also, thank you to The Nature Conservancy for the opportunity to collect my own data at Ellsworth Creek Preserve and the access their vast dataset to develop my model. Thank you to Mike Ruth and Nathan Chapman for their insight on GIS and model development.

I would like to thank my thesis reader Erin Martin who welcomed the challenge of this study with open arms. Her invaluable perspective and willingness to push the boundaries to explore new options in this complex field strengthened this study enormously.

Thank you, Amy Salyer, Steve Brand, Jesse Bahr, Holly and Poppy, for joining me in the rain and snow to collect soil samples on the steep slopes of the preserve.

And most importantly, I extend my deepest gratitude to my wife, Jennifer Brand, for your endless support, understanding and love throughout this thesis. Thank you for encouraging me to dig deeper in the hard moments and standing by my side on this wonderful and excruciating adventure.

Chapter 1. Introduction

Worldwide, soils contain twice as much carbon as the atmosphere (Batjes & Sombroek, 1997). Facing impending climate change, preserving carbon in soils is of the utmost importance, as small percentage losses could have drastic impacts on the greenhouse gas effect (Smith, 2012). Soil organic carbon (SOC) represents the portion of carbon in the soil that has been contributed by organic matter inputs (Broadbent, 1965). In the soil, organic matter is broken down over time by soil fauna, where it will either be decomposed and respirated as CO2, stabilized and stored, or leached away (Flaig et al., 1975). As such, the SOC content of the soil is dependent on the variables that affect organic matter inputs and decomposition rates. Organic matter inputs are generally controlled by the local flora's type and amount (Schlesinger & Bernhardt, 2013). Two major controls on decomposition are soil temperature and moisture (Meyer et al., 2018, Rey et al., 2005).

SOC storage varies greatly geographically. Global decomposition rates can range from almost nothing in very dry and cold regions, to extremely high in areas where it is warm and moist. Generally, decomposition is faster in lower latitudes, where there are higher temperatures and precipitation amounts (Zhang et al., 2008). Pacific Northwest forests, particularly in Western Washington and Oregon, are characterized by having extremely high stores of carbon, both above and belowground (Carpenter et al., 2014). This is due in part to their slow rate of organic matter decomposition. SOC content varies on many spatial scales ranging from continental to the single forest stand (Antos et al., 2003).

While efforts to consider the distribution of forest soil carbon on regional scale are important, many of the decisions that lead to SOC losses will occur at small scales by local land and forest managers because of timber harvest. Historically, the largest amount of soil carbon that

has been lost was due to clearcutting native forests to produce agricultural spaces (Jackson et al., 2017). Timber harvest has been shown to directly reduce soil carbon stores due to changes in organic matter input and decomposition rates from temperature and moisture exposure (James & Harrison, 2016). In 2019, it was estimated that 47% of Washington State forests are being utilized for timber, meaning that the carbon stored in those soils is facing increased decomposition and release into the atmosphere (Palmer et al., 2019). Any efforts to attempt to preserve that carbon would require a detailed understanding of where carbon is being stored at the scale of a single forest stand. Processing and collecting samples for all Washington forests to identify regions of high carbon would have a prohibitive cost. Instead, this study proposes the use of digital soil modelling to map carbon distribution at a small scale. A methodology for modelling SOC in a forest stand would help inform timber harvesting and agricultural development operations by identifying regions of high, or low, carbon, and further identify the relationships that the carbon has with topological variables.

To combat the extreme heterogeneity of SOC in forests (Rodrigo-Comino et al., 2020), for this thesis research, six variables have been selected to develop an accurate prediction of SOC storage. Additionally, for the model to have the highest potential impact on small-scale forest owners, each variable was selected because it could either be measurable remotely or easily measured in the field. This would reduce labor and lab costs for forest managers in model generation by removing the need for lab analysis for the predictor variables. For ease of discussion these variables have been separated into two groups, topographical and ecological predictors. The topographical predictors include (1) elevation, (2) slope, and (3) aspect. The ecological predictors include (4) overstory cover, (5) understory cover, and (6) stand age.

Digital soil mapping has been widely utilized to varying success in forest settings (Khaledian & Miller, 2020). Developing a statistically effective model to predict where and how carbon is stored in forest soils poses a particular challenge. One novel approach to this challenge is the utilization of multivariate machine learning (Padarian et al., 2020). If machine learning can effectively produce a soil carbon model for a small-scale forest stand using variables that are measurable in the field, then it may reduce the financial and environmental impact of protecting soil carbon stores.

1.1 Study Overview

This study seeks to evaluate and model the relationships between soil organic carbon (SOC) in the organic horizon with the six field-measurable predictor variables established above in the highly variable soils of a coastal temperate rainforest. We ask the following research question: *How effectively can a random forest model predict soil organic carbon in a Pacific Northwest coniferous temperate rainforest using field measurable variables at a single watershed basin scale?*

To answer this question, we collected soil samples in Ellsworth Creek Preserve near Willapa Bay, Washington. The study area was a 318-acre watershed basin within the northern portion of the preserve that had been historically managed for timber harvest. SOC data was gathered by collecting 54 10 cm soil samples that were measured in the lab using CHN analysis. We used previously collected SOC data to bolster this dataset to 66 points within the basin. Ecological predictor data was contributed by The Nature Conservancy and topographical predictor data was gathered using LIDAR from USGS. A multivariate predictive model was developed using random forest regression. This model was then used to generate a digital SOC storage map of the basin and analyzed for regional SOC hotspots.

Overall, the model predictive performance was found to be low, though comparable to previous research at this spatial scale. Our model highlighted the complexity of the SOC system in subject forests, particularly in the topmost soil layer. Though quantitative predictive performance was low, the model successfully replicated previous findings on the relationship between SOC and stand age following timber harvest. The model challenges were likely due in part to data limitations, sample bias and overfitting. This study serves as a conceptual proof of concept for developing SOC models on small spatial scales in coastal forests. We believe that with higher data quantities and additional model tuning, this tool would provide the opportunity to quickly evaluate SOC stores ahead of management decisions.

Chapter 2: Literature Review

2.1 Introduction

Soils hold more carbon than any other terrestrial sink (Lal et al., 2021). With the current and impending impacts of climate change, developing an understanding of how, and why, carbon is being stored in the soil is of utmost importance. Recent research has focused on the Soil Organic Carbon (SOC) dynamics of wooded areas, particularly with a focus on protecting their SOC stores (Sedjo & Sohngen, 2012; Carpenter et al., 2014). Due to the spatial and temporal variability of soil carbon in forest settings, developing a statistically effective model to predict where and how carbon is stored in forest soils poses a particular challenge (Rodrigo-Comino et al., 2020). One novel approach to this modelling challenge is the utilization multivariate machine learning to develop a digital soil map (DSM, Padarian et al., 2020). This chapter will seek to consider the relevant literature on SOC and DSM.

To begin, a consideration of the development of SOC within forest ecosystems will be used to find the most relevant controls to the project goals. This will involve describing how soil is formed and typically characterized, the ecological relationship between forests and soils, and how environmental and topological variables affect SOC. Once the relevant variables have been identified, a description of machine learning modelling will follow including recent advances in their use in digital soil mapping.

2.2 Soil Characterization

2.2.1 Soil Formation & Evolution

Soil formation begins with the weathering of geological minerals into very small fragments, typically caused by wind or water (Hillel, 1998). These materials react to this

weathering in a range of ways – resistant primary minerals, such as quartz, will continue to break down into very small particles forming sandy soils, while other more chemically reactive minerals will begin chemically reacting and forming into secondary minerals, such as clay (Barton, 2002). These minerals combine to form the soils that are found across the world. In general, soil is made up of all three phases of matter, gases, liquids, and solids, and is built from many foundational chemical elements and minerals. At it's most broad, soil formation and development is controlled by variations in temperature and moisture (Hilgard, 1882). These factors play a role in every stage of soil development and can cause both continent-scale and local-scale variation in soil properties.

The physical and chemical structure of soil is highly dynamic, depending on countless factors, which causes it to be notoriously difficult to characterize (Simonson, 1968). Soil formation and evolution, known as pedogenesis, depends on a few primary controlling factors: parent material, climate, topography, and biota (Hillel, 1998; Bockheim et al., 2005). These factors each play a role in the relationship between the soil and the moisture and temperature characteristics present.

The parent material represents the original weathered geological minerals that form the soil's foundation (Bockheim et al., 2005). This fundamental material make-up dictates how the soil reacts to variations in temperature and moisture and can affect a variety of chemical and physical properties of the soil. One example of a parent material factor is acidity, which has been shown to have a significant effect on the chemical reactions that occur within it (Penn & Camberato, 2019).

Climate describes the long-term evolutions of weather on the timescale of soil formation. Hilgard emphasized in 1882 the inseparability of soil formation and climate. He describes how climate affects large-scale variations in temperature and moisture and can have enormous effects

on soil formation and characteristics (Hilgard, 1882). Soils vary widely across the world, due in large part to global climatic variation.

At the landscape-scale, topography has a significant effect on pedogenesis through water run-off and drainage patterns (Bockheim et al., 2005). Small-scale depressions or basins can cause accumulation, which affects the water penetration of the soil (Jenny, 1994). Topographically flat regions, such as wetlands, can have shallow water tables causing increased duration of soil saturation and leading to variation in soil properties.

Regional biotic factors from organisms also play a significant role in petrogenesis (Comber, 1938). Comber described the complexity of the relationships between organisms and soil as so great that it can be difficult to form any base conclusions on their relations (1938). Organisms on a local scale drive changes in regional temperature and moisture while also contribute to the evolution of soil parent material (Bockheim et al., 2005).

Topography and local soil parent material make-up contribute significantly to small scale variations in soil pedogenesis (Bockheim et al., 2005). This study will focus on these factors as a tool to understand soil organic carbon. See *2.4 Environmental & Topographical Controls on Soil Organic Carbon*.

2.2.2 Soil Organic and Inorganic Carbon

Carbon in the soil represents only 10% of the mass of the entire chemical and physical soil make-up on Earth (Rice et al., 2023). It is stored in two primary forms: soil organic carbon (SOC) and soil inorganic carbon (SIC) (Hillel, 1998). SOC is the portion of soil carbon that is contributed by, and constructed primarily of, organic compounds, while SIC consists of mineral fragments such as carbonate (Bai et al., 2017). For most soil types, a majority of the carbon is stored in the

form of SOC, though in arid regions SIC can dominate (Nelson & Sommers, 1996). SOC has been the primary focus of recent research in response to climate change due, in part, to its instability compared to SIC (Han et al., 2016).

Soil organic carbon is stored entirely in soil organic matter (SOM) within the soil profile (Nelson & Sommers, 1996). Generally, SOM consists of organic material from dead plants, animals, fungi and microbial communities. The exact form and nature of SOM depends greatly on the biotic system where the soil is present (Kononova, 2013). In section *2.3 Forest and Soil Ecological Relationship* we describe SOM evolution in the context of forest soils. SOM is a generally unstable form of organic matter that is easily accessed by biotic communities through decomposition and respiration (Schlesinger & Bernhardt, 2013). SOC represents between 40-60% of the SOM fraction by mass, as the biological make-up of organic material varies depending on the organism (Howard, 1965). As SOM is decomposed and respirated, this carbon is transitioned into CO_2 or more stable forms such as humus. The rate of respiration and decomposition is broadly controlled primarily by temperature, moisture, and access to organic material (Lenton & Huntingford, 2003). The changes in these controls are caused by environmental and topological factors and are the subject of this study (Bockheim et al., 2005). Additionally, there is consideration for the effect of the chemical structure of the SOC present in the soil on respiration rates (Killops & Killops, 2013). Previous research indicates that the form of the SOM that establishes SOC content can control chemical properties such as pH that change how that material is respirated and decomposed (Macko & Estep, 1984). This study does not consider these properties as they are difficult or impossible to quantify in the field, and are outside the bounds of the project.

SOC preservation has been identified as a natural climate solution in response to rising greenhouse gas emissions (Minx et al., 2018). Bossio et al. established that worldwide soil carbon stocks represent 25% of the global potential for natural climate solutions (2020). Approximately half of that potential is in the protection of currently present soil carbon stocks. An estimated 2400 Pg of carbon in the form of SOC is stored in the top to 2 meters of soil globally (Han et al., 2016).

2.2.3 Soil Structure & Texture

Soil texture has been well correlated with SOC content through both physical make-up and stability (Peng Xinhua et al., 2013; Burke et al., 1989; Nichols, 1984). Soil texture refers to the relative presence of three soil particle sizes – clay, silt and sand (Hillel, 1998). The sizes of these particles range from 1 mm (sand) to 10^{-3} mm (clay). There is no consistent texture classification schema globally (Duarte et al., 2018), but one example can be found in figure 1 by the Natural Resource Conservation Service from the United States Department of Agriculture.

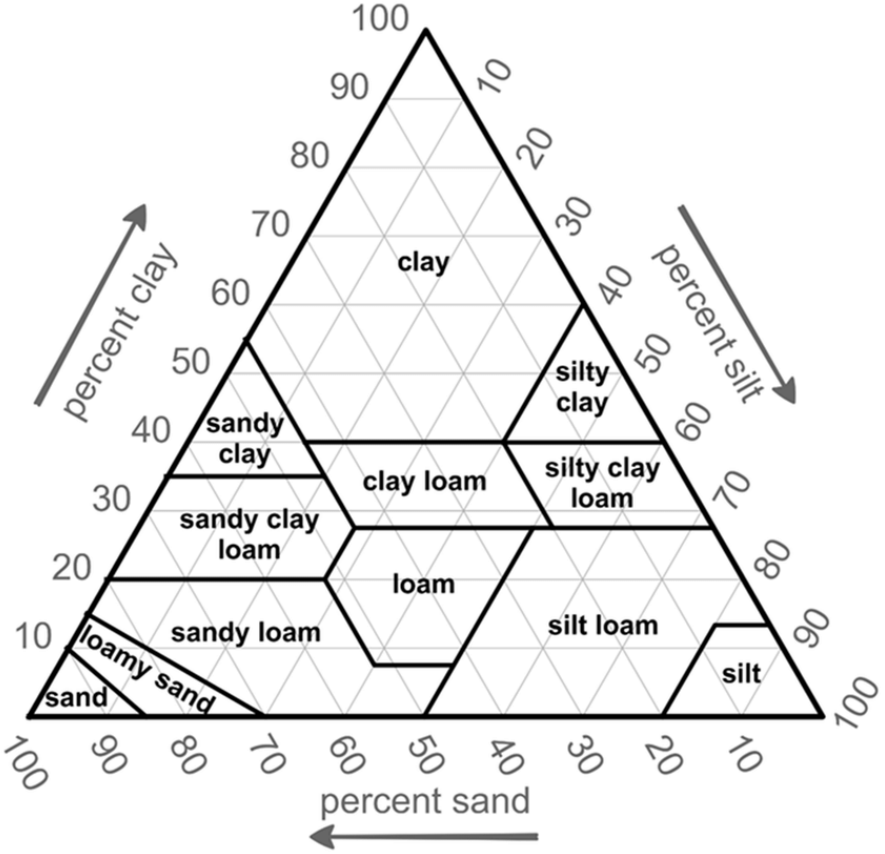


Figure 1. Soil texture classification triangle (USDA, n.d.).

Soil texture is also associated with water retention, aeration, and biotic movement through soil which can directly affect the stability of SOM present in the soil (Weil & Brady, 2017). This relationship is not consistent with all forms of SOM, so previous studies have proposed a schema for distinct SOM fractions based on their decomposition rates (Wander, 2004). These include inert organic matter, dissolved organic matter and particulate organic matter (Chan, 2001). Texture has been shown to directly relate to the stability of SOM, though the specifics of the relationship are not well understood (Nciizah & Wakindiki, 2012). Nciizah & Wakindiki found that SOM content increases in soils with higher clay contents. Microbial mineralization of organic matter has also been shown to slow in soils with a greater clay content causing more of the carbon to be stored in the soil (McLauchlan, 2006). Soil texture is susceptible to land use disturbance, which has been strongly associated with an increase in SOM respiration during tillage and timber harvest events (Minx et al., 2018; James & Harrison, 2016).

2.2.4 Soil Profile & Depth

As soil develops over time, its parent material and physical make-up begins to go through transitions that form discrete layers (Hillel, 1998). These layers are known as horizons and they make up the soil profile (Schlesinger & Bernhardt, 2013). Development into a profile is caused by a combination of physical and chemical weathering, water erosion and biological activity that occur over long timescales. Each horizon contains unique biogeochemical processes that relate to SOC – but the most relevant to this study is the O-horizon.

The topmost layer or horizon in the soil profile is known as the organic, or O-horizon (Hartemink et al., 2020). The O-horizon is characterized by a high organic carbon concentration of whole, partially, or completely decomposed organic matter. This layer is the topmost organic litter found on the soil surface that is the direct result of organic matter inputs. Typically, the O-

horizon contains small amounts of mineral soil, less than half by weight and far less by volume. The thickness of the O-horizon is highly variable depending on the regional characteristics, soil taxonomy, and anthropogenic disturbance (Hillel, 1998; Solomatova & Sidorova, 2008). Overall, it can range from 0 cm (absent) to greater than 50 cm (Seyfried et al., 2021; Stutter et al., 2009).

In the context of SOC, the O-horizon stores the vast majority of organic carbon in the soil profile (Zhang & Hartemink, 2019). This is due to the organic matter that makes up its taxonomy. As a function of depth, SOC depletes quickly due to the age of deep soils providing longer opportunity for decomposition and respiration (figure 2). Additionally, carbon retained in deeper horizons is significantly more stable (Jobbágy & Jackson, 2000). As such, the O-horizon not only contains high amounts of SOC but that carbon is more susceptible to release through respiration and decomposition.

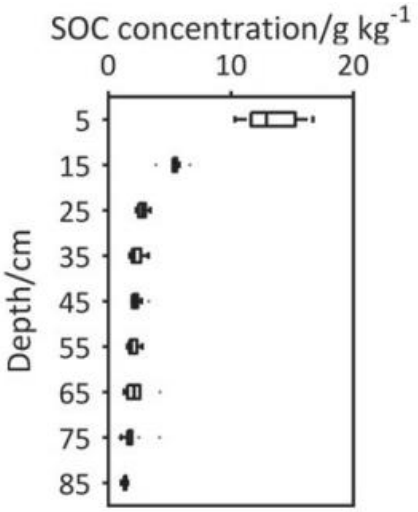


Figure 2. Soil texture classification triangle (Zhang & Hartemink, 2019).

2.2.5 Regional Soil Characterization

Soils of North American Pacific Northwest (NAPC) coastal forests are challenging to succinctly characterize (Carpenter et al., 2014). They are known for having extreme temporal and spatial heterogeneity which can lead to oversimplifications when characterization is attempted. Carpenter et al. found that across NAPC forests, there is a high diversity in soil types reaching across 8 orders and 31 suborders (2014). Specific characterizations of coastal forests in the NAPC are lacking, with much of the research efforts being focused on boreal and northern forests (McNicol et al., 2019).

NAPC coastal forests were found to have high SOC storage, with Washington forests storing a median of 211 MgC/ha (Carpenter et al., 2014). This can be compared to an overall storage of 143 MgC/ha for the entire NAPC region or 60-100 MgC/ha across the entire United States (Homann et al., 2007). The high carbon storage of the region has been associated with both the temperate and precipitation climate which leads to high moisture content of the soils, which produces ideal growing conditions for the characteristic conifers of the area (Waring & Franklin, 1979). Additionally, Shuur et al. attributed this carbon storage to the anerobic environment caused by the high moisture in the soils which reduces the respiration rate of stored carbon (2001). The soils were found to have a high moisture storage capacity due to the restricted drainage associated with thick, carbon dense organic horizons in the soil.

While overall findings showed high SOC storage and moisture content, Carpenter et al. also found a high variability in soil characteristics across the NAPC, indicating that pedogenesis was spatially inconsistent in the region (2014). This indicates an opportunity for further research into the high spatial variability of carbon storage in Pacific Northwest coastal forests.

2.3 Forest and Soil Ecological Relationship

Soil and plants form the foundation of forest ecosystems (Wardle et al., 2004). Their interactions drive the cycling of nutrients and water throughout the environment, both of which are pivotal to ecological success (Attiwill & Adams, 1993). Plants contribute biomass that is the primary source of carbon in forest soils, they move and utilize water present in the soil, and their roots affect the soil structure. The ecosystem of soil organisms mineralizes nutrients organic matter provided by plants and animals into accessible forms required for plant growth. Due to this relationship, a full consideration of the interactions between soil and plants is essential to understand patterns of soil organic carbon (SOC) in Pacific Northwest temperate rainforests.

2.3.1 Nutrient and Carbon Cycling

Soil nutrients, including carbon and nitrogen, cycle through forest ecosystems by undergoing many physical and chemical transitions (Ebermayer, 1876; Attiwill & Adams, 1993). The complexity of these cycles are significant enough to warrant entire fields of research. For the purposes of this study, a macroscopic picture is sufficient to understand how certain factors may drive soil carbon content.

Floral, faunal, and fungal communities drive cycling transitions through decomposition, respiration, and photosynthesis (Foster & Bhatti, 2005). Nutrients are accumulated into living plant biomass through uptake from the soil and are then returned through dead plant matter that falls to the forest floor before being mineralized by microbial and fungal organisms (Gower, 2003; Schlesinger & Bernhardt, 2013).

In forest ecology research this carbon cycle relationship is conveyed in terms of the systems gross primary production (GPP, Schlesinger & Bernhardt, 2013). GPP is the total amount of carbon that is captured by vegetation in the target area through photosynthesis. But, plant organisms also convert oxygen into carbon dioxide to produce energy for biomass development and cell maintenance, in a process known as plant respiration. When plant respiration of carbon is considered and subtracted from GPP, a picture of the total carbon captured as biomass remains, known as net primary production (NPP).

Eventually, the plant biomass accumulated by NPP will fall to the forest floor and begin its interaction with the soil ecosystem. Soil heterotrophic organisms including bacteria and fungi will begin to decompose available plant biomass, releasing a significant portion of the stored carbon through respiration, often symbolized as Rh. The remaining carbon that was not respired

into the atmosphere is stored in the soil as soil organic carbon. SOC can come in many different forms, though a majority potion is stored in humus, a biochemically amorphic substance and byproduct of decomposition (Flaig et al., 1975; Schnitzer, 2015). Other nutrients that were stored in the plant biomass, such as nitrogen and potassium, are also mineralized as part of decomposition into a form that is accessible for plant growth (Schlesinger & Bernhardt, 2013).

Additionally, nutrients can be transported within and without the system in water through leaching (Lehmann & Schroth, 2002). Water from rainfall or other inputs passes through the soil and captures soluble nutrients before leaving the system below the root zone, becoming unreachable for plant growth. The strength of this effect is related to the cation exchange capacity (CEC) of the soil, and as such the available anions (Johnson et al., 1982). In general, leaching is only significant in humid environments with large water inputs, such as the Pacific Northwest (Cole, 1995).

2.3.2 Belowground Plant Growth

Input of organic matter from plants through litterfall and root growth is one of the primary sources of carbon in soils – as such, the success and growth of flora has direct ties to SOC content (Yanai et al., 2003; Wang et al., 2016). Plant and root growth has been shown to respond to environmental factors such as the physical, chemical, and biological conditions of the soil (Passioura, 2002). Plant systems will undergo physiological responses to changes in environmental factors to ensure survival that are driven by signals from root systems.

The impact of these root systems on the soil carbon pool is relatively understudied, though their impact cannot be understated (Finér et al., 2011). Soil hardness, a measure of a soil's resistance to compression, and soil moisture content have been related with changes root growth

(Bengough & Mullins, 1990). Root penetrative depth generally decreases as soil hardness increases to a maximum of 1 MPa of hardness, where root penetration stops. Moisture content is interconnected with soil hardness and has strong correlation to root growth (Passioura, 2002; Lalnunzira et al., 2019). Root uptake of water will cause the soil to become more hardened and thus reduce following root growth. Additionally, moisture release from roots will loosen the soil, increasing root penetration (McCully, 1995). This triangular relationship between soil hardness, moisture content and root growth fuels the subterranean carbon cycle in the soil. Fine root mass provides up to 20% more carbon than inputs from litterfall in forests (Wang et al., 2016). The majority of the contribution from these root systems resides in the top 20 cm of the soil layer, where 90% of total fine root mass occurs. As such, measuring soil moisture and carbon content within the top 20 cm of soil should accurately account for the impact of subterranean plant growth on soil carbon.

Plant biomass has also been shown to influence soil moisture, which is also directly tied to carbon content (Keenan et al., 2013; Schlesinger & Bernhardt, 2013). This occurs through two mechanisms: photosynthesis and transpiration. During photosynthesis, plant organisms convert water and carbon dioxide into organic carbons and oxygen (Schlesinger & Bernhardt, 2013). To fuel this biological process, plants utilize their stomal openings and intake carbon dioxide from the air and water from the soil or air. When the stomata are left open to collect carbon dioxide, stored water within the plant begins to evaporate in a process called transpiration. Transpiration, in effect, acts as a pathway for soil moisture to be removed from the forest-soil system and released instead into the atmosphere (Bittelli et al., 2015). The relationship between the carbon intake from photosynthesis and water loss from transpiration is quantified as water-use-efficiency. In general, transpiration accounts for a majority of moisture loss in the soil (Gardner & Ehlig, 1963).

Plants and their associated biophysical processes orchestrate soil carbon, and in many ways act as a pathway for atmospheric carbon to reach the soil. As such, plant growth and density acts as a major predictor of both soil carbon itself and soil moisture and will be a vital variable in developing a carbon model.

2.3.3 Timber Harvest

Active timberlands have been heavily researched for how harvest changes soil organic carbon (James & Harrison, 2016). Timber harvest not only removes aboveground carbon stored in tree biomass, but also has been shown to strongly effect belowground soil carbon (Yanai et al., 2003). This finding was made notable in 1981 by Wallace Covington who developed a temporal model of soil organic matter following timber harvest that has since become known as the Covington curve shown in figure 3 (Covington et al., 1981).

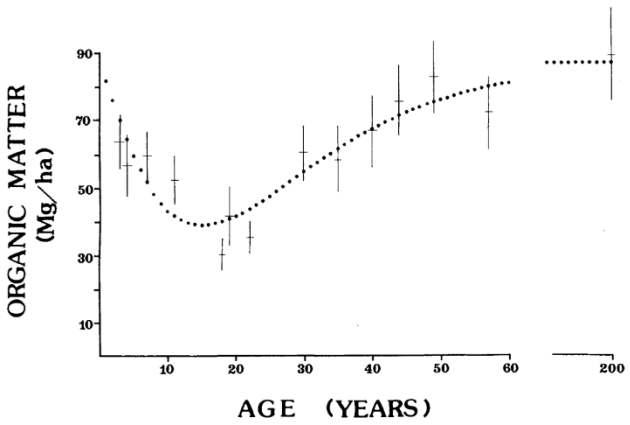


Figure 4. Forest floor organic matter following harvest in a second growth hardwood forest (Covington et al., 1981).

While there is not a research consensus as to the magnitude of the net effect of harvest on SOC, James & Harrison's meta-analysis showed that harvest reduces a forest stand's soil carbon content by 11.2% in total, and 30.2% in the O-horizon with a recovery period of 60 years, consistent with Covington's original findings. This loss of carbon has been attributed to an increase in decomposition rates caused by soil disturbance and increased environmental exposure following harvest (Yanai et al., 2003). During harvest methods where stump systems are removed or destroyed, the soil surrounding each stump is disturbed similarly to an agricultural till. Soil

disturbance has been heavily researched for its connection to soil nutrient and carbon loss (Govindasamy et al., 2021). Additional analysis of decomposition rates in pacific northwest forests following timber harvest have shown a decrease of decomposition rate shortly after harvest (Binkley, 1984; Prescott, 1997). Instead, loss of input from litterfall or contribution of woody debris may instead be the driving factor behind the SOC losses associated with timber harvest (Clarke et al., 2015; Yanai et al., 2003).

