Fall 12-15-2023

Modeling Forest Growth Using Sentinel-2-Derived Variables and Site Data

Peter G. Larson
University of Maine, peter.larson@maine.edu

MODELING FOREST GROWTH USING SENTINEL-2-DERIVED VARIABLES AND SITE DATA

By

Peter G. Larson

B.A. University of Maine, 2001

M.A. University of Alberta, 2003

Ph.D. University of Alberta, 2009

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Ecology and Environmental Sciences)

The Graduate School

The University of Maine

December 2023

Advisory Committee:

Parinaz Rahimzadeh-Bajgiran, Associate Professor of Remote Sensing of Natural Resources, Advisor

Aaron Weiskittel, Professor of Forest Biometrics and Modeling

Michael Premer, Assistant Professor of Forest Management

MODELING FOREST GROWTH USING SENTINEL-2-DERIVED VARIABLES AND SITE DATA

By Peter G. Larson
Thesis Advisor: Dr. Parinaz Rahimzadeh-Bajgiran

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(Ecology and Environmental Sciences)
December 2023

Growing stock volume (GSV) is an important metric for determining economic yield, carbon sequestration and other ecosystem services. GSV has traditionally been estimated *in situ* by measuring individual trees in a stand. This process is slow and expensive, and, as a result, is not a viable means to estimate GSV on a large scale. It is also not feasible in places that are difficult to access and in places that do not have reliable management records. Multispectral optical sensors mounted on satellites are an important technology for monitoring forest resources because they offer the possibility of measuring forest resources quickly and over large areas. In this study, forest potential productivity was estimated by evaluating 65 variables including several remotely sensed optical variables and site and climate data. Optical variables were Sentinel-2 band 3, band 8a, the Normalized Difference Vegetation Index using bands 4 and 5 (NDVI45) and the Sentinel-2 red-edge position index (S2REP). The variables were used as inputs in a random forest machine learning algorithm. The response variable was constructed using the tree height differences estimated using the National Agricultural Imagery Program (NAIP) orthographic imagery data derived from the NAIP 2018 and NAIP 2021 (ΔNAIP) data. This study was conducted in Maine, USA, where 89% of the land is covered by forests and forest product industry is a significant contributor to the state economy. The best-performing final model to estimate forest productivity

(growth), which incorporated Sentinel-2 band 3, the NDVI45, and the S2REP as well as seven site variables, achieved an R^2 value of approximately 0.56.

DEDICATION

This thesis is dedicated to my wife and my parents.

ACKNOWLEDGEMENTS

I would like to thank my wife and parents for their support. I would also like to thank Professor Parinaz Rahimzadeh-Bajgiran for her great investment in time and effort in helping me to achieve my goals.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

APAR	Absorbed Photosynthetically Active Radiation
AVHRR	Advanced Very High Resolution Radiometer
BGI	Biomass Growth Index
CHM	Canopy Height Model
DBH	Diameter at Breast Height
DSM	Digital Soil Model
ESA	European Space Agency
ETR	Electron Transfer Rate
EVI	Enhanced Vegetation Index
fAPAR	Fraction of PAR Absorbed by Vegetation
FPAR	Fraction of Photosynthetically Active Radiation
GSV	Growing Stock Volume
iBGI	improved Biomass Growth Index
LAI	Leaf Area Index
LST	Land Surface Temperature
LUE	Light Use Efficiency (ε)
LUEp	Photosynthetic Light Use Efficiency

ML	Machine Learning
MNDREP	Modified Normalized Difference Red Edge Position Index
MSI	MultiSpectral Instrument
NEE	Net Ecosystem Carbon Dioxide Exchange
NOAA	National Oceanic and Atmosphere Administration
NDVI	Normalized Difference Vegetation Index
NPQ	Non-Photochemical quenching
OLI	Operational Land Imager (aboard LandSat-8)
P	Carbon fixed through photosynthesis
PAR	Photosynthetically Active Radiation
PPFD	Photosynthetic Photon Flux Density
PRI	Photochemical Reflectance Index
PSN_{net}	Daily Net Photosynthesis
PSP	Permanent Sample Plots
PVI	Perpendicular Vegetation Index
R_e	Daytime Ecosystem Respiration
REP	Red Edge Position (index)
RMSE	Root Mean Square Error

S2	Sentinel-2
S2REP	Sentinel-2 Red Edge Position
sPRI	Scaled Photochemical Reflectance Index
TG	Temperature and Greenness
TGM	Temperature and Greenness Model
USGS	United States Geological Survey
VPD	Vapor Pressure Deficit

1 CHAPTER ONE: INTRODUCTION

Human society increasingly relies heavily on wood products despite the proliferation of non-wood alternative materials. As of 2022, the rate of increase in demand for wood worldwide was double that of the increase in population (FAO UN 2022). It is necessary to manage forest resources carefully on a large scale in order to meet the material needs of human society and to preserve natural forests and the ecosystem services they provide. Satellite remote sensing can increase the efficacy of sustainable resource management, so that ecosystem services are not compromised and forest resources remain available. Remote sensing can help increase industrial output and the amount of forest habitat preserved by gathering more information with greater cost-efficiency. The scale of forest resource management is one of the reasons that satellite remote sensing can make a significant contribution to increasing management efficiency. As of 2022, 31% of the terrestrial earth is covered by primary or secondary forest, which equals approximately 4.06 billion hectares or 1,568,000 square miles (FAO UN 2022). This work will examine the application of satellite remote sensing to the problem of determining which forested areas are maximally potentially productive. Productivity is usually measured in terms of biomass per unit area of land per unit time, such as in $kg \cdot ha^{-1} \cdot year^{-1}$. Targeting forest areas of maximum potential productivity for the application of intensive sustainable management techniques will increase production of forest-derived resources and minimize negative impacts to the ecosystem, as the forest will be the most capable of regenerating quickly and effectively.

The concept of *productivity* here is actually *potential productivity*, or the maximum possible volume of biomass that a given area can support given ideal climatic conditions. This is not necessarily the same as the observed productivity of that area. Current biomass does not equal potential productivity. Biomass quantity as well as quality can be increased by human

interventions such as thinning, which directs the plot's resources into a few trees that are likely to thrive, and also by controlling invasive insect, tree, and other plant species (Crow *et al.* 2006). In forest management, Crow *et al.* point out that, due to the volume of forests, even incremental increases can yield large gains. Historically, site index has been the most-used indicator of the potential productivity of a unit of forested land (Weiskittel *et al.* 2011). Site index is determined by measuring the height of the dominant tree in a given area and using a species-specific table to arrive at a potential productivity estimate (Hennigar *et al.* 2017). However, site index is a metric that is expensive and time-consuming to obtain because of the need for direct, in-person measurement of individual trees. Furthermore, site index decreases in accuracy when the stands being measured are multicohort (different ages) or comprised of different. Hennigar *et al.* (2017) explore other ways to gauge productivity, developing a biomass growth index (BGI), which is the asymptote of a model quantifying above-ground dry biomass potential of a given area. The researchers use observations of current above-ground biomass as one of several inputs to the BGI.

1.1 Contribution of Remote Sensing to Vegetation Studies

Plants produce biomass through the process of photosynthesis. Chlorophyll is a pigment that absorbs the light necessary to provide energy for photosynthesis. The amount of chlorophyll present in a plant's leaves has a direct linear relationship with the amount of biomass produced (Fleischer 1934). Chlorophyll is a pigment which absorbs most wavelengths of radiation, particularly red and blue wavelengths, but not the green wavelengths, which chlorophyll reflects, thus appearing 'green' (*Encyclopedia Britannica*, "Chlorophyll"). The pigment's interaction with solar radiation is the basis for both plant physiology and optical multispectral remote sensing of plants, which detects those interactions with solar radiation. The leaves, which are the tree's 'solar panels,' function to expose chlorophyll to sunlight. The chlorophyll absorbs some of the energy

from the sunlight and uses that energy to produce food for the tree, in the form of carbohydrates from atmospheric carbon. There are several types of chlorophyll, with slightly differing chemical formulae. The two most common types in tress are Chlorophyll A ($C_{55}H_{72}MgN_4O_5$) and Chlorophyll B ($C_{55}H_{70}MgN_4O_6$). Productivity and carbon flux are closely related through the physiological processes of respiration and growth. Carbon is taken into the plant, in this case a tree, through respiration and used to store energy for the tree in the form of carbohydrates. The carbohydrates can then be used to generate the mass of the tree. Table 1-1 describes the main components of wood and their relative proportions. The ratio of lignin to cellulose and to hemicellulose is one of the chemical features that distinguishes hardwood from softwood (Fleischer 1934; Hirons and Thomas 2018).

Table 1-1: Chemical composition of wood.

Substance	Approximate Fraction of Wood Mass	Formula
Lignin	26%-34% (softwood); 23%-30% (hardwood)	$C_{18}H_{13}N_3Na_2O_8$
Cellulose	40%-45% (softwood); 38%-49% (hardwood)	$C_6H_{10}O_5$
Hemicellulose	7%-14% (softwood); 19%-26% (hardwood)	$C_5H_8O_4$
Other	varies	

The high carbon content of wood is one of the features that makes wood such an environmentally friendly material. The sequestration of carbon in the built environment and the renewable nature of this resource mean that, if properly managed, wood has a significant role to play in the built environment and the economy of the future.

Satellite remote sensing is a method of natural resource observation and quantification that offers several significant benefits, such as cost-efficiency and high temporal resolution. Satellites such as moderate resolution imaging spectroradiometer (MODIS) and Landsat have a long history of being used to estimate vegetation properties ranging from chlorophyll content to productivity (Running *et al.* 2004). Specific to this project, combined with other site and environmental observations, satellite remote sensing can improve potential productivity estimation accuracy. Rahimzadeh-Bajgiran *et al.* (2020) made progress towards this realization with the development of a new productivity index. They improved upon the BGI by proposing the improved Biomass Growth Index (iBGI), based on the Biomass Growth Index (BGI) model that utilizes only field measurements as inputs to generate a statistical model of the entire forest (Rahimzadeh-Bajgiran *et al.* 2020). Among the improvements that Rahimzadeh-Bajgiran *et al.* (2020) bring are increased temporal resolution, increased accuracy, and the ability to account for variability of site physical characteristics in different zones of the same forest. The improvements are the result of the inclusion of remote sensing data from the European Space Agency's (ESA) Sentinel-2 (SENTINEL-2) mission, which collects multispectral images of the entire earth's surface approximately once every five days (Rahimzadeh-Bajgiran *et al.* 2020). Of particular interest are the red-edge bands, which were found to be useful in predicting potential productivity (Rahimzadeh-Bajgiran *et al.* 2020). Rahimzadeh-Bajgiran *et al.* (2020), for the first time, used several Sentinel-2 spectral vegetation indices, such as the Sentinel-2 red-edge position index (S2REP) to estimate total volume (TV), height (HT) and productivity. This study attempts to improve upon Rahimzadeh-Bajgiran *et al.* (2020) by including the most recent site variables as well as National Agriculture Imagery Program (NAIP) -derived canopy height model (CHM) data,

United States Geological Survey (USGS) digital soil model (DSM), and other data, to estimate potential productivity for parts of the state of Maine, USA.

1.2 Literature Review

There is already a large body of literature describing the uses of satellite remote sensing for agricultural purposes. Contemporary specialists in the subject are fortunate because they are at the cusp of an exciting era when sensors are improving and the large amount of data they generate can be analyzed using machine learning (ML) algorithms, extending the realm of the possible beyond what humans could achieve without computers. In this time of 'bigger and bigger' data, we have only begun to explore what satellite remote sensing and machine learning can do together.

1.2.1 Remote Sensing, Productivity and Vegetation Traits

Foody and Curran (1994) studied using optical satellite remote sensing from the advanced very-high-resolution radiometer (AVHRR) instrument on board the National Oceanic and Atmospheric Administration (NOAA) satellite series (program active 1979-2019) to estimate regeneration in tropical forests. The NOAA satellite series had a coarse 1.1 km spatial resolution, but a relatively frequent temporal resolution. They found that increased tree density is correlated to decreased red reflectance, it has a very mild positive correlation with near-infrared reflectance, and the normalized difference vegetation index (NDVI) is positively correlated with tree density.

MODIS instruments were launched by the National Aeronautics and Space Administration (NASA) aboard the Terra vehicle (1999) and the Aqua vehicle (2002). One of the first studies of optical satellite remote sensing using MODIS was based on mechanistic modeling of radiation with leaf structure. The MOD17 product based on MODIS imagery was developed from

Monteith's (1972) principles of measurement of absorbed photosynthetically active radiation (APAR) and estimations of light use efficiency (LUE) to estimate global productivity (Guyot 1992).

Gitelson *et al.* (2006) found that gross primary production (GPP) has a strong positive correlation with chlorophyll content in maize and soybean where water was not a limiting factor. They developed a model for estimating chlorophyll concentration in those crops solely using remotely sensed data. Carbon dioxide flux, leaf area index (LAI), absorbed photosynthetically active radiation (PAR), leaf chlorophyll content, total canopy chlorophyll content, and spectral reflectance of the crop in the 400-900 nm range were used as model variables. They found that LUE fluctuates throughout the growing season, declining during periods of moisture stress. Their model for insolation is: GPP ($mg/m^2 \cdot s$) = NDVI × sPRI × PAR ($mmol/m^2 \cdot s$), where *sPRI* is the scaled photochemical reflectance index. There is a wide range of dispersion about the best fit, resulting in a high root mean square error (RMSE) for the model. They found that GPP can be estimated from chlorophyll content, but that content varies throughout the day.

Confounding factors can also include physical structure of the canopy, leaf area index (LAI), and soil (Gitelson *et al.* 2006). In terms of optical reflectance, the model becomes: GPP in units of $mg/m^2 \cdot s$ is approximately equal to: $[(R_{NIR}/R_{720\text{-}740}) - 1] \times PAR$ in maize, both irrigated and rainfed (Gitelson *et al.* 2006). To improve the strength of their observations, researchers created vegetation indices (VI) to gain more information about the target. Vegetation indices, also known as *spectral indices* in this thesis (as no non-vegetation spectral indices are considered), are algebraic manipulations of the percentages of reflectance at various bands of the electromagnetic spectrum that can highlight information that may otherwise have been missed.

Using vegetation indices to estimate GPP through vegetation indices continued to be an important goal in the remote sensing academic community. Running *et al.* (2004) sought to quantify net primary production (NPP) as opposed to GPP, which refers to the total amount of mass produced from all the energy photosynthetically absorbed from solar radiation, including carbon dioxide that is respirated out of the plant. NPP is the total biomass produced from photosynthesis, and can be defined in terms of mass as *the net change in carbon dioxide or the net increase in biomass,* as well as energy, as cumulative energy generated by photosynthesis from which the energy expended from respiration has been subtracted. Running *et al.* (2004) remained with the convention of measuring GPP, but with the eventual purpose of deriving NPP from that. However, because the two are positively related, that distinction is not important for this research.