2.4 Environmental & Topographical Controls on Soil Carbon

The primary predictors for SOC content in forest soils are temperature, moisture, topography, and organic matter inputs (Meyer et al., 2018; Tsui et al., 2004; Winkler et al., 1996). These factors have been shown to be correlated with carbon cycling processes such as respiration, mineralization, leaching and pedogenesis. In generation of a predictive model as part of this study, due to the temporal scale, it would not be feasible to measure temperature, moisture, or organic matter inputs directly as there would be little variation. Instead, we investigate additional variables that serve as analogs for one of the primary predictors that have been shown to be correlated with SOC content.

2.4.1 Temperature

One of the primary relationships between soil temperature and soil organic carbon is through respiration or mineralization (Winkler et al., 1996). Generally, soil respiration tends to increase with temperature due to an increase of biological activity (Onwuka, 2018; Rey et al., 2005). This increase in respiration rate eventually plateaus and begins to decrease for very warm temperatures, above what would occur naturally in the field. Many studies have highlighted the importance of considering temperature in soil organic carbon modelling (Fang et al., 2005; Meyer

et al., 2018; Onwuka, 2018). Temperature varies on two temporal scales: a daily cycle with the sun, and an annual cycle with the seasons. Daily temperature fluctuations are relatively consistent and occur on a more rapid scale than the rate of respiration and mineralization so have very little effect on SOC.

2.4.2 Moisture

Moisture has been shown to have a significant effect on soil respiration and mineralization rates (Onwuka, 2018). Unlike temperature, respiration tends to reach a maximum rate in an optimal moisture saturation range (around 60%), and generally decreases outside of that range (Howard & Howard, 1993). As the soil becomes more saturated, it begins to become anoxic which causes a dramatic decrease in respiration rates as anerobic microorganisms are generally less biologically active (Szafranek-Nakonieczna & Stêpniewska, 2014). This means that for areas on the extreme ranges of moisture content soil mineralizes more slowly. For Pacific Northwest temperate rainforest soils, the upper range of moisture content is most relevant. Soils with a high moisture content have been shown to also have large carbon stores (Carpenter et al. 2014). This is due in large part to the lower decomposition rate causing more carbon to remain immobilized in the soil. Additionally, for areas that experience severe droughts, Rey et al. (2005) found that sudden increases of moisture content will cause a spike in mineralization. Washington state is predicted to experience an increase of drought events as the region is impacted by anthropogenic climate change (Mote & Salathé, 2010). These drought events will further drive down the carbon mineralization rate and cause immobilized carbon stores to continue to grow.

2.4.3 Organic Matter Inputs

Organic matter inputs provide a direct relationship with SOC content in the soil through an increase of SOM (Yanai et al., 2003; Wang et al., 2016). One of the primary forms of organic matter input is through above- and below-ground plant growth. The relationship between organic matter inputs and SOC is not well understood, and there is some disagreement in the literature. Generally, it has been considered that for regions of high primary productivity the increased quantity of litterfall would be associated with increased SOC storage (Carvalhais et al., 2014; Cotrufo et al., 2015). However, recent findings suggest that increased litterfall is not necessarily associated with higher SOC storage, rather it depends on a variety of environmental and biological factors (Xiong et al., 2020). The material that makes up the SOC stored as SOM is primarily a byproduct of litter decomposition, which involves significant mass loss. This researched uncertainty identifies a need for further investigation in the role organic matter inputs on SOC.

2.4.4 Ecological Variables

Due to the infeasibility of directly measuring the primary SOC controlling variables of moisture, temperature, and organic matter inputs, this study will utilize three analog ecological variables. These variables include (1) Overstory Cover, (2) Understory Cover, and (3) Stand Age.

(1) Overstory, or canopy, cover describes the magnitude of the forest canopy and has been shown to be correlated with SOC content on the forest floor (Boča et al., 2014; Maraseni & Pandey, 2014; Saimun et al., 2021; Liu et al., 2014). Maraseni & Pandey found that forest soils underneath dense canopies (>70% cover) stored more SOC than sparse (<70%) canopies (2014). This is due, in part, to the increased organic matter input provided by the litterfall of a denser overstory.

Additionally, increased SOC storage is likely associated with a change in respiration rates as a response to a temperature and moisture microclimate (Liu et al., 2014; McCarthy & Brown, 2006; Cahoon et al., 2012). Underneath tree canopies, temperature is regulated through shade cover and evapotranspiration, which would slow respiration and store more carbon. Overstory cover provides a spatially variable analog for temperature, moisture and organic matter input.

(2) Understory, or vegetative, cover describes the distribution of herb and shrub plants underneath the forest canopy. Increased understory cover has been shown to increase SOC storage as well as protect the soil from erosion (Ruiz-Colmenero et al., 2013). Additionally, Zhang et al. found that understory cover may increase SOC storage, and its absence can cause increased carbon loss, though this effect is not as prevalent in coniferous forests (2022). The removal of understory vegetation causes multiple changes to the soil system by reducing organic matter inputs, reducing SOC release through root respiration, and increasing the mean annual temperature of the soil by removing shade cover. This established that understory cover can serve as an analog for both temperature and organic matter input.

(3) Stand age and time since last harvest have been associated with changes in SOC stocks (Yanai et al., 2003; Chen & Shrestha, 2012; Deng et al., 2022). In active or historic timberlands, the age of the forest stand can be considered equivalent as the age since last timber harvest. As described in the previous section, timber harvest can cause changes in decomposition rates, litterfall input and environmental exposure (Binkley, 1984; Prescott, 1997). James & Harrison found that SOC recovery following timber harvest can take between 60-100 years, with an overall resulting 17% loss of SOC storage (2016). Additionally, stand age helps consider temporal changes in overstory and vegetative cover as the new growth of the forest developed. When paired with

direct overstory and understory cover measurement, it forms a more comprehensive perspective on the temperature, moisture and organic matter input variation caused by forest development.

2.4.5 Topographical Variables

Topography is not only directly associated with pedogenesis, but also has an effect on temperature, moisture and organic matter inputs. To evaluate the effect of topography on SOC, this study will utilize three variables: (1) Elevation, (2) Slope, and (3) Aspect. The three variables can characterize the overall topography of the study region.

(1) Elevation has an inverse relationship with soil temperature and moisture, as temperatures tend to decrease at higher elevations, and water leaches downhill (Franzmeier et al., 1969; Bhardwaj et al., 2016). This causes respiration rates to typically slow at higher elevations increasing SOC storage overall (Tsui et al., 2004). Additionally, the quantity and type of vegetation present can vary greatly depending on elevation, especially in areas of high topographical variability (Whittaker & Niering, 1975). This also impacts SOC storage by changing the nature of organic matter inputs.

(2) Slope directly affects moisture runoff and erosion rates in soil (Hall, 1983). On slopes, water-runoff rate increases and collects in downslope regions of low gradient. This leads to areas of high slope having decreased moisture holding capacity which changes respiration rates depending on the average moisture content (Tsui et al., 2004). Additionally, high gradient slopes lead to an increase of soil erosion, particularly on the organic horizon (Olson, 2010). As such, in regions of high gradient slope there is lower SOC storage due to the downslope erosion of the organic layer and changes to respiration rates in the soil (Li et al., 2019).

(3) Aspect describes the cardinal angle of a slope and has been found to influence pedogenesis and SOC storage (Zhu et al., 2017). Franzmeier et al. established a relationship between aspect with temperature and organic matter inputs (1969). For steep slopes, aspect can dictate sun exposure and associated temperature fluctuations. In the Pacific Northwest, the sun primarily lies in the southern region of the sky, which means that south-facing slopes would experience increased sun exposure than north-facing. While research is limited in the Pacific Northwest on the connection between aspect and SOC storage, studies in the Mediterranean have demonstrated that north-facing slopes tend to store more SOC (Jakšić et al., 2021; Lozano-García et al., 2016). Jakšić et al. established that this was likely due to increased vegetative cover and overall biomass on the more shaded Northern slopes (2021). While the ecology and climate of the Mediterranean is considerably different than the Pacific Northwest, they lie on similar latitudes and their aspect relationships should be comparable.

2.5 Machine Learning Modelling

2.5.1 Machine Learning Overview

As one of the most quickly growing fields, machine learning (ML) algorithms and methods have improved significantly over the last three decades (Jiang, 2022). This field of study found its inception in the 1950s, made famous by Allan Turing's computer intelligence test (Turing, 1950). Frank Rosenblatt has been attributed to the first practical example of ML, where a letters machine was taught to recognize that of the alphabet (Rosenblatt, 1957). Since then, the work of simulating human intelligence and learning using computers has been attributed the blanket term *artificial intelligence*. Ultimately, ML is a mathematical and statistical analysis tool and as such shares many similarities to traditional statistics (Jiang, 2022). ML features both parametric and non-parametric methodologies that can be used for both continuous regression

analysis and discrete classification. It accomplishes this analysis using a series of knowledge rules that can either be manually provided by the user, or more recently developed using learning methods (Nandi & Pal, 2022). Learning, in this context, is the result of mathematical optimization of how incorrect the model is compared to measured data, known as a cost function (Yi et al., 2020). In essence, ML methods seek to minimize how wrong its predictions are using some form of learning. Most common ML methods utilize supervised learning. During supervised learning, before the model is constructed, a subset of the data (typically between 15-30%) is set aside for model evaluation, known as the testing subset (Jiang, 2022). The ML model will then evaluate its developed predictions from the training subset of the data (70-85%) and adjust the model accordingly.

2.5.2 Decision Tree Algorithms

The first published use of decision trees (DT) for predictive modelling was in 1984 by Leo Breiman in *Classification and Regression Trees* (Breiman et al., 1984; Genuer & Poggi, 2020). In this publication, Breiman describes a new non-parametric supervised ML method to generate DTs which is known today as Classification and Regression Tree (CART) analysis. These DTs are built from many nodes, which form branches and ultimately reach terminal nodes or leaves (James et al., 2014). They are constructed top-down, beginning at the root node which includes all

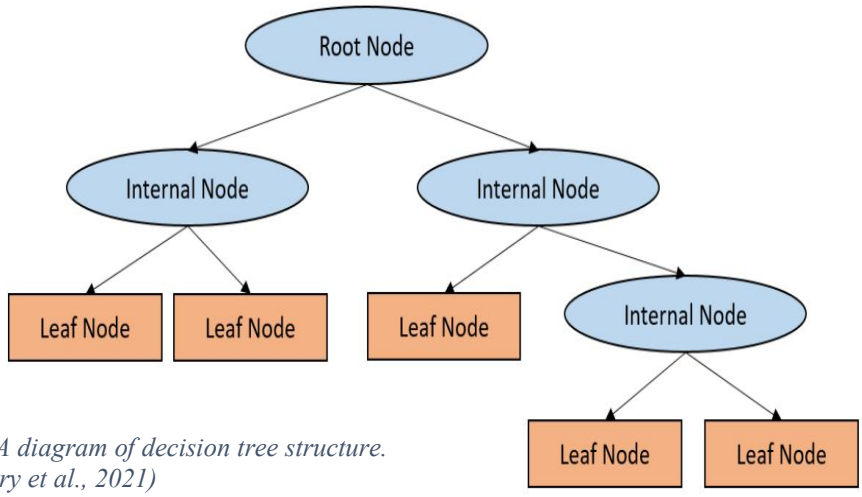


Figure 4. A diagram of decision tree structure. (Chowdhury et al., 2021)

data, and terminating at a leaf node when a stopping criterion is met (Figure 4). Each leaf node can be considered a specific subset of the final predictor space - the entirety of which would be able to produce a prediction with any input data of matching dimensions. This means that all potentially predictable values by the DT model are represented by exactly one leaf node above. DTs grow in complexity extremely quickly, both in terms of computational weight as well as the number of nodes and leaves (James et al., 2014). To make the generation of DTs computationally feasible, CART utilizes *greedy recursive binary splitting*, meaning that each non-terminal node is split into exactly two groups without consideration of any future splits.

Nodes are split using some evaluation criteria which the model utilizes to determine the split that results in the best performance (Breiman et al. 1984). The evaluation criterion is what classifies DT algorithms as machine learning methods, as it represents their cost function (Yi et al., 2020). DT algorithms determine splits which minimize their evaluation criteria to produce a high performing predictive result. CART can use any evaluation criteria to generate splits, but most commonly in regression problems it uses the residual sum of squares (RSS) (James et al. 2014). RSS for a given predictor variable region in this context calculates the sum of the squared difference between the dependent variable training data and the mean dependent variable value for the entire region. The predictor variable which produces the lowest RSS for a given split is then used to generate that split and the process is repeated for all potential cutoffs for the selected variable. The cutoff which has the lowest RSS is then selected and two new nodes are produced. These nodes are then split using the same method and the process continues until a stopping criterion is met for each node. Common stopping criterions include restrictions on tree size or a minimum limit on the number of data points represented in a node, referred to as *nodesize* (Breiman et al. 1984, Genuer & Poggi, 2020).

DTs individually are highly interpretable models that can be quickly understood using visual diagrams similar to figure 4. They are a well-established tool in many scientific fields including medical research, wildlife management and meteorology (Vayssières et al., 2000). Frequently encountered drawbacks for DT algorithms include the potential for overfitting and high variability (Hastie et al., 2009). If we consider the extreme case where each provided training datapoint is associated with a single terminal node, then the model will be perfectly fitted to the input data which is likely a model simplification and is known as overfitting. Additionally, CART and other DT algorithms have generally higher variance and lower performance compared to other predictive modelling methods, which is improved when many trees are aggregated together into a forest (Genuer & Poggi, 2020; James et al., 2014).

2.5.3 Random Forest Regression

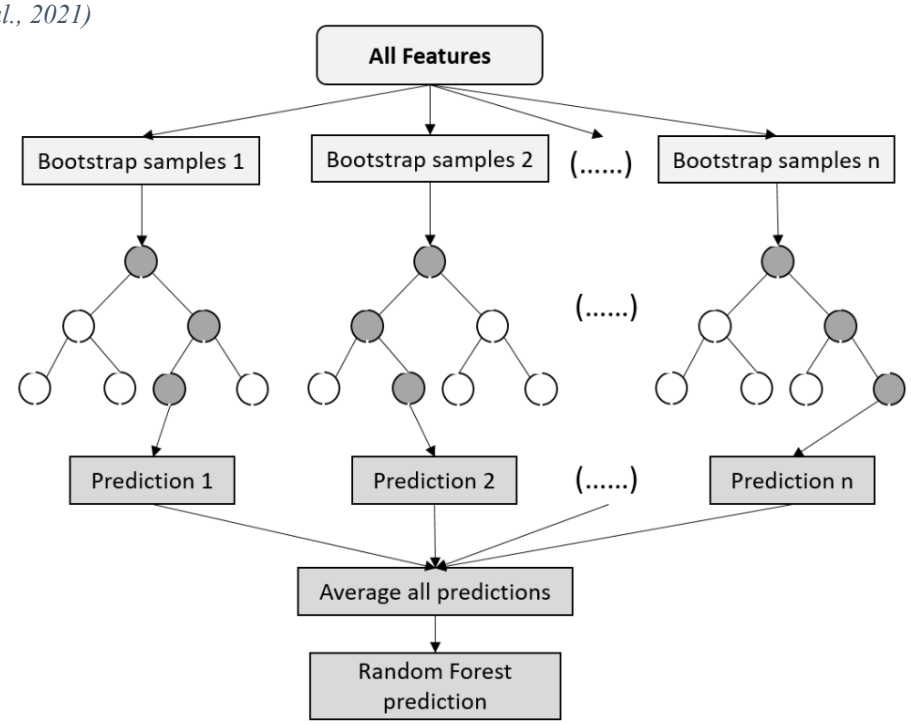
To combat the high variance involved with CART and other DT algorithms, Breiman established a new method in 2001 that utilizes an ensemble of DTs to produce a predictive result known as Random Forest (RF) shown in figure 5 (Breiman, 2001). Ensemble learning is the process of generating many independently developed models, each of which contribute to the final predicted result (Hastie et al., 2009; Liu et al., 2000). In the context of DT regression, this consists of generating many decision trees using multiple input datasets and averaging their individual predicted results to generate the final model prediction. This collection of DTs is known as a forest. In RF, the number of DTs to include the model's forest is a parameter known as *ntree* (Genuer & Poggi, 2020.

Ensemble learning reduces the variance of a model and provides a generally more accurate final prediction but has a high data demand as each model requires an independent dataset (James et al., 2014). RF takes advantage of the benefit of ensemble learning from a single dataset through

bagging. Bagging, originally introduced by Breiman in 1996, refers to bootstrap aggregation, or the generation of many models into an ensemble that each uses a random subset of the training data with repetition (Breiman, 1996). For each DT generated in RF the provided training data is split into individual points and collected with repetition to form the tree. This means that a single data feature may be used multiple times and treated as separate samples for the context of DT construction. Bagging improves model accuracy and simulates the benefits of ensemble learning while only utilizing a single dataset (Genuer & Poggi, 2020). To generate a final prediction with many bagged trees, RF produces the mean predicted value of each tree without weighting.

Bagging can cause inter-correlation between DTs when data features are repeated (James et al., 2014). RF decorrelates the bagged DTs that make up its forest by also changing how each split in the tree is generated (Breiman, 2001). In RF, when a node is evaluated for splitting, a random subset of the predictor variables are used instead of the whole set. The number of randomly selected predictor variables to consider is a parameter known as *mtry* (Genuer & Poggi, 2020). The

Figure 5. A diagram of random forest regression model structure using n bootstrapped decision trees. (Chowdhury et al., 2021)



ideal value for *mtry* is dependent on the individual model and is often evaluated as part of the study (Díaz-Uriarte & Alvarez de Andrés, 2006; Sreenivas et al., 2014).

In summary, random forest regression uses decision tree analysis to generate *ntree* number of DTs to produce a predicted result through bagging. Each DT consists of nodes that are split into two branches by selecting the one variable out of *mtry* predictor variables that minimizes the RSS following the split. Splitting continues until the node has *nodesize* number of represented data features, at which point it becomes a leaf node and is used for prediction. The predicted result of that leaf node is the mean value of the input variable of the *nodesize* remaining features.

2.6 Digital Soil Mapping

Digital soil maps (DSM) have been a fast-growing aspect of soil research in the 21st century (Boettinger et al., 2010; McBratney et al., 2003; Minasny & McBratney, 2016; Grunwald & Lamsal, 2006). It has been utilized widely to monitor and build predictive models for various chemical and physical properties of soil. As spatial information technology, such as ESRI's GIS, have grown in their capabilities, so has the precision of DSMs (Grunwald & Lamsal, 2006). McBratney first developed the general DSM generation framework, which establishes the foundational concept that soil properties can be modelled using environmental variables and spatial associations (2003). From this framework, many branching disciplines have been characterized across DSM research (Scull et al., 2003; Wadoux et al., 2020). Soil organic carbon (SOC) mapping has come to the forefront of discussion due to the high carbon storage of soils and the associated challenges with modelling it (Lamichhane et al., 2019).

In the context of DSM random forest (RF) predictive modelling for SOC has been widely utilized to statistically significant success (Khaledian & Miller, 2020). RF has been found to

outperform many other common methods, including kriging and support vector regression (Lamichhane et al., 2019). Specifically, RF has shown competence in areas with high landscape diversity even with comparatively few samples. Where there has been comparatively little research consideration is the use of RF to predict SOC on small spatial scales.

2.7 Summary

Organic carbon in forest soils is the product of a complex series of biological, chemical, and physical interactions within the ecosystem. In general, SOC stores are made up of organic matter inputs from litterfall and belowground root growth. Soil organisms then begin to decompose the organic matter, producing humus compounds containing carbon and respirating the rest. Due to the immense impact that respiration and decomposition have on stored SOC, the variables that significantly control their rates are the most relevant to carbon storage. While there are many, SOC storage is primarily controlled by moisture, temperature, and organic matter inputs. Overstory cover, understory cover, stand age, elevation, slope, and aspect each have significant large-scale effects on those controls and their relationship with SOC will be the focus of this study.

To analyze the relationships between SOC and its controllers, many studies have found statistically significant results using predictive machine learning models to develop a digital soil map. Random forest regression models have found recent success and are well suited for this method of predictive modelling. The majority of these models have been developed in large scale regions between county and country sized. There is a research opportunity to consider whether a smaller-scale model is possible using variables that can be easily measured at the field scale.

Chapter 3: Methods

3.1 Ellsworth Creek Study Design

This study will focus on the N1 subbasin of the Ellsworth Creek Preserve (ECP) (figure 6). ECP is a 2,235-hectare experimental preserve owned by The Nature Conservancy (TNC) near Willapa Bay in Washington (Chamberlain et al. 2021). Topologically, ECP is made up of high ridges, up to 365 m, and steep watershed basins that flow north to the Naselle River. The area is a characteristic example of a coastal temperate climate that experiences wet, cool winters with high rainfall and strong windstorms (Beck et al., 2018). In the summer, conditions are generally warm and dry during which time, due to the elevation, fog banks overtake much of the preserve. From the mid-1900s until the purchase of the property by TNC in 2001, ECP was active timberland and heavily managed, and as such the forest stands represent secondary forests up to 90 years old.

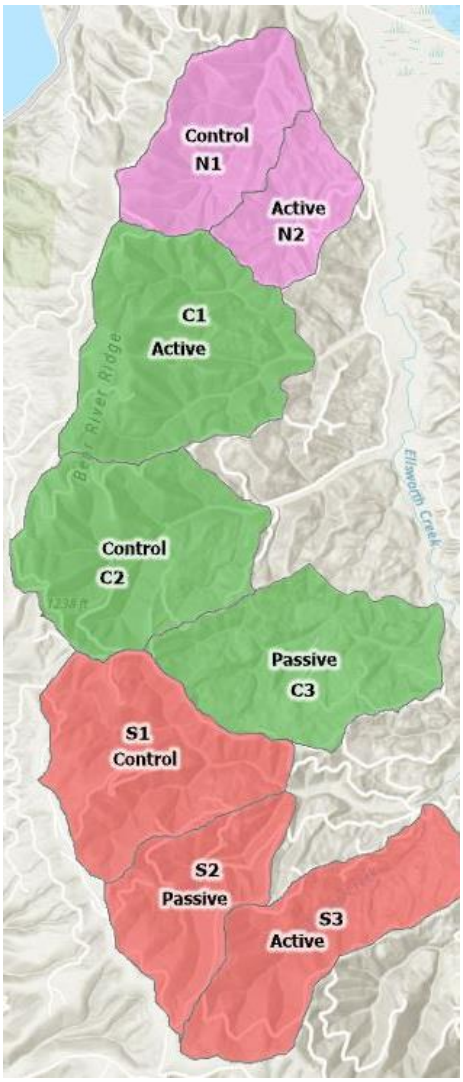


Figure 6. Map of Ellsworth Creek Preserve. Basins labeled by name and management type.

ECP has been operated as an experimental forest since 2007 and contains 8 subbasins that have been exposed to a variety of forest management methods including restorative thinning and clear cutting (Chamberlain et al., 2021). The eight subbasins were divided into three categories of forest management methods – active, passive and control. In the active basins, 30% of the trees were thinned between 2009 and 2013 which provides room for the remaining trees to grow faster and

stronger (The Nature Conservancy, 2020; Chamberlain et al., 2021). Passive basins have had their roads decommissioned and have been completely naturally isolated. Control basins have had their forest service roads maintained but are otherwise unmanaged.

The N1 subbasin is a 128-hectare (318 acre) control basin and, as per TNC study design, has received no specific forest management since the property was purchased in 2005. The subbasin is dominated by *Tsuga heterophylla* (western hemlock) and *Pseudotsuga menziesii* (douglas-fir) and has generally high density and canopy cover representative of a second growth forest (Chamberlain et al., 2021). Within the N1 subbasin, 26 vegetation plots of size 0.1 hectare were randomly placed by TNC in 2006 for data collection purposes (Case et al., 2023). Plots were measured and slope-corrected using a Haglof Vertex Laser 400 and identified using PVC markers with an orange flag at the center and the two nearest trees spray-painted orange at chest height. Each plot was then categorized into four 0.002-hectare subplots 9 meters horizontally and slope-corrected from center in the four cardinal directions (N,E,S,W). These subplots were marked with shorter PVC markers and pink flags. In 2007, data was collected by TNC at all 26 plots and 19 were resampled in 2020, this data is described below in *3.2 Initial Data Description*. This study focused only on the 19 plots that were resampled by TNC in 2020 as that provides the most modern picture for the ecological variables measured.

3.2 Initial Data Description

To construct a more comprehensive digital soil model this study has built upon previously collected data by Steven Quick and Dylan Fischer (Quick & Fischer) as well as The Nature Conservancy (TNC). This section describes the two datasets and how they were utilized in this study. Table 1 summarizes the source of all data referenced in this thesis.

Variable	*Soil Organic Carbon*	*O-horizon Mass*	*O-horizon Depth*	*Overstory Cover*	*Understory Cover*	*Stand Age*	*Elevation*	*Slope*	*Aspect*
Source	Thedell, Quick & Fischer		Thedell, TNC	TNC			USGS		

Table 1. Sources for all data utilized in this study (Thedell). Previously collected data gathered from Quick & Fischer, The Nature Conservancy (TNC), and the United States Geological Survey (USGS)

3.2.1 Quick & Fischer Soil Data (SOC)

Quick & Fischer collected soil samples from 10 vegetation plots across ECP during the period of winter through spring 2022. The goal of their study was to consider the effect of each basin management method (active, passive, control) on topsoil organic carbon. They specifically collected data from the #16, #24, and #26 plots within the N1 subbasin. In the field they collected soil samples to a depth of 10 cm, separating the O- and A-horizons in situ, using the methods described below in *3.3 Field Sample Collection and 3.4 Carbon Analysis*. This study replicated their sample collection process as closely as possible to ensure compatibility within a single model. Quick & Fischer's 12 data points on soil organic carbon content for the four subplots of N1-16, N1-24, and N1-26 were used in the construction of this study's model.

3.2.2 TNC Ecological Data (Overstory Cover, Understory Cover, Stand Age)

TNC conducted expansive surveys of vegetation plots in both 2007 and 2020 to collect ecological data on forest characteristics. 19 plots in the N1 subbasin were surveyed in 2020 by TNC, including those sampled for SOC by Quick & Fischer, and are the focus of this study (Figure 7 – Map of plots below in section 3.3). Of the vast amount of data collected by TNC, this study utilized the following measurements (1) overstory cover, (2) understory cover, (3) stand age and (4) O-horizon depth. TNC utilized the following field protocols to measure these variables. (1) Overstory cover at each subplot was measured using a convex densiometer from four readings surrounding the center point (upslope, left, downslope and right). These readings were then

averaged into a single overstory cover value for use in this study. (2) Understory vegetation cover was measured by dividing the subplot into quarters and estimating percent cover of each plant species found by two technicians. Understory plants were considered as all shrubs, forbs and fern species and included all specimens that extended into the plot even if they were rooted outside. Total understory cover was quantified by using the sum of percent cover for all species. (3) Stand age was measured for each plot by collecting tree core data from dominant trees. Age was considered equivalent at all subplots within the vegetation plot for the purposes of this study. (4) O-horizon depth was measured every 0.5 m along a 2.5 m transect within each subplot oriented in the corresponding cardinal direction. Measurements were made by using a garden trowel to disturb the organic layer and measured to the nearest tenth of a centimeter.

3.3 Soil Sample and Data Collection

As part of this study, 54 soil samples were gathered in the field from 14 vegetation plots within the N1 subbasin that were unsampled by Quick & Fischer. Four soil samples were gathered at 12 of those plots within each cardinal subplot. At N1-25 and N1-27, only 3 subplots were sampled due to safety concerns. Figure 7 shows the geospatial location of each vegetation plot utilized in this study – pink locations were sampled and measured in this study as per the description below, green locations were sampled by Quick & Fischer.

Plots were navigated to using a Geode GNS3 GPS and identified visually according to the attributes mentioned above in *3.1 Ellsworth Creek Study Design.* Four soil cores were collected at

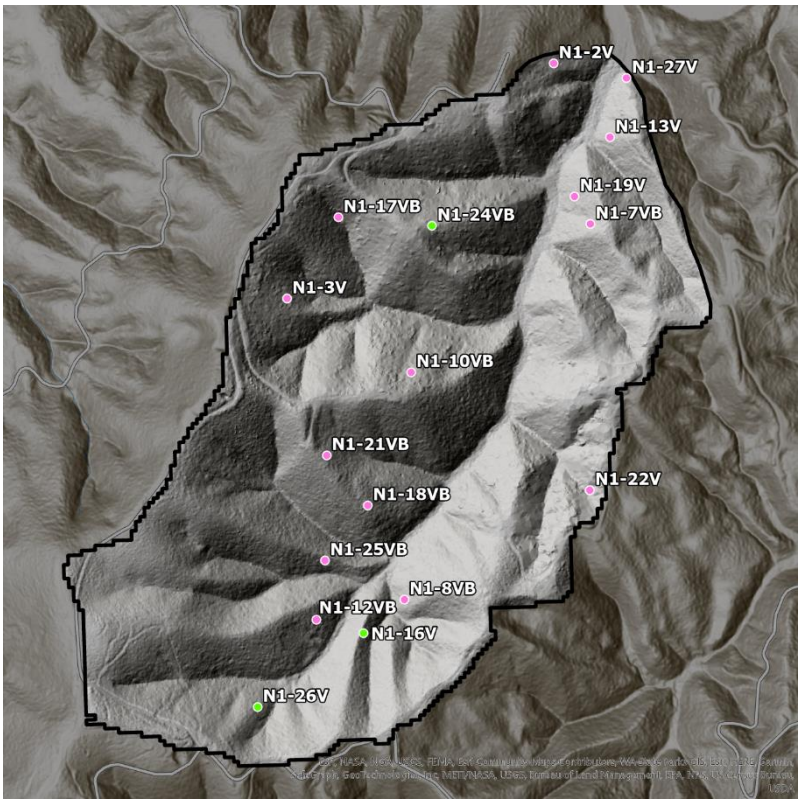


Figure 7. Bare earth map of the N1 subbasin of Ellsworth Creek Preserve. Pink sites were sampled as part of this study and green sites were sampled by Quick & Fischer.

each plot, one in each of the cardinal subplots (N, S, W, E) within 1 meter of the marker in the direction of the cardinality. If significant obstruction was encountered, a sample was taken from the nearest possible location within 0.5 m. Samples were taken to a depth of 10 cm using a homemade 5 cm diameter PVC soil augur (figure 8). To aid in sample collection, the top layer of organic material was broken up using a garden trowel before the augur was pressed firmly by hand into the soil.

Figure 8. Image of PVC soil augur in the field next to a collected sample location.

For 12 of the 14 sites, the combined O- and A-horizon soil was stored in pre-labeled paper bags. This method was identified to have the potential to lose organic material from the O-horizon during lab preparation due to its brittle nature once dried. After identifying the potential shortfalls of the previous sample collection method, 2 of 14 sites, the O- and A-horizons were separated in the field and stored in separate paper bags. This discrepancy was a result of efforts to replicate Quick & Fischer's sampling

process, in which the O and A horizon were stored separately, to ensure data compatibility. The 2 sites with separated O- and A-horizons were used as reference in lab to assist in identifying the unique visual characteristics of the O-horizon for hand pruning.

Once a sample was collected at each subplot, the depth of the O-horizon was measured in inches using a 6-inch ruler by eye from the core opening. Sample location, date, time, and O-horizon depth were recorded using FieldMaps for ArcGIS and the Geode GPS. Additionally, the three previously sampled vegetation plots (N1-16, N1-24, N1-26) were remeasured for their subplot GPS location for consistent and improved spatial precision. Previous sample locations were identified by eye and for sites that were no longer clearly identifiable.

3.4 Data Collection & Calculation

3.4.1 Sample Preparation & Pretreatment

Before carbon analysis, samples collected in this study underwent the following preparation process. First, within 36 hours of field collection each sample bag was placed in a drying oven held at 70°C. Each sample was dried for a minimum of 24 hours before analysis. Samples were stored within their original labeled paper bag sealed in a gallon zip top plastic bag with 2 silicate packets to prevent moisture buildup. After drying, samples were passed through a two-tiered sieve with a 2 mm and 500 µm layer to aid in visual OM determination. For each sieve layer the O-horizon material (OM) was collected, separating it from mineral soil, large mineral debris (rocks, etc.), and A-horizon roots by eye using steel forceps. The remaining soil, debris and roots were gathered and stored for future analysis beyond this study. Separated OM portions from both the sieve layers were then combined and consisted primarily of organic and plant fragments from litterfall in various states of decay – representing the O-horizon layer of the soil. Each OM

sample was weighed and stored in an individual quart plastic zip top bag. All OM was then passed through a Wiley THOMAS mill for 60 seconds until it reached homogenous grain size. Milled OM was then stored in a lidded glass vial for carbon analysis (see *3.4.3 Carbon Analysis)*.

3.4.2 Bulk Density

For each sample, bulk density (ρ) was calculated using the dry mass of the organic material (M_o) and volume of the O-horizon portion of the sample (V_o).

$$\rho = \frac{M_o}{V_o} \tag{3.1}$$

O-horizon sample volume was calculated using the area of the soil augur ($A_o = \pi(2.5^2) = 19.635$ cm^2) and the O-horizon depth (D).

$$V_o = D \cdot A_o \tag{3.2}$$

This provides a final bulk density (g/cm^3) equation of the following:

$$\rho = \frac{M_O}{(D)(19.635)} \tag{3.3}$$

3.4.3 Carbon Analysis

Samples were measured for percent carbon content by weight using a PerkinElmer 2400 CHN analyzer via the following process. Before samples were analyzed, an initialization sequence was utilized for calibration. This involved processing a 10-sample sequence which included 2 ± 0.2 mg of an Acetanilide (C_8H_9NO) K-factor, 2 ± 0.2 mg of known Alderwood sandy-loam conditioner passed through a 500 µm sieve, and empty tin container blanks. Once calibrated, 2 ±

0.2 mg of the dried and milled samples were collected into 5x8 mm tin containers and weighed. Then samples were placed in sequence to be measured in the CHN analyzer following standard inter-sequence calibration practices the above K-factor and blanks (Nelson & Sommers, 1996). Measurements were delivered in weight percent organic carbon. Nitrogen and hydrogen content were also measured, but unused in this study.

Carbon content values ($OC_{\%}$) were then converted to carbon density (OC_{ρ} g/m^2) using the dry mass of the sample (M_o) and the area of the sample (A_o = 19.635 cm^2 = 1.9635x10^{-3} m^2). Then, SOC was converted to the common units of Mg/ha for comparison across studies.

$$OC_{\rho} = \frac{(OC_{\%})(M_o)}{(1.9635\text{x}10^{-3})} \ \ (g/m^2) \tag{3.4}$$

$$SOC_{Mg/ha} = \frac{SOC_{g/sqm}}{10} \ \ (Mg/ha) \tag{3.5}$$

3.4.4 Topological Variables from DEM

The topological variables involved in this study were collected using a digital surface model (DSM) constructed from a LIDAR dataset from the USGS 3D Elevation Program (U.S. Geological Survey, 2020). The DSM was utilized in ArcGIS Pro and further analyzed for slope and aspect using the integrated *Surface Parameters* tool. Slope and aspect were generated using a 3-meter neighborhood averaging and quantified in degrees from horizontal and north respectively. Each sample location was then queried from each raster (figure 9) to provide elevation (1), slope (2) and aspect (3) for the sample.

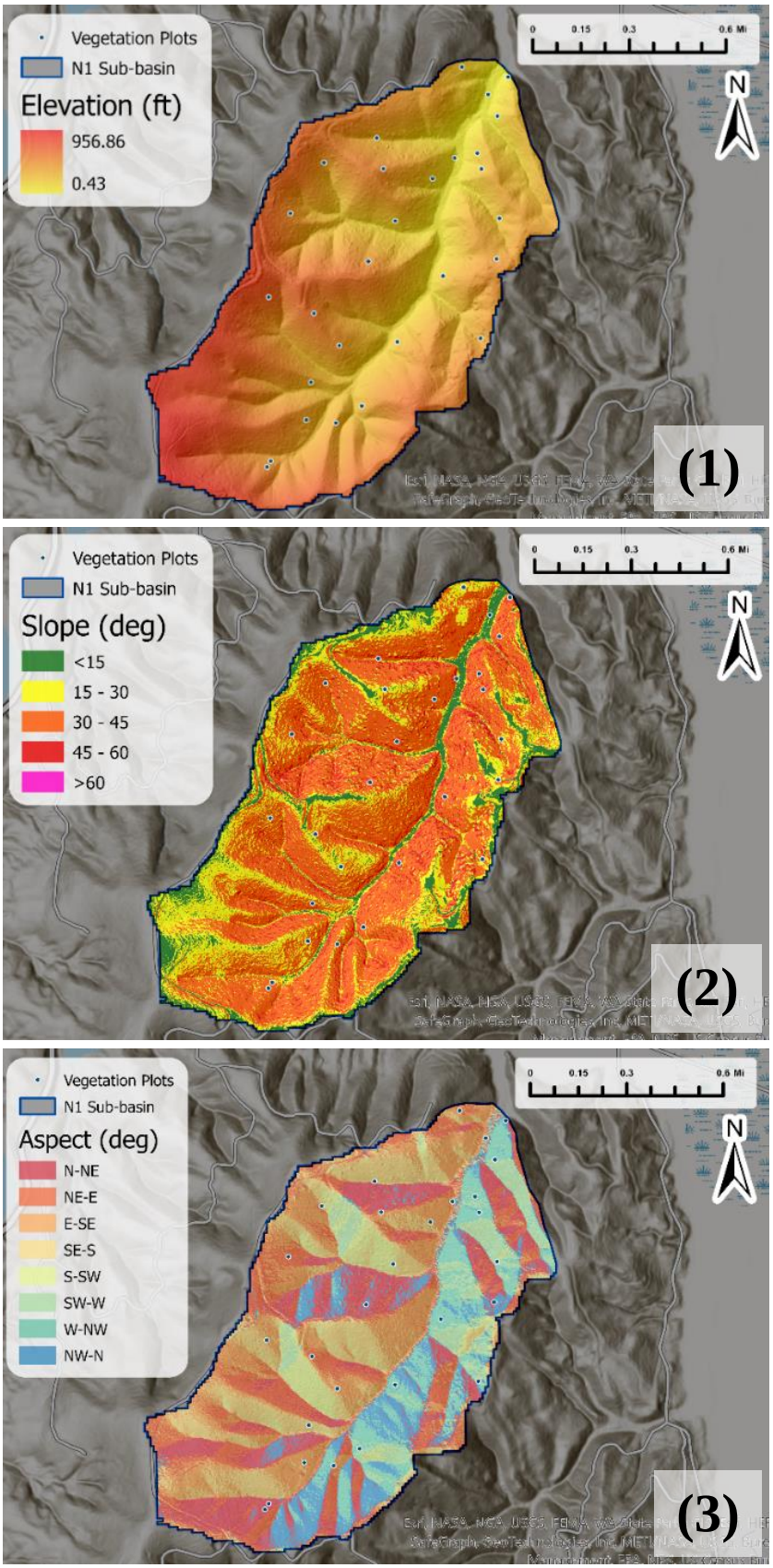


Figure 9. Map series of topological rasters generated from USGS lidar data. (1) Elevation in feet, (2) Slope in degrees from horizontal, & (3) Aspect in degrees from north. Blue points are all vegetation plots within the N1 sub-basin.

3.5 Model Construction & Parameter Optimization

All the above data was utilized in the construction of a predictive model for SOC. This model was built using the *RandomForestRegressor* function in version 1.3.0 of the scikit-learn package for Python 3.11.1. The code was written and compiled within a Jupyter Notebook. Random forest regression (RF) is a machine learning algorithm that utilizes an ensemble of decision trees (DT) to predict results (Breiman 2001). RF is used widely for predicting soil properties and for digital soil mapping (Grimm et al., 2008; Lamichhane et al., 2019; Suleymanov et al., 2023; Zhou et al., 2022).

The RF model included all six of the above independent variables (overstory cover, understory cover, stand age, elevation, aspect, slope) in addition to a plot dummy variable to consider the effect of pseudo-replication (PR) (Urban, 2005). Each subplot is spatially correlated within each plot (9 m horizontally separated in each cardinal direction), so the associated SOC values may also share plot-correlation. To quantify this potential effect, plots were label encoded as integers ranging from 0-16. Each data feature was then assigned an integer variable that represented the plot the sample was located in, from here forward referred to as plot dummy. For each value of plot dummy, there are 3 or 4 data features that were taken within that plot. Plot dummy was then included in the RF model as an additional predictor variable. Urban establishes that a strong effect by this dummy variable would imply that SOC's relevant spatial scale is larger than a single plot and subplot samples are pseudo-replicated (2005).

Parameter (*Python Name*)	**Default**
Mtry (*Max_features*)	1
Ntree (*N_estimators)*	100
Nodesize (*Min_samples_leaf*)	1

Table 2. The default values in the RandomForestRegressor function of the three parameters, mtry, ntree, and nodesize.

Model development typically considers three parameters, sometimes referred to as *hyperparameters*, for model optimization

(Khaledian & Miller, 2020): (1) The number of randomly selected variables to utilize in each split of a DT ($mtry$), (2) The number of DT to generate in the RF ($ntree$), and (3) The minimum number of datapoints included in terminal nodes ($nodesize$). The optimized value for each of these parameters varies greatly with input datasets.

To find the optimized value for each of these parameters, we first consider them separately. For the following sections regarding one specific parameter (*3.5.1, 3.5.2, 3.5.3)*, the parameters not being considered were held at their default values, shown in table 2. To evaluate the model's success, we used the coefficient of determination (R^2) according to the following equation (Gotelli & Ellison, 2012; Pedregosa et al., 2012):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3.6}$$

Where y_i is any measured value for SOC, $\hat{y}_i$ is the predicted value for the same point, and $\bar{y}$ is the mean measured SOC value across all points.

Then the target parameter will be varied on a relevant range and the corresponding coefficient of determination will be compared along that range (Díaz-Uriarte & Alvarez de Andrés, 2006; Sreenivas et al., 2014). To correct for the inherit randomness in model generation the following evaluation will use a predetermined random state, using the *random_state* parameter, to assure comparability between models (Pedregosa et al., 2012).

3.5.1 $Mtry$

The number of variables to include in each split of the DT in a RF model can affect the final model in a variety of ways, both in terms of quantitative success and interpretability (Breiman, 2001). As the Random Forest model builds decision trees, each split, or node, will

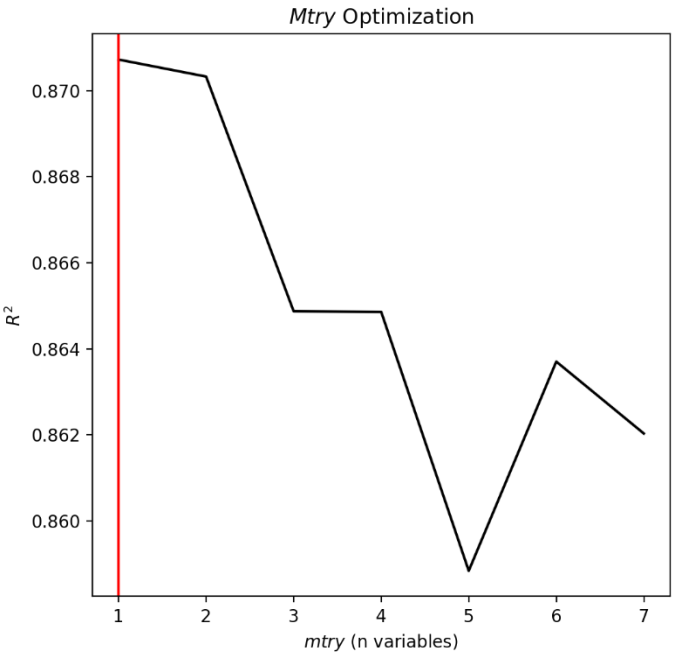


Figure 10. Optimization of the mtry parameter for random forest model development by evaluating the model's R^2 for all values of mtry. The red line indicates the selected optimal value.

consider a random subset of variables to establish the split. $Mtry$ represents the number of randomly selected variables that the model will select at the node. If we consider the default value of $mtry = 1$, this means that exactly one variable will be used to determine the split. The maximum value for $mtry$ is the number of predictor variables used to build the model – in this case, seven.

To optimize *mtry* we evaluated all values from one to seven. Generally, the R^2 of the model decreased as *mtry* increased, with a small bump in precision beyond $mtry = 5$. This indicates that model performance is highest when *mtry* is minimized. Figure 10 shows the general trend of precision with *mtry*. For model development, a parameter value of $mtry = 1$ will be used.

3.5.2 Ntree

The total number of DTs generated as part of the RF model is a dial to control the model performance but at the cost of increased computational load during model generation. Considering the extremes of this value, for $ntree = 1$, the model would generate exactly one DT and would no longer be RF at all, but rather traditional decision tree analysis (Breiman, 2001). There is no upper limit to the value of $ntree$, though the theory involved with RF modelling dictates that there are diminishing returns as $ntree$ increases. There is an approximate theoretical limit to the model's success, and continuing to generate trees beyond that limit would no longer increase the R^2

of the model. As such, to reduce the computing cost of running and using the model, selecting a value of $ntree$ that optimizes the model accuracy without adding unnecessary trees is important.

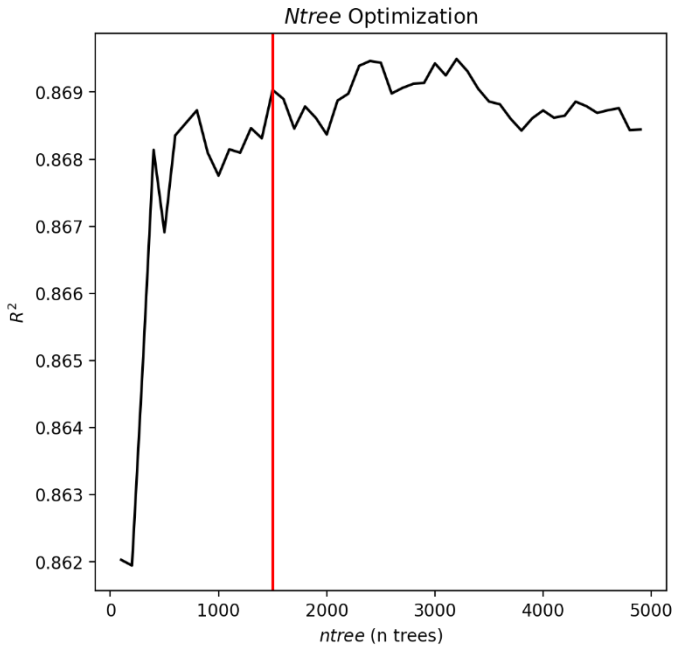


Figure 11. Optimization of the ntree parameter for random forest model development by evaluating the model's R^2 for all values of ntree from 100 to 5000. The red line indicates the selected optimal value.

To optimize the value for $ntree$ we evaluated values ranging from the default 100 DTs to 5000 DTs in steps of 100 (Figure 11). As expected by the theory, the model's R^2 increased with higher values of $ntree$, though, the rate that R^2 increased drops off at values much larger than $ntree = 2000$. The computational cost increases dramatically with higher values of $ntree$, so in the interest of a conservative selection, for model development a value of $ntree = 1500$ will be utilized.

3.5.3 Nodesize

Varying the value of *nodesize*, the minimum number of datapoints included in DT terminal nodes, leads to a change in the typical depth of each DT (Breiman, 2001). For low values, DTs tend to be larger as they can continue splitting the data into smaller portions. The inverse is true for large values of *nodesize*. On the extremes, increasing *nodesize* to its maximum, the number of data points in the set, leads to the model being unable to make any unique prediction at all as it becomes unable to produce any splits.

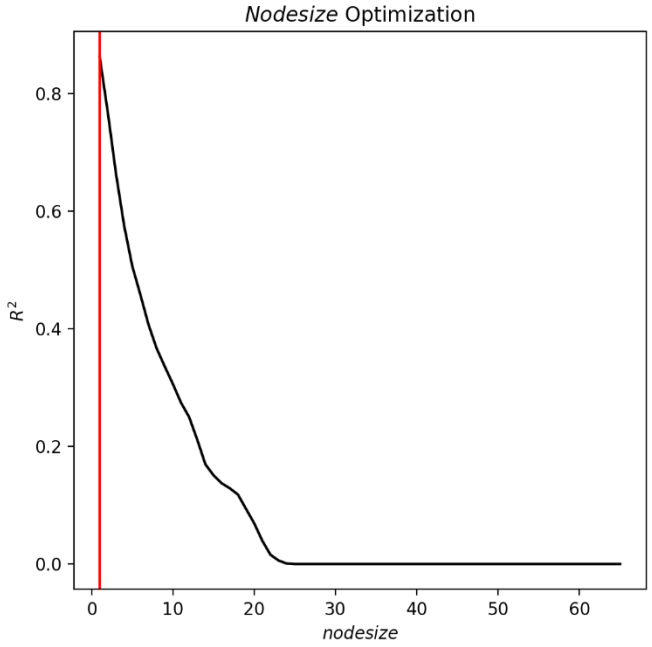


Figure 12. Optimization of the nodesize parameter for random forest model development by evaluating the model's R^2 for all values of nodesize. The red line indicates the selected optimal value.

To test the optimal value for *nodesize* in our study, we considered values ranging from 1-*n*, where n is the total number of samples in this study (66). As before, we used the coefficient of determination (R^2 – eq. 3.6) to evaluate model success. A total of 10 models were generated for each value of *nodesize* to eliminate any effect of randomness on parameter evaluation. Figure 12 shows a sharp decrease in model success for increased values of *nodesize*. The peak performance of the model was utilizing the default value of *nodesize* = 1. This default value was established by Breiman during his initial description of decision tree algorithms and is frequently used in other studies which utilize RF (Breiman et al. 1984; Díaz-Uriarte & Alvarez de Andrés, 2006; Sreenivas et al., 2014).

3.5.4 Model Parameter Summary

After individual evaluation, the optimal values for mtry, ntree and nodesize were selected as 1, 1500 and 1 respectively. For all future RF model generation and evaluation, these values will be utilized. For further details regarding the error evaluation of the model, go to *Chapter 4: Model & Results.*

3.6 GIS Raster Analysis

Random forest models are known for their lack of interpretability due to the final model function and need for thousands of decision trees (Breiman, 2001; Khaledian & Miller, 2020). To remedy this, following model development, a predicted SOC raster was generated across the N1 sub-basin. This raster is a digital soil map (DSM) and serves as a visual interpretation of the model's prediction across the sub-basin. The DSM was built using ArcGIS Pro with an integrated Jupyter notebook running Python 3.11.1. Functions from the sci-kit learn package as mentioned in the previous section in addition to the ArcPy package developed by ESRI were utilized in the DSM construction (Pedregosa et al., 2012)

The DSM used six predictor variable rasters that stretched the N1 subbasin (shown in figure 13). These include the three topological rasters generated using the USGS DTM mentioned above, as well as three additional rasters that were created from the overstory cover, understory cover and stand age data provided by TNC. To create the overstory and understory cover rasters from point data the *Kriging* tool from the spatial analyst toolset in ArcGIS Pro was utilized. This tool interpolates the point data across a given area using ordinary kriging, a mathematical spatial prediction method (Chilès & Desassis, 2018; Cressie, 1988). The rasters shown in figure 13 are generated using a spherical semi-variogram with the nearest 30 neighboring points. To generate a raster for stand age, we combined point stand age data collected by TNC at each plot with a polygon dataset developed by TNC for the forest stands. The resulting raster shown in figure 14 represents the historical timber stands across the N1 subbasin and their corresponding age. The resolution of all rasters was set to 3 square meters due to the high computational cost of higher resolution calculations.

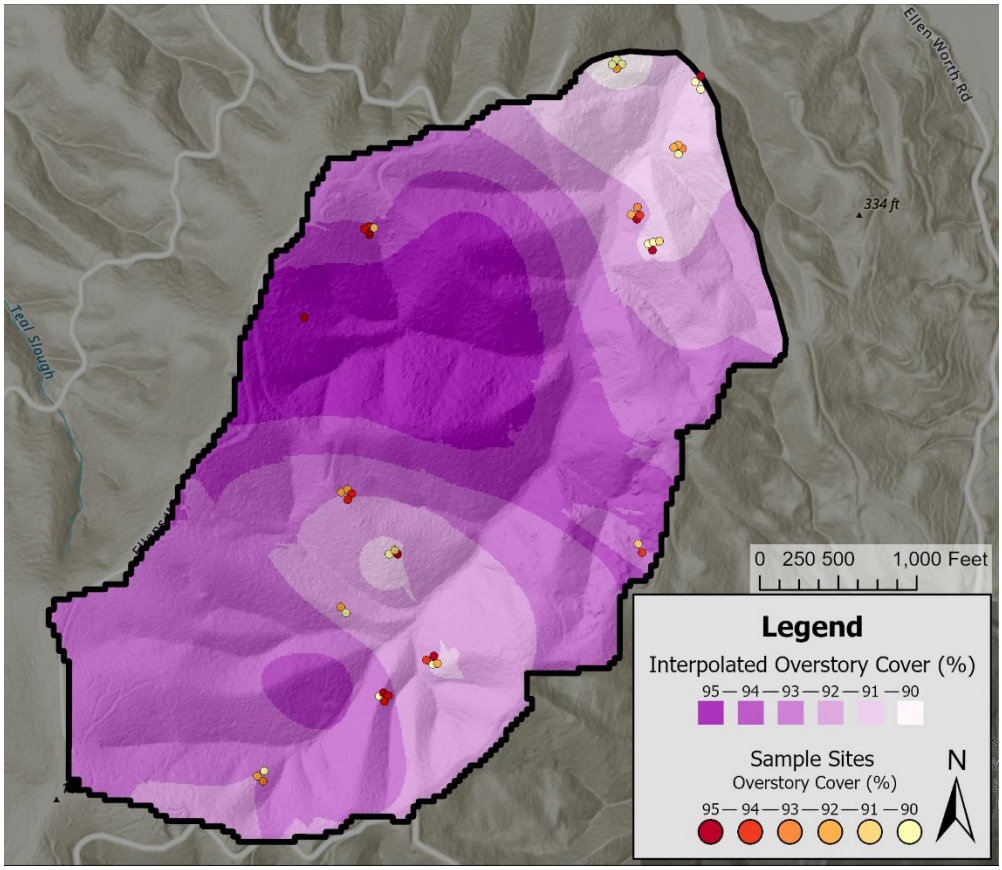
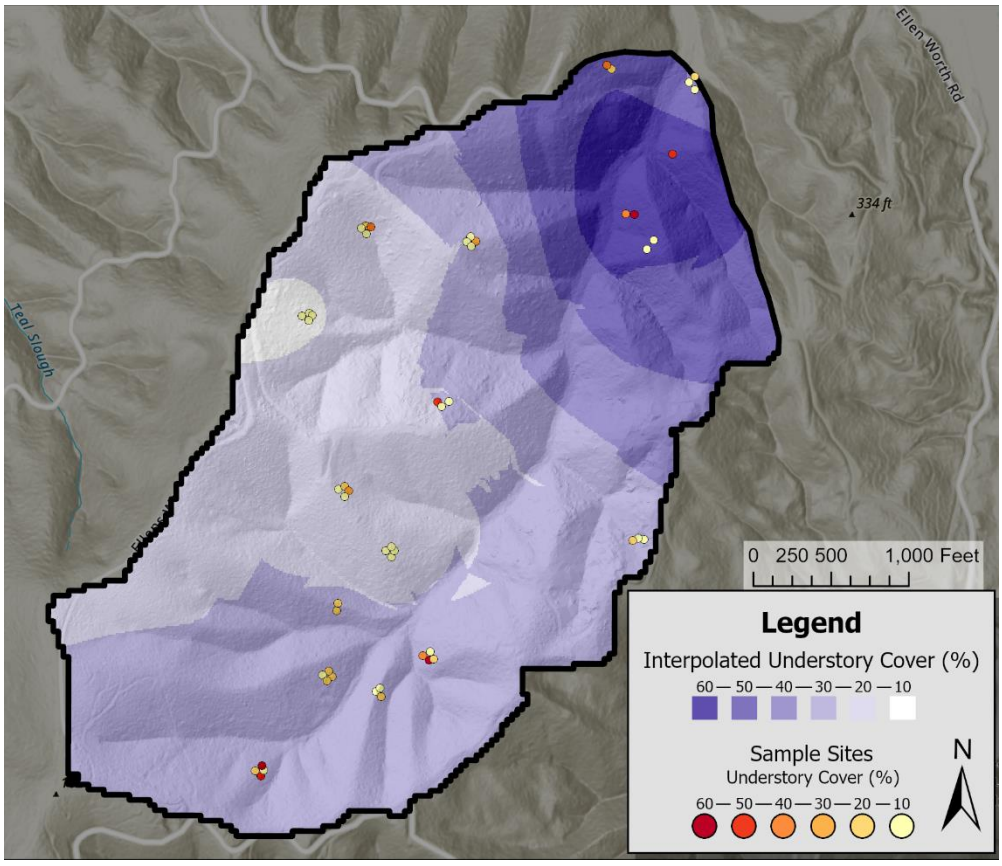


Figure 13. Interpolated rasters for overstory cover (top) and understory cover (bottom) across the N1 sub-basin using Ordinary Kriging from field-collected point data by TNC.