Running *et al.* (2004) stated that the science of estimating productivity based on spectral indices rests on three fundamental principles, which are: (1) NPP is a factor of utilized solar energy (2), vegetation indices, and (3) the difference between actual productivity and potential productivity caused by limiting factors (Running *et al.* 2004). Climatological constraints can be measured more easily, but measuring leaf area is more problematic (Running *et al.* 2004). This constraint is quantified by the LAI. The NDVI is directly related to photosynthetically active radiation, in that $APAR/PAR \approx NDVI$ and $GPP = \varepsilon \times FPAR \times PAR \approx \varepsilon \times NDVI \times PAR$, where *FPAR* is *fraction of photosynthetically active radiation* (Running *et al.* 2004). The value of the conversion efficiency coefficient, ε, is dependent on the biophysical characteristics of each individual species and on environmental factors, such as variations in water and nutrient availability, as well as a high vapor pressure deficit which forces plant stomata to close (Running *et al.* 2004).

Sims *et al.* (2008) pointed out that, while the MOD17 product was, as of 2008, the most important product for the satellite remote sensing of plant productivity, it does have limitations. Among these are the accuracy of the interpolated climate data and of the LUE coefficient for individual plant species (Sims *et al.* 2008). Sims *et al.* (2008) produced a new model replacing the environmental and LUE inputs solely with the enhanced vegetation index (EVI), but this, too, had shortcomings, including reduced accuracy when monitoring evergreens and periods of reduced growth caused by low or high temperatures and high vapor pressure deficits (Sims *et al.* 2008). The researchers added data relating to temperature and stress caused by drought, but kept remotely sensed optical variables as the only other input (Sims *et al.* 2008). The researchers developed a 'temperature and greenness model' (TGM) based on temperature and greenness (TG) using the MODIS green and infrared channels. As a result, the TGM achieved highly accurate results far surpassing the MODIS GPP product, which also relied on many other complicated factors (Sims *et al.* 2008).

Wu *et al.* (2009) used red-edge indices to estimate GPP in wheat, also relating GPP to LUE. They confirmed the previous findings of other researchers that vegetation indices are directly related to LUE and chlorophyll content. The researchers worked towards the same goal as Sims *et al.* (2008) of correlating remotely sensed vegetation indices with mass of chlorophyll production. The researchers studied the red-edge NDVI, the maximum chlorophyll absorption ration index ($MCARI_{710}$), the chlorophyll index of red edge ($CI_{red\ edge}$) and the MERIS Terrestrial Chlorophyll Index (MTCI) (Wu *et al.* 2009). Their findings strengthened the theory that GPP and PAR are proportional to chlorophyll production (Wu *et al.* 2009). Furthermore, the researchers found that Red Edge NDVI and $MCARI_{710}$ were highly corelated with GPP and PAR, with R^2 values of 0.70

and 0.71, respectively, as well as strong correlation between canopy chlorophyll content and LUE (Wu *et al.* 2009).

Donmez *et al.* (2010) used MERIS data from the ENVISAT platform to model NPP, which was measured using a carbon cycle model taking satellite data and meteorological data observed from 50 stations as inputs. They conceptualize LUE as a scaling factor of the fraction of photosynthetically active radiation (FPAR) and PAR (Donmez *et al.* 2010). They built upon the premise of Running *et al.* (2004) that LUE holds the key to quantifying NPP. One of the goals of their project was to use satellite remote sensing to quantify productivity over a smaller spatial resolution for use in ecosystems with high variability of conditions. They found that NPP could be fully accounted for by the addition of solar radiation and water availability (Donmez *et al.* 2010).

Several studies have sought to apply optical remote sensing to productivity estimation in order to study the possibility of carefully controlling nitrogen fertilizer application and to estimate nitrogen mass within the foliage, on the principle that greater nitrogen mass in the leaves denotes greater biomass overall. Sharma *et al.* (2015) studied the use of commercially available active optical agricultural sensors combined with various levels of nitrogen application to predict maize yield in order to determine if the combination of remote sensing and targeted nitrogen application increased yield. The researchers distinguished between red NDVI, which at the time of publication was the standard index for commercial agricultural optical sensors, and red-edge NDVI, which they tested as a novel alternative, as the red-edge is responsive to a wider range of chlorophyll contents (Sharma *et al.* 2015). The researchers found that results varied significantly at some growth stages depending on the brand of sensor used. However, they did not report any significant differences between the effectiveness of the red NDVI and the red-edge NDVI (Sharma *et al.* 2015, 27848).

Kanke *et al.* (2012) investigated the lower sensitivity of the NDVI to chlorophyll relative to the REP index and the responses of the indices to various levels of nitrogen using a commercially available handheld spectrometer. They found that NDVI sensitivity decreased as foliage reached maturity, likely due to increased nitrogen content (Kanke *et al.* 2012). They found that, although promising, the red edge position index technology was not mature enough for exploitation (Kanke 2012).

Bandyopadhyay *et al.* (2017) investigated using the red edge index, recorded using the Hyperion instrument on NASA's Earth Observing-1 satellite, as well as instruments on ESA's Rapid Eye satellite, for the purpose of nitrogen estimation. They found that the modified normalized difference red edge position index was highly correlated with nitrogen mass in vegetation, with a coefficient of determination (in their work written as r^2) of 0.89 (Bandyopadhyay *et al.* 2017).

Bulut and Günlü (2016) sought to assess carbon storage of mixed-species forests. Assessment of carbon storage is, in essence, the same as estimation of GPP. The researchers began by first using the Erdas Imagine software to develop a supervised classification algorithm to distinguish between coniferous stands, broadleaf stands, and non-forest groundcover types using training and testing polygons, which were then verified on the ground by physical visual inspection (Bulut and Günlü 2016). Carbon storage was calculated on the basis of biomass and growing stock volume, which was measured using inventory data from an anonymous forest management source (Bulut and Günlü 2016). The researchers created their model on the basis of tree type and area covered by that type (Bulut and Günlü 2016). Their modeling attempts were not successful in terms of estimating GPP, but the introduction of tree type as a variable may have been an important step.

1.2.2 Sentinel-2 Satellites and Vegetation Studies

The launch of the Sentinel-2 vehicles and their MultiSpectral Instruments (MSI) in 2016 provided a significant new tool for remote sensing scientists. Chrysafis *et al.* (2017) were among the first researchers to explore utilizing this tool for quantifying primary production. They sought to use the MSI to quantify forest growing stock volume. Their work is of particular interest to this thesis as it is the first known paper in which researchers (1) jettison their focus on carbon or nitrogen to focus explicitly on biomass; (2) utilize vegetation indices as their primary means of investigation; and (3) utilize the Sentinel-2 platform. These three features form the basis of this thesis, as their approach embodies several advantages, such as simplicity, economy, and efficiency. The researchers compared the performance of the Sentinel-2 vegetation indices to that of Landsat-8 Operational Land Imager (OLI).

The researchers' study area—Mediterranean forests—is similar to Maine, USA, in its heterogeneity of landscape features, species, and age. The researchers collected field data as their ground-truth variable. Tree volumes, based on the commonly used diameter at breast height (DBH) metric, were collected from 112 square plots 0.1 ha in area (Chrysafis *et al.* 2017). Then, "Linear regression models were developed with individual spectral bands and vegetation indices in order to explore the relationship with Growing stock volume (GSV).

The researchers turned to ML and used the random forest algorithm to fit a model of productivity using remotely sensed optical data as inputs. The researchers found that vegetation indices incorporating red-edge position significantly increased accuracy. (Chrysafis *et al.* 2017). In those indices, the short wave infrared band (SWIR1) and the red edge band (B6; RE2) had the highest correlations, with R^2 values of 0.46 and 0.37, respectively (Chrysafis *et al.* 2017). The modified spectral vegetation indices showed higher correlations, with R^2 values of 0.52 for the

Enhanced Vegetation Index (EVI_{RE1}) and 0.51 for the red-edge Non-Linear Index (NLI_{RE1}) (Chrysafis *et al.* 2017). When a random forest regression model was constructed with all ten of the spectral bands from the Sentinel-2 MSI, the R^2 value obtained was 0.63 and the RMSE was 64.40 $m^3 \cdot ha^{-1}$ of growing stock volume (Chrysafis *et al.* 2017). The researchers noted that replacing the NIR band with band five (RE1 on the SENTINEL-2 platform) led to increased discrimination at higher chlorophyll levels (Chrysafis *et al.* 2017).

Astola *et al.* (2019) compared Sentinel-2 and Landsat-8 imagery for predicting forest parameters, including volume and height in mixed boreal forest in conjunction with two machine learning algorithms: multilayer perceptron and regression tree with brute force forward selection method. For all three models, reference plots were used that had been imaged by Sentinel-2 and Landsat 8. Equally sized sets of training, validation, and test data were used for 100 iterations of each algorithm. Astola et al. (2019) modeled volume, stem diameter, tree height, and basal area for both needle and broadleaf trees. They found that using all Sentinel-2 bands produced the highest accuracy. The characteristic that could be most accurately measured was tree height, with a RMSE of 30.4%. Of all the bands, the red-edge 1, the shortwave infrared, and the green (band 3) bands contributed the most to the accuracy of the models. They concluded that using SENTINEL-2 bands as the only inputs was a promising method of quantifying forest variables.

Schumacher *et al.* (2019) attempted to model timber volume using Sentinel-2 imagery combined with point clouds derived Schumacher et al. (2019) attempted to model timber volume using Sentinel-2 imagery combined with point clouds derived from stereoscopic aerial photographs for both broadleaf and conifer trees. Ten-fold cross validation was used to train a support vector machine (SVM) model for predicting timber volume. Inventory information from forest plots were used for ground-truth. The model obtained reasonably good results with an RMSE

of 31.7%. (Schumacher *et al.* 2019, 15). This study did not address the use of any vegetation indices,—red-edge or otherwise—and also did not use the Sentinel-2 imagery for anything other than differentiation between broadleaf trees and conifers. The value of this paper for our study is the use of photogrammetric data in their CHM, which is the same concept as that of NAIP, used in our own study.

Lin et al. (2019) also evaluate using vegetation indices that incorporate red-edge indices from SENTINEL-2. Their ground truth variables were readings from instruments measuring photosynthetically active radiation and carbon flux measured by tower-mounted sensors (Lin et al. 2019). They constructed a linear model where GPP = a × VI × PAR + b (Lin et al. 2019). The most effective VI was shown to be the EVI, with an R^2 value of 0.76, RMSE of 1.67 gC·m-2·day-1 and relative RMSE (rRMSE) of 0.18 (Lin et al. 2019).

Hu *et al.* (2020) used plot data combined with Sentinel-2 imagery and machine learning algorithms to estimate GSV in Hunan Province, China. They used Google Earth Engine as their tool for pixel extraction and tested the random forest, support vector regression, and multiple linear regression algorithms for predicting GSV (Hu *et al.* 2020). The researchers extracted 24 variables for each Sentinel-2 pixel from a series of images taken over the periods May October in 2017 and 2018(Hu et al. 2020). A summary of their results is provided in Table 1.2, using a 70%/30% split.

Table 1-2:A summary of Hu et al. (2020) results.

Algorithm	Number of Training Samples	Training R^2	Training RMSE ($m^3 \cdot ha^{-1}$)	Number of Testing Samples	Testing R^2	Testing RMSE ($m^3 \cdot ha^{-1}$)
RF	321	0.91	35.13	138	0.58	65.03
SVR	321	0.54	65.60	138	0.54	66.00
MLR	321	0.38	75.74	138	0.49	70.22

The variables used in the ML models (Hu *et al.* 2020) were:

1. Blue band (B2);
2. Green band (B3);
3. Red band (B4);
4. SWIR2 band (B12);
5. Tasseled Cap Wetness (TCW);
6. Normalized Difference Vegetation Index using band 5 ($NDVI_{B5}$);
7. Tasseled Cap Brightness (TCB)

Hu *et al.* (2020) tested their models on stands of ten different species and on three stands of mixed species, which included both coniferous and broad-leaved trees. The different stands had different mean tree heights, but within the stand tree height was not highly disparate. The ground truth data was obtained by site survey, taking measurements and using those measurements with a volume estimation formula specific to each species to estimate the volume of each tree, and then added to calculate the GSV of each hectare. The researchers then used R software to transform the data to a normal distribution ($f(y) = y^{\lambda}$). The researchers then used the variable selection algorithm from the VSURF package in R, which resulted in only B5 and MSI being used for the MLR model, the others having been discarded due to multicollinearity. For the RF model, the smallest error rate was found to be when mtry = five and ntree = 257, and the algorithm was run both with the full range of variables and with only the B5 and MSI variables, which resulted in no significant difference in results. The researchers found that, in the SVR model—the worst-performing of the three models attempted—that use of all variables in the training phase resulted in a slightly better performance, while the use of only the B5 and MSI variables in the testing phase resulted in a slightly better performance. The researchers found that the RF model often over- or under-estimated GSV due to factors including ground vegetation and the difficulty of discerning volume

at high saturation points, but could be balanced out with the right forest composition (Hu *et al.* 2020).

The researchers stated that, in conclusion, there were four major difficulties in the study, which were: (1) processing the satellite imagery to eliminate distortions; (2) the over- and under-estimation problem of the RF algorithm; and (3) the transformation of the data to a normal distribution did not eliminate bias (Hu *et al.* 2020); and, (4) the Google Earth Engine software was limited in its machine learning functionality, resulting in the necessity of turning to R to perform advanced analysis (Hu *et al.* 2020).

Dalen *et al.* (2020) did specifically examine the red-edge for the purpose of biomass estimation, using nitrogen uptake and observed fiber mass to correlate with the Simple Ratio using red band (SR_{red}), $SR_{red\text{-}edge}$, $NDVI_{red}$, and $NDVI_{red\text{-}edge}$ vegetation indices, using Microsoft Excel for regression analysis. The study is partly motivated by the NDVI's poor performance in saturated canopies, meaning that it becomes less reliable towards to the peak of the growing season (Dalen *et al.* 2020). Nitrogen fertilizer application is here also used as a predictor variable.

Imran *et al.* (2020) used field inventory data including DBH and height, from which volume was calculated and "basic wood density (BWD; kg/m^3) and biomass expansion factor (BFF) were taken from literature of Pakistan Forest Institute" and then calculated biomass using the formula: Biomass = V × BWD × BEF (Imran *et al.* 2020). The red-edge normalized difference vegetation index, version two (RENDVI-2: (b7-b4)/(b7+b4)) index as the single independent variable in linear model produced the best results, with an R^2 of 0.64 and RMSE of 31.29 tons per hectare (Imran *et al.* 2020).

Peng *et al.* (2021) continued along the path of calculation GPP on the basis of available photosynthetically active radiation and LUE. They sought to combine two different traditions in

remotely sensed productivity estimation: vegetation indices and physical productivity models (Peng *et al.* 2021). They settled on light use efficiency models as the most compatible with remote sensing data. They also noted a significant increase in performance by substituting red-edge variables into vegetation indices.

Rees *et al.* (2021) sought to quantify GSV in boreal forest using the Sentinel-2 MSI and land cover classification. They studied boreal forest in Siberia, which is an example of a large GSV for which detailed records are not available. Ground truth data was collected in 20-m^2 sample plots of approximately homogenous heights by calculating tree volume on the basis of DBH, height, and species. The researchers then generated land cover maps for the area based on a Russian refinement of MODIS imagery, in which point objects were created for plots that were surrounded on all sides by pixels of the same land cover type to ensure homogeneity within the pixel (Rees *et al.* 2021). Then Sentinel-2 images using bands 2, 3, 4, and 8 were taken in two series: one in summer and one in winter (Rees *et al.* 2021). The Sentinel-2 imagery is not used to directly quantify the GSV, but rather fed into an equation in which GSV of a plot is calculated as a function of Sentinel-2 imagery and class counts. The R^2 values of the two study areas were 0.787 and 0.679 (Rees *et al.* 2021).

Moradi *et al.* (2022) worked to exclusively use Sentinel-2 data to quantify dense above-ground biomass for dense forests, which traditionally have been a problem for indices such as the NDVI which decrease in accuracy as vegetation density increases. They tested both parametric (multiple regression) and non-parametric (artificial neural network, K-nearest neighbor and random forest; all machine learning) methods. In the case of the parametric method, normality was confirmed via the Kolmogorov-Smirnov test (Moradi *et al.* 2022). In the case of the artificial neural network, 70% of the data were used for training, 15% for validation and 15% for testing; while in

the case of the random forest algorithm, two-thirds of the data were used for training and one-third for evaluation (Moradi *et al.* 2022). The best results were obtained from the artificial neural network, with a training R^2 of 0.89, a training RMSE of 8.79%, a validation R^2 of 0.65, and a validation RMSE of 19.93%. The researchers conclude that research into red-edge variables should be continued (Moradi *et al.* 2022).

1.3 Goals and Objectives

The overarching goal of this work was to continue the line of inquiry into the best method for estimating forest potential productivity in Maine, United States using Sentinel-2 optical imagery as the primary input in conjunction with site variables and determining the optimal variables and a ML algorithm using the caret package in R. Specific objectives were:

1. Evaluating NAIP-derived CHM data as a predictor of forest growth/potential productivity
2. Determining if using Sentinel-2 based variables (Rahimzadeh-Bajgiran et al., 2020) and updated site variables can improve forest growth model performance

2 CHAPTER TWO: STUDY AREA AND METHODS

The choice of study area was simplified by the availability of data relating to Maine, United States. The choice of methodology was more complex, due to the large variety of machine learning algorithms available. The random forest algorithm was chosen because it is usually found to be the most effective algorithm in predicting forest variables (Chrysafis *et al.* 2017; Moradi *et al.* 2022; Rahimzadeh-Bajgiran *et al.* 2020; Bhattarai *et al.* 2022).

2.1 Study Area

The study area is located in the north-central part of the State of Maine, United States. The bounding coordinates in degrees of latitude and longitude are:

1. Northwest: (46.30478°, -69.472431°) or (46°18′17.208″, -69°28′20.752″);
2. Northeast: (46.303634°, -68.303445°) or (46°18′13.0824″, -68°18′12.402″);
3. Southwest: (45.494582°, -69.465609°) or (45°29′40.4952″, -69°27′56.192″);
4. Southeast: (45.493468°, -68.313503°) or (45°29′36.4848″, -68°18′48.611″).

The valid data locations are those raster grid cells for which data is available for all variables. Each square grid cell is 20 m × 20 m, making the area of each grid cell 400 m². The gird cells are arranged in a square over the study area, with the dimensions of 4,501 rows × 4,501 columns. The majority of these grid cells are missing values for one or more variables. There are 155 grid cells that are not missing any values. They are the 'valid data locations.' All the geospatial layers used in our study were rasterized at a grid cell resolution of 20 m × 20 m, which corresponds to the grid of the Sentinel-2 data, from which it was constructed, in all aspects, including resolution, projection, and placement (the grids fit squarely on top of one another with no shift in position).

The study area—the total combined area of the raster cells—is depicted in Figure 2.1 (below). The fuchsia crosses indicate the positions of the 155 grid cells for which all variables have measurements. Figure 2-1 illustrates the study area at the large scale (left) and the small scale (right), with fuchsia crosses marking the 155 data locations in the small scale portion of the figure.

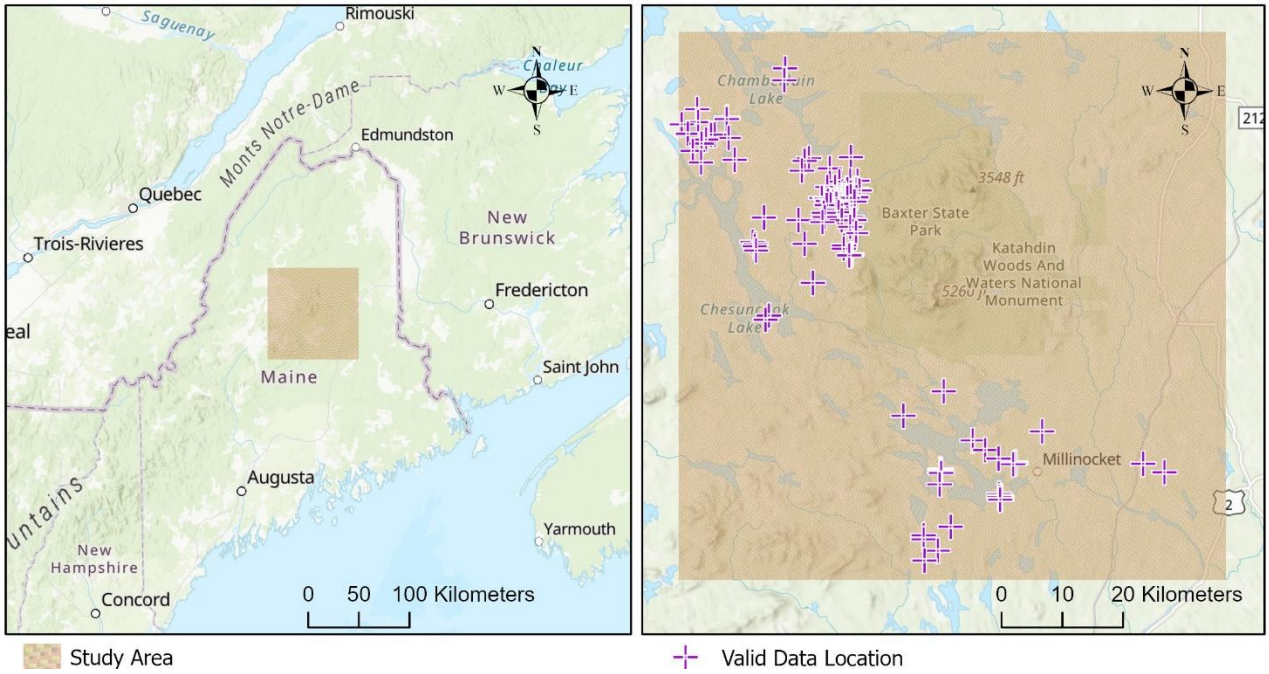

Figure 2-1:The location of the study area in Maine; markers indicate valid data locations.

In Figure 2.2, each grid cell has been randomly assigned one of a series of colors in order to make it visibly distinguishable from its neighbors; in this figure grid cell colors do not represent values, only demarcation. Each grid cell has a unique number assigned to it for identification. The grid cells for which data exist for all variables are marked with a fuchsia cross—they are the valid data locations. Those grid cells which are not so marked do not contain data for all variables, and are not used in any model, and are not valid data locations.

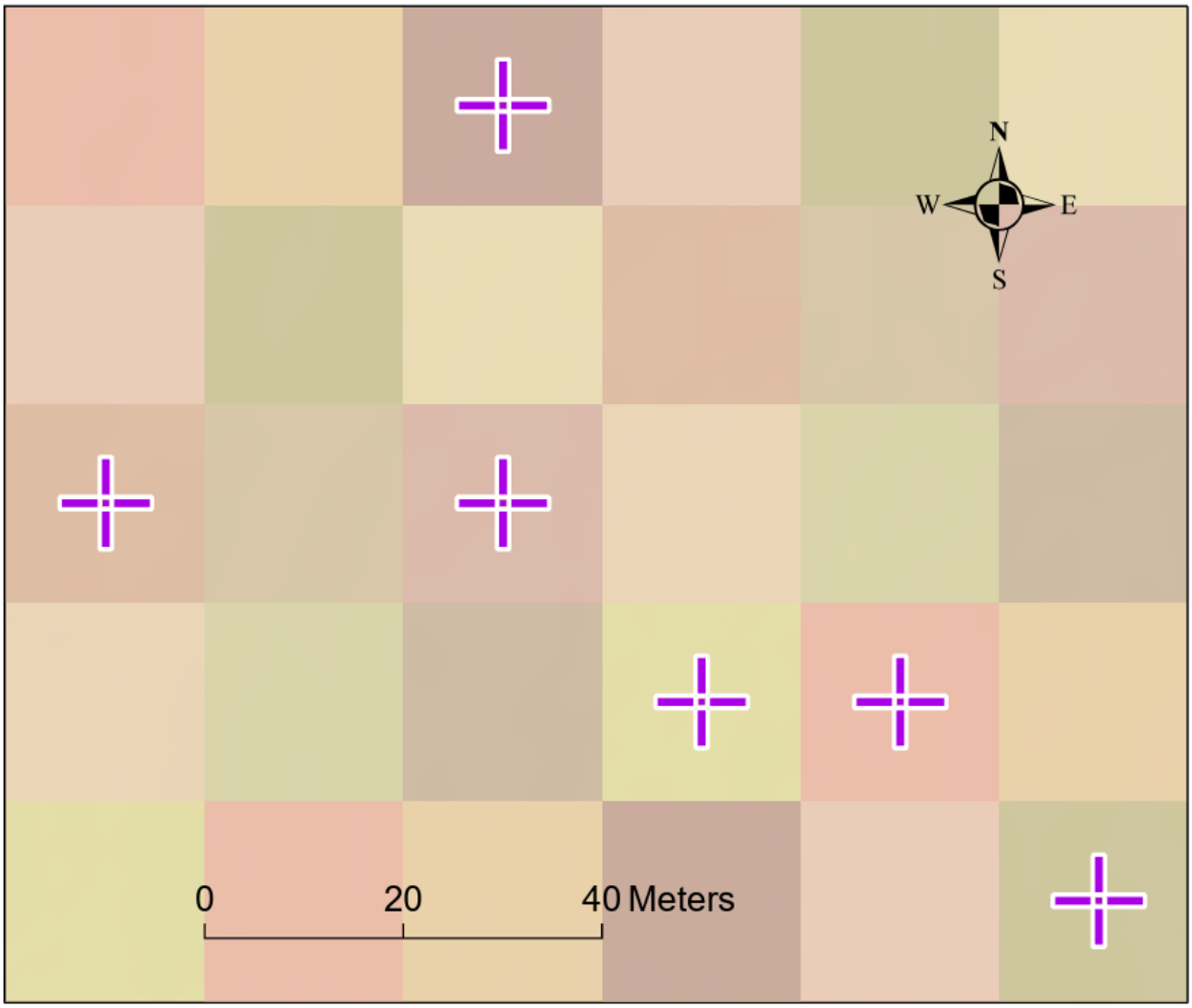

Figure 2-2: Enlarged section of study area.

In this project the projection and datum used are UTM Zone 19N and World Geodetic Survey (WGS) 1984, respectively. The study area is 8,103.60 km². Baxter State Park, the Katahdin Woods and Waters National Monument, and the town of Millinocket are located within the boundaries of the study area. The town of Millinocket is the largest town in the study area, followed by the town of Patten.

The region is mountainous, with elevations from zero to 374 meters above sea level. The arboreal makeup of the region is a mix of conifer and broadleaf forests. Conifer types include red spruce (*Picea rubens* Sarg.), balsam fir (*Abies balsamea* (L.) Mill.), eastern hemlock (*Tsuga canadensis* (L.) Carr.), eastern white pine (*Pinus strobus* L.) and northern white-cedar (*Thuja occidentalis L.*). Hardwoods include sugar maple (*Acer saccharum* Marsh.), American beech (*Fagus grandifolia* Ehrh.), yellow birch (*Betula alleghaniensis* Britton), and red maple (*Acer rubrum* L.) (Rahimzadeh-Bajgiran *et al.* 2020).

Figure 2-3 describes the range of elevation in the study area, with markers indicating valid data locations for reference. From the figure one can see that the valid data locations are generally located at similar elevations.

The climate of the study area, and predominantly of the state, is Köppen climate type Dfb: Warm-summer humid continental (Wikipedia). Monthly average precipitation, in inches (1902-2023) for the Millinocket area is presented in Table 2-1 (Source: National Oceanic and Atmospheric Administration (NOAA)). Table 2-2 displays average temperature (1902-2023), in degrees Fahrenheit (°F) for the Millinocket area.

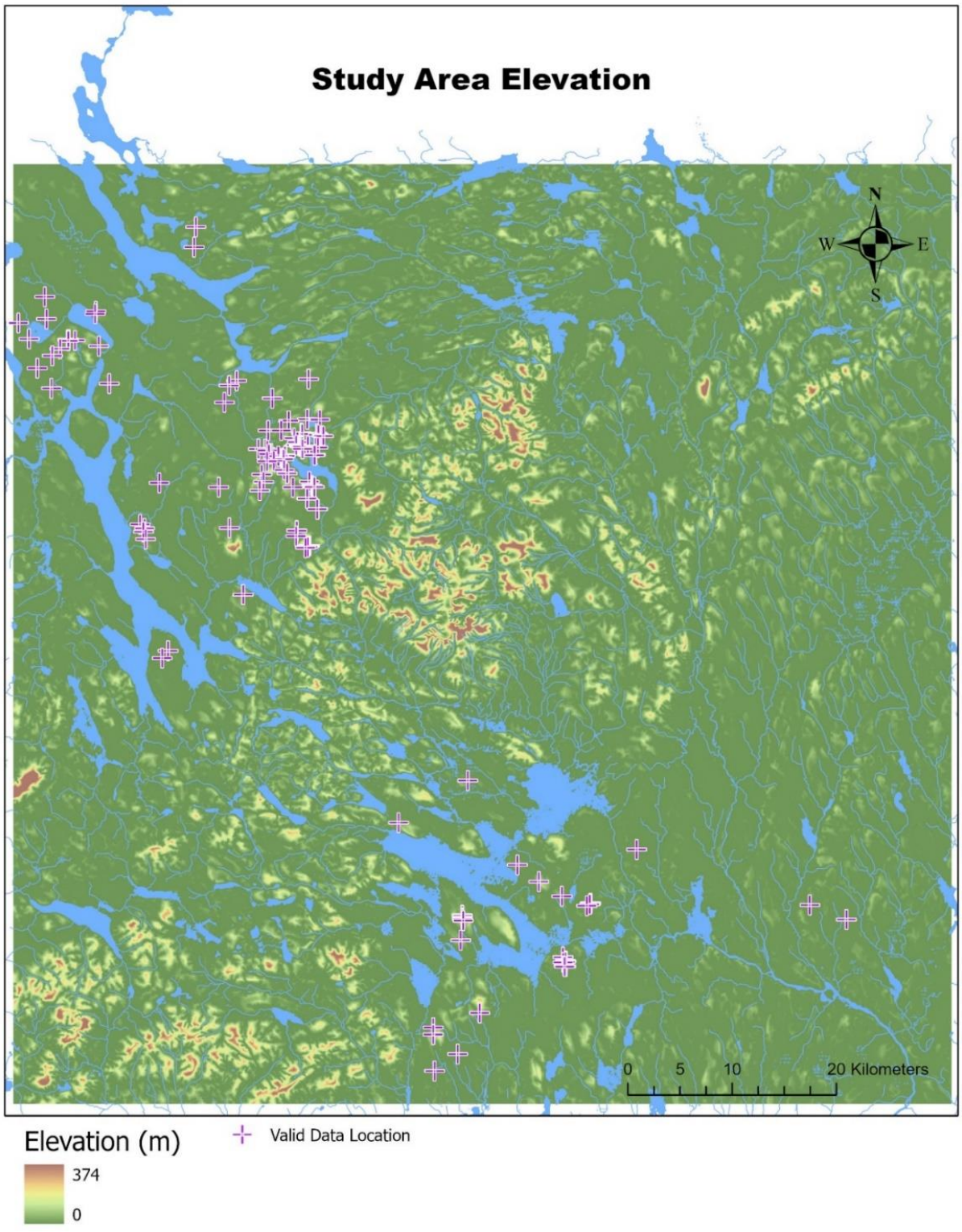

Figure 2 3: The elevation range in the study area.