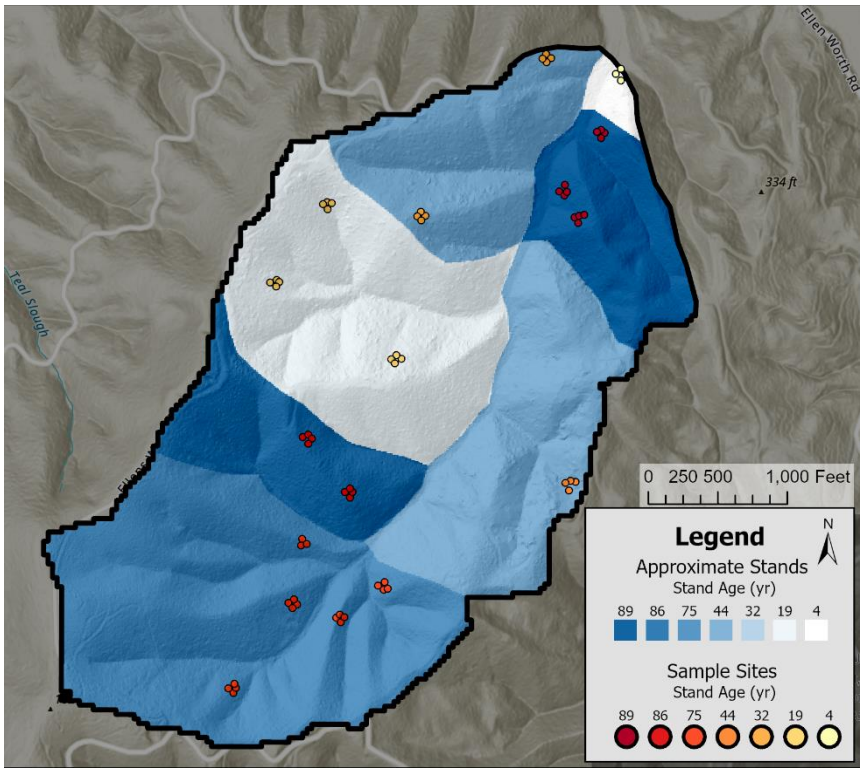


Figure 14. Approximate stand age raster generated using point-data collected by TNC and historical timber stand polygons provided by TNC.

Once all six predictor variable rasters were generated, the DSM was built using a variety of functions in the ArcPy and Numpy packages for Python 3.11.1 (Harris et al., 2020). Each predictor raster was first converted to an array of values each representing a 3 square meter patch of ground. Then, a new predicted SOC array was constructed by passing the six predictor values at each point through an RF model built out of the 66 collected datapoints above (*mtry = 1, ntree = 1500, nodesize = 1)*. The resulting predicted SOC array was then converted into the DSM where each value represented a 3 square meter patch of ground. This resulting raster was finally converted into the same coordinate system as the predictor rasters to align each location spatially.

Once a DSM was generated, to identify areas of statistically high and low SOC values, hot-spot analysis was performed to determine the Getis-Gi and Gi* statistics (Ord & Getis, 1995). This was performed using the *Hot Spot Analysis (Getis-Ord Gi*)* tool in the spatial statistics toolkit in ArcGIS Pro. The raster was first converted into a point dataset, where each point represents the center of a 3 square meter area, then analyzed for statistical hotspots. Due to the imprecision of

raster development, the final carbon hot-spot analysis is not necessarily ecologically representative, but rather an approximation of the model for interpretation. For further investigation on the raster results and their interpretability, see *Chapter 4: Model & Results.*

3.7 Summary

The model constructed in the following *Chapter 4: Modelling & Results* was built using field collected data as part of this study as well as external data sourced from The Nature Conservancy, the USGS 3D Elevation Program, and Quick & Fischer. The subject area of this study is a watershed basin near Willapa Bay Washington and is a characteristic coastal temperate second growth forest. This basin has been unmanaged for 21 years following its purchase by TNC. Soil samples were collected in the field and measured for carbon content in lab using elemental analysis. Ecological data gathered by TNC, including overstory cover percent, understory cover percent, and stand age, was used for model development. USGS elevation data through raster analysis was also used to construct point measurements for all three of the topological variables (elevation, slope and aspect). The sources for each variable in this study are described in table 1. In total, 66 points were sampled for soil organic carbon, overstory cover, understory cover, stand age, elevation, slope, and aspect across the subbasin.

With the collected data, a predictive SOC model was generated using Random Forest modelling. Each relevant parameter for the model was optimized using statistical comparisons. Once a model was developed, to create an interpretable result, it was utilized to generate a SOC digital soil map for the watershed basin. This was performed through raster and hotspot analysis in ArcGIS Pro.

Chapter 4: Modelling & Results

This chapter will begin by summarizing all variables measured by this study (section 4.1 and 4.2). This will be carried out by considering their range, mean, median, and normality. Their normality will be determined using the Shapiro-Wilk Normality Test with $\alpha = 0.05$ (Whitlock & Schluter, 2008). To evaluate the variation of each variable across the sub-basin, the standard deviation of each variable overall will be compared to the average standard deviation of that variable within each plot. Stand age was not considered for this variation due to plots residing in single forest stands and having no age deviation. All summary results are collated in Table 3 prior to section 4.1. Additionally, each predictor variable will also be evaluated on their relationship with this study's final dependent variable, Soil Organic Carbon (SOC) Pool (Mg/ha), using non-parametric spearman's rank correlation and simple linear regression ($\alpha=0.05$) (Spearman, 1904; Gotelli & Ellison, 2012). These relationships will be used to consider potential broad impacts of each variable on the final model but are not necessarily representative of model contributions as the complexity of soil systems causes a trend towards non-linear relationships (Attiwill & Adams, 1993).

Following individual variable characterization, the final model results will be evaluated and visualized. The random forest (RF) model is evaluated using the coefficient of determination and root mean square error (Matinfar et al., 2021, Grimm et al., 2008). The individual effect of each variable on predicted SOC is considered by isolating that variable across its measured range. To visualize the RF model, a predicted SOC digital soil map is generated using variable interpolation across the sub-basin. The resulting soil map is then utilized to perform hot spot analysis to demonstrate regions of significantly high and low carbon.

Variable	Unit	Min	Max	Mean	Median	SD_{total}	SD_{plot}	Normal?
O-Horizon Depth (D)	cm	1.27	10	6.9	7.2	1.6	1.9	No
O-Horizon Mass (M_O)	g	0.748	34.057	12.70	12.03	2.95	6.03	No
Bulk Density (ρ)	g/cm^3	0.011	0.516	0.096	0.085	0.073	0.051	No
Percent Carbon (%OC)	%	31.8	59.3	46.2	47	4.91	3.93	Yes
Soil Organic Carbon Pool (SOC)	Mg/ha	17.1	896.9	308.5	284.1	220.3	159.4	No
Overstory Cover	%	81.75	98.25	93.03	93.75	3.51	2.50	No
Understory Cover	%	0	150	29.55	12	34.81	24.12	No
Stand Age (Age)	yr	4	89	58.38	75	29.48	NA	No
Elevation	m	17.11	580.58	343.08	369.34	140.90	17.81	No
Slope	deg	13.49	48.62	33.58	35	8.14	5.31	Yes
Aspect	deg	17.59	357.01	171.66	121.80	106.88	35.92	No

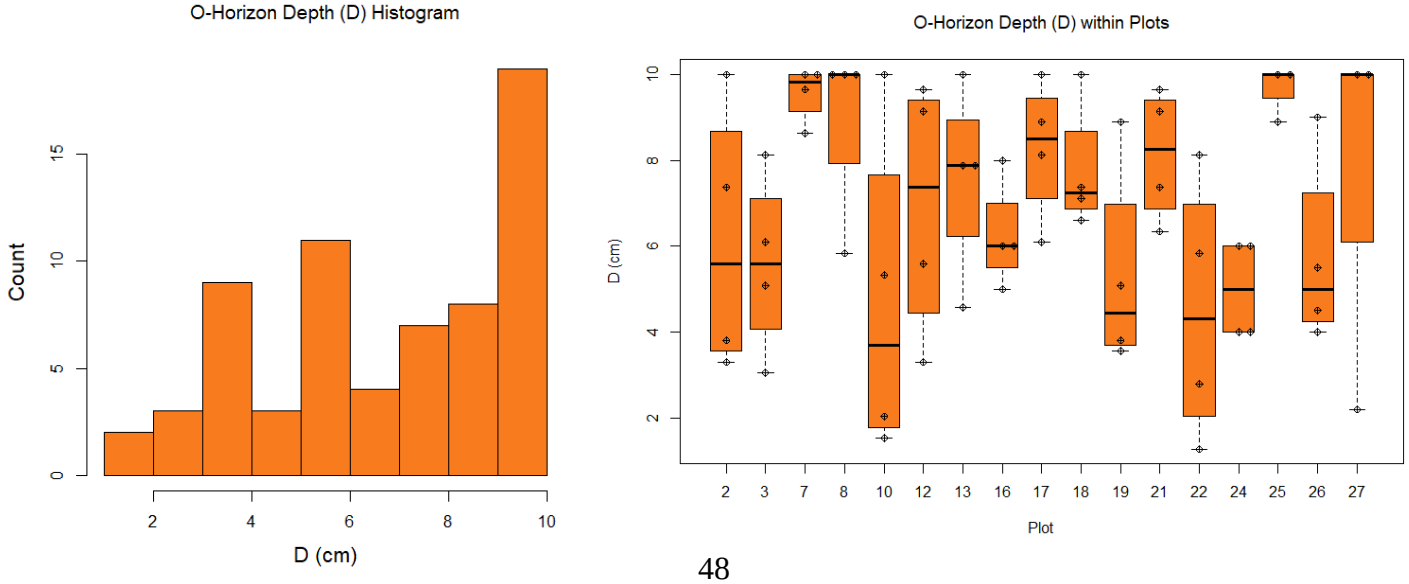
Table 3. Summary statistics for all data utilized and collected in this study.

4.1 Carbon Variable Summary

4.1.1 O-Horizon Depth (D)

The depth (D) of the organic horizon (O-horizon) at each site was measured as part of field collection for this study (see *Chapter 3: Methods* for data collection details). D ranged from 1.27 cm to 10 cm (figure 15). D was not significantly normally distributed ($W = 0.92,\ p \ll 0.05$) with a mean of 6.9 cm and a median of 7.2 cm. The overall standard deviation of D was 1.6 cm, while the standard deviation within each plot was 1.9 cm (figure 15).

Figure 15. Summary plots of O-horizon depth (D). (Left) Histogram distribution. (Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.



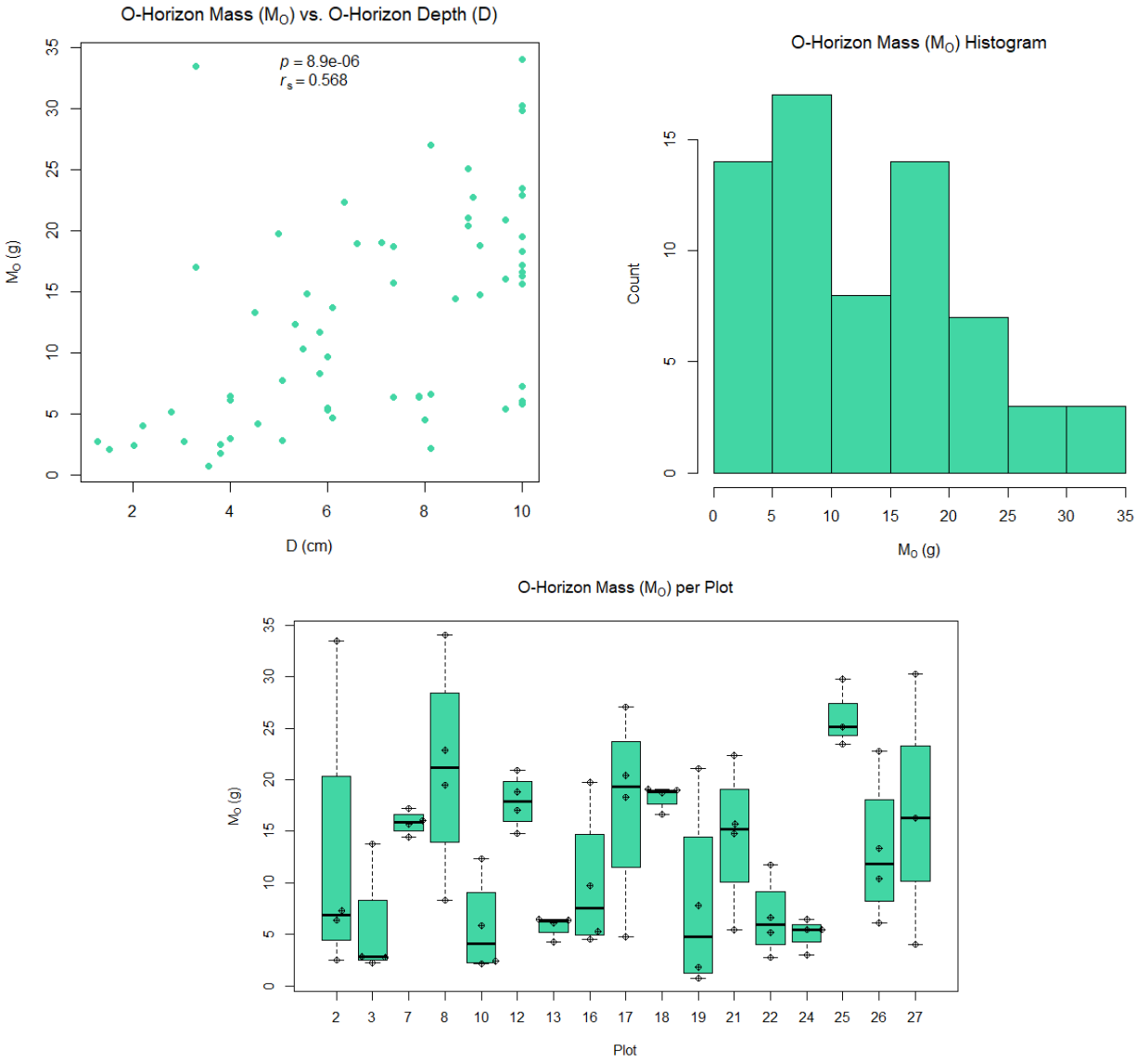


Figure 16. Summary plots of O-horizon mass (M_O). (Top Left) Scatterplot of M_O against O-horizon depth (D). Included are the spearman's correlation coefficient (r_s) and the resulting p-value (Top Right) Histogram distribution. (Bottom) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.

4.1.2 O-Horizon Mass (M_O)

The mass of the O-horizon, M_O, was collected from each sample to calculate bulk density (see *Chapter 3: Methods*). Of the data collected, M_O was not normally distributed ($W = 0.93$, $p \ll 0.05$) and ranged from 0.75 g to 34.06 g with a mean of 12.70 g and a median of 12.03 g (figure 16). The overall standard deviation of M_O was 2.95 g, while the standard deviation within each plot was 6.03 g (figure 16). M_O showed a positive correlation with D with a statistically significant linear relationship ($r_s = 0.57, p \ll 0.01$, figure 16)

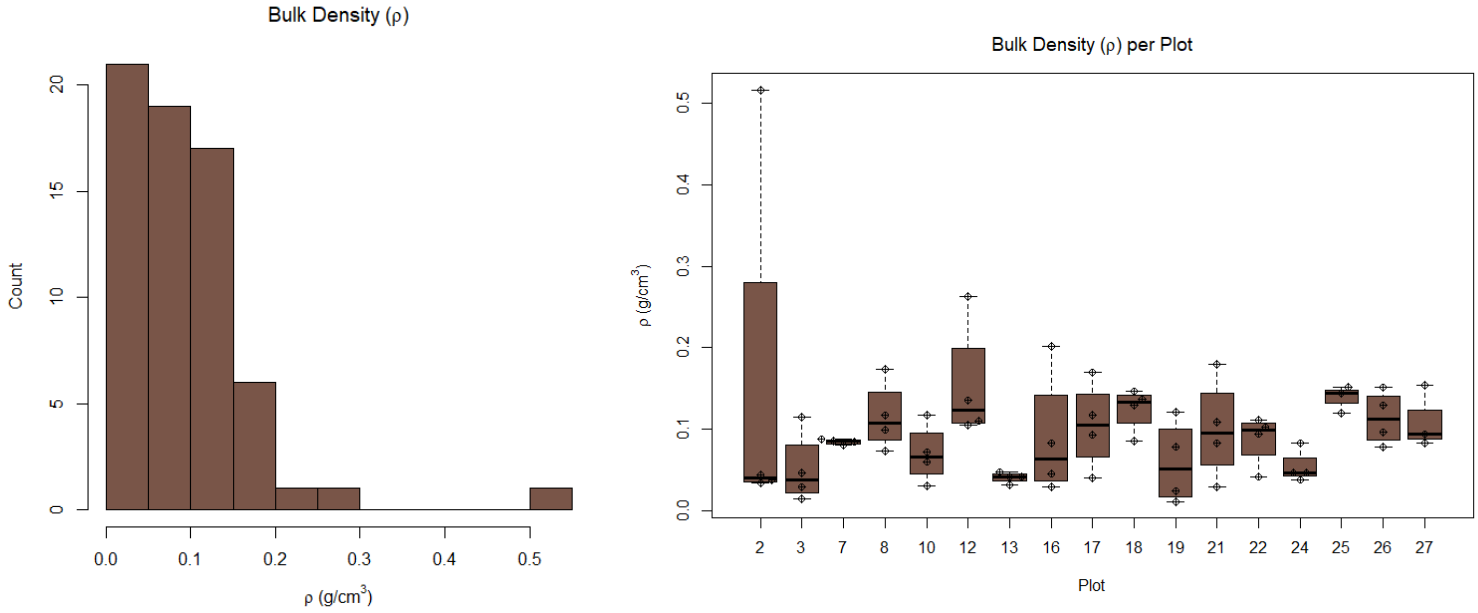
4.1.3 Bulk Density (ρ)

The bulk density (ρ) of the O-horizon was calculated using the O-horizon depth (D) and mass (M_O) of each sample according to equation 3.3 from *Chapter 3: Methods*.

$$\rho = \frac{M_O}{(D)(19.635)} \tag{3.3}$$

ρ ranged from 0.011 g/cm^3 to 0.516 g/cm^3 with a mean of 0.096 g/cm^3 and median of 0.085 g/cm^3 (figure 17). ρ was not significantly normally distributed ($W = 0.74,\ p \ll 0.05$) with an overall standard deviation of 0.073 g/cm^3 compared while the mean standard deviation within plots was 0.051 g/cm^3 (figure 17).

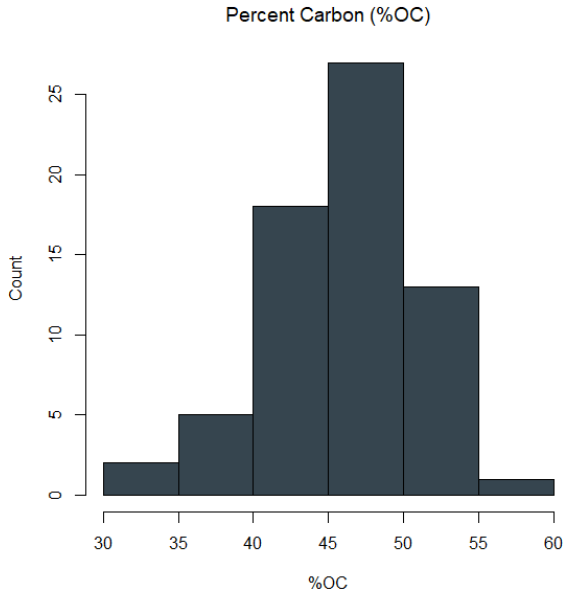
Figure 17. Summary plots of bulk density (ρ). (Left) Histogram distribution. (Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.

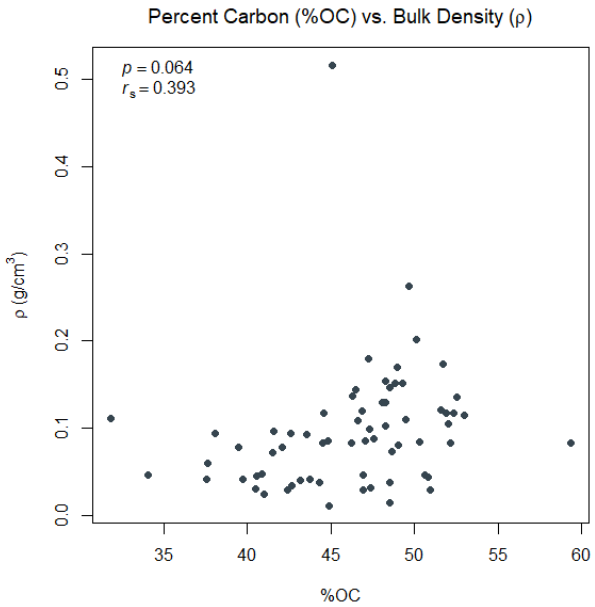


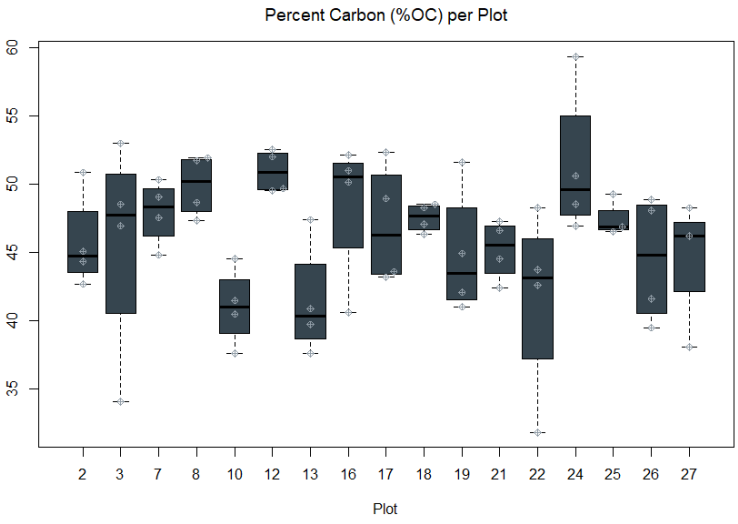
4.1.4 Percent Organic Carbon (%OC)

Percent carbon by mass ($\%OC$) of the organic portion of the soil samples ranged from 31.8% to 59.3% with a mean of 46.2% and median of 47% (figure 18). This value represents the by-mass carbon percentage of the O-horizon of the soil up to 10 cm from the surface. $\%OC$ was significantly normally distributed ($W = 0.97,\ p = 0.13$) with an overall standard deviation of 4.91% and an average inter-plot standard deviation of 3.93% (figure 18). $\%OC$ showed a weak positive correlation with ρ with a near-significant linear relationship ($r_s = 0.39, p = 0.06$, figure 18)

Figure 18. Summary plots of percent organic carbon (%OC). (Top Right) Histogram distribution. (Bottom Left) Scatterplot of %OC against bulk density (ρ). Included are the spearman's correlation coefficient (r_s) and the resulting p-value. (Bottom Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.







4.1.5 Soil Organic Carbon Pool (SOC Mg/ha)

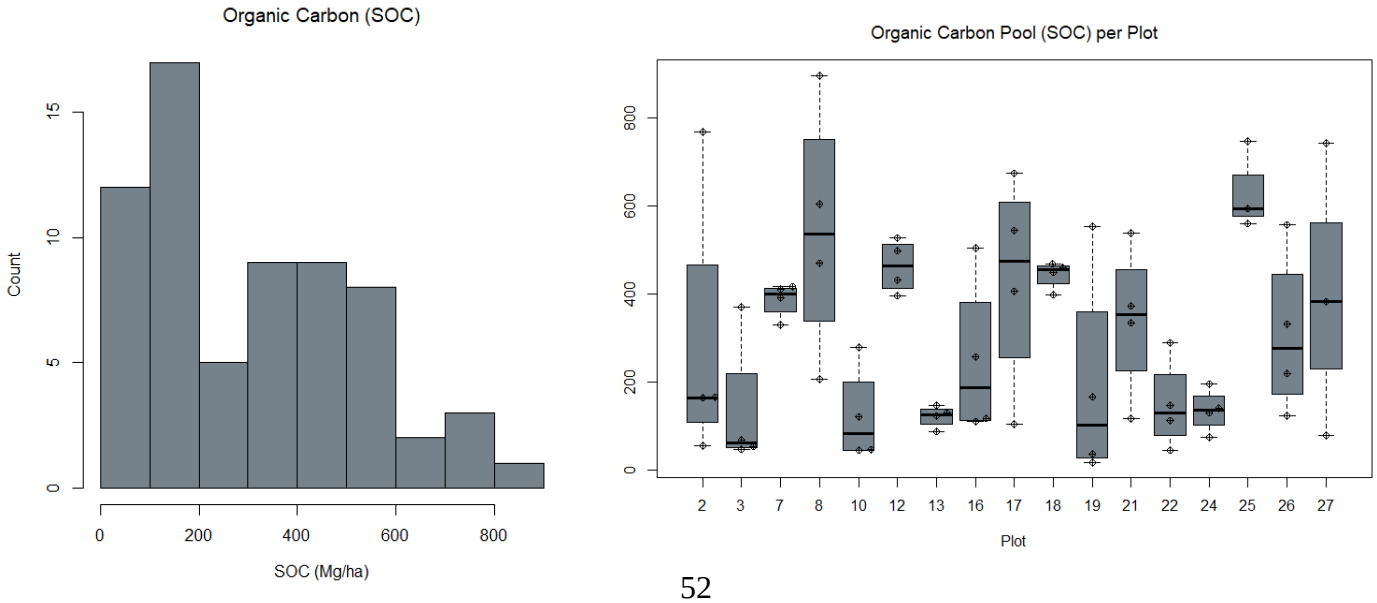
The total soil organic carbon pool (SOC) represents the soil organic carbon spatial density at each sample location. It was calculated using $\%OC$ and M_O and expressed in units of Mg/ha according to the following equations 3.4 and 3.5 from *Chapter 3: Methods.*

$$SOC_{g/sqm} = \frac{(\%OC)(M_o)}{(1.9635\text{x}10^{-3})} \ \frac{g}{m^2} \tag{3.4}$$

$$SOC_{Mg/ha} = \frac{SOC_{g/sqm}}{10} \ \frac{Mg}{ha} \tag{3.5}$$

SOC ranged from 17.1 MgC/ha to 896.9 MgC/ha with a mean of 308.5 MgC/ha and median of 284.1 MgC/ha (figure 19). SOC was tested for outliers on the highest and lowest values using the Grubbs-test (Grubbs, 1950) with $\alpha = 0.05$. It was found that the maximum and minimum values were not significant outliers ($p \gg 0.05$). SOC was not significantly normally distributed ($W = 0.93,\ p \ll 0.05$) with an overall standard deviation of 220.25 MgC/ha and an average inter-plot standard deviation of 156.39 MgC/ha (figure 19). In the following section *4.2 Predictor Variables* the relationship between SOC and each predictor variable in this study will be described.