Table 2-1: Study area precipitation (units are in inches).

	Mean	Max	Max Year	Min	Min Year
January	3.02	7.11	1979	0.02	2003
February	2.56	6.97	1960	0.46	1964
March	2.94	8.11	1936	0.11	1915
April	3.21	8.91	2005	0.35	1941
May	3.34	9.38	1989	0.38	2003
June	3.77	10.82	1922	0.93	1965
July	3.69	7.38	1972	0.65	1952
August	3.86	9.42	1991	0.73	1987
September	3.57	9.70	1909	0.60	2014
October	3.97	10.02	2005	0.22	1947
November	4.05	10.28	1950	0.23	1939
December	3.49	10.41	1973	0.72	1943

Table 2-2: Study area temperature (in °F).

	Mean	Max	Max Year	Min	Min Year
January	14.5	24.5	1956	2.7	1994
February	16.4	25.6	2010	5.1	1993
March	27.4	37.0	1903	20.4	2014
April	40.1	46.9	2010	34.6	1914
May	52.9	59.6	1904	44.7	1967

Table 2-2 continued

June	62.3	68.3	1930	56.0	1958
July	67.9	73.3	1952	61.6	1962
August	65.7	71.7	1937	59.8	1964
September	57.2	64.1	2015	51.7	1963
October	46.0	52.9	2017	37.5	1925
November	34.0	43.5	1903	26.2	1933
December	20.4	32.8	2015	5.2	1989

Figure 2-4 identifies the bedrock formations from which was derived the soil parent materials. Formation IDs are given in Table 2-3.

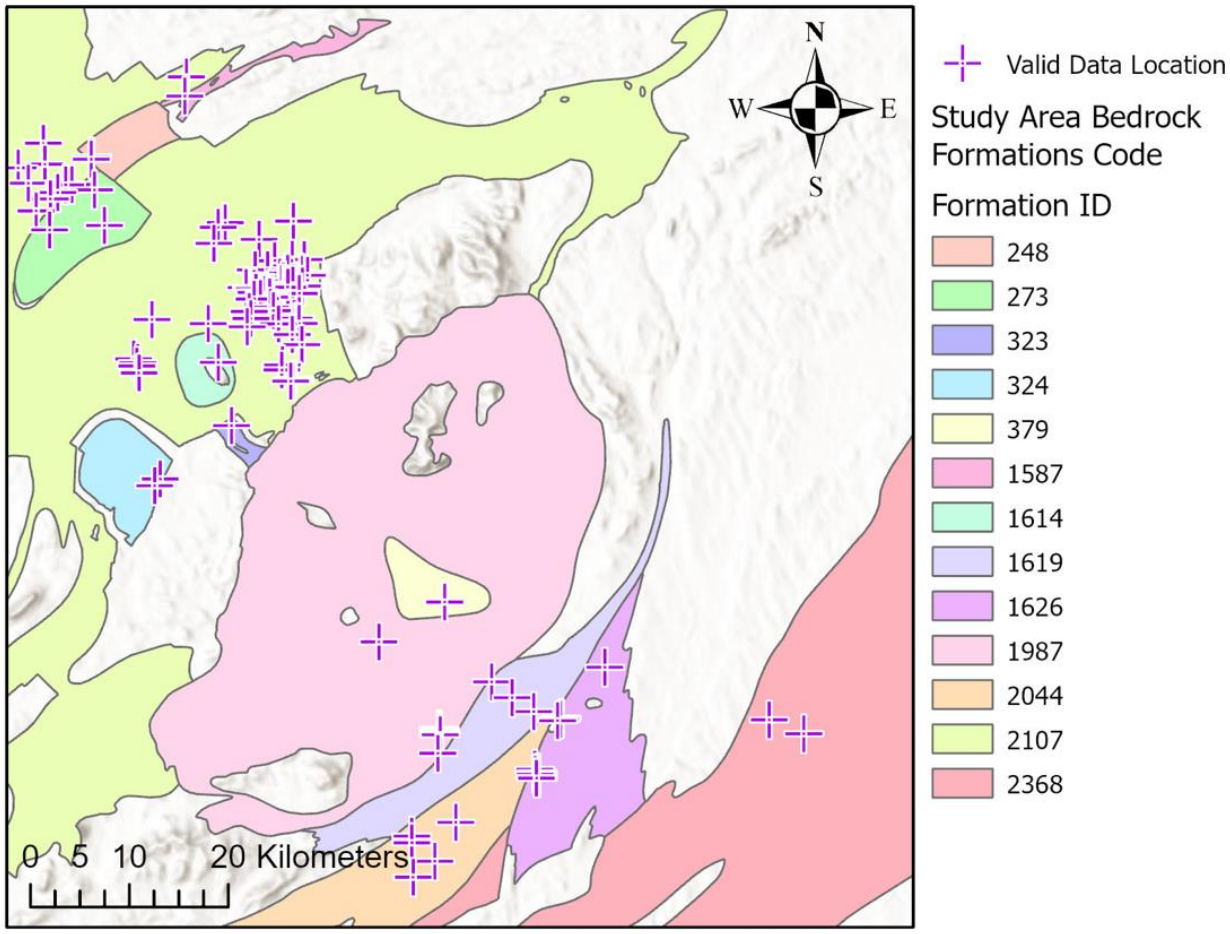

Figure 2-3: Soil parent material in the study area.

Table 2-3 lists identifier codes for bedrock formation. (Source: Maine GIS database)

Table 2-3: Explanation of soil parent material (https://maine.hub.arcgis.com/datasets/74bfaad5358444dda1d242e2846a56ed/explore)

Formation Number	Count of Locations	Formation Description
248	2	Ordovician-Devonian with mixed volcanic rocks in north of study area.
273	3	Ordovician-Devonian with mixed volcanic rocks in north of study area.
323	3	Devonian: Marine sandstone and slate in northern and central Maine. Gneiss and schist in southwest.
324	2	Cambrian-Ordovician: Mostly volcanic rocks and related sedimentary rocks in north and east. Schist, marble, and gneiss in central coast.
379	1	Devonian: Granite, granodiorite, and gabbro throughout state.

Table 2-3 continued

1587	1	Silurian-Devonian: Marine sandstone and slate in east grading to gneiss and schist in southwest. Some volcanic rocks in central coast.
1614	1	Devonian: Marine sandstone and slate in northern and central Maine. Gneiss and schist in southwest.
1619	3	Devonian: Marine sandstone and slate in northern and central Maine. Gneiss and schist in southwest.
1626	16	Silurian-Devonian: Marine sandstone and slate in east grading to gneiss and schist in southwest. Some volcanic rocks in central coast.
1987	9	Devonian: Granite, granodiorite, and gabbro throughout state.
2044	8	Devonian: Marine sandstone and slate in northern and central Maine. Gneiss and schist in southwest.
2107	104	Devonian: Marine sandstone and slate in northern and central Maine. Gneiss and schist in southwest.
2368	2	Silurian-Devonian: Marine sandstone and slate in east grading to gneiss and schist in southwest. Some volcanic rocks in central coast.

Figure 2-5 depicts soil hydrologic group. Because precipitation across the study area is not highly variable, variations in productivity may be explained by differing soil regimes. The soil hydrologic groups are described in Table 2-4.

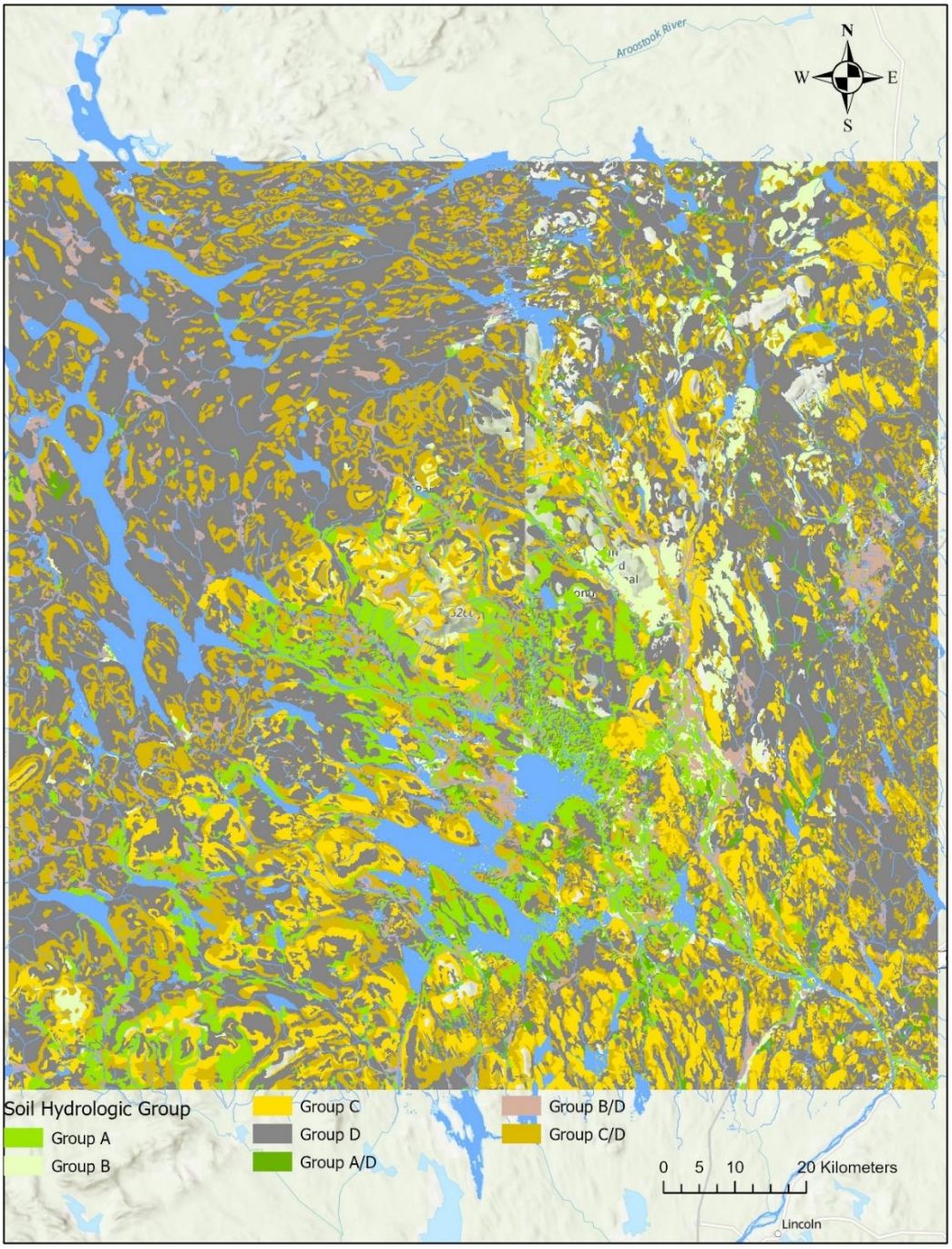

Figure 2-4: Hydrologic soil group (Source: USA Soil Survey Geographic Database (SSURGO), Soil Hydrologic Group layer from ArcGIS Pro Living Atlas)

Table 2-4: Soil hydrologic group descriptors (Source: Soil Survey Geographic Database (SSURGO)).

Group Designator	Count of Locations	Group Description
A	20	"Group A soils consist of deep, well drained sands or gravelly sands with high infiltration and low runoff rates."
C	9	"Group C consists of soils with a layer that impedes the downward movement of water or fine textured soils and a slow rate of infiltration."
C/D	46	"Group C/D soils naturally have a very slow infiltration rate due to a high water table but will have a slow rate of infiltration if drained."
D	76	"Group D consists of soils with a very slow infiltration rate and high runoff potential. This group is composed of clays that have a high shrink-swell potential, soils with a high water table, soils that have a clay pan or clay layer at or near the surface, and soils that are shallow over nearly impervious material."

A contingency table of bedrock parent material set against soil hydrologic group is presented in Table 2.5 (data sources: Maine GIS Database and USA SSURGO Soil Hydrologic Group layer from ArcGIS Pro Living Atlas):

Table 2-5: Soil hydrologic group and bedrock parent material contingency table.

	Hydrologic Group			
Bedrock	A	C	C/D	D
248	0	0	0	2
273	0	0	1	2
323	0	0	0	3
324	0	0	1	1
379	1	0	0	0
1587	0	0	0	1
1614	0	0	0	1

Table 2-5 continued

1619	1	0	1	1
1626	10	1	1	0
1987	8	0	1	0
2044	0	8	0	0
2107	0	0	40	64
2368	0	0	1	1

The study grid cells are located in areas of average insolation for Maine (Figure 2.6). Insolation is one of the limiting factors of vegetation growth, the others being water and nutrient availability. In Maine, it is likely that insolation is a limiting factor of greater importance than water, as water is generally abundant throughout the state. Differences in insolation may also account for local variability among trees in regions with stark differences in elevation, as some trees will be shaded by terrain features more than others. Insolation was included as a variable in the all-variables model run. Most study locations receive similar levels of insolation.