Figure 19. Summary plots of Soil Organic Carbon Pool (SOC). (Left) Histogram distribution. (Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.

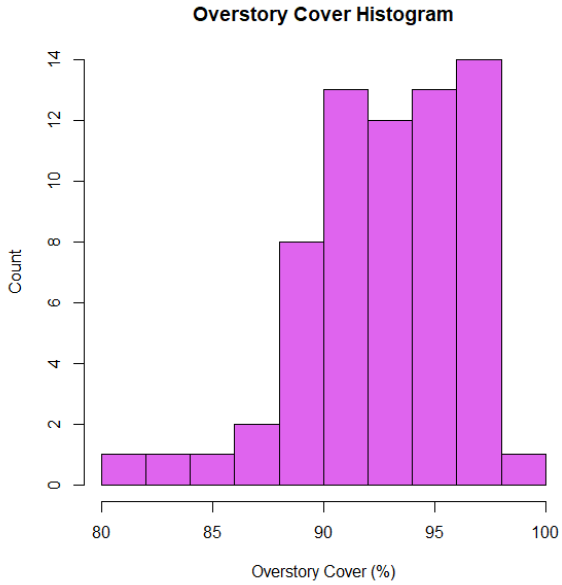


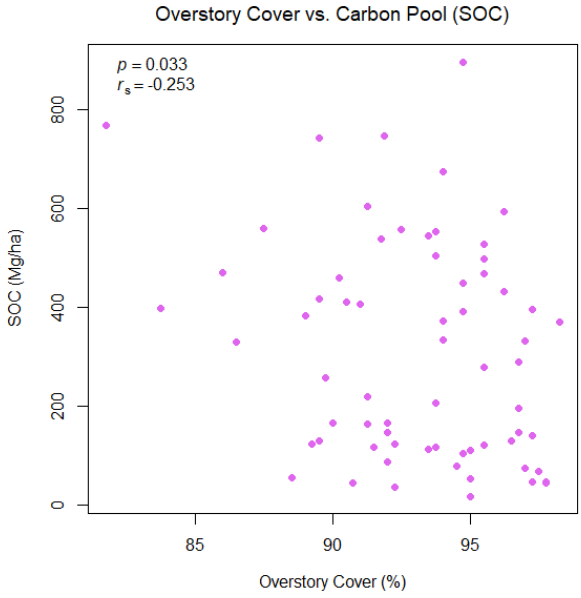
4.2 Predictor Variable Summary

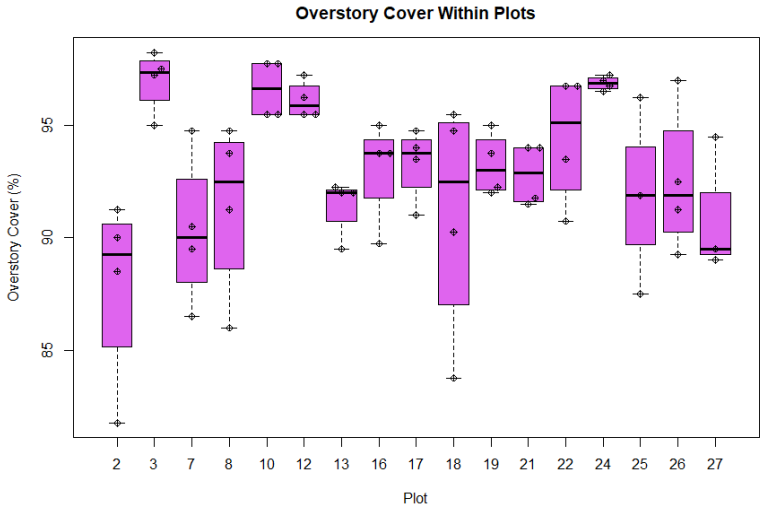
4.2.1 Overstory Cover

Overstory cover, measuring the percentage of light-blocking forest canopy cover, ranged from 81.75% to 98.25% across all sites (figure 20). Of the data collected, overstory cover was not significantly normally distributed ($W = 0.94,\ p < 0.05$) with a mean cover of 93.03% and median cover of 93.75%. Within each plot, the mean standard deviation of overstory cover was 2.50% (figure 20) compared to a total standard deviation across all plots of 3.51%. Overstory cover was found to have a negative correlation and significant linear relationship with SOC ($r_s = -0.253, p = 0.033$, figure 20).

Figure 20. Summary plots of overstory cover. (Top Right) Histogram distribution. (Bottom Left) Scatterplot of overstory cover against SOC pool. Included are the spearman's correlation coefficient (r_s) and the resulting p-value. (Bottom Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.



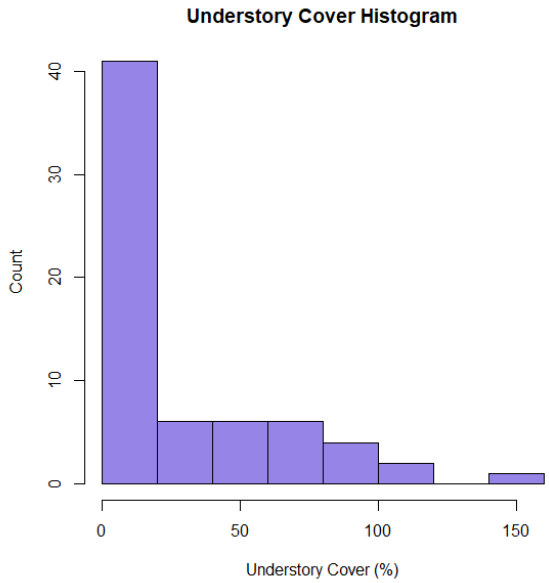


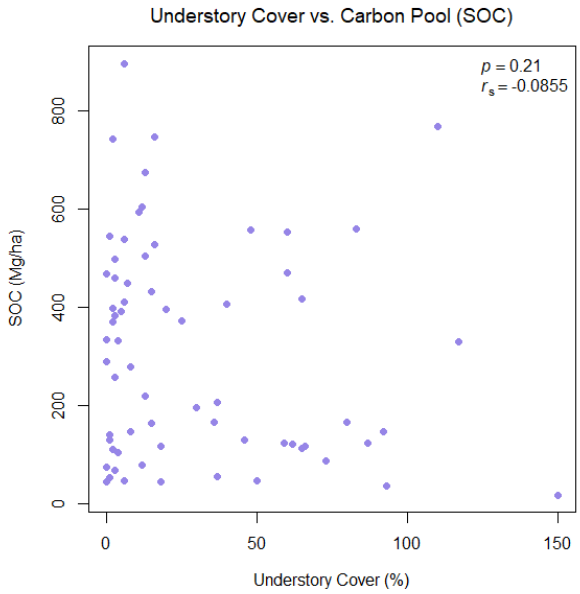


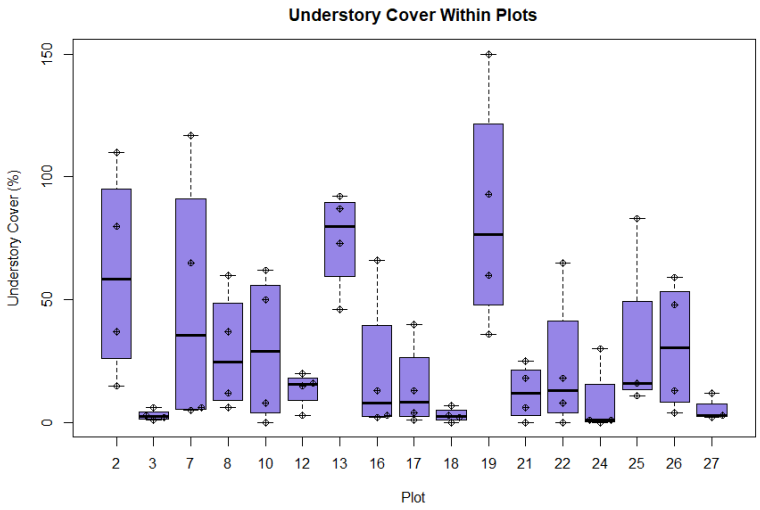
4.2.2 Understory Cover

Understory cover, measuring the total percent coverage by all understory plan species, ranged from 0% to 150% across all sites (figure 21). This value can reach over 100% because multiple plants may occupy the same space in dense areas. Of the data collected, understory cover was not significantly normally distributed ($W = 0.81,\ p \ll 0.05$) with a mean cover of 29.55% and a median cover of 13%. Within each plot, the mean standard deviation of understory cover was 24.12% (figure 21) compared to a total standard deviation across all plots of 34.81%. Understory cover was not significantly correlated with SOC ($r_s = -0.086$, $p = 0.21$, figure 21).

Figure 21. Summary plots of understory cover. (Top Right) Histogram distribution. (Bottom Left) Scatterplot of understory cover against SOC pool. Included are the spearman's correlation coefficient (r_s) and the resulting p-value. (Bottom Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.



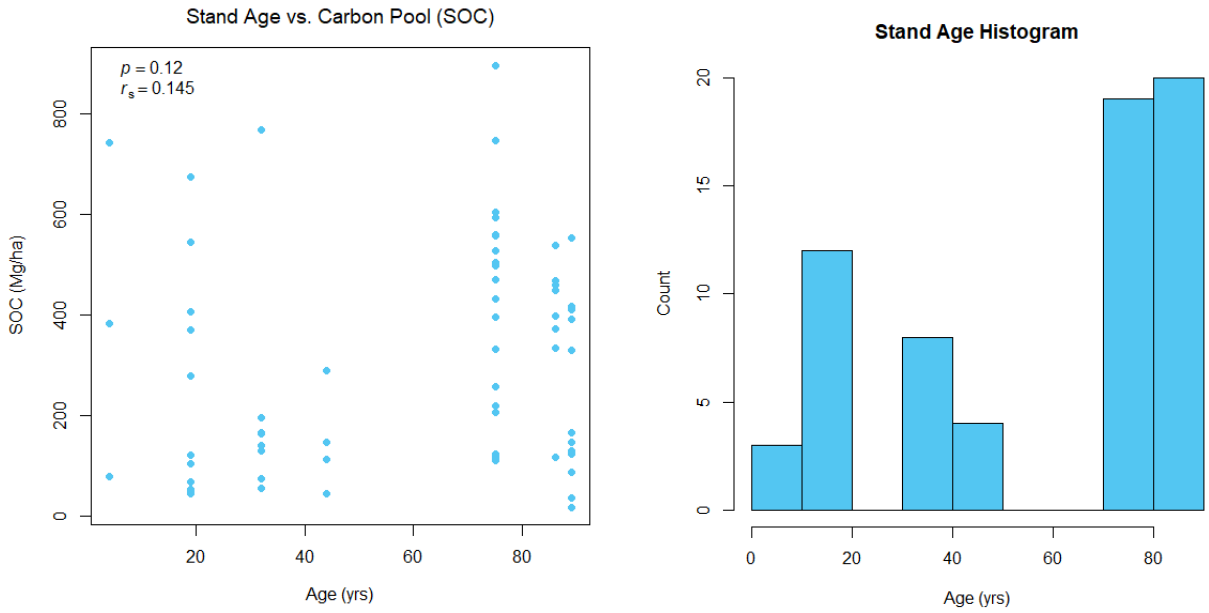




4.2.3 Stand Age (Age)

Stand age (age), measuring the time in years since the forest stand was last harvested, ranged from 4 to 89 years across all sites (figure 22). Of the data collected, age was not significantly normally distributed ($W = 0.82$, $p \ll 0.05$) with a mean of 58.38 years and a median of 75 years. Age was not found to have any linear relationship with SOC ($r_s = 0.145$, $p = 0.12$, figure 22).

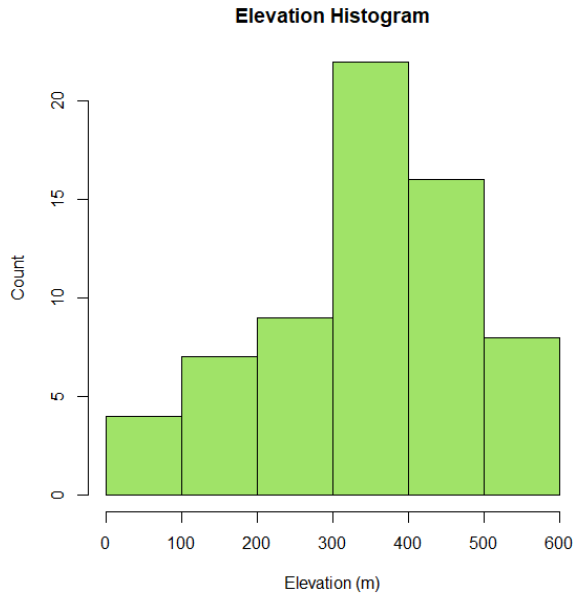
Figure 22. Summary plots of stand age. (Right) Histogram distribution. (Left) Scatterplot of stand age cover against SOC pool. Included are the spearman's correlation coefficient (r_s) and the resulting p-value.

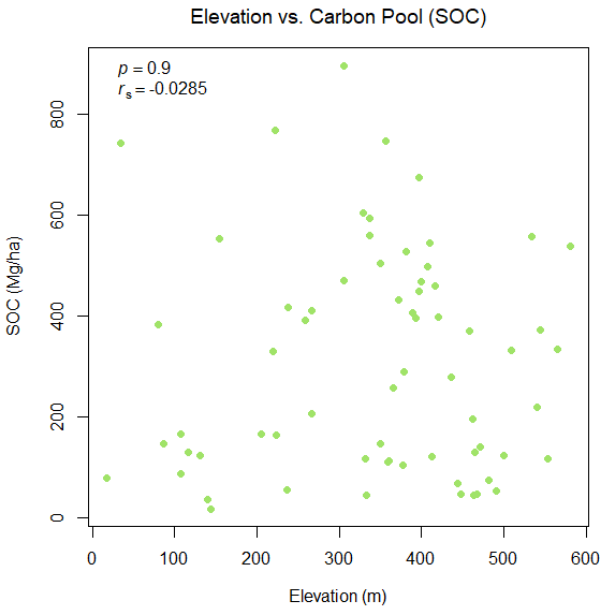


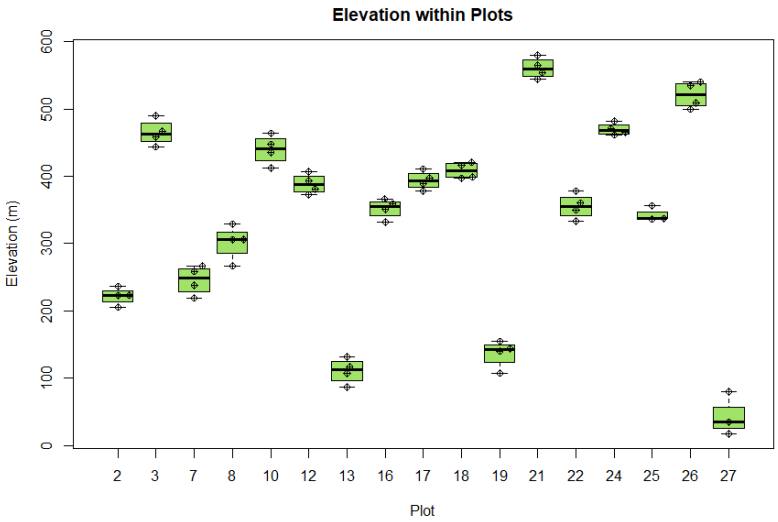
4.2.4 Elevation

Elevation, measuring the height in ft above sea level, ranged from 17.11 m to 580.58 m across all sites (figure 23). Of the data collected, elevation was not normally distributed ($W = 0.96,\ p < 0.05$) with a mean of 343.08 m and a median of 369.34 m. Within each plot, the mean standard deviation of elevation was 17.81 m (figure 23) compared to a total standard deviation across all plots of 140.90 m. Elevation was found to have no significant correlation or linear relationship with SOC ($r_s = -0.03$, $p = 0.9$, figure 23)

Figure 23. Summary plots of elevation. (Top Right) Histogram distribution. (Bottom Left) Scatterplot of elevation against SOC pool. Included are the spearman's correlation coefficient (r_s) and the resulting p-value. (Bottom Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.







4.2.5 Slope

Slope, measuring the angle of the terrain in degrees from horizontal, ranged from 13.49° to 48.62° across all sites (figure 24). Of the data collected, slope was normally distributed ($W = 0.97,\ p = 0.076$) with a mean of 33.58° and a median of 35°. Within each plot, the mean standard deviation of slope was 5.31° (figure 24) compared to a total standard deviation across all plots of 8.14°. Slope was not significantly correlated with SOC ($r_s = -0.09,\ p = 0.81$, figure 24)

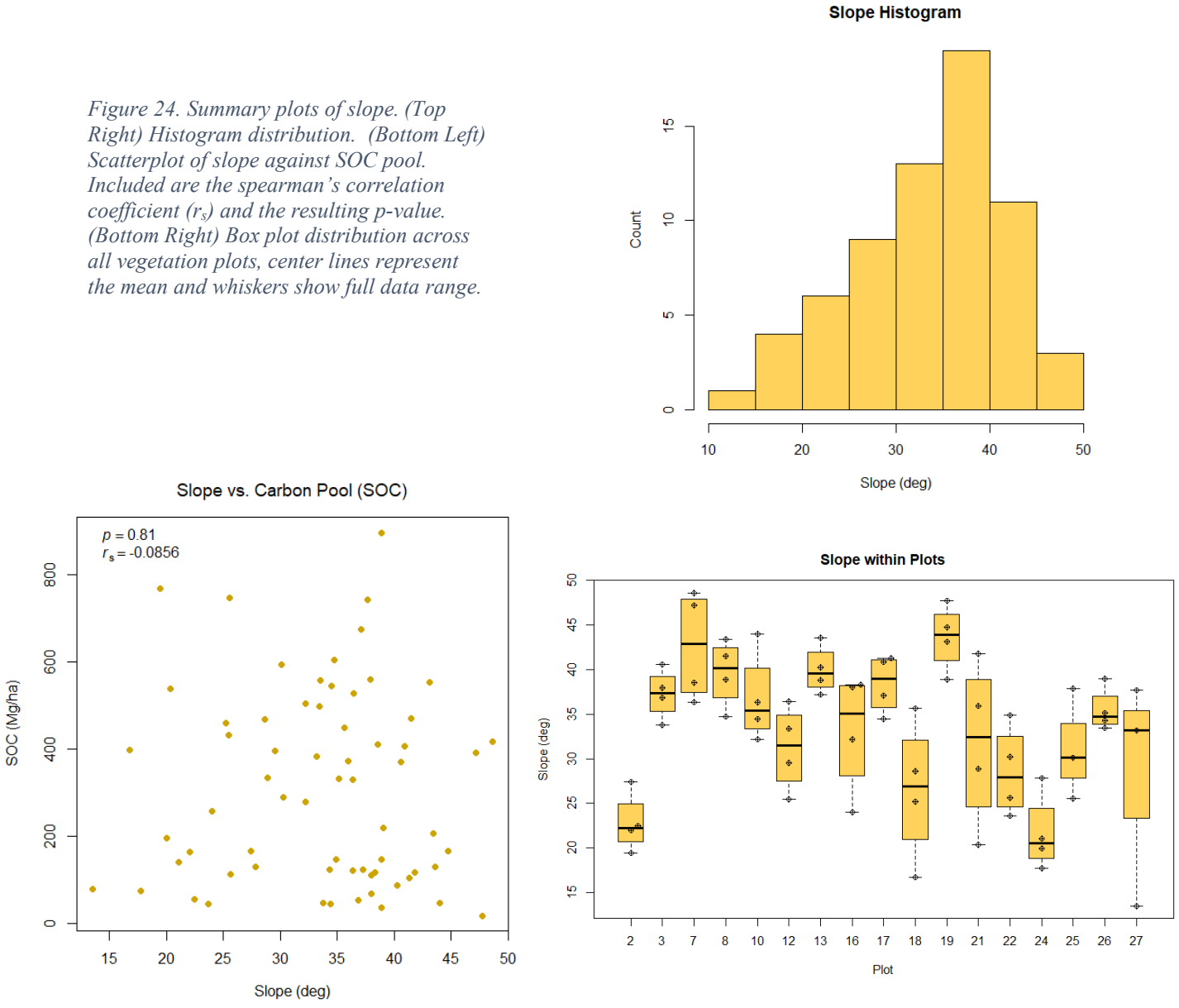


Figure 24. Summary plots of slope. (Top Right) Histogram distribution. (Bottom Left) Scatterplot of slope against SOC pool. Included are the spearman's correlation coefficient (r_s) and the resulting p-value. (Bottom Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.

4.2.6 Aspect

Aspect, measuring the cardinal angle of the terrain in degrees from north, ranged from 17.59° (NNE) to 357.01° (NNW) across all sites (figure 25). Of the data collected, aspect was not significantly normally distributed ($W = 0.88,\ p \ll 0.05$) with a mean of 171.66° (SSE) and a median of 121.80° (ESE). Sample sites primarily faced more strongly East and West (figure 25) due to the ravine structure of the N1 sub-basin. Within each plot, the mean standard deviation of aspect was 35.92° (figure 25) compared to a total standard deviation across all plots of 106.88°. Aspect was not significantly correlated SOC ($r_s = 0.07, p = 0.95$, figure 25)

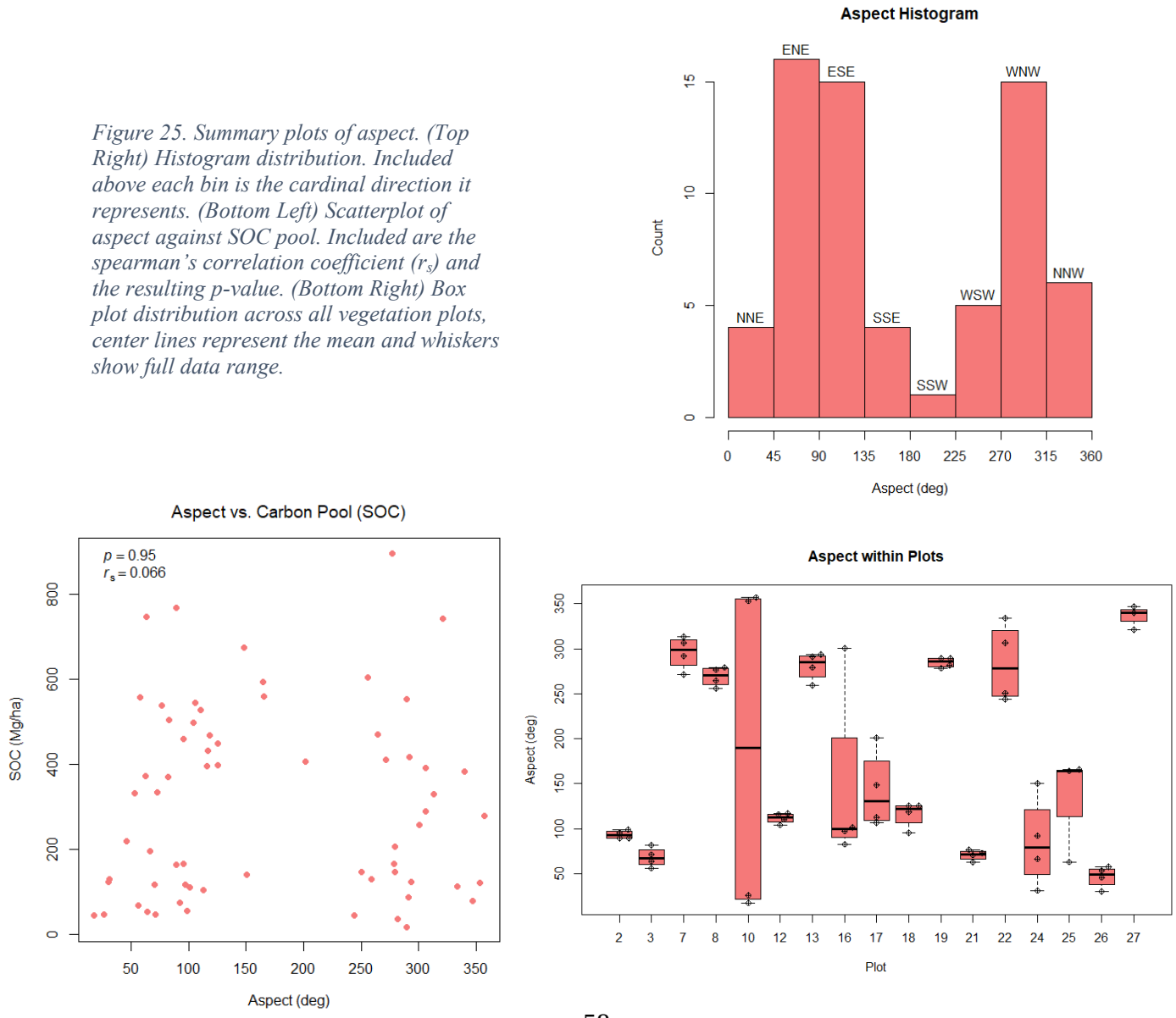


Figure 25. Summary plots of aspect. (Top Right) Histogram distribution. Included above each bin is the cardinal direction it represents. (Bottom Left) Scatterplot of aspect against SOC pool. Included are the spearman's correlation coefficient (r_s) and the resulting p-value. (Bottom Right) Box plot distribution across all vegetation plots, center lines represent the mean and whiskers show full data range.

4.3 Model Results & Evaluation

4.3.1 Model Performance

The above data was combined into a single random forest (RF) regression model (*mtry = 1, ntree = 1500, nodesize = 1)* with overstory cover, understory cover, stand age (age), elevation, slope and aspect as predictor variables and soil organic carbon pool (SOC) as the dependent variable. The remaining variables described in section *4.1 Carbon Variable Summary*, O-horizon depth, O-horizon mass, Bulk Density and Percent Organic Carbon, were not utilized for model construction. Percent organic carbon and O-horizon mass and depth used to calculate SOC storage. Bulk density is intercorrelated with SOC due to also using O-horizon mass and depth in its calculation. Additionally, a label-encoded dummy variable for the sample's vegetation plots was included to measure the potential effect of pseudo-replication within these plots. For further details on the model development, see *Chapter 3: Methods.*

To evaluate the RF model's performance, we used the coefficient of determination (R^2) and the root mean square error (RMSE), shown in table 4. The coefficient of determination can be interpreted in the same fashion as classical statistics method as the equation is the same. An R^2 of 0.874 indicates that the model explains 87.4% of the variation in SOC storage using the provided predictor variables. R^2 was calculated using equation (3.6) and the RMSE was calculated using leave-one-out cross-validation (LOOCV). Cross-validation in model evaluation typically involves splitting the dataset into training and testing subsets (Lachenbruch & Mickey, 1968). These subsets are used to calculate the error between predicted and actual values of the response variable. For datasets with a high number of data features (*n*), the testing subset is often built by taking a random

RF Evaluation Parameter	Result
R^2	0.874
RMSE (MgC/ha)	165.54

Table 4. The evaluation parameters (R^2 and RMSE) of the predictive RF regression model.

portion of features, usually 10-20%, assigning it to the testing subset and not utilizing it in model generation (Refaeilzadeh et al., 2009). When *n* is low, removing up to 20% of the data may limit model performance so LOOCV can be utilized instead (Cheng et al., 2017). In LOOCV, the model is generated by removing a single datapoint from the dataset and constructing the model using *n-1* values. Then this process is repeated for each datapoint in the dataset, providing *n* total predicted vs. actual error values. LOOCV is rarely used in studies with high *n* (*n* ≫ *100)* due to the computational cost of constructing the model many times repeatedly (Cheng et al., 2017).