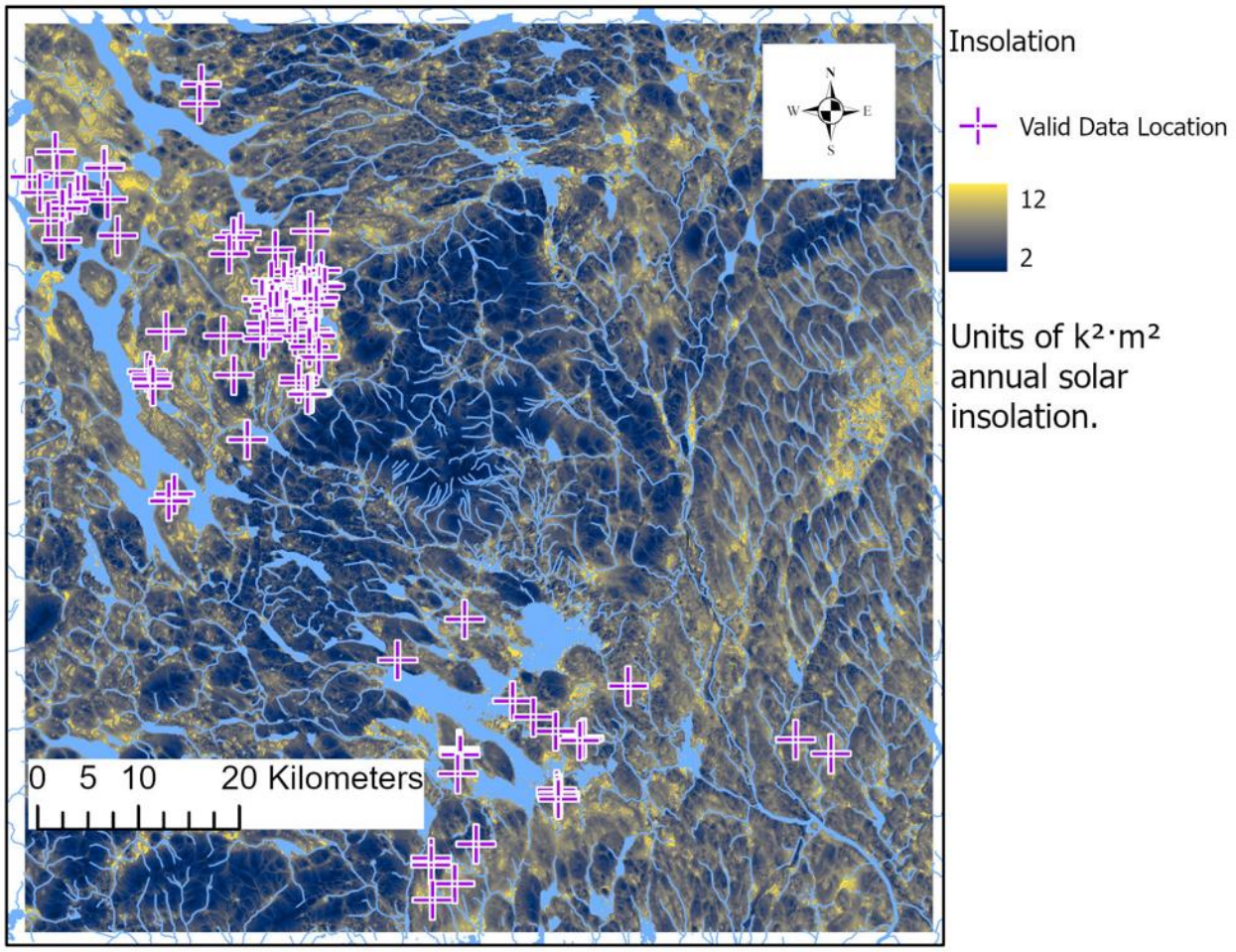

Figure 2-5: Study area insolation (Hargraves and Samani 1985; Smith and Metcalfe 2018).

Soil nitrogen content is one of the very important nutritional factors for vegetation growth (Kanke *et al.* 2012; Sharma *et al.* 2015; Bandyopadhyay *et al.* 2017) and can be an indicator of vegetation vitality (Dalen et al. 2020). Soil nitrogen content was included as a variable in our model. Soil nitrogen distribution is more dense in the northwest quadrant of the study area, and less dense in the southeast quadrant. Figure 2-7 describes the soil nitrogen concentrations within the study area.

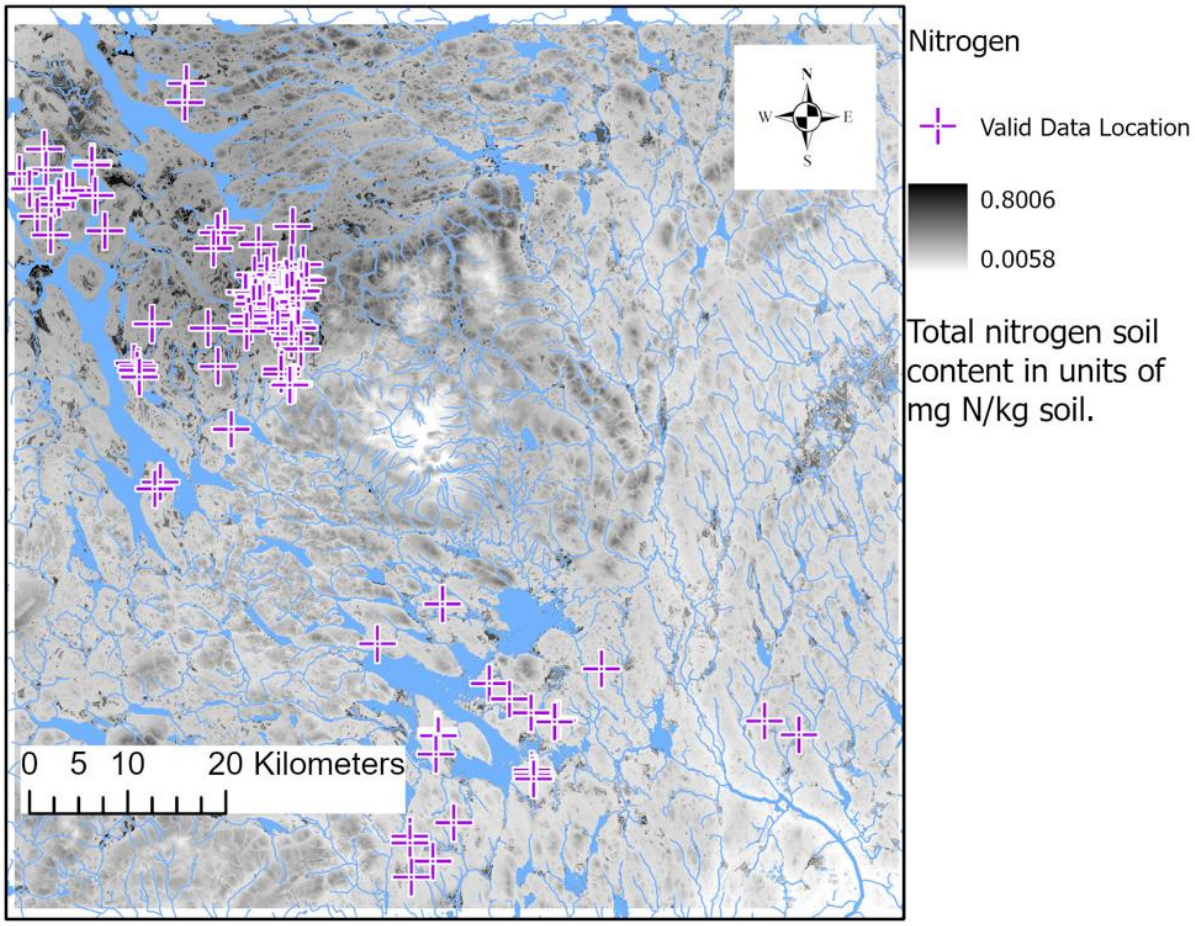

Figure 2-6: Study area nitrogen content.

2.2 Data

The data used in this study can be divided into three categories. The first data set is the remotely sensed bands and indices provided through Dr. Rahimzadeh-Bajgiran Remote Sensing lab, School of Forest Resources, University of Maine (Rahimzadeh-Bajgiran et al., 2020). The second data are the 2018 and 2021 NAIP-derived CHM maps provided through Wheatland Geospatial Lab, University of Maine School, Forest Resource which were used to construct the ΔNAIP response variable. The third data set is the suite of site and climate variables. The list and source of these data were provided in Table 2.6.

2.2.1 Sentinel-2 Remotely Sensed Variables

Remote sensing data are a subset of Sentinel-2 cloud-free mosaic image produced for the entire state of Maine for the peak growing season in 2018 at 20 m spatial resolution (Rahimzadeh-Bajgiran *et al.* 2022). The remotely sensed variables were selected based on their performance in Rahimzadeh-Bajgiran *et al.* (2022) (Table 2.6):

The images were passed through median filter algorithms to reduce noise. All other variables were resampled to the native resolution and coordinate reference system of the Sentinel-2 variables in order to prevent any distortion of the remotely sensed optical data.

2.2.2 National Agriculture Imagery Program (NAIP)

The NAIP is a government program administered by the United States Department of Agriculture's Farm Service Agency. The program creates and disseminates orthographically rectified aerial imagery at a spatial resolution of 1 m (0.6 m starting 2018) for the United States since 2003 during the agricultural growing season. NAIP data collected in 2018 and 2021 from state of Maine were used to estimate canopy height by the Wheatland Geospatial Laboratory, the School of Forest Resources, University of Maine. The NAIP CHM data are in the form of a raster, with the raster's value indicating the average canopy height within that pixel in meters. The 2018 and 2021 NAIP CHM maps have a resolution of 6 m and 1 m, respectively. The valid range of NAIP CHM data are zero to 50 meters. The NAIP 2018 data were contaminated with other features such as water bodies, with original values between -718.87 and 2023.32 meters and therefore needed to be cleaned. A 0.25 m threshold of error was used for each year (0.5 m in total). All data outside the range of -0.5 m to 45.5 m were assigned a 'no data' value. Both datasets were then reprojected to the same projection as the Sentinel-2 data, and resampled to a spatial resolution of 400 m^2. To

create the ΔNAIP variable, the NAIP 2018 values were subtracted from the NAIP 2021 values. The result was a raster layer with a range from approximately -8.50 to 8.37. The negative value could be accounted for either by trees that had been cut or blown down, or a simple data error given the wide range of possible errors inherent in this project. The mean ΔNAIP value is approximately 1.22 m. As can be seen from Figure 2-8, the preponderance of values are between zero and five, as would be expected in a valid data set. In order to learn as much as possible about the newly created variable, no values were removed.

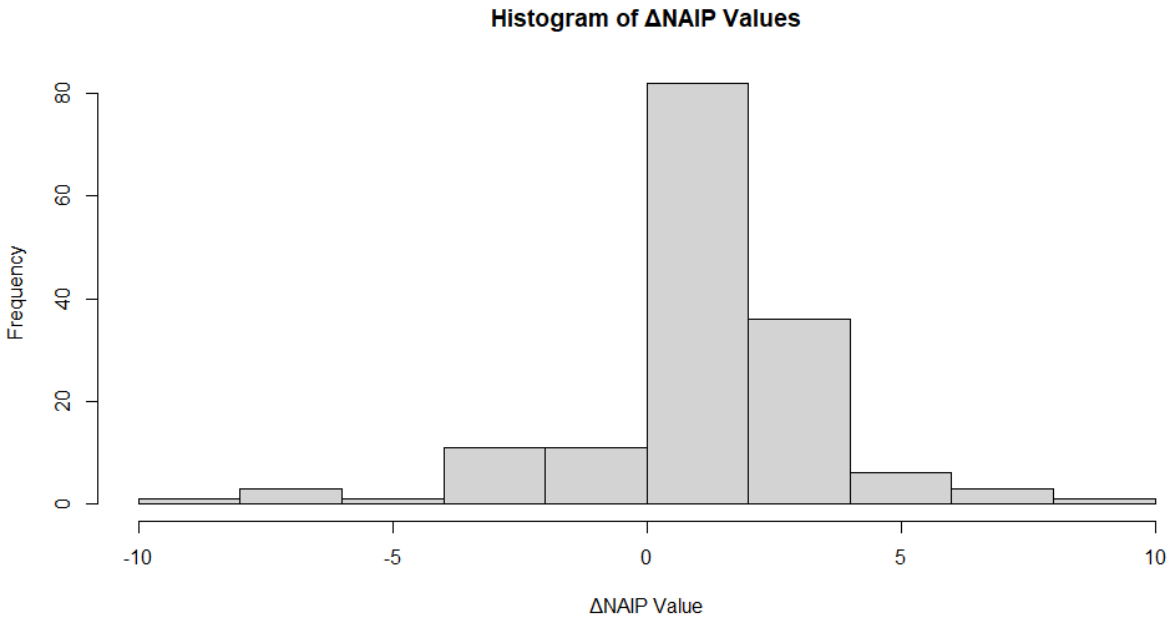

Figure 2-7: Histogram of ΔNAIP values.

As stated earlier in this chapter, a pixel is considered a 'valid' observation if values exist for all variables after the NAIP 2018 data was cleaned and resampled. Of 20,259,001 grid cells (4501 rows × 4501 columns at a resolution of 20 m), 155 grid cells meet the 'valid' criteria. Figure 2.10 is an example of the same pixels with either their ΔNAIP or the Sentinel-2 Red Edge Position Index values, as an illustration of this project's goal, which is to determine if Sentinel-2 Red Edge

Position Index can be used to estimate ΔNAIP, using a variety of helper variables and machine learning algorithms. A model that successfully correlates ΔNAIP, as a proxy for biomass, with Sentinel-2 Red Edge Position Index, would substantially increase the efficiency of monitoring of forest productivity. Figure 2-9 illustrates ΔNAIP and Sentinel-2 red-edge index values for the same pixels in some of the 155 valid data point locations.

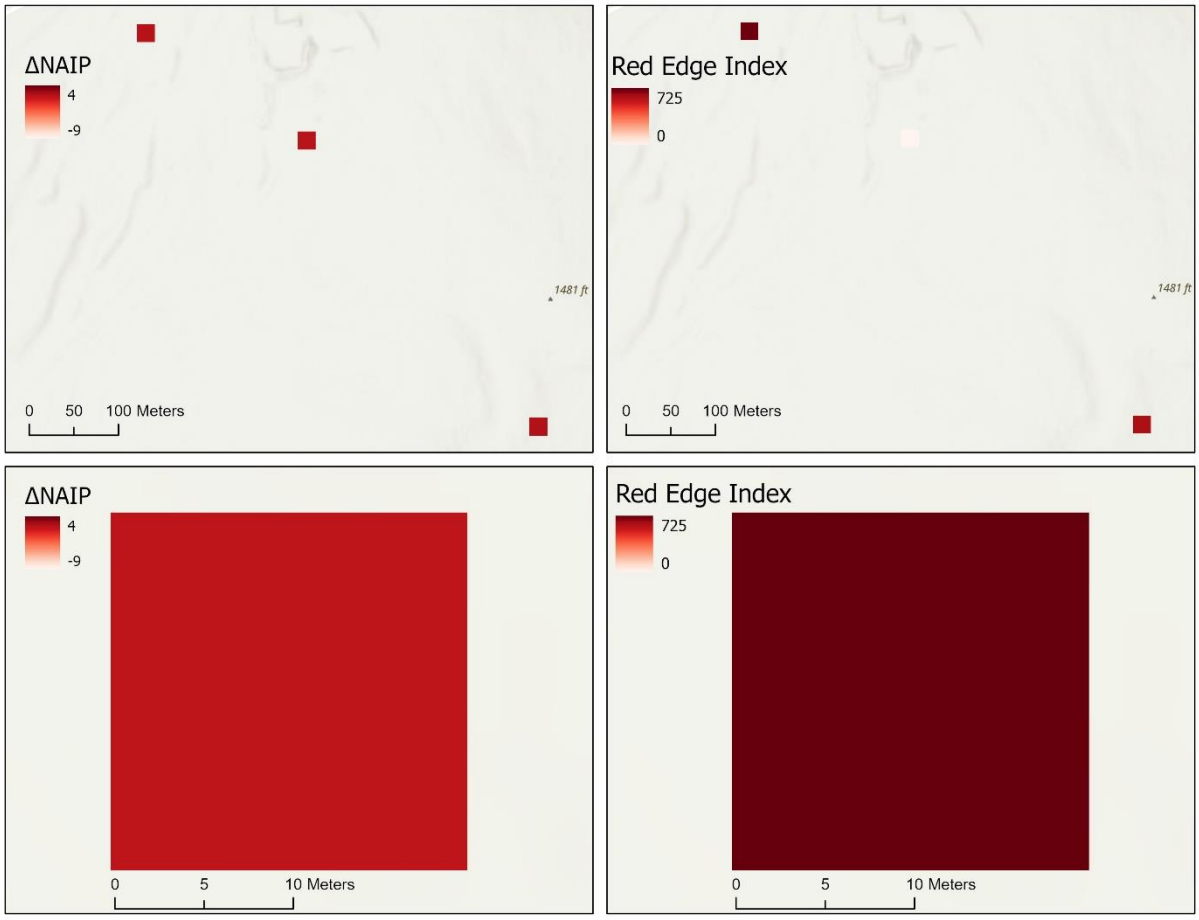

Figure 2-8: Pixel comparison of ΔNAIP and Sentinel-2 red-edge index values.

2.2.3 Site and Climate Variables

Site and climate variables used in the analysis were collected from several different sources and are presented in Table 2.6. All variables were used in the initial model. Table 2-6 also indicates if a variable was used in the select-variable model.

Table 2-6: Variables used for modeling.

Predictor	Used in 'Select Variables' Model	Reference/Data Provider
Sentinel-2 Optical		
Band 3 (Green)	Yes	Rahimzadeh-Bajgiran *et al.* 2020
Band 8a (SWIR)	Yes	
NDVI45	Yes	
S2REP	Yes	
NAIP 2018	Yes	Prior *et al.* 2022; Schroeder *et al.* 2022
NAIP 2021	Response	
Old or young trees predominant	No	Bhattarai *et al.* 2020
Climate		
Thirty-year average precipitation	No	Global Wind Atlas; Hargraves and Samani 1985; Smith and Metcalfe 2018
Annual solar insolation	No	
Thirty-year average maximum temperature	No	
Thirty-year average mean temperature	No	
Thirty-year average minimum temperature	No	
Thirty-year average maximum vapor pressure deficit	No	
Thirty-year average minimum vapor pressure deficit	No	
Water deficit index for year 2000	No	
Water deficit index for year 2020	No	
Mean windspeed at 10 m altitude above terrestrial surface	No	
Mean windspeed at 50 m altitude above terrestrial surface	No	

Table 2-6 continued

Predictor	Used in 'Select Variables' Model	Reference/Data Provider
Site (Terrain)		
Aspect	No	Beven and Kirkby 1979; Bolstand, Swank and Vose 1998; De Reu *et al.* 2013
Bolstad's topographic index	No	
Elevation	No	
Flow direction	No	
McNab's topographic index	No	
Planform curvature	No	
Profile curvature	No	
Topographic roughness	No	
Slope	No	
Topographic wetness index	No	
Site (Soil)		
Soil depth	No	Natural Resource Conservation Service 1999; PRISM Climate Group 2013
Depth to densic soil layer	No	
Depth to lithic soil layer	No	
Depth to redox soil layer	No	
Bedrock parent material	No	
Soil water holding capacity	No	
Total soil rooting depth	No	
Soil exchangeable calcium	No	
Soil exchangeable Potassium	Yes	
Soil exchangeable Magnesium	Yes	
Soil exchangeable Nitrogen	Yes	

Table 2-6 continued

Predictor	Used in 'Select Variables' Model	Reference/Data Provider
Site (Soil, *continued*)		
Depth to glacially compacted soil horizon 0.05 quantile which is the lower 90 percent prediction interval.	No	Brungard and Hennigar 2022
Depth to glacially compacted soil horizon 0.5 quantile, which is the median value.	No	
Depth to glacially compacted soil horizon 0.95 quantile which is the upper 90 percent prediction interval.	No	
Depth to glacially compacted soil horizon prediction interval width which is the 0.95 layer minus the 0.05 layer and indicates the width or distance between the upper and lower bounds of the prediction interval.	No	
Depth to glacially compacted soil horizon relative prediction interval width which is the prediction interval width divided by the 90 percent interquartile range of the observed soil properties.	No	
Depth to bedrock soil horizon 0.05 quantile which is the lower 90 percent prediction interval.	No	
Depth to bedrock soil horizon 0.5 quantile which is the median value.	No	
Depth to bedrock soil horizon 0.95 quantile which is the upper 90 percent prediction interval from DSM.	No	

Table 2-6 continued

Predictor	Used in 'Select Variables' Model	Reference/Data Provider
Site (Soil, *continued*)		
Depth to bedrock soil horizon prediction interval width which is the 0.95 layer minus the 0.05 layer and indicates the width or distance between the upper and lower bounds of the prediction interval.	No	
Depth to glacially compacted soil horizon relative prediction interval width which is the prediction interval width divided by the 90 percent interquartile range of the soil properties.	No	
Depth to redox soil horizon 0.05 quantile which is the lower 90 percent prediction interval	No	
Depth to redox soil horizon 0.5 quantile which is the median value.	No	Brungard and Hennigar 2022
Depth to redox soil horizon 0.95 quantile which is the upper 90 percent prediction interval.	No	
Depth to redox soil horizon prediction interval width which is the 0.95 layer minus the 0.05 layer and indicates the width or distance between the upper and lower bounds of the prediction interval.	No	
Depth to glacial redox soil horizon relative prediction interval width which is the prediction interval width divided by the 90 percent interquartile range of soil properties	No	

2.3 Methods

Satellite remote sensing provides the ability to monitor and measure phenomena at the global scale for targeted management of natural resources, no matter how inaccessible the terrain is. Because most developments of events to which one can attach a hierarchical valuation is a double-edged sword, one of the advantages of satellite remote sensing—the ability to efficiently collect vast amounts of data—has also historically been one of its drawbacks. The advent of machine learning, made possible by advances in computer processing power, is the answer to the previously-stated problem. Using algorithms programmed through an analytical programming language, vast amounts of data can be quickly processed and interpreted. In this work the programming is done in the R computing language (See 'R Core Team' in bibliography).

2.3.1 Satellite Optical Remote Sensing

The goal of this work is to increase the utility of remotely sensed optical variables in modeling tree growth rates. Rahimzadeh-Bajgiran *et al.* (2020) applied a number of remotely sensed bands and indices acquired from the Sentinel-2 mission to this problem, and found that the red-edge bands and indices were the most useful for forest productivity prediction. The *red-edge* bands also known as the *visible and near infrared (VNIR)* bands, which are located at the border between the visible red and the near infrared sections of the electromagnetic spectrum. Chlorophyll generally absorbs red radiation but reflects infrared radiation, resulting in a sudden increase in the percent of radiation reflected by the plant (Guyot and Jacquemoud 1992; Foody and Curran 1994; Gitelson *et al.* 2006).

Figure 2-10 depicts a generic spectral signature of a generic plant based on three comparative chlorophyll densities. The top curve is the signature of low chlorophyll, the middle curve is the signature of medium chlorophyll, and the bottom curve is the signature of high chlorophyll (Source: https://seos-project.eu/agriculture/images/rededge_large.jpg).

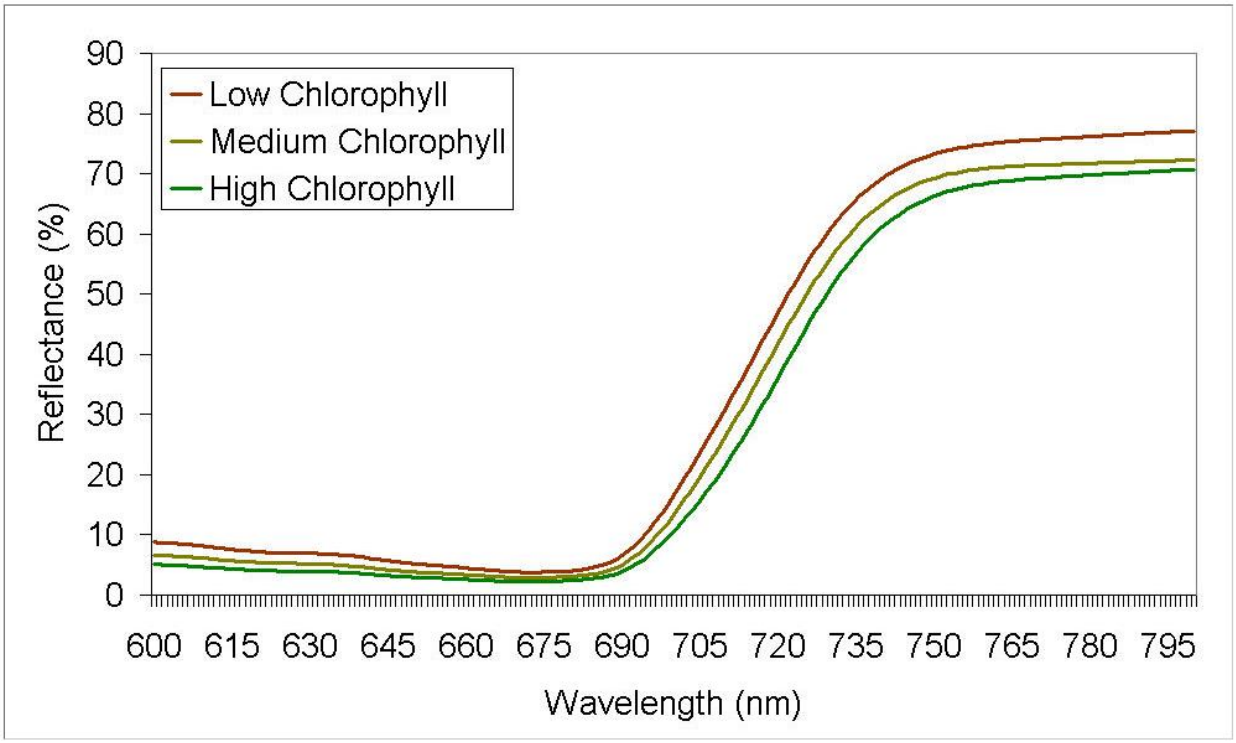

Figure 2-9: Typical vegetation reflectance profile as affected by chlorophyll content.

The boundary between the two is known as the *red edge.* In the Sentinel-2 mission, bands 5 (centered at 705 nm), 6 (centered at 740 nm) and 7 (centered at 783 nm) are red-edge bands while bands 8 (centered at 842 nm) and 8a (centered at 865 nm) are the *Near infrared bands.* The multitude of red-edge bands is one of the features that makes the Sentinel-2 mission unique and interesting compared to other optical missions such as MODIS and Landsat. It allows observers to quantify the differences in plant qualities by quantifying the relative chlorophyll contents as well as vegetation amount.

Spectral indices are algebraic manipulations of the reflectance percentages of individual bands, which capture the radiation from specific wavelengths of the electromagnetic spectrum. Spectral indices are powerful tools that enable researchers to observe and record features that are not inherently visible from the basic bands themselves. Vegetation indices in remote sensing are exactly the equivalent of feature engineering in machine learning: two or more base features are manipulated to create a new feature that, because of its new relationships, exposes some quality hitherto unexposed. The red-edge position index (REP) was based on work performed by Guyot et al. (1988) and, in terms of generalized band names can be expressed as:

$$\mathrm{REP} = \frac{\frac{\mathrm{Red}+\mathrm{VNIR3}}{2}-\mathrm{VNIR}}{\mathrm{VNIR2}-\mathrm{VNIR}} \qquad (2.1)$$

where Red, VNIR, VNIR2 and VNIR 3 are reflectances at their respective wavelengths.

For Sentinel-2 mission known as Sentinel-2 red-edge position index (S2REP), this formula is presented as :

$$\mathrm{S2REP} = 705 + 35 \times \frac{\frac{\mathrm{b4}+\mathrm{b7}}{2}-\mathrm{b5}}{\mathrm{b6}-\mathrm{b5}} \qquad (2.2)$$

where b4, b5, b6 and b7 are reflectances at 665 nm, 705 nm, 740 nm and 783 nm, respectively. The normalized difference vegetation index used in this work utilizes bands 4 and 5 of the Sentinel-2 mission, and thus is designated the 'NDVI45' which is also a red-edge index. The NDVI45 was developed by Delegido *et al.* (2011) based on the NDVI, proposed by Rouse *et al.* (1974). The formula for the NDVI45 in terms of Sentinel-2 bands is:

$$\mathrm{NDVI45} = \frac{\mathrm{b5}-\mathrm{b4}}{\mathrm{b5}+\mathrm{b4}} \qquad (2.3)$$

Because spectral reflectances and indices are influenced by biophysical substances, the premise, as in all satellite optical remote sensing of vegetation, is that changes in the quantity and quality

of these substances are reflected by changes in the quantity and quality of the reflected radiation, which becomes a marker of biophysical change (Jones and Vaughan, 2010).

2.3.2 Machine Learning

The machine learning algorithm most widely used for quantifying forest attributes is the random forest (Astola *et al.* 2019; Čabravdić and Balić 2019; Bhattarai *et al.* 2020; Rahimzadeh-Bajgiran *et al.* 2020; Moradi *et al.* 2022). This algorithm is a member of the 'classification and regression tree' (CART) family of models, which is a brute force ensemble method. In many research papers it consistently outperforms other models in accuracy of predicted forest models, using the R^2 and RMSE metrics as the criteria (Astola *et al.* 2019; Čabravdić and Balić 2019; Bhattarai *et al.* 2020; Rahimzadeh-Bajgiran *et al.* 2020; Moradi *et al.* 2022). In this study several other machine learning algorithms also were evaluated such as boosted tree, extreme gradient boosting method (xgbDART), second extreme gradient boosting method (xgbTree) as well as model averaged neural network. Because these algorithms underperformed compared to random forest method, these algorithms were not exploited further. However more explanations and some of the results are presented in the next chapter.