Once LOOCV is performed, the RMSE of the model were calculated using the following equation (Gotelli & Ellison, 2012; Chai & Draxler, 2014):

$$RMSE = \sqrt{\frac{1}{n}\sum\left(Y_i - \widehat{Y}_i\right)^2} \tag{4.1}$$

Where *n* is the number of datapoints utilized in the model, and Y_i and $\widehat{Y}_i$ are the true and predicted values respectively of SOC using a single test datapoint each model construction. RMSE can be interpreted as the standard deviation of the model's residuals, which indicates how dispersed the true values are from the predicted values (James et al., 2014; Gotelli & Ellison, 2012). Our RMSE indicates that the model's prediction for the single left-out variable was, on average, 165.54 MgC/ha off of the true value.

4.3.2 Model Results

Producing interpretable results from an RF regression model is typically very challenging due to the high complexity of the prediction model (Breiman, 2001; Genuer & Poggi, 2020). Figure 27 shows an example of a single prediction decision tree (DT) out of 1500 utilized for model predictions. This figure only shows the initial three nodes of DT development and continues further

for another seven nodes before entirely terminating. Visualized inside each non-terminal node are the following characteristics: (1) the variable that was selected to generate the split and the split boundary value. For all node branches, left represents below or equal to and right represents above the value displayed. (2) The number of remaining data features (*samples*) represented at that node. Note that the topmost node shows that only 40 data features were used to construct this DT, representing the way that RF utilizes bagging. (3) The median true value for SOC in the remaining test sample that is utilized for splitting (*value*). Splitting continues until the number of remaining features equals the predetermined *nodesize* = 1, at which point it forms a terminal node. The terminal node's *value* is the predicted SOC when all the above branch conditions are met. For further description on DT generation and RF model background, see *Chapter 2: Literature Review*.

To provide a more visually interpretable result, this study produced a predicted digital soil map (DSM) for SOC using ArcGIS Pro (*see Chapter 3: Methods*). The resulting DSM, shown in figure 29, demonstrates an approximation of the models' predictions across the N1 sub-basin using interpolated values for overstory cover and understory cover.

The predicted values for SOC across the N1 sub-basin ($\widehat{SOC}$) ranged from 140.31 MgC/ha to 516.78 MgC/ha. The mean and median $\widehat{SOC}$ were 312.52 MgC/ha and 306.15 MgC/ha respectively. Figure 26 shows the

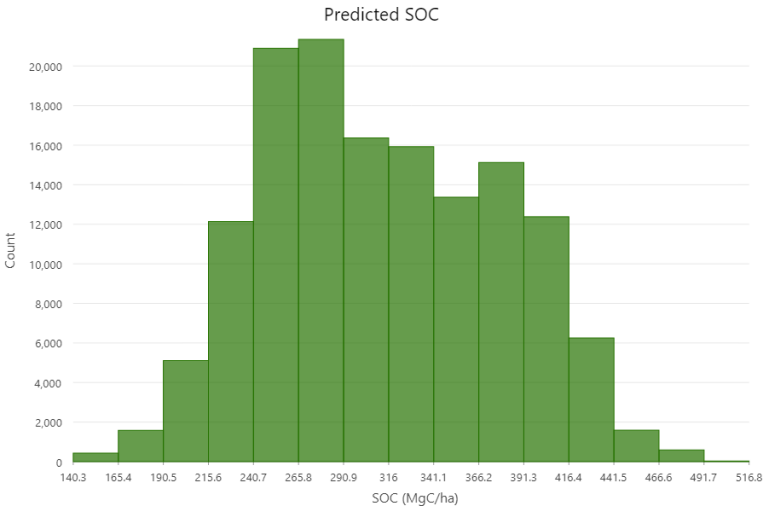


Figure 26. Histogram distribution of the DSM raster of predicted SOC values across the N1 sub-basin.

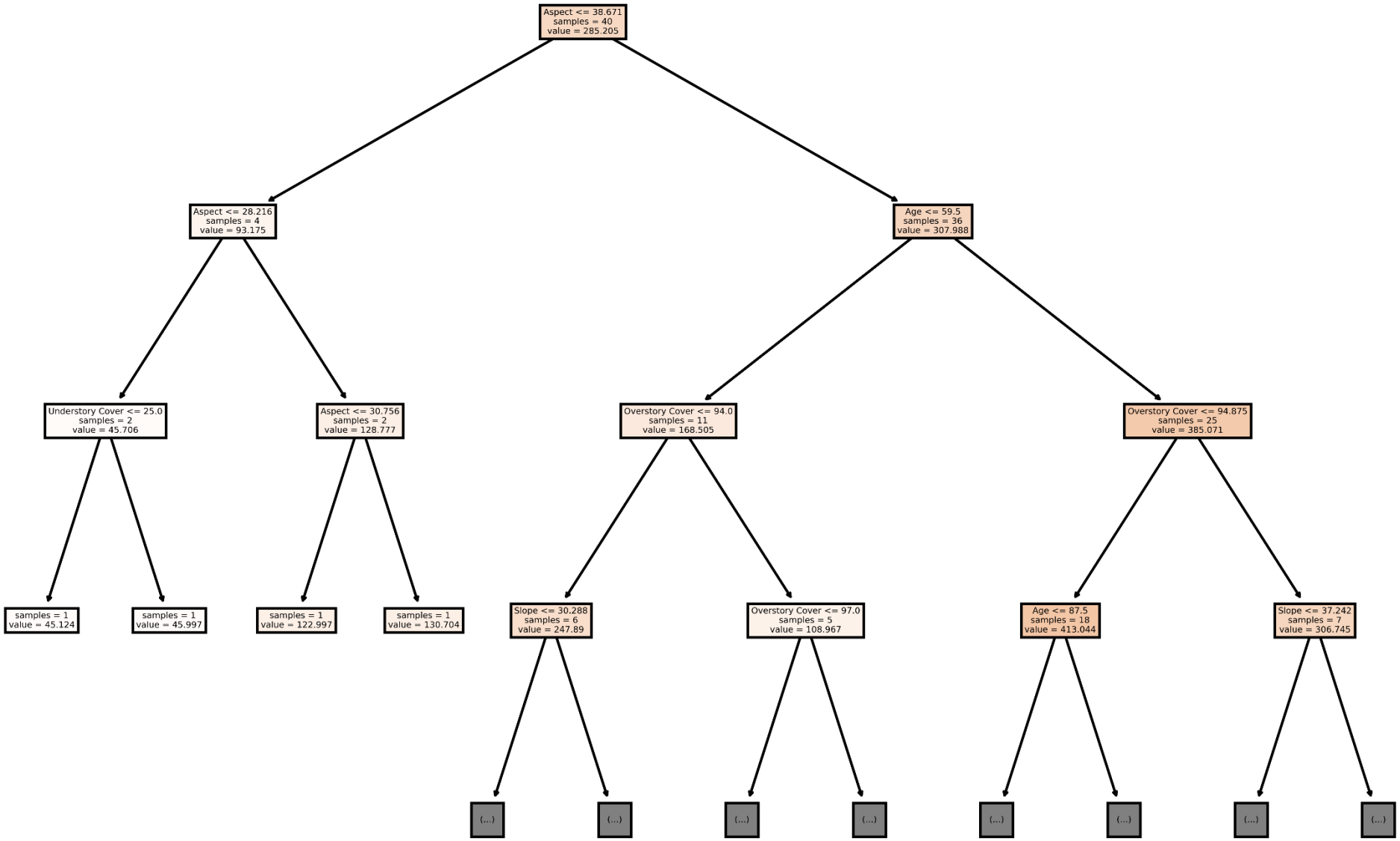


Figure 27. An example decision tree of the 1500 decision tree forest that makes up the random forest regression model utilized in this study. The variable name and boundary used for splitting is shown at the top of each node. Following is the number of remaining data features (samples). Soil organic carbon predicted values ("value", units: MgC/ha) are shown on the bottom of each node. Nodes that predict higher values of carbon are shaded in darker shades of orange.

distribution of $\widehat{SOC}$ across the N1-subbasin in addition to the measured values of SOC at the sample sites.

An additional tool to interpret the performance of the RF model is to consider how each variable individually affects $\widehat{SOC}$. Figure 28(A-G) shows these relationships for all six predictor variables as well as the plot dummy variable. To generate these plots, each other non-represented variable was held at their median (solid), maximum (dotted), and minimum (dashed) values, and the target variable was varied across their measured range. These plots do not convey the full complexity of each variable's contribution to the model due to the mutli-dimensional nature of DTs. Further discussion on the model's performance and dependence on each variable can be found in *Chapter 5: Discussion*

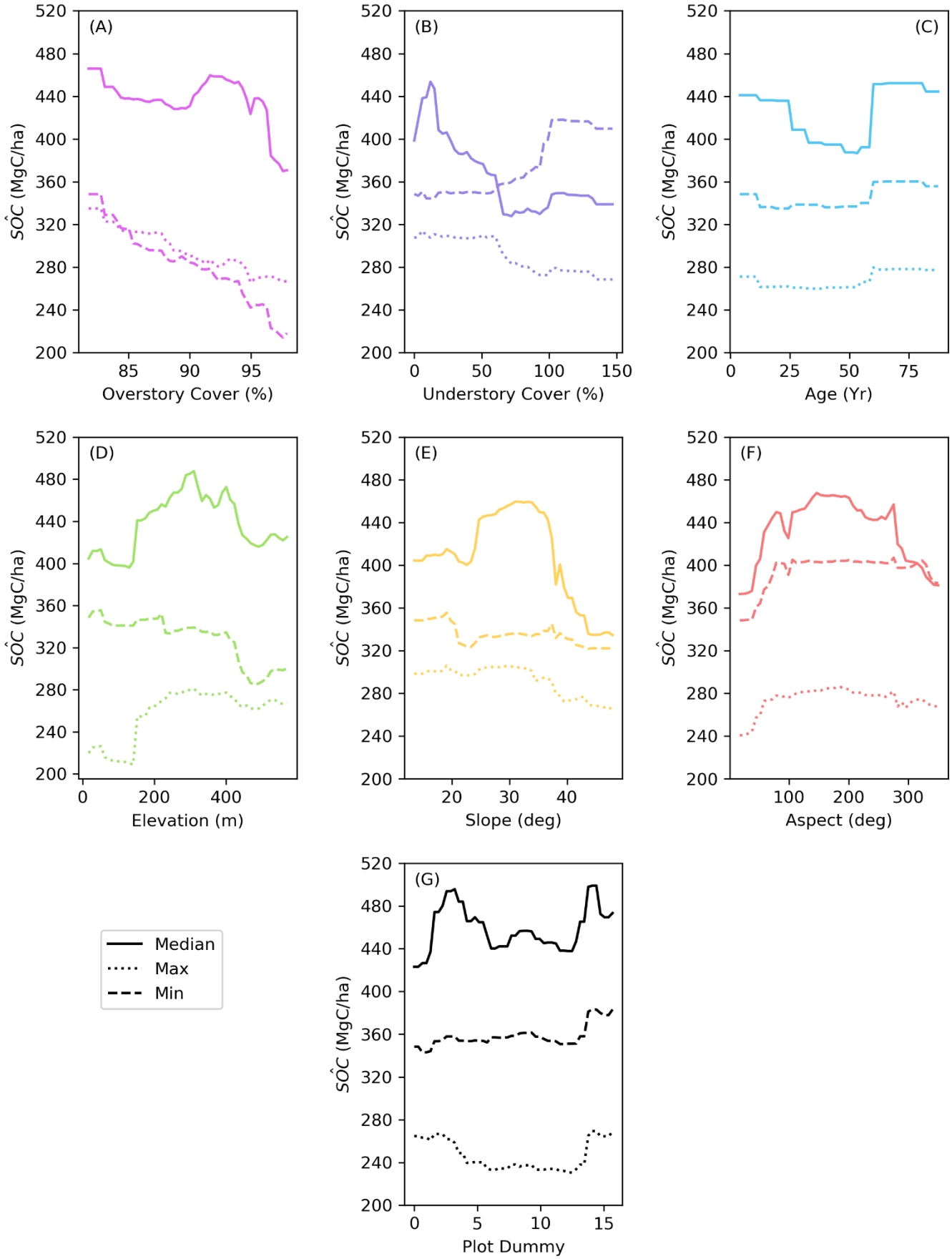
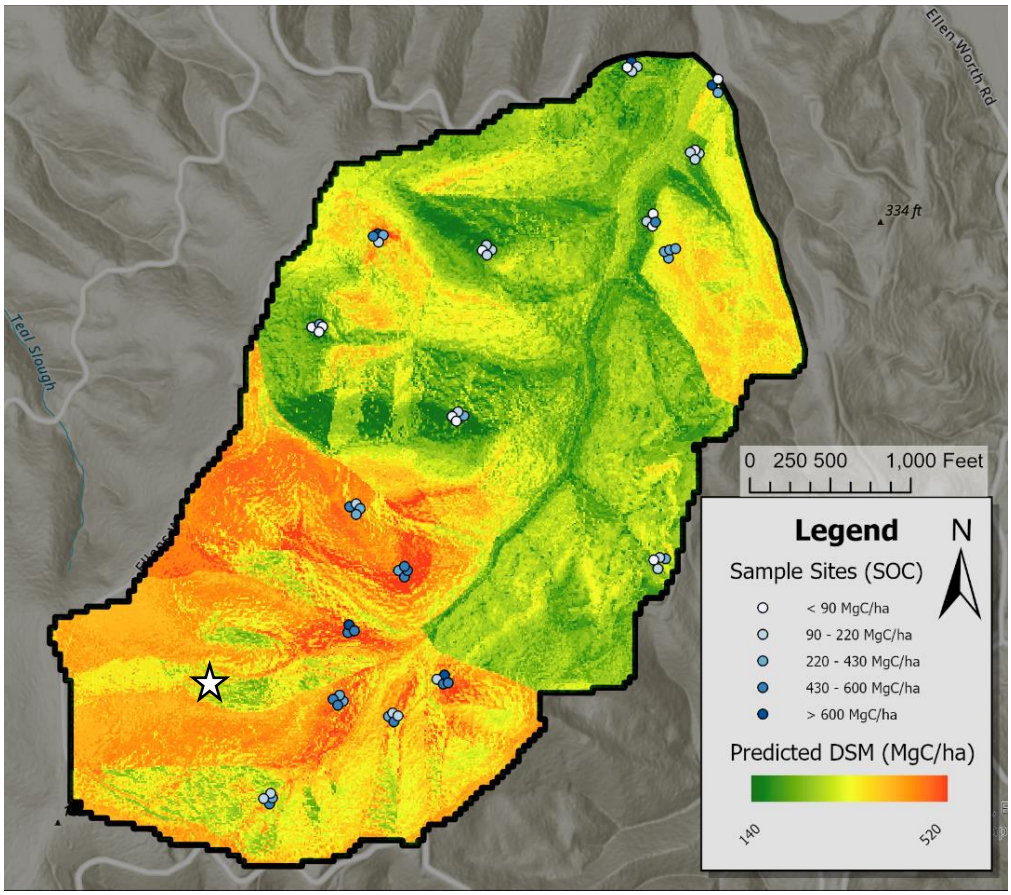


Figure 28. Predicted SOC using the RF model and adjusting only one variable across its measured range: (A) Overstory Cover, (B) Understory Cover, (C) Age, (D) Elevation, (E) Slope, (F) Aspect, (G) Plot Dummy. For each plot, all variables not shown are held at their median measured value.



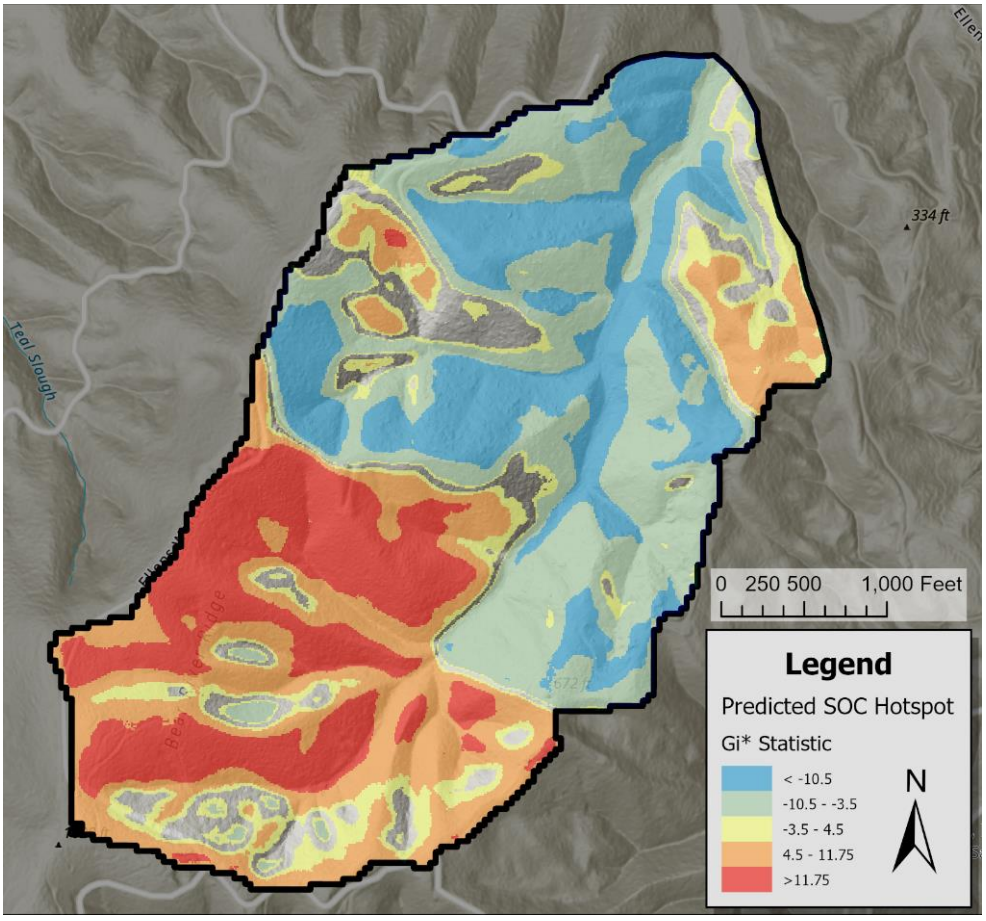


Figure 29. (Top) Predicted soil organic carbon digital soil map of the N1 sub-basin. Overlayed are sample points and their measured SOC (MgC/ha).

(Bottom)Hotspot map of predicted SOC visualizing the Gi statistic ($p<0.05$). Large positive or negative values of Gi* represent regions of significantly high or low SOC respectively.*

4.3.3 Hot Spot Analysis

Using the predicted soil organic carbon ($\widehat{SOC}$) digital soil map (DSM), hot spot analysis was performed to evaluate areas of statistically high and statistically low concentrations in $\widehat{SOC}$. Due to the interpolation necessary to generate the DSM, the hot-spot map serves more as an additional visualization tool rather than a direct ecological prediction across the sub-basin. See *Chapter 3: Methods* for a full description of how hot-spot analysis was performed.

Figure 29 shaded regions show areas with statistically significant ($\alpha = 0.05$) $\widehat{SOC}$ hot spots using the Getis Ord Gi* statistic (Ord & Getis, 1995). Visualized is the Gi* statistic, representing the strength of the hotspot, which ranged from -25.82 to 27.14. This value can be considered equivalent to a z-score which is interpreted as deviation above and below the mean as a ratio of the standard deviation (Gotelli & Ellison, 2012). Warm-color shaded (orange and red) areas represent significant positive GI* values. The $\widehat{SOC}$ values in those areas are greater than 4.5 standard deviations above from the median, representing a $\widehat{SOC}$ hot spot. Cool-color (light-blue and blue) shaded areas represent significant negative GI* values. The $\widehat{SOC}$ values in those areas are greater than 3.5 standard deviations below from the median, representing a $\widehat{SOC}$ cold-spot.

Chapter 5: Discussion

Second-growth coastal temperate forests similar to those present in the N1 sub-basin at Ellsworth Creek Preserve are known for their high above- and below-ground carbon storage, representing important carbon sinks within the region (Carpenter et al. 2014). Our investigations into the soil organic carbon (SOC) dynamics of the organic horizon in this forest demonstrated a highly complex and variable system. Previous studies that investigated carbon storage on equivalent spatial scales have highlighted results of similar complexity (Matinfar et al., 2021; McCarthy & Brown, 2006; Stutter et al., 2009). The N1 sub-basin is particularly unique due to its lack of management for the last two decades following a century-long intensive timber management history. Our model findings showed a consistent SOC relationship with forest age that has been historically confirmed and highlights the impact of timber harvest on SOC (Covington et al. 1981). Many of our findings highlighted a need for additional sample points to identify significant relationships with each variable individually. But utilization of the machine learning modelling method extracted similar relationships to previous findings when all the data is considered together in a single model. The following section will discuss the individual findings for each predictor variable, then how they contributed to the broader model. Additionally, we consider the capability of the model to generate a digital soil map (DSM) and discuss limitations and future work.

5.1 Soil Organic Carbon and Predictor Variables

The median measured value of SOC in the organic horizon (O-horizon) for the N1 sub-basin was 284.1 MgC/ha, which is also characteristic of the region. Carpenter et al. found a median measurement of SOC in coastal forests in Oregon and Washington of 211 MgC/ha compared to

143 MgC/ha across the entirety of the Pacific Northwest (2014). The relative increase reported here as well as in other coastal forests could be attributed to the lack of management causing the forest to become denser, slowing decomposition rates in the soil and thus storing more organic carbon in the O-horizon across the sub-basin (Liu et al., 2014).

Data provided on overstory cover, understory cover, and stand age by The Nature Conservancy (TNC) conveyed the complexity of the study forest system. This data was collected in 2020 so the current actual field values may be different in the sub-basin today. Overstory cover was overall high across the sub-basin, characteristic of coastal rainforests in Washington (Carpenter et al., 2014). High values for canopy cover will affect both the volume of organic matter inputs into the soil but also the temperature and moisture systems in the forest (Liu et al., 2014). Previous studies have found significant positive correlations between overstory cover percentage and SOC stock (Maraseni & Pandey, 2014; Saimun et al., 2021). This is contrasted by the significant, albeit weak, negative correlation between overstory and SOC observed in this study (Fig. 20). This is likely attributed to sample size and confounding forest development factors such as tree-species make-up and topology.

Understory cover showed high variability across the sample sites, ranging from 0% to 150%. This is to be expected in a variable, second growth forest, where post-harvest conditions can affect the capability of understory plants to grow (Zhang et al., 2022). The median understory cover was 13%, showing that the forests were generally very sparse with a few regions having high cover. Only 16 samples had understory cover values over 50%, showing a strong bias of this data towards low understory cover. Many of the stands were dominated by dense, small trees, that did not allow a significant understory to develop. Zhang et al. found that understory cover may increase soil carbon storage, and its absence can cause increased carbon loss, though this effect is

not as prevalent in coniferous forests (2022). This study observed no relationship between SOC and understory cover, which is potentially attributed to the strong bias of the data towards low values of understory cover.

Stand age ranging from 4 years to 89 years across the sub-basin caused the individual forest stands to display distinctly different characteristics. Figure 30 shows images of two vegetation plots where the ecological characteristics of each are noticeably different. N1-27 (top) is a 4-year-old stand full of dense narrow trees and a high volume of downed woody debris. N1-25 (bottom) is located on the southern portion of the sub-basin in a 71-year-old stand. Here the forest is noticeably less dense with almost no understory. With a median age of 75 years, 70% of the sample points will have crossed the 60-year threshold for soil organic carbon recovery (Covington et al., 1981). During and directly after timber harvest, the top layer of the soil is disturbed, and SOC has been found to decrease drastically due to changes in respiration rates and litterfall inputs. Following a steep decline in SOC there is a long period of recovery until after 60-100 years (James & Harrison, 2016). This may explain why there was very little relationship between SOC and stand age during direct analysis. We would expect stands that are over the age of 60 would no longer depend on tree age for their SOC variation.

The topological variables of elevation, slope, and aspect characterized the sub-basin as extremely topologically varied with steep, sharp slopes and large elevation changes. Elevation varied over 550 meters across the sub-basin, with the lowest samples being only 17.11 m above

Figure 30. Two images in the N1 sub-basin of Ellsworth Creek Preserve. (Top) A young 4-year-old stand with high downed debris and dense trees. (Bottom) An old 71-year-old stand with low understory and less tree density

sea level. Generally, the samples were biased to East and West facing slopes, which was expected due to the geological ravine-like structure of the study area. Runoff tended to flow North causing all slopes to reside on either the eastern or western sides of the resulting drainage feature. None of the topological variables showed any strong or significant relationships with measured SOC values, which is inconsistent with previous study findings (Szatmári et al., 2021; Bhardwaj et al., 2016; Tsui et al., 2004; Zhu et al., 2017). We would expect SOC to strongly relate to topography due to the influence of elevation, slope and aspect on pedogenesis and soil evolution (Bockheim et al., 2005). Tsui et al. found that respiration and decomposition rates slowed with increased elevation due to low temperatures and high-water runoff (2004). This would lead to an increase in SOC storage for high elevations if all other variables were controlled. They also found that steep gradient slopes increase leaching and lead to changes in respiration. Additionally, Olson found that high slope values lead to increased erosion, particularly in the less compact topmost layers of the soil (2010). On such slopes, we would expect a decrease in SOC values as those top layers are the highest in SOM content (Hartemink et al., 2020).

The lack of statistically significant findings is due to sample size, plot designation methodology and topography intercorrelation. Vegetation plots, while randomly placed, had to be located in areas that were accessible by foot to manage data collection. As such, this bias may have caused the forest characteristics or the topological variables to not show true randomization. Additionally, due to the small sample size, topographic measurements within the sub-basin were highly intercorrelated, with central elevation values often having high slopes.

5.2 Model Results

A predictive model for SOC was generated using random forest (RF) regression to generate many decision trees (DT) that use bagging to determine a predicted result. The above variables in addition to label encoded plot dummy variable to highlight potential pseudo-replication were utilized in this model construction. The model was evaluated using R^2 and RMSE which were found to be 0.84 and 165.24 MgC/ha respectively. Compared to the median measured SOC of 284.1 MgC/ha, the RMSE represents approximately a 58% error from median value, showing high variability in the success of predicted results. By viewing the R^2 and RMSE value together, we see potential for model over-fitting. The data is able to explain the variance of the input data well, but when required to make a new prediction, the resulting prediction has a high error rate.

This result is comparable to some previous studies but was also far less successful compared to other studies that predicted SOC using RF (Matinfar et al., 2021, Grimm et al., 2008). Mantifar et al. developed a predictive SOC model in a grassland in Iran using a variety of machine learning methods, including RF (2021). Their model, which considered the top 15 cm of soil, used percent SOC by weight for their dependent variable and had a resulting R^2 and RMSE of 0.84 and 0.24% respectively. Their RMSE represented 43% error from mean predicted SOC percentage. The primary differences between this study and Mantifar et al. were their use of entirely remotely sensed variables as well as the nature of the study area.