In this thesis two sets of random forest models were constructed in the R language. All packages used, and the core R language, are cited in this work's bibliography. Modeling was performed using the 'caret' package (Kuhn 2022), which was developed as a wrapper for a myriad of machine learning algorithms and packages. One of the primary features of this thesis is the attempt to combine many large and disparate datasets into a single model that can be used effectively. One of the ironies of this project is that out of a very large dataset of 65 variables only 155 raster grid cells contained data for each variable, and only the observations for these 155 grid cells were used for modeling. At the recommendation of Professor Aaron Weiskittel, all

observations were used for training and no observations were reserved for validation or testing. After training, the models were tested on that same, full dataset. During training, ten-fold cross validation was used for each model training control. The use of ten-fold cross validation, which continuously recombines variables for training and validation within the training sequence, makes it possible to use all data for training and for testing.

2.3.2.1 The First Model Set

For this project a large dataset of 65 variables was obtained for analysis. In the first model set, also called the *all-variables model,* 155 observations of all variables were used as input into five iterations of the random forest model, with each observation corresponding to the values of a raster cell. All five iterations were run using the same train control function. This was necessitated by the fact that the random forest method is random, and because it is constructed randomly by the algorithm, different results can occur each time. However, the results of several model runs should generally converge, which is why five model runs were performed in this work.

2.3.2.2 The Second Model Set

After running the first model set, the VSURF algorithm (Bhattarai et al., 2022) was used to for variable selection and importance ranking. The variables selected by the random forest algorithm were:

1. Sentinel-2 band 3;
2. Exchangeable soil magnesium;
3. NAIP 2018 CHM;
4. Exchangeable soil potassium;
5. McNab roughness index;
6. Sentinel-2 red-edge position index (S2REP);
7. Water deficit for year 2020;

8. NDVI45;
9. Soil water index.

Because all the variables selected except for the McNab roughness index that were not optical data were soil and water data, it made sense to include soil hydrologic group as a tenth variable. Thus 155 observations of these ten variables were used to construct the second-model set, known as the select-variable model. It was theorized that the input simplification would make the random forest algorithm's results more accurate due to less noise and a clearer signal in the inputs.

To summarize, two sets of models were constructed and run:

1. The all-variables model;
2. The select-variables model.

3 CHAPTER THREE: RESULTS AND DISCUSSION

In this study sixty-five variables and five machine learning algorithms (random forest, boosted tree, extreme gradient boosting method (xgbDART), second extreme gradient boosting method (xgbTree) as well as neural network) were used to model forest growth. The random forest model performed the best and was evaluated in the all-variables model and select-variables model as presented below.

3.1 Results

After training the model set of 155 observations of 65 variables, two series of random forest models were fitted. The first series utilized all variables (Figure 3-1). For the second series, the VSURF algorithm was used to determine the most important variables and their relative contributions to the predictive capacity of the second model set (the select-variables model set). Table 3-1 ranks the relative contributions of each variable to the accuracy of the select-variables model. The nine variables which contributed the most to model efficacy can be divided into three groups: (1) optical data; (2) soil nutrient data; (3) hydrological data (even though soil hydrologic group was not important and could have been omitted).

The Sentinel-2 green band (b3) was found to be the most important variable of all, the Sentinel-2 red-edge position index had a variable importance score of 59.38%, and the NDVI45 had a variable importance score of 42.48%. When only Sentinel-2 optical variables in combination with NAIP 2018 CHM were used in the same setup as previously, the highest R^2 value achieved was approximately 0.32, with an associated RMSE of approximately 2 m and an MAE of approximately 1.39 m. When Sentinel-2 optical variables were used as the sole input variables

without the NAIP 2018 CHM, the best R^2 value was approximately 0.22 with an associated RMSE of 2.19 m and MAE of 1.50 m.

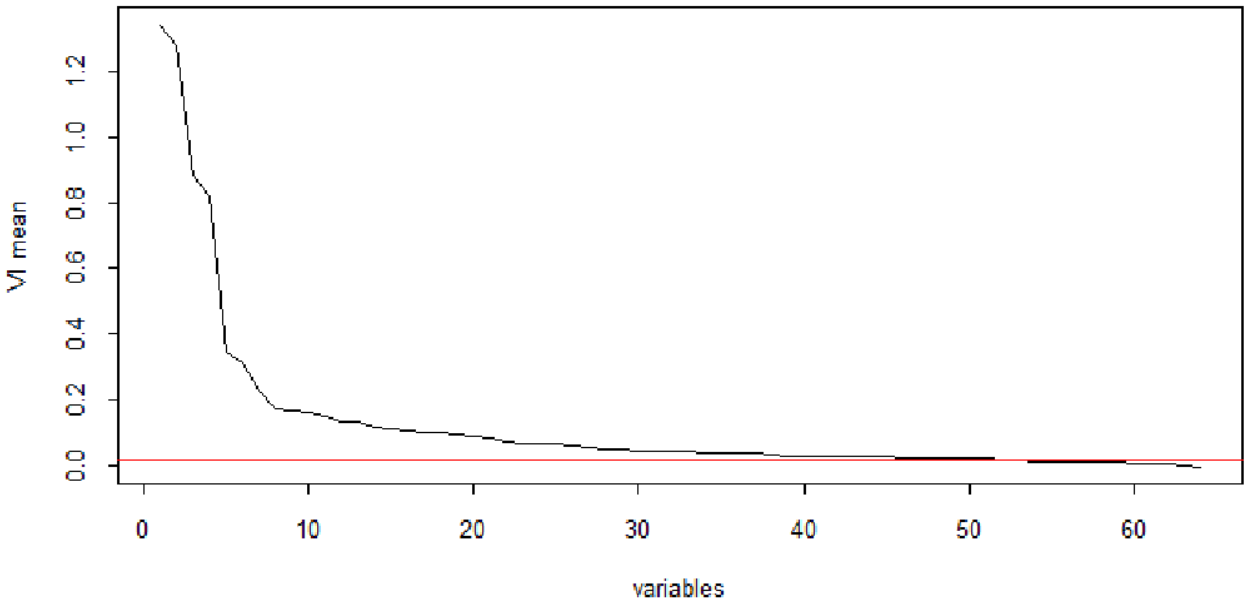

Figure 3-1: Variable Importance.

Table 3-1 Variable importance of the select-variables model as ranked by VSURF.

Variable	Importance Score
Sentinel-2 band 3	100.00
Exchangeable soil magnesium	99.63
NAIP 2018 CHM	93.59
Exchangeable soil potassium	91.25
McNab roughness index	65.16
Sentinel-2 red edge position index	59.38
Water deficit 2020	49.08
NDVI45	42.48
Soil water index	37.96
Hydrologic soil group	0.00

Each model series was run five times. One train control was used for all models regardless of series. Ten-fold cross validation, repeated ten times, was used in a hyperparameter grid search based on the caret default settings. Table 3-2 presents the results for both model series: the all-variables model and the select-variables model. The best result is highlighted. As seen from the table, the select-variables model performed the best. When the 35 most extreme values of ΔNAIP were removed and the model trained again on the remaining 120 samples, the best training R^2, on mtry = 3 and ntree = 65, was 0.49 with a training RMSE of 0.50 and a test RMSE of 0.20.

Table 3-2: Model results for the two model series.

Model Input	Min R^2	Associated RMSE	Associated MAE	Max R^2	Associated RMSE	Associated MAE
All-variables	0.42	1.86	1.20	0.49	1.87	1.21
Select-variables	0.47	1.78	1.11	0.56	1.63	1.04
Select-variables minus 35 most extreme values	--	--	--	0.49	0.50	0.38

Figure 3-2 is a map of the results of the select-variables model, indicating the predicted range of Δheight between NAIP 2018 and NAIP 2021. The values for the predicted difference fall within the predicted range, with the vast majority occurring between zero and five meters. Negative values may be attributed to errors in the model, errors in the data, or a reflection of tree cutting or forest

damage. Figure 3-3 displays the frequencies of the Δheight measurements obtained by the select-variables model. The mean ΔNAIP for the entire study area was 1.22

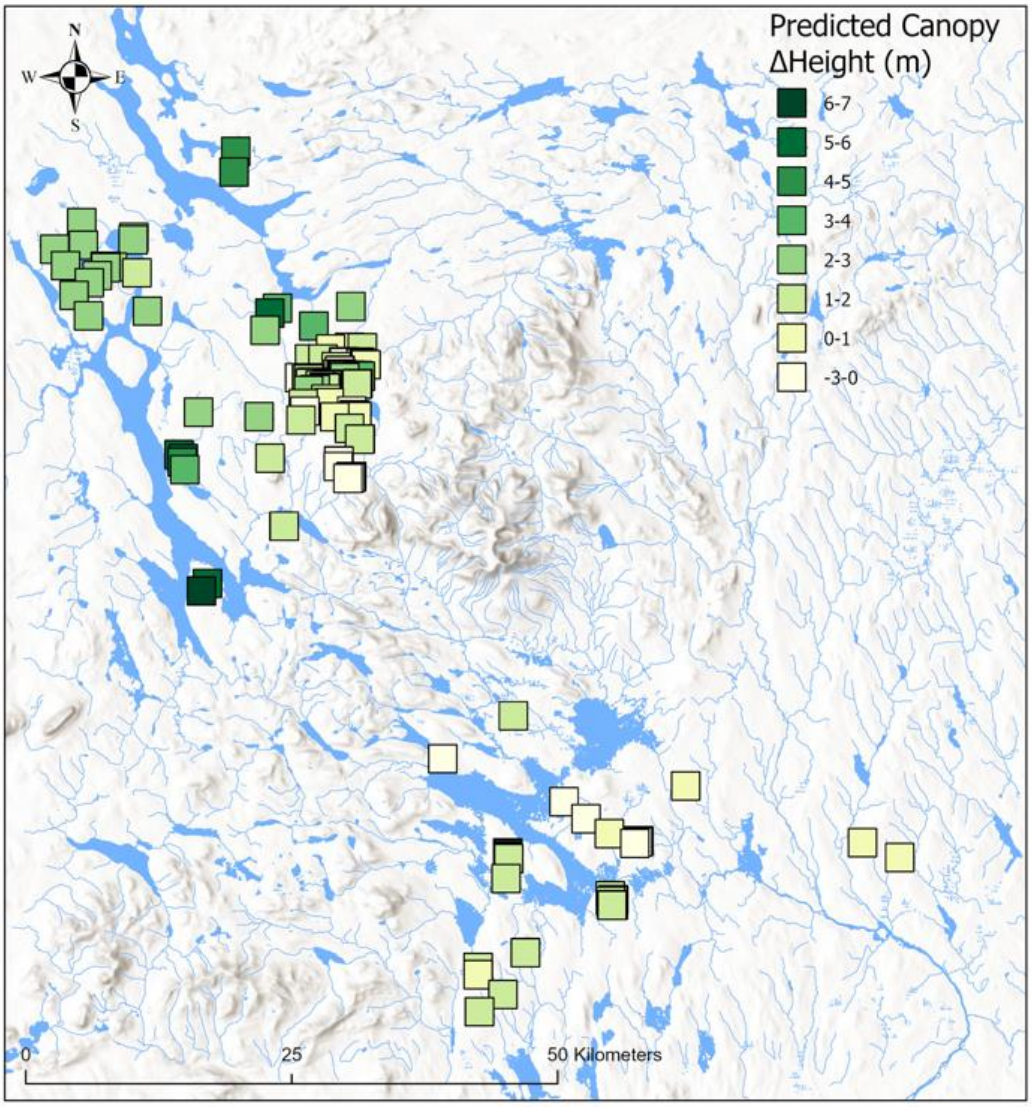

Figure 3-2: Predicted canopy Δheight (growth) based on the random forest select-variables model

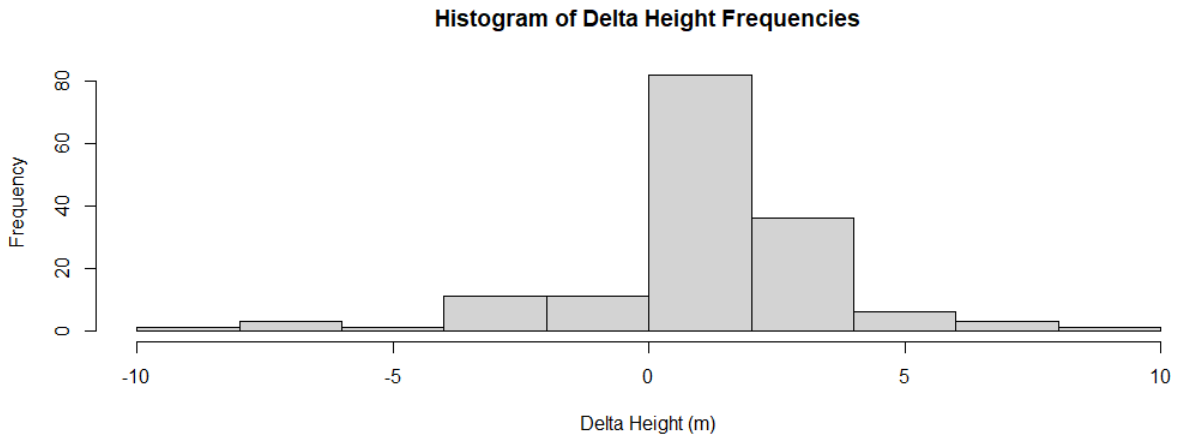

Figure 3-3: Frequency distribution of the Δheight measurements obtained by the select-variables model

The fact that most predictions fall roughly between one and two meters is expected, as this is within the likely tree growth values over a three-year period, and so suggests that the model has performed well. Figure 3-4 illustrates the absolute error of the select-variables model results.