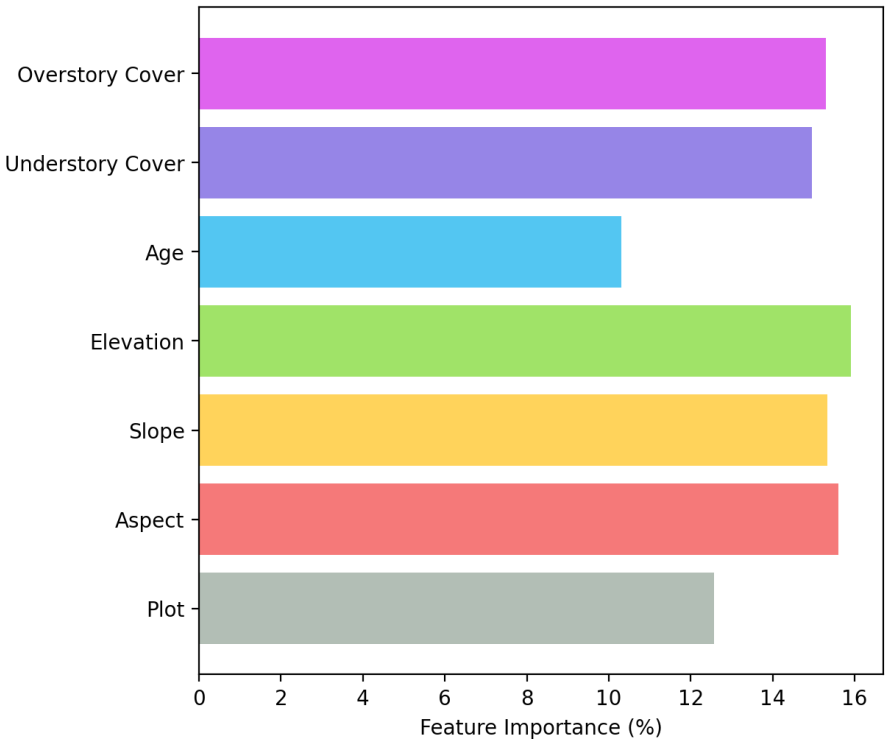
Grimm et al. built a predictive RF model for SOC in Panama and had a median RMSE in the top 10 cm of 0.55 MgC/ha, with a measured SOC mean of 38.05 MgC/ha (2008). Their model showed only a 1.5% error from the mean measured value. Their model's success may be attributed to a significantly larger sample size (n=161) and inclusion of many predictor variables not considered in this study such as soil texture.

5.2.1 Variable Importance

In addition to overall model performance, we considered the contribution of each variable to the model's predictions. This is performed using Gini importance, or mean decrease in impurity, which evaluates the normalized total decrease in impurity that each variable provides to model predictions across all DTs (Breiman et al., 1984). This method for evaluating feature contribution is commonly used in RF and other machine learning models (Grimm et al., 2008; Matinfar et al., 2021; Padarian et al., 2020). Figure 31 shows the feature importance of each variable to the model's performance. The values presented can be interpreted as percentage direct comparisons to contribution to the model overall.

We found that all included variables showed strong contribution to the model. Elevation provided the most significant contribution with 15.9% compared to the lowest contribution of 10.3% by Stand Age. The plot dummy variable also underperformed compared to other variables

Figure 31. Variable contribution as measured by the gini index for all predictor variables in the random forest model utilized in this study.



at 12.6%. The remaining four variables, overstory cover, understory cover, slope, and aspect, provided approximately equivalent contributions. This result indicates that the RF model was able to identify relationships between many of the predictor variables and SOC that the preliminary analysis was not.

Additionally, the approximately equivalent contribution of all seven variables indicates that no variable was a dominant contributor to model performance. Exclusion of any single variable would result in a similar loss of model purity. This highlights the benefit of including multiple variables that have an effect on the large-scale controllers of SOC storage – temperature, moisture, and organic matter input. Temperature variation, for example, would not have been equivalently captured without consideration from overstory's evapotranspiration, understory's shade cover, elevations air temperature, and slope & aspect's sun exposure (Liu et al., 2014; Ruiz-Colmenero et al., 2013; Tsui et al., 2004; Zhu et al., 2017).

5.2.2 Variable Individual Relationships

In addition to an overall predictive model for SOC, we considered how varying each feature individually affected predicted SOC ($\widehat{SOC}$) by holding all other variables constant to approximate a one-dimensional model. Figure 28 shows this result using three cases – when all not considered variables are held at their median, maximum and minimum. To aid in discussion, consider the following ecological interpretations of these cases:

- Median case: a mature, structurally complex forest on a moderate south-facing slope.
- Minimum case: a young, sparse, lowland forest with no slope gradient.
- Maximum case: an old, dense, high-elevation noth-facing forest on a steep ridge.

These three cases are not necessarily ecologically representative of any specific location in the N1 sub-basin. Rather, they are necessary to consider how each variable relates to SOC due to the multivariate nature of the RF model, which causes the predictions of varying a single feature to still depend on the status of the remaining variables.

There were many prominent patterns identified from this analysis. Nearly all variables predicted higher SOC values when the remaining features were held at their median, illustrating that generally at the extremes there is lower $\widehat{SOC}$. Overstory cover showed a negative relationship with $\widehat{SOC}$ across all three cases, which is consistent with preliminary analysis with measured SOC (Fig 20). As mentioned above, previous research has indicated that SOC stocks would increase under canopies with higher density – the inverse of what we demonstrate here (Maraseni & Pandey, 2014; Saimun et al., 2021). This inconsistency in findings may be due, in part, to the extremely high values of overstory cover sampled in the N1 sub-basin. The relationship established by Maraseni & Pandey (2014) was considering overstory cover thresholds at or near 70%. The median overstory cover measured in this study was 93.75% with no measurement below 80%. This indicates a need for more numerous and diverse samples in regions with canopies that are significantly less dense.

Understory cover also showed a general negative relationship with $\widehat{SOC}$, though the minimum case deviated from this behavior. We would expect overall increased SOC storage with higher values of understory cover (Ruiz-Colmenero et al., 2013). The behavior of understory cover in the minimum case is representative of the expected behavior as the vegetation will serve as a direct source of plant residue (Liu et al., 2013). Our finding in the minimum case is validated by previous studies, which provides perspective on the relative impact of understory cover – particularly in relation with overstory cover. In the minimum case overstory cover was near 80%

which may have provided the opportunity for understory cover to contribute to the SOC system more significantly. We believe this also indicates a need for more diverse samples in areas of low overstory cover to consider this relationship.

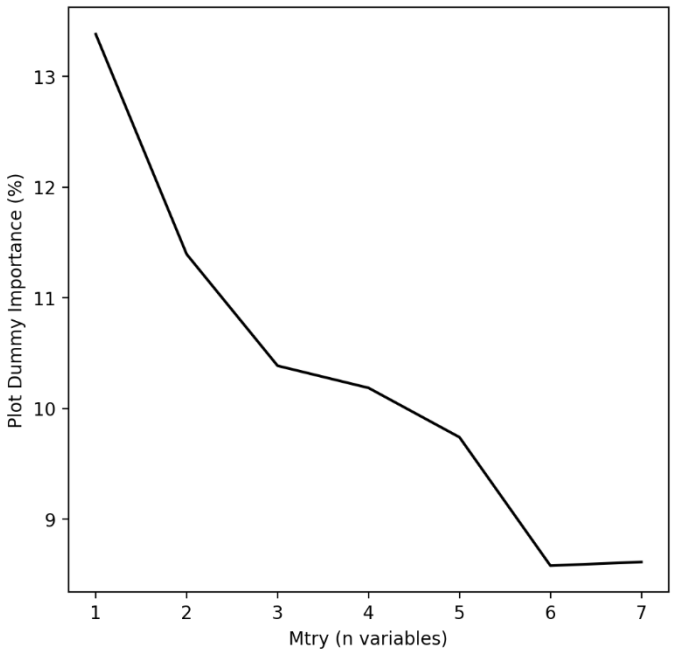
Stand age showed perhaps the most interesting behavior when isolated in the RF model. For each of the median, maximum and minimum plots we observed a similar shape with a progressive decrease in $\widehat{SOC}$ until around 15-30 years, and then a period of recovery until approximately 60 years where there is a near-entire recovery of $\widehat{SOC}$. Timber harvest has long been known to affect SOC in the topsoil due to increased elemental exposure and soil disturbance. Figure 3 shows the life of SOC following timber harvest known as the Covington curve (Covington et al. 1981). The general behavior of SOC observed by Covington and confirmed by many studies has been captured in this RF model (Yanai et al., 2003; Chen & Shrestha, 2012; Deng et al., 2022). Notably, the two results do not explicitly align. Rather, our finding shows a shallower decline following harvest compared to Covington. This does not necessarily indicate disagreement between the two results as it may have been a result of sample bias. Due to sample plot designation and the size of timber forest stand, only seven distinct values for age were measured in this study. As such we cannot reach any specific conclusions on a fine-scale relationship between SOC and Age and would require further data across a diversity of ages. Overall, this result highlights the capability of RF to capture large-scale physical relationships with sparse datasets, as this relationship was not observed during preliminary analysis using classical statistics.

Each topological variable showed strongly chaotic behavior with $\widehat{SOC}$ when isolated in the RF model. Elevation, which was the strongest contributor to model purity as discussed in the previous section, displayed very different behavior across all three model states. The geological nature of this sub-basin causes a strong inter-relationship between elevation and slope. Generally,

the extreme values of elevation are regions with low slopes, at the top or bottoms of ravines. For the central values of elevation, we observed high slope. As such, the chaotic behavior of elevation and slope in the isolated models is potentially related to the inter-relationship of those variables. For Aspect, each isolated model showed similar behavior with a relative central maximum on south facing slopes (180°-200°). This finding opposes what would be expected as south-facing slopes would have increased sun exposure in the northern hemisphere, resulting in higher temperatures and increased decomposition and respiration rates (Franzmeier et al., 1969; Onwuka, 2018; Rey et al., 2005). Additionally, this is inconsistent with behavior observed on a similar latitude near the Mediterranean, where north-facing slopes had higher measured values of SOC (Jakšić et al., 2021; Lozano-García et al., 2016). While there are many confounding factors such as climate, plant species, and geological structure in making this cross-continental comparison, it highlights a need for further measurements on a greater range of aspect values. One potential explanation for the observed behavior is the general topography of the sub-basin which faces generally North, with high western and southern ridges. This region of coastal Washington is known for strong wind and storm events, which can affect forest development (Beck et al., 2018). The south-facing slopes would then be more wind protected, which would result in less erosion and an associated increase in SOC storage (Li et al., 2019; Wang et al., 2023). The data collected as part of this study, due to the nature of the sub-basin, were biased on eastern and western dominated slopes, so there is a need for increased sample size for northern and southern slopes to better characterize the relationship.

5.2.3 Pseudo-replication

Figure 32. Plot dummy variable model importance for increasing values of mtry, the number of variables to include in each model split.



Due to the site design at Ellsworth by The Nature Conservancy, samples were collected in groups of 3-4 centered around a 0.1 ha vegetation plot central point, 9 meters horizontally from center. This is cause for concern for pseudo-replication (PR) with data analysis and model construction. Prior to model construction, PR was considered by evaluating the standard deviation (SD) of each measured variable overall and within their vegetation plot. To compare and normalize these values, we considered the standard deviation ratio of $SD_{plot}/SD_{overall}$ for each variable. Figure 32 shows this ratio for all measured values (See *Chapter 4: Model & Results*). Nearly all variables showed a lower SD within each plot compared to their overall deviation. This indicates a potential effect of PR where samples within vegetation plots are not truly randomly sampled compared to the overall dataset. Although, O-horizon depth and O-horizon mass showed larger SDs within each plot compared to overall, which indicates that the nature of the O-horizon highly variable, even on small spatial scales.

For all model construction and analysis prior to utilizing spatial visualization, a dummy variable for vegetation plot (plot dummy) was included to quantify the effect of PR (Urban, 2005).

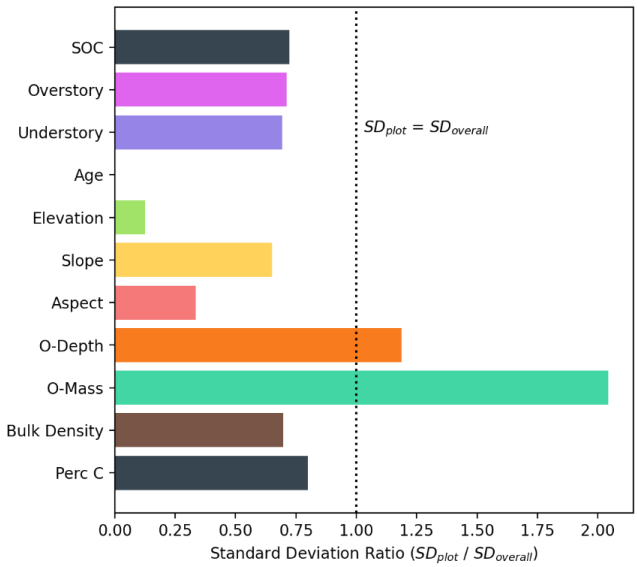


Figure 33. Ratio of standard deviation of each measured variable within each vegetation plot and overall. Values of 1 represent equal deviations.

Plot dummy was the second worse contributing variable to the model's performance at 12%, which can be interpreted as any effects associated with PR in this model is non-dominant. Additionally, in the isolated models shown in Figure 28, we would expect the plot dummy plots to produce consistent $\widehat{SOC}$ across the entire plot range. This is particularly untrue in the median plots, which have highly dynamic SOC predictions, suggesting that PR may be present in this study. Comparatively, $\widehat{SOC}$ is consistent in the maximum and minimum plots, showing that PR may not have as strong of an affect on the extremes of the dataset. During model optimization, we also found that plot dummy's contribution to the model decreased with higher values of *mtry*, though overall model performance decreased (figure 32). This result indicates that any spatial inter-relationship of variables is less significant when more variables are considered. Intuitively, we expect that when the model can consider more variables it can better isolate samples by gathering more detail.

The effect of PR within this study is primarily due to the spatial inter-relationship of samples due to the study design of Ellsworth. Though PR does not appear to dominate the model's

performance overall, its presence is opaque throughout field data collection and model construction.

5.2.3 Model Wrap-Up

Overall, this study built a successful predicted SOC ($\widehat{SOC}$) model with a few significant flaws. Model performance aligned with previous studies in forests at the field-scale. All collected variables were strong contributors to model success. Some individual variables, such as stand age, produced results that are ecologically representative of previous historical findings. Model flaws include evidence of over-fitting with a large disparity between R^2 and RMSE as well as a strong presence of pseudo-replication.

5.3 Spatial Analysis

5.3.1 Digital Soil Map

Following model construction, this study considered the spatial distribution of predicted soil organic carbon ($\widehat{SOC}$) across the sub-basin through the construction of a digital soil map (DSM). In addition to providing model insights, this process also created a visually interpretable random forest model result. To construct this DSM, overstory and understory cover were interpolated across the sub-basin using point data and stand polygons were estimated from historical timber stands (see Figures 9,13,14 for the interpolated and estimated rasters). As such, this DSM is an approximation and not necessarily ecologically representative (See *Chapter 3: Methods)*. Additionally, this DSM was limited to a 3 square meter resolution due to computational complexity. This limits the interpretability of fine-scale variation, though sample sites were approximately 6 meters apart leading to no raster bin overlap between samples. Figure 29 shows the resulting DSM in addition to the sample locations and their measured SOC.

This DSM demonstrates the multivariate nature of the predictive model. $\widehat{SOC}$ was comparatively low on north facing slopes and low drainage streams. Generally, we observe that the southern portion of the sub-basin has high $\widehat{SOC}$ likely due to the old, 71-year-old stand that covers most of that region in addition to the low understory cover present in that region. The DSM was able to capture small pockets of $\widehat{SOC}$ variance that are likely due to small regions of high, steep slopes (indicated with a white star in figure 29). This is ecologically representative as steep slopes in regions of high rainfall and extreme weather can increase the rate of erosion and sediment loss (Wang et al., 2023). The first sediment to lose is the O-horizon which would shed downslope, reducing SOC on the slope and increasing SOC downslope.

5.3.2 Hot Spot Analysis

In addition to generating an overall DSM for $\widehat{SOC}$, we considered regions of significantly high and low $\widehat{SOC}$ using hot spot analysis and the Gi* statistics. Figure 29 shows a map of $\widehat{SOC}$ hotspots where the Gi* statistic can be interpreted as strength of regional deviation. This map provides a more macroscopic picture of where SOC is predicted to be located in the sub-basin. The oldest stands on the southern portion of the sub-basin measured extremely high in $\widehat{SOC}$, which is consistent with our model findings. Downstream, the north facing slopes hold lower $\widehat{SOC}$, including some densely forested regions. Generally, for both maps in Figure 29 we observe higher relative $\widehat{SOC}$ on shallow ridge peaks compared to the steep ravine sides. Though this behavior was dominated by the effect of the southern region of the preserve almost entirely taken up by an old, 71-year-old, stand. One consistent behavior across both maps is a reduction in $\widehat{SOC}$ along drainage features such as troughs and streams. In those areas, expect a low canopy and understory cover, and relative slope which would indicate a higher predicted $\widehat{SOC}$ considering our one-dimensional

relationships. As such, these maps highlight the multi-dimensionality of the SOC system in this forest and the necessity to consider all variables simultaneously to make ecologically representative predictions.

5.4 Limitations & Future Work

5.4.1 Limitations

Sample Size & Diversity

The timeline of this study did not allow ample opportunity to collect the wealth of soil samples necessary to entirely characterize soil organic carbon and its relationships with field-measurable variables. As a result, many of our findings prior to model development were not ecologically representative. Additionally, due to the vegetation plot layout and movement limitations in a steep dense forest, the samples were not properly diverse to evaluate relationships with many variables. These included overstory cover which was biased to be very high, and aspect, which carried high eastern and western bias. With a higher sample size in more diverse regions of the forest, we would expect more representative results.

Field Data Collection

Due to the complexity of forest systems and an ambitious timeline, the field data collection of this study was not consistent. A majority of field soil samples (12 of 14) had their O- and A-horizons stored together and separated in lab, while 2 of 14 were separated in field. This inconsistency likely had a small effect on overall model results, as measured values for each were comparable and dried samples were compared by eye in lab for significant discrepancies. Considering these lab precautions, all samples were considered together as if they were collected equivalently.

Additionally, field data collection for O-horizon depth varied from sample site to sample site. This is due to the subjectivity of the boundaries of shallow soil horizons. Samples were collected during the winter in heavy rain, and field delineation of the O-horizon was frequently difficult. This inconsistency had no effect on model construction or DSM construction as O-horizon depth is not used in the calculation of soil organic carbon storage. This may have caused any results associated with O-horizon depth to be not reliable beyond the confines of this study.

Model Evaluation & Optimization

There are many methods to evaluate random forest model performance. One common method not utilized in this study due to time constraints is the out-of-bag error. This method calculates error by testing the data points not utilized in model construction against model predictions. This is widely considered a quality tool to evaluate random forest model performance but is not utilized by every study. This work is built upon previous research that used R^2 and RMSE calculated using leave-one-out cross validation to evaluate model performance. In similar future work involving predictive SOC modelling, we recommend considering out-of-bag error as an additional tool to evaluate model performance.

Additionally, during model optimization, this study did not take precaution to overfitting. As such, the final model was likely overfit with a high coefficient of determination as well as high residual error. To combat this, we recommend using error estimation as a tool to select optimal parameters. This may result in a lower explained variance, but an overall more ecologically representative model.

5.4.2 Future Work

Future studies that seek to model SOC variation and construct a DSM on this spatial scale would benefit by considering additional field-measurable variables. In particular, those that are able to. be continuously sampled such as NDVI. Building a robust set of continuous predictor variables would allow the resulting DSM to be able to provide more significant ecological representation. Additionally, our findings highlight a need for additional consideration for overstory cover and understory cover and their relationship with SOC in forests. There are few studies who have established these relationships, particularly with modern model analysis.

Chapter 6: Conclusion

Soil organic carbon (SOC) in forests represents an exceptional opportunity for in depth and comprehensive modelling due to is enormous spatial heterogeneity, general complexity, and vulnerability in a warming climate (Rodrigo-Comino et al., 2020; Smith, 2012; Carpenter et al., 2014). This study sought to build upon the immense wealth of scientific research on these SOC dynamics by constructing a small-scale multivariate organic-horizon SOC model using only variables that were measurable in the field. The subject area of this study was a single watershed basin in Ellsworth Creek Preserve, an experimental nature preserve near Willapa Bay, Washington. Using field gathered data on SOC bolstered with previously collected SOC and ecological data, we built a predictive model using random forest regression. Despite our limited sample size, the model showed a prediction capability that compares to other similar studies with a strong ability to explain variance but a moderately high error (Matinfar et al., 2021). Using this predictive SOC model, we constructed a digital soil map and considered predicted SOC hot- and cold-spots.

Despite limited sample size, we confirmed previous findings on the relationship between SOC and timber harvest, in particular the recovery of SOC as the forest regrows. Our model predicted that SOC stores reduce dramatically following timber harvest and continue falling until the stand reaches approximately 50-years-old. This confirmation only continues to highlight the importance of careful forest management in the carbon-dense forests of the Pacific Northwest. Additionally, our model highlighted the multi-dimensionality of O-horizon SOC and the capability of machine learning algorithms to characterize the associated relationships using limited data. Many individual relationships between SOC and the predictor variables were not significant or ecologically representative. But, when brought together into the multivariate model, SOC predictions began to show expected ecological relationships.

This study serves as a conceptual proof of concept for developing small-scale O-horizon SOC models in forests. With this toolset, forest managers would be able to better preserve the unstable SOC found on the forest floor. This would, in-turn, reduce the amount of carbon entering the atmosphere and reduce the impact of forest management on climate change. The great forests of the Pacific Northwest have stored vast amounts of carbon for generations and will continue to do so for many more to come with proper management.

References

Antos, J. A., Halpern, C. B., Miller, R. E., Jr..Cromack, K., & Halaj, M. G. (2003). Temporal and spatial changes in soil carbon and nitrogen after clearcutting and burning of an old-growth Douglas-fir forest. *Res. Pap. PNW-RP-552. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 19 p*, *552*. https://doi.org/10.2737/PNW-RP-552

Attiwill, P. M., & Adams, M. A. (1993). Nutrient cycling in forests. *New Phytologist*, *124*(4), 561–582. https://doi.org/10.1111/j.1469-8137.1993.tb03847.x

Bai, S. G., Jiao, Y., Yang, W. Z., Gu, P., Yang, J., & Liu, L. J. (2017). Review of progress in soil inorganic carbon research. *IOP Conference Series: Earth and Environmental Science*, *100*(1), 012129. https://doi.org/10.1088/1755-1315/100/1/012129

Barton, C. D. (2002). Clay Minerals. *In: Rattan Lal, Comp., Ed. Encyclopedia of Soil Science. New York, New York: Marcel Dekker: 187-192.* https://www.fs.usda.gov/research/treesearch/7016

Batjes, N. h., & Sombroek, W. g. (1997). Possibilities for carbon sequestration in tropical and subtropical soils. *Global Change Biology*, *3*(2), 161–173. https://doi.org/10.1046/j.1365-2486.1997.00062.x

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, *5*, 180214. https://doi.org/10.1038/sdata.2018.214

Bengough, A. G., & Mullins, C. E. (1990). Mechanical impedance to root growth: A review of experimental techniques and root growth responses. *Journal of Soil Science*, *41*(3), 341–358. https://doi.org/10.1111/j.1365-2389.1990.tb00070.x

Bhardwaj, D. R., Banday, M., Pala, N. A., & Rajput, B. S. (2016). Variation of biomass and carbon pool with NDVI and altitude in sub-tropical forests of northwestern Himalaya. *Environmental Monitoring and Assessment*, *188*(11), 635. https://doi.org/10.1007/s10661-016-5626-3

Binkley, D. (1984). Does forest removal increase rates of decomposition and nitrogen release? *Forest Ecology and Management*, *8*(3), 229–233. https://doi.org/10.1016/0378-1127(84)90055-0

Bittelli, M., Campbell, G. S., & Tomei, F. (2015). Transpiration and Plant–Water Relations. In M. Bittelli, G. S. Campbell, & F. Tomei (Eds.), *Soil Physics with Python: Transport in the Soil–Plant–Atmosphere System* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199683093.003.0014

Boča, A., Van Miegroet, H., & Gruselle, M.-C. (2014). Forest Overstory Effect on Soil Organic Carbon Storage: A Meta-analysis. *Soil Science Society of America Journal*, *78*(S1), S35–S47. https://doi.org/10.2136/sssaj2013.08.0332nafsc

Bockheim, J. G., Gennadiyev, A. N., Hammer, R. D., & Tandarich, J. P. (2005). Historical development of key concepts in pedology. *Geoderma*, *124*(1), 23–36. https://doi.org/10.1016/j.geoderma.2004.03.004

Boettinger, J. L., Howell, D. W., Moore, A. C., Hartemink, A. E., & Kienast-Brown, S. (2010). *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. Springer Science & Business Media.

Bossio, D. A., Cook-Patton, S. C., Ellis, P. W., Fargione, J., Sanderman, J., Smith, P., Wood, S., Zomer, R. J., von Unger, M., Emmer, I. M., & Griscom, B. W. (2020). The role of soil carbon in natural climate solutions. *Nature Sustainability*, *3*(5), Article 5. https://doi.org/10.1038/s41893-020-0491-z

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis.

Broadbent, F. E. (1965). Organic Matter. In *Methods of Soil Analysis* (pp. 1397–1400). John Wiley & Sons, Ltd. https://doi.org/10.2134/agronmonogr9.2.c41

Burke, I. C., Yonker, C. M., Parton, W. J., Cole, C. V., Flach, K., & Schimel, D. S. (1989). Texture, Climate, and Cultivation Effects on Soil Organic Matter Content in U.S. Grassland Soils. *Soil Science Society of America Journal*, *53*(3), 800–805. https://doi.org/10.2136/sssaj1989.03615995005300030029x

Cahoon, S. M. P., Sullivan, P. F., Shaver, G. R., Welker, J. M., & Post, E. (2012). Interactions among shrub cover and the soil microclimate may determine future Arctic carbon budgets. *Ecology Letters*, *15*(12), 1415–1422. https://doi.org/10.1111/j.1461-0248.2012.01865.x

Carpenter, D. N., Bockheim, J. G., & Reich, P. F. (2014). Soils of temperate rainforests of the North American Pacific Coast. *Geoderma*, *230–231*, 250–264. https://doi.org/10.1016/j.geoderma.2014.04.023

Carvalhais, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mu, M., Saatchi, S., Santoro, M., Thurner, M., Weber, U., Ahrens, B., Beer, C., Cescatti, A., Randerson, J. T., & Reichstein, M. (2014). Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature*, *514*(7521), Article 7521. https://doi.org/10.1038/nature13731

Case, M. J., Ettinger, A. K., & Pradhan, K. (2023). Forest restoration thinning accelerates development of old-growth characteristics in the coastal Pacific Northwest, USA. *Conservation Science and Practice*, *n/a*(n/a), e13004. https://doi.org/10.1111/csp2.13004

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chamberlain, C. P., Kane, V. R., & Case, M. J. (2021). Accelerating the development of structural complexity: Lidar analysis supports restoration as a tool in coastal Pacific Northwest forests. *Forest Ecology and Management*, *500*, 119641. https://doi.org/10.1016/j.foreco.2021.119641

Chan, K. y. (2001). Soil particulate organic carbon under different land use and management. *Soil Use and Management*, *17*(4), 217–221. https://doi.org/10.1111/j.1475-2743.2001.tb00030.x

Chen, H. Y. H., & Shrestha, B. M. (2012). Stand age, fire and clearcutting affect soil organic carbon and aggregation of mineral soils in boreal forests. *Soil Biology and Biochemistry*, *50*, 149–157. https://doi.org/10.1016/j.soilbio.2012.03.014

Cheng, H., Garrick, D. J., & Fernando, R. L. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of Animal Science and Biotechnology*, *8*, 38. https://doi.org/10.1186/s40104-017-0164-6

Chilès, J.-P., & Desassis, N. (2018). Fifty Years of Kriging. In B. S. Daya Sagar, Q. Cheng, & F. Agterberg (Eds.), *Handbook of Mathematical Geosciences: Fifty Years of IAMG* (pp. 589–612). Springer International Publishing. https://doi.org/10.1007/978-3-319-78999-6_29

Clarke, N., Gundersen, P., Jönsson-Belyazid, U., Kjønaas, O. J., Persson, T., Sigurdsson, B. D., Stupak, I., & Vesterdal, L. (2015). Influence of different tree-harvesting intensities on forest soil

carbon stocks in boreal and northern temperate forest ecosystems. *Forest Ecology and Management*, *351*, 9–19. https://doi.org/10.1016/j.foreco.2015.04.034

Cole, D. W. (1995). Soil nutrient supply in natural and managed forests. *Plant and Soil*, *168*(1), 43–53. https://doi.org/10.1007/BF00029312

Comber, N. M. (1938). Pedology. *Science Progress (1933-)*, *33*(129), 106–110.

Cotrufo, M. F., Soong, J. L., Horton, A. J., Campbell, E. E., Haddix, M. L., Wall, D. H., & Parton, W. J. (2015). Formation of soil organic matter via biochemical and physical pathways of litter mass loss. *Nature Geoscience*, *8*(10), Article 10. https://doi.org/10.1038/ngeo2520

Covington, W. W. (1981). Changes in Forest Floor Organic Matter and Nutrient Content Following Clear Cutting in Northern Hardwoods. *Ecology*, *62*(1), 41–48. https://doi.org/10.2307/1936666

Cressie, N. (1988). Spatial prediction and ordinary kriging. *Mathematical Geology*, *20*(4), 405–421. https://doi.org/10.1007/BF00892986

Deng, W., Wang, X., Hu, H., Zhu, M., Chen, J., Zhang, S., Cheng, C., Zhu, Z., Wu, C., & Zhu, L. (2022). Variation Characteristics of Soil Organic Carbon Storage and Fractions with Stand Age in North Subtropical Quercus acutissima Carruth. Forest in China. *Forests*, *13*(10), Article 10. https://doi.org/10.3390/f13101649

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3

Duarte, I. M. R., Rodrigues, C. M. G., & Pinho, A. B. (2018). Classification of Soils. In P. T. Bobrowsky & B. Marker (Eds.), *Encyclopedia of Engineering Geology* (pp. 125–133). Springer International Publishing. https://doi.org/10.1007/978-3-319-73568-9_52

Ebermayer, E. (1876). *Die gesammte Lehre der Waldstreu mit Rücksicht auf die chemische Statik des Waldbaues. Unter Zugrundlegung der in den Königl. Staatsforsten Bayerns angestellten Untersuchungen*. Springer. https://doi.org/10.1007/978-3-642-91491-1

Fang, C., Smith, P., Moncrieff, J. B., & Smith, J. U. (2005). Similar response of labile and resistant soil organic matter pools to changes in temperature. *Nature*, *433*(7021), Article 7021. https://doi.org/10.1038/nature03138

Finér, L., Ohashi, M., Noguchi, K., & Hirano, Y. (2011). Factors causing variation in fine root biomass in forest ecosystems. *Forest Ecology and Management*, *261*(2), 265–277. https://doi.org/10.1016/j.foreco.2010.10.016

Flaig, W., Beutelspacher, H., & Rietz, E. (1975). Chemical Composition and Physical Properties of Humic Substances. In J. E. Gieseking (Ed.), *Soil Components: Vol. 1: Organic Components* (pp. 1–211). Springer. https://doi.org/10.1007/978-3-642-65915-7_1

Foster, N., & Bhatti, J. (2005). Forest Ecosystems: Nutrient Cycling. In R. Lal, *Encyclopedia of Soil Science, Second Edition*. CRC Press. https://doi.org/10.1201/NOE0849338304.ch145

Franzmeier, D. P., Pedersen, E. J., Longwell, T. J., Byrne, J. G., & Losche, C. K. (1969). Properties of Some Soils in the Cumberland Plateau as Related to Slope Aspect and Position. *Soil Science Society of America Journal*, *33*(5), 755–761. https://doi.org/10.2136/sssaj1969.03615995003300050037x

Gardner, W. R., & Ehlig, C. F. (1963). The influence of soil water on transpiration by plants. *Journal of Geophysical Research (1896-1977)*, *68*(20), 5719–5724. https://doi.org/10.1029/JZ068i020p05719

Genuer, R., & Poggi, J.-M. (2020). *Random Forests with R*. Springer International Publishing. https://doi.org/10.1007/978-3-030-56485-8

Gotelli, N. J., & Ellison, A. M. (2012). *A Primer of Ecological Statistics* (Second Edition). Oxford University Press.

Govindasamy, P., Liu, R., Provin, T., Rajan, N., Hons, F., Mowrer, J., & Bagavathiannan, M. (2021). Soil carbon improvement under long-term (36 years) no-till sorghum production in a sub-tropical environment. *Soil Use and Management*, *37*(1), 37–48. https://doi.org/10.1111/sum.12636

Gower, S. T. (2003). Patterns and Mechanisms of the Forest Carbon Cycle. *Annual Review of Environment and Resources*, *28*(1), 169–204. https://doi.org/10.1146/annurev.energy.28.050302.105515

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma*, *146*(1), 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Grunwald, S., & Lamsal, S. (2006). Emerging geographic information technologies and soil information systems. In *Environmental Soil-Landscape: Geographic Information Technologies and Pedometrics* (pp. 127–154). https://doi.org/10.1201/9781420028188.sec2

Hall, G. F. (1983). Chapter 5—Pedology and Geomorphology. In L. P. Wilding, N. E. Smeck, & G. F. Hall (Eds.), *Developments in Soil Science* (Vol. 11, pp. 117–140). Elsevier. https://doi.org/10.1016/S0166-2481(08)70600-7

Han, L., Sun, K., Jin, J., & Xing, B. (2016). Some concepts of soil organic carbon characteristics and mineral interaction from a review of literature. *Soil Biology and Biochemistry*, *94*, 107–121. https://doi.org/10.1016/j.soilbio.2015.11.023

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array

programming with NumPy. *Nature*, *585*(7825), Article 7825. https://doi.org/10.1038/s41586-020-2649-2

Hartemink, A. E., Zhang, Y., Bockheim, J. G., Curi, N., Silva, S. H. G., Grauer-Gray, J., Lowe, D. J., & Krasilnikov, P. (2020). Chapter Three - Soil horizon variation: A review. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 160, pp. 125–185). Academic Press. https://doi.org/10.1016/bs.agron.2019.10.003

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hilgard, E. W. (1882). *A Report on the Relations of Soil to Climate*. Weather Bureau.

Hillel, D. (1998). *Environmental Soil Physics: Fundamentals, Applications, and Environmental Considerations* (1st edition). Academic Press.

Homann, P. S., Kapchinske, J. S., & Boyce, A. (2007). Relations of mineral-soil C and N to climate and texture: Regional differences within the conterminous USA. *Biogeochemistry*, *85*(3), 303–316. https://doi.org/10.1007/s10533-007-9139-6

Howard, D. M., & Howard, P. J. A. (1993). Relationships between co2 evolution, moisture content and temperature for a range of soil types. *Soil Biology and Biochemistry*, *25*(11), 1537–1546. https://doi.org/10.1016/0038-0717(93)90008-Y

Howard, P. J. A. (1965). The Carbon-Organic Matter Factor in Various Soil Types. *Oikos*, *15*(2), 229–236. https://doi.org/10.2307/3565121

Jackson, R. B., Lajtha, K., Crow, S. E., Hugelius, G., Kramer, M. G., & Piñeiro, G. (2017). The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls. *Annual Review of Ecology, Evolution, and Systematics*, *48*(1), 419–445. https://doi.org/10.1146/annurev-ecolsys-112414-054234

Jakšić, S., Ninkov, J., Milić, S., Vasin, J., Živanov, M., Jakšić, D., & Komlen, V. (2021). Influence of Slope Gradient and Aspect on Soil Organic Carbon Content in the Region of Niš, Serbia. *Sustainability*, *13*(15), Article 15. https://doi.org/10.3390/su13158332

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer New York.

James, J., & Harrison, R. (2016). The Effect of Harvest on Forest Soil Carbon: A Meta-Analysis. *Forests*, *7*(12), Article 12. https://doi.org/10.3390/f7120308

Jenny, H. (1994). *Factors of Soil Formation: A System of Quantitative Pedology*. Courier Corporation.

Jiang, H. (2022). *Machine Learning Fundamentals: A Concise Introduction* (New edition). Cambridge University Press.

Jobbágy, E. G., & Jackson, R. B. (2000). The Vertical Distribution of Soil Organic Carbon and Its Relation to Climate and Vegetation. *Ecological Applications*, *10*(2), 423–436. https://doi.org/10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2

Johnson, D. W., Cole, D. W., Bledsoe, C. S., Cromack, K., Gessel, S. P., Grier, C. C., & Richards, B. N. (1982). *Nutrient Cycling in Forests of the Pacific Northwest*. 47.

Keenan, T. F., Hollinger, D. Y., Bohrer, G., Dragoni, D., Munger, J. W., Schmid, H. P., & Richardson, A. D. (2013). Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise. *Nature*, *499*(7458), Article 7458. https://doi.org/10.1038/nature12291

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, *81*, 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Killops, S. D., & Killops, V. J. (2013). *Introduction to Organic Geochemistry*. John Wiley & Sons.

Kononova, M. M. (2013). *Soil Organic Matter: Its Nature, Its Role in Soil Formation and in Soil Fertility*. Elsevier.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics*, *10*(1), 1–11. https://doi.org/10.1080/00401706.1968.10490530

Lal, R., Monger, C., Nave, L., & Smith, P. (2021). The role of soil in regulation of climate. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1834), 20210084. https://doi.org/10.1098/rstb.2021.0084

Lalnunzira, C., Brearley, F. Q., & Tripathi, S. K. (2019). Root growth dynamics during recovery of tropical mountain forest in North-east India. *Journal of Mountain Science*, *16*(10), 2335–2347. https://doi.org/10.1007/s11629-018-5303-9

Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma*, *352*, 395–413. https://doi.org/10.1016/j.geoderma.2019.05.031

Lehmann, J., & Schroth, G. (2002). Nutrient leaching. In G. Schroth & F. L. Sinclair (Eds.), *Trees, crops and soil fertility: Concepts and research methods* (1st ed., pp. 151–166). CABI Publishing. https://doi.org/10.1079/9780851995939.0151

Lenton, T. M., & Huntingford, C. (2003). Global terrestrial carbon storage and uncertainties in its temperature sensitivity examined with a simple model. *Global Change Biology*, *9*(10), 1333–1352. https://doi.org/10.1046/j.1365-2486.2003.00674.x

Li, T., Zhang, H., Wang, X., Cheng, S., Fang, H., Liu, G., & Yuan, W. (2019). Soil erosion affects variations of soil organic carbon and soil respiration along a slope in Northeast China. *Ecological Processes*, *8*(1), 28. https://doi.org/10.1186/s13717-019-0184-6

Liu, Y., Liu, S., Wang, J., Zhu, X., Zhang, Y., & Liu, X. (2014). Variation in soil respiration under the tree canopy in a temperate mixed forest, central China, under different soil water conditions. *Ecological Research*, *29*(2), 133–142. https://doi.org/10.1007/s11284-013-1110-5

Liu, Y., Yao, X., & Higuchi, T. (2000). Evolutionary ensembles with negative correlation learning. *IEEE Transactions on Evolutionary Computation*, *4*(4), 380–387. https://doi.org/10.1109/4235.887237

Lozano-García, B., Parras-Alcántara, L., & Brevik, E. C. (2016). Impact of topographic aspect and vegetation (native and reforested areas) on soil organic carbon and nitrogen budgets in Mediterranean natural areas. *Science of The Total Environment*, *544*, 963–970. https://doi.org/10.1016/j.scitotenv.2015.12.022

Macko, S. A., & Estep, M. L. F. (1984). Microbial alteration of stable nitrogen and carbon isotopic compositions of organic matter. *Organic Geochemistry*, *6*, 787–790. https://doi.org/10.1016/0146-6380(84)90100-1

Maraseni, T. N., & Pandey, S. S. (2014). Can vegetation types work as an indicator of soil organic carbon? An insight from native vegetations in Nepal. *Ecological Indicators*, *46*, 315–322. https://doi.org/10.1016/j.ecolind.2014.06.038

Matinfar, H. R., Maghsodi, Z., Mousavi, S. R., & Rahmani, A. (2021). Evaluation and Prediction of Topsoil organic carbon using Machine learning and hybrid models at a Field-scale. *CATENA*, *202*, 105258. https://doi.org/10.1016/j.catena.2021.105258

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

McCarthy, D. R., & Brown, K. J. (2006). Soil respiration responses to topography, canopy cover, and prescribed burning in an oak-hickory forest in southeastern Ohio. *Forest Ecology and Management*, *237*(1), 94–102. https://doi.org/10.1016/j.foreco.2006.09.030

McCully, M. E. (1995). Water efflux from the surface of field-grown grass roots. Observations by cryo-scanning electron microscopy. *Physiologia Plantarum*, *95*(2), 217–224. https://doi.org/10.1111/j.1399-3054.1995.tb00830.x

McLauchlan, K. K. (2006). Effects of soil texture on soil carbon and nitrogen dynamics after cessation of agriculture. *Geoderma*, *136*(1), 289–299. https://doi.org/10.1016/j.geoderma.2006.03.053

McNicol, G., Bulmer, C., D'Amore, D., Sanborn, P., Saunders, S., Giesbrecht, I., Arriola, S. G., Bidlack, A., Butman, D., & Buma, B. (2019). Large, climate-sensitive soil carbon stocks mapped with pedology-informed machine learning in the North Pacific coastal temperate rainforest. *Environmental Research Letters*, *14*(1), 014004. https://doi.org/10.1088/1748-9326/aaed52

Meyer, N., Welp, G., & Amelung, W. (2018). The Temperature Sensitivity (Q10) of Soil Respiration: Controlling Factors and Spatial Prediction at Regional Scale Based on Environmental Soil Classes. *Global Biogeochemical Cycles*, *32*(2), 306–323. https://doi.org/10.1002/2017GB005644

Minasny, B., & McBratney, Alex. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Minx, J. C., Lamb, W. F., Callaghan, M. W., Fuss, S., Hilaire, J., Creutzig, F., Amann, T., Beringer, T., Garcia, W. de O., Hartmann, J., Khanna, T., Lenzi, D., Luderer, G., Nemet, G. F., Rogelj, J., Smith, P., Vicente, J. L. V., Wilcox, J., & Dominguez, M. del M. Z. (2018). Negative emissions—Part 1: Research landscape and synthesis. *Environmental Research Letters*, *13*(6), 063001. https://doi.org/10.1088/1748-9326/aabf9b

Mote, P. W., & Salathé, E. P. (2010). Future climate in the Pacific Northwest. *Climatic Change*, *102*(1), 29–50. https://doi.org/10.1007/s10584-010-9848-z

Nandi, A., & Pal, A. K. (2022). The Evolution of Machine Learning. In A. Nandi & A. K. Pal (Eds.), *Interpreting Machine Learning Models: Learn Model Interpretability and Explainability Methods* (pp. 1–14). Apress. https://doi.org/10.1007/978-1-4842-7802-4_1

Nciizah, A., & Wakindiki, I. (2012). Particulate organic matter, soil texture and mineralogy relations in some Eastern Cape ecotopes in South Africa. *South African Journal of Plant and Soil*, *29*(1), 39–46. https://doi.org/10.1080/02571862.2012.688882

Nelson, D. W., & Sommers, L. E. (1996). Total Carbon, Organic Carbon, and Organic Matter. In *Methods of Soil Analysis* (pp. 961–1010). John Wiley & Sons, Ltd. https://doi.org/10.2136/sssabookser5.3.c34

Nichols, J. D. (1984). Relation of Organic Carbon to Soil Properties and Climate in the Southern Great Plains. *Soil Science Society of America Journal*, *48*(6), 1382–1384. https://doi.org/10.2136/sssaj1984.03615995004800060037x

Olson, K. R. (2010). Impacts of Tillage, Slope, and Erosion on Soil Organic Carbon Retention. *Soil Science*, *175*(11), 562. https://doi.org/10.1097/SS.0b013e3181fa2837

Onwuka, B. (2018). Effects of Soil Temperature on Some Soil Properties and Plant Growth. *Advances in Plants & Agriculture Research*, *8*(1). https://doi.org/10.15406/apar.2018.08.00288

Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, *27*(4), 286–306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. *SOIL*, *6*(1), 35–52. https://doi.org/10.5194/soil-6-35-2020

Palmer, M., Kuegler, O., & Christensen, G. (2019). Washington's forest resources, 2007–2016: 10-year Forest Inventory and Analysis report. *Gen. Tech. Rep. PNW-GTR-976. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 79 p.*, *976*. https://doi.org/10.2737/PNW-GTR-976

Passioura, J. B. (2002). Soil conditions and plant growth. *Plant, Cell & Environment*, *25*(2), 311–318. https://doi.org/10.1046/j.0016-8025.2001.00802.x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*.

Peng Xinhua, Z. B. and Z. Q. (2013). A REVIEW ON RELATIONSHIP BETWEEN SOIL ORGANIC CARBON POOLS AND SOIL STRUCTURE STABILITY. *ACTA PEDOLOGICA SINICA*, *41*(4), 618–623. https://doi.org/10.11766/trxb200308110419

Penn, C. J., & Camberato, J. J. (2019). A Critical Review on Soil Chemical Processes that Control How Soil pH Affects Phosphorus Availability to Plants. *Agriculture*, *9*(6), Article 6. https://doi.org/10.3390/agriculture9060120

Prescott, C. E. (1997). Effects of clearcutting and alternative silvicultural systems on rates of decomposition and nitrogen mineralization in a coastal montane coniferous forest. *Forest Ecology and Management*, *95*(3), 253–260. https://doi.org/10.1016/S0378-1127(97)00027-3

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems* (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Rey, A., Petsikos, C., Jarvis, P. G., & Grace, J. (2005). Effect of temperature and moisture on rates of carbon mineralization in a Mediterranean oak forest soil under controlled and field conditions. *European Journal of Soil Science*, *56*(5), 589–599. https://doi.org/10.1111/j.1365-2389.2004.00699.x

Rice, C. W., Pires, C. B., & Sarto, M. V. M. (2023). Carbon cycle in soils: Dynamics and management. In M. J. Goss & M. Oliver (Eds.), *Encyclopedia of Soils in the Environment (Second Edition)* (pp. 219–227). Academic Press. https://doi.org/10.1016/B978-0-12-822974-3.00154-3

Rodrigo-Comino, J., López-Vicente, M., Kumar, V., Rodríguez-Seijo, A., Valkó, O., Rojas, C., Pourghasemi, H. R., Salvati, L., Bakr, N., Vaudour, E., Brevik, E. C., Radziemska, M., Pulido, M., Di Prima, S., Dondini, M., de Vries, W., Santos, E. S., Mendonça-Santos, M. de L., Yu, Y., & Panagos, P. (2020). Soil Science Challenges in a New Era: A Transdisciplinary Overview of Relevant Topics. *Air, Soil and Water Research*, *13*, 1178622120977491. https://doi.org/10.1177/1178622120977491

Rosenblatt, F. (1957). *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Cornell Aeronautical Laboratory.

Ruiz-Colmenero, M., Bienes, R., Eldridge, D. J., & Marques, M. J. (2013). Vegetation cover reduces erosion and enhances soil organic carbon in a vineyard in the central Spain. *CATENA*, *104*, 153–160. https://doi.org/10.1016/j.catena.2012.11.007

Saimun, Md. S. R., Karim, Md. R., Sultana, F., & Arfin-Khan, M. A. S. (2021). Multiple drivers of tree and soil carbon stock in the tropical forest ecosystems of Bangladesh. *Trees, Forests and People*, *5*, 100108. https://doi.org/10.1016/j.tfp.2021.100108

Schlesinger, W. H., & Bernhardt, E. S. (2013). *Biogeochemistry: An Analysis of Global Change*. Academic Press.

Schnitzer, M. (2015). Organic Matter Characterization. In A. L. Page (Ed.), *Agronomy Monographs* (pp. 581–594). American Society of Agronomy, Soil Science Society of America. https://doi.org/10.2134/agronmonogr9.2.2ed.c30

Schuur, E. A. G. (2001). The Effect of Water on Decomposition Dynamics in Mesic to Wet Hawaiian Montane Forests. *Ecosystems*, *4*(3), 259–273. https://doi.org/10.1007/s10021-001-0008-1

Scull, P., Franklin, J., Chadwick, O. A., & McArthur, D. (2003). Predictive soil mapping: A review. *Progress in Physical Geography: Earth and Environment*, *27*(2), 171–197. https://doi.org/10.1191/0309133303pp366ra

Sedjo, R., & Sohngen, B. (2012). Carbon Sequestration in Forests and Soils. *Annual Review of Resource Economics*, *4*, 127–144.

Seyfried, G. S., Canham, C. D., Dalling, J. W., & Yang, W. H. (2021). The effects of tree-mycorrhizal type on soil organic matter properties from neighborhood to watershed scales. *Soil Biology and Biochemistry*, *161*, 108385. https://doi.org/10.1016/j.soilbio.2021.108385

Simonson, R. W. (1968). Concept Of Soil. In A. G. Norman (Ed.), *Advances in Agronomy* (Vol. 20, pp. 1–47). Academic Press. https://doi.org/10.1016/S0065-2113(08)60853-6

Smith, P. (2012). Soils and climate change. *Current Opinion in Environmental Sustainability*, *4*(5), 539–544. https://doi.org/10.1016/j.cosust.2012.06.005

Solomatova, E., & Sidorova, V. (2008). Spatial variability of forest litters in bilberry spruce forests of Fennoscandia. *Soil Geography and Geostatistics*, *26*.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Sreenivas, K., Sujatha, G., Sudhir, K., Kiran, D. V., Fyzee, M. A., Ravisankar, T., & Dadhwal, V. K. (2014). Spatial Assessment of Soil Organic Carbon Density Through Random Forests Based Imputation. *Journal of the Indian Society of Remote Sensing*, *42*(3), 577–587. https://doi.org/10.1007/s12524-013-0332-x

Stutter, M., Lumsdon, D., Billett, M., Low, D., & Deeks, L. (2009). Spatial Variability in Properties Affecting Organic Horizon Carbon Storage in Upland Soils. *Soil Science Society of America Journal - SSSAJ*, *73*. https://doi.org/10.2136/sssaj2008.0413

Suleymanov, A., Gabbasova, I., Komissarov, M., Suleymanov, R., Garipov, T., Tuktarova, I., & Belan, L. (2023). Random Forest Modeling of Soil Properties in Saline Semi-Arid Areas. *Agriculture*, *13*(5), Article 5. https://doi.org/10.3390/agriculture13050976

Szafranek-Nakonieczna, A., & Stêpniewska, Z. (2014). Aerobic and Anaerobic Respiration in Profiles of Polesie Lubelskie Peatlands. *International Agrophysics*, *28*, 219–229. https://doi.org/10.2478/intag-2014-0011

Szatmári, G., Pásztor, L., & Heuvelink, G. B. M. (2021). Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. *Geoderma*, *403*, 115356. https://doi.org/10.1016/j.geoderma.2021.115356

The Nature Conservancy. (2020, January 7). *Restoring Old Growth on the Coast—Using Science to Measure Our Success*. The Nature Conservancy in Washington. https://www.washingtonnature.org/fieldnotes/2019/12/18/restoring-old-growth-on-the-coast-using-science-to-measure-our-success

Tsui, C.-C., Chen, Z.-S., & Hsieh, C.-F. (2004). Relationships between soil properties and slope position in a lowland rain forest of southern Taiwan. *Geoderma*, *123*(1), 131–142. https://doi.org/10.1016/j.geoderma.2004.01.031

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Urban, D. L. (2005). Modeling Ecological Processes Across Scales. *Ecology*, *86*(8), 1996–2006. https://doi.org/10.1890/04-0918

U.S. Geological Survey, 2020. 20200224, USGS Lidar Point Cloud WA Olympic Peninsula C2 2017 46123D8302: U.S. Geological Survey.

Vayssières, M. P., Plant, R. E., & Allen-Diaz, B. H. (2000). Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, *11*(5), 679–694. https://doi.org/10.2307/3236575

Wadoux, A. M. J.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, *210*, 103359. https://doi.org/10.1016/j.earscirev.2020.103359

Wander, M. (2004). *3 Soil Organic Matter Fractions and Their Relevance to Soil Function*. https://doi.org/10.1201/9780203496374.ch3

Wang, C., Han, S., Zhou, Y., Zhang, J., Zheng, X., Dai, G., & Li, M.-H. (2016). Fine root growth and contribution to soil carbon in a mixed mature Pinus koraiensis forest. *Plant and Soil*, *400*(1), 275–284. https://doi.org/10.1007/s11104-015-2724-x

Wang, L., Li, Y., Wu, J., An, Z., Suo, L., Ding, J., Li, S., Wei, D., & Jin, L. (2023). Effects of the Rainfall Intensity and Slope Gradient on Soil Erosion and Nitrogen Loss on the Sloping Fields of Miyun Reservoir. *Plants*, *12*(3), 423. https://doi.org/10.3390/plants12030423

Wardle, D., Bardgett, R., Klironomos, J., Setälä, H., Putten, W., & Wall, D. (2004). Ecological Linkages Between Aboveground and Belowground Biota. *Science (New York, N.Y.)*, *304*, 1629–1633. https://doi.org/10.1126/science.1094875

Waring, R. H., & Franklin, J. F. (1979). Evergreen Coniferous Forests of the Pacific Northwest. *Science*, *204*(4400), 1380–1386. https://doi.org/10.1126/science.204.4400.1380

Weil, R., & Brady, N. (2017). *The Nature and Properties of Soils. 15th edition*.

Whitlock, M. C., & Schluter, D. (2008). *The Analysis of Biological Data* (First Edition, 1st). Roberts and Company Publishers.

Whittaker, R. H., & Niering, W. A. (1975). Vegetation of the Santa Catalina Mountains, Arizona. V. Biomass, Production, and Diversity along the Elevation Gradient. *Ecology*, *56*(4), 771–790. https://doi.org/10.2307/1936291

Winkler, J. P., Cherry, R. S., & Schlesinger, W. H. (1996). The Q10 relationship of microbial respiration in a temperate forest soil. *Soil Biology and Biochemistry*, *28*(8), 1067–1072. https://doi.org/10.1016/0038-0717(96)00076-4

Xiong, X., Zhou, G., & Zhang, D. (2020). Soil organic carbon accumulation modes between pioneer and old-growth forest ecosystems. *Journal of Applied Ecology*, *57*(12), 2419–2428. https://doi.org/10.1111/1365-2664.13747

Yanai, R. D., Currie, W. S., & Goodale, C. L. (2003). Soil Carbon Dynamics after Forest Harvest: An Ecosystem Paradigm Reconsidered. *Ecosystems*, *6*(3), 197–212. https://doi.org/10.1007/s10021-002-0206-5

Yi, D., Ahn, J., & Ji, S. (2020). An Effective Optimization Method for Machine Learning Based on ADAM. *Applied Sciences*, *10*(3), Article 3. https://doi.org/10.3390/app10031073

Zhang, D., Hui, D., Luo, Y., & Zhou, G. (2008). Rates of litter decomposition in terrestrial ecosystems: Global patterns and controlling factors. *Journal of Plant Ecology*, *1*(2), 85–93. https://doi.org/10.1093/jpe/rtn002

Zhang, S., Yang, X., Li, D., Li, S., Chen, Z., & Wu, J. (2022). A meta-analysis of understory plant removal impacts on soil properties in forest ecosystems. *Geoderma*, *426*, 116116. https://doi.org/10.1016/j.geoderma.2022.116116

Zhang, Y., & Hartemink, A. E. (2019). Digital mapping of a soil profile. *European Journal of Soil Science*, *70*(1), 27–41. https://doi.org/10.1111/ejss.12699

Zhou, Y., Zhao, X., Guo, X., & Li, Y. (2022). Mapping of soil organic carbon using machine learning models: Combination of optical and radar remote sensing data. *Soil Science Society of America Journal*, *86*(2), 293–310. https://doi.org/10.1002/saj2.20371

Zhu, M., Feng, Q., Qin, Y., Cao, J., Li, H., & Zhao, Y. (2017). Soil organic carbon as functions of slope aspects and soil depths in a semiarid alpine region of Northwest China. *CATENA*, *152*, 94–102. https://doi.org/10.1016/j.catena.2017.01.011