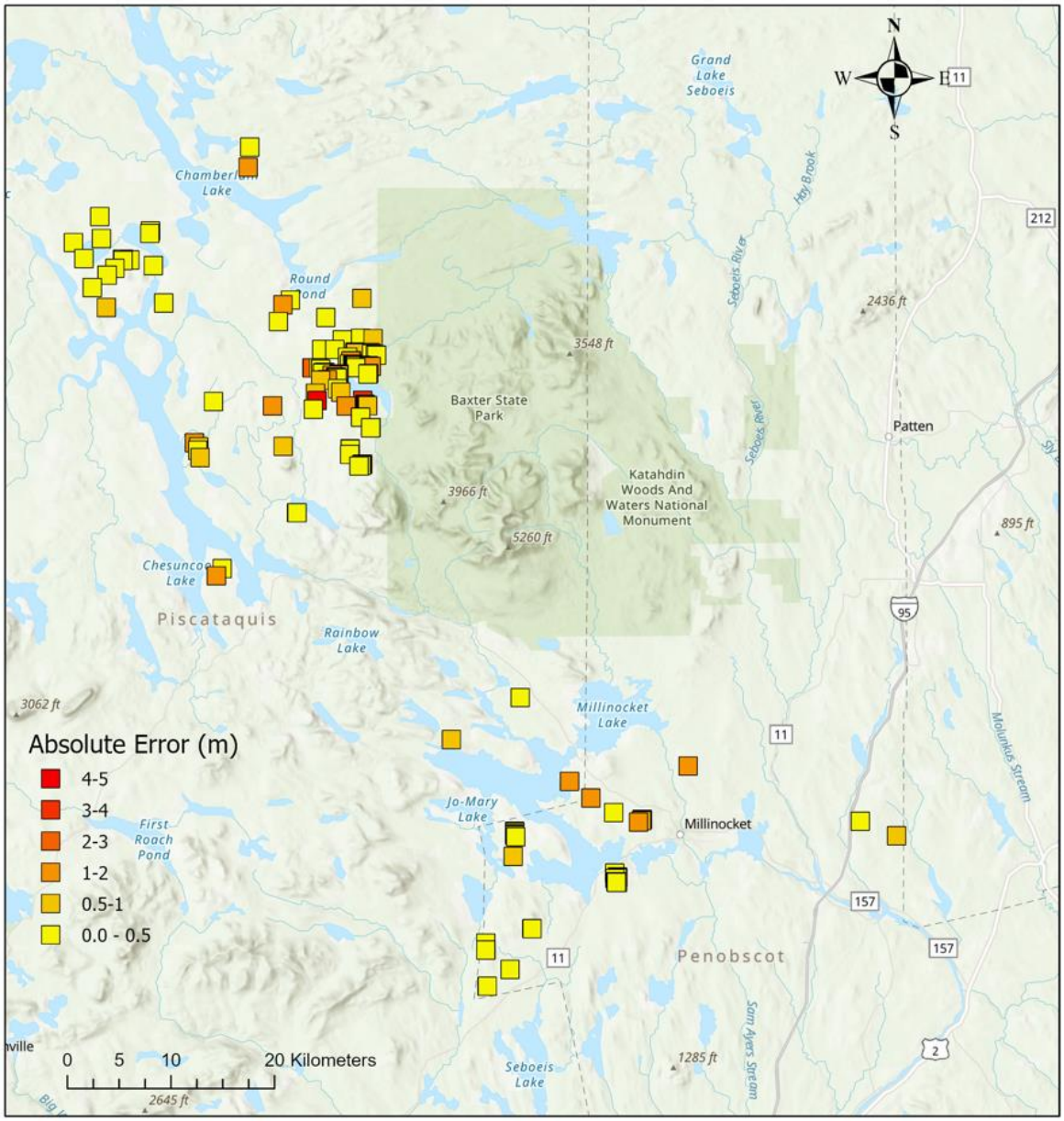

Figure 3-4: Absolute model error of the select-variables model.

Another indication that the model has performed well is, as shown in Figure 3-4, the fact that the majority of absolute error values were within zero to 0.5 meters, and the second largest category of error values was from 0.5 to one. The fact that there is a range of error values from 0.5 to five is actually an encouraging indication of the model's efficacy, as their presence suggests that the model was not overfit, which was the danger posed by using all observations for training.

One of the goals of this research was to compare the performance of the random forest model with that of other models. Four other algorithms were used to predict ΔNAIP values using the select-variables as inputs, and the results were compared with the random forest model using select variables. The four algorithms tested were: (1) model averaged neural network (Venables and Ripley 2002); (2) boosted tree (Hothorn *et al.* 2006a; Hothorn *et al.* 2006b; Zeileis *et al.* 2008; Strobel *et al.* 2007; Strobel *et al.* 2008); (3) conditional inference tree (Hothorn *et al.* 2006a; Hothorn *et al.* 2006b; Zeileis *et al.* 2008; Strobel *et al.* 2007; Strobel *et al.* 2008); and (4) extreme gradient boosting, using two different variants (Chen *et al.* 2023). These models were implemented using the caret (Kuhn 2022) package.

The best training R^2 results for the model averaged neural network was approximately 0.21 and the associated RMSE approximately 1.52 m. The best training R^2 results for the boosted tree method was approximately 0.26 and associated RMSE 1.31 m. The best training R^2 results for the conditional inference tree was approximately 0.23 and associated RMSE 2.19 m. The best training R^2 results for the first extreme gradient boosting method (xgbDART) was approximately 0.40 and associated RMSE 1.90 m. The best training R^2 results for the second extreme gradient boosting method (xgbTree) was approximately 0.35 and associated RMSE 1.86 m. Because these training results were significantly lower than with the random forest method, these algorithms were not exploited further. It is possible that with hyperparameter tuning and data reorganization they could achieve higher scores.

3.2 Discussion

This project builds on the work of Rahimzadeh-Bajgiran *et al.* (2020), who sought to model forest potential productivity at landscape scale with fine spatial resolution by increasing reliance on satellite remotely-sensed optical wavelengths and indices, and decreasing reliance on site variables, which are slow and expensive to obtain. Remotely-sensed optical data can include varying wavelengths of electromagnetic radiation as well as indices created from algebraic manipulations of those wavelengths. This technique can highlight biophysical characteristics of vegetation that otherwise may not be apparent. They used data from the Sentinel-2 sensor to obtain a map of two single spectral bands—b3 (green), b8a (near infra-red)—and two spectral vegetation indices based on the red-edge position—S2REP and NDVI45. The authors used these variables and inputs to a random-forest machine learning algorithm, and were able to detect differences in GSV with an R^2 value of 49%. This achievement has the potential to make GSV estimation much more efficient. The project described in this thesis uses the same optical variables in a more focused study area and with the addition of other site variables. The premise behind this approach is that, if Sentinel-2 optical variables can be combined with other variables that are easier to obtain than the data required for site indices, then the advantage of Sentinel-2 optical variables can be further extended. This project combined the Sentinel-2 variables with a very large dataset of other variables and compared the results.

Initially, 65 variables, including the four optical variables mentioned above, were used as inputs to a random forest model. The model was run five times using all variables, and the best result was an R^2 value of approximately 49%. When the same procedure was repeated with only the ten most important variables—which included the optical NDVI45, band three (green), and Sentinel-2 red-edge position index—the R^2 value increased to approximately 56%, indicating

that a random forest model with too many variables creates noise due to autocorrelation among variables (Rahimzadeh-Bajgiran et al., 2020; Bhattarai et al. 2022) which negatively impacts the accuracy of the results. It further indicates the high importance of optical variables for predicting growth similar to Rahimzadeh-Bajgiran et al., (2020). Although the top site variables identified in this study differ from those used in Rahimzadeh-Bajgiran et al. (2020) and Hennigar et al. (2017), they identified to be important for forest growth and productivity. Magnesium is a core element in chlorophyll molecule structure (REF) and regulates stomata conductance (Hirons and Thomas 2018). Water is essential in a delicate balance; either too much or too little water will be detrimental to vegetation growth. Therefore it was expected that variables related to water would be among the most important.

Optical spectral bands can be affected by confounding factors such as leaf area index (LAI) and leaf angle distribution (LAD), reducing their effectiveness for vegetation studies, so spectral indices have been developed that incorporate a difference between two closely-related bands in order to arrive at an objective ratio of reflectance rather than an absolute percentage (Jones and Vaughan 2010). The NDVI45 and S2REP are two such indices, whose values are attributable to the red-edge shift. The red-edge shift is the name used to describe the sharp increase in reflectance exhibited by green vegetation and 700 nm. It has been the premise of this thesis that changes in the red-edge shift correspond to changes in the volume of vegetation. It is already accepted that spectral indices are positively correlated with canopy density and other biophysical markers of vegetation health (Jones and Vaughan 2010). The R^2 value obtained by this study suggests that the premise is valid, but requires further refinement.

One possible avenue for improvement is to return to the beginning of satellite optical remote sensing, when estimates of productivity were closely tied to the volume of

photosynthetically active radiation (PAR) and, more specifically, to the fraction of absorbed photosynthetically active radiation (fAPAR). If the efficiency factor ε can be reliably calculated and multiplied by the volume of PAR multiplied by the fAPAR, (Monteith 1992; Gitelson *et al.* 2006) then a basic physical model could be joined with optical data, providing a stronger and more robust model. The inclusion of the 'radiation budget' combined with optical data as inputs for machine learning algorithms could help the algorithms to reach more accurate results. In the full raft of 65 input variables, the only radiation-related variable was that of net insolation. However, net insolation is not able to account for relative differences of productivity. Furthermore, the inclusion of species percentage (Bhattarai et al., 2022) would benefit such a model, as each species has a different characteristic value of ε. Although a basic map of species distribution was included into the raft of 65 variables, the data was not further processed to account for fraction of vegetation, and thus a corresponding value of ε. The 65 variables used in the initial runs of the model contained many hydrological and soil variables on the premise that these held the key to explaining variation in levels of productivity. However, Maine, USA, is apparently not a place where water is a significant limiting factor. Therefore, this study has made a contribution in illustrating that, in areas with high water availability, hydrological inputs do not yield useful results in combination with optical data.

The NAIP orthographic imagery data are a proven means of constructing a canopy height model (CHM) that is in regular use. Use of these data further increases the efficiency of the method of using remote sensing to quantify forest values as it reduces the reliance on site surveys and physical interactions with the environment.

4 CHAPTER FOUR: CONCLUSIONS, LIMITATIONS and FUTURE DIRECTIONS

From this work it is possible to conclude that optical satellite remote sensing and machine learning techniques have the potential to improve forest growing stock measurement, which can, in turn, significantly improve both economic and ecosystem services outcomes in an efficient manner. When combined with certain hydrologic and soil nutrient measurements, remotely sensed data can contribute to further increase model accuracy.

One limitation of this study was that due to the unavailability of the 2021 NAIP CHM data for most of Maine, we had to focus on a smaller study area, limiting the number of training samples that could be used for training and validation of the model. As these data will become available in near future, it can be recommended to run the same model for the entire state of Maine. Another limiting factor was that the site variables input dataset was very large and from heterogeneous sources, which likely affected the accuracy of the final model. Finally, this project considered a 3-year interval for forest growth as mandated by NAIP data availability. However, it is recommended to evaluate the model for a 5-year interval to understand if a larger growth period could affect model performance.

5 BIBLIOGRAPHY

Astola, Heikki *et al.* 2019. "Comparison of Sentinel-2 and Landsat 8 Imagery for Forest Variable Prediction in Boreal Region." *Remote Sensing of Environment* 223: 257-273. DOI: https://doi.org/10.1016/j.rse.2019.01.019

Bandyopadhyay, Debmita *et al.* 2017. "Red Edge Index as an Indicator of Vegetation Growth and Vigor Using Hyperspectral Remote Sensing Data." *Proceedings of the National Academies of Science, India, Section A: Physical Science* 87 (4): 879-888. DOI: https://doi.org/10.1007/s40010-017-0456-4

Beven, K.J., and Kirkby, M.J. 1979. A Physically Based Variable Contributing Area Model of Basin Hydrology. *Bulletin of the International Association of Scientific Hydrology* 24: 43-69. doi:10.1080/02626667909491834.

Bhattarai, Rajeev, Parinaz Rahimzadeh-Bajgiran, Aaron Weiskittel, and David A. MacLean. 2020. "Sentinel-2 Based Prediction of Spruce Budworm Defoliation Using Red-Edge Spectral Vegetation Indices." *Remote Sensing Letters* 11 (8): 777-786. DOI: https://doi.org/10.1080/2150704X.2020.1767824

Bolstad, P.V., Swank, W., and Vose, J. 1998. Predicting Southern Appalachian Overstory Vegetation with Digital Terrain Data. *Landscape Ecology* 13: 271-283. doi:10. 1023/A:1008060508762.

Brungard, C. and Hennigar, C. 2022. Interdisciplinary Spatial Modeling of Terrain, Wetness Soils, and Productivity: New Tools for Forest Management. In: Smith, R. K., (Ed) 2022. *Cooperative Forestry Research Unit: 2022 Annual Report.* Center for Research on

Sustainable Forestry, University of Maine. Orono, ME. p 35-40. https://umaine.edu/cfru/wp-content/uploads/sites/224/2023/04/AR-2022-PDF.pdf.

Bulut, S., and A. Günlü. 2019. "Determination of Total Carbon Storage Using Sentinel-2 and Geographic Information Systems in Mixed Forests." *Anatolian Journal of Forest Research* 5 (2): 127-135. DOI: NA.

Čabravdić, Azra and Besim Balić. 2019. "Modelling Stand Variables of Beech Coppice Forest Using Spectral Sentinel-2A Data and the Machine Learning Approach. *Southeast Eurasian Forestry* 10 (2): 137-144. DOI: https://doi.org/10.15177/seefor.19-21

Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; Li, M.; Xie, J.; Lin, M.; Geng, Y.; Li, Y.; Yuan, J. 2023. xgboost: Extreme Gradient Boosting. R package version 1.7.5.1, <https://CRAN.R-project.org/package=xgboost>.

Chrysafis, Irene *et al.* 2017. "Assessing the Relationships between Growing Stock Volume and Sentinel-2 Imagery in a Mediterranean Forest Ecosystem." *Remote Sensing Letters* 8 (6): 508-517. DOI: http://dx.doi.org/10.1080/2150704X.2017.1295479

Crow, Thomas. 2006. "Moving to the Big Picture: Applying Knowledge From Landscape Ecology to Managing U.S. National Forests." *Forest Landscape Ecology: Transferring Knowledge to Practice,* 157-180. DOI: https://doi.org/10.1007/978-0-387-34280-1_7

Dalen, Marilyn *et al.* "Relationship of Red and Red-Edge Reflectance-Based Vegetation Indices with Stalk and Fiber Yield of Energy Cane Harvested at Different Dates." *Archives of Agronomy and Soil Science* 66 (13): 1888-1907. DOI: https://doi.org/10.1080/03650340.2019.1701658

Delegido, J. *et al.* 2011. "Evaluation of Sentinel-2 Red-Edge Bands for Empirical Estimation of Green LAI and Chlorophyll Content. *Sensors* 11: 7063-7081. DOI: https://doi.org/10.3390/s110707063

De Reu, J., Bourgeois, J., Bats, M., Zwertvaegher, A., Gelorini, V., De Smedt, P., *et al.* 2013. Application of the Topographic Position Index to heterogeneous landscapes. Geomorphology, 186: 39-49. doi:10.1016/j.geomorph.2012.12.015

Donmez, Cenk, *et al.* 2011. "Modelling the Current and Future Spatial Distribution of NPP in a Mediterranean Watershed." *International Journal of Applied Earth Observation and Geoinformation* 13, 336-345. DOI: https://doi.org/10.1016/j.jag.2010.12.005

Fleischer, Walter E. 1934. "The Relation Between Chlorophyll Content and Rate of Photosynthesis." *Journal of General Physiology* 18 (4): 573-597. DOI: https://doi.org/10.1085/jgp.18.4.573

FAO UN. 2022. *The State of the World's Forests: Forest Pathways for Green Recovery and Building Inclusive, Resilient and Sustainable Economies.* Rome. Web address: https://www.fao.org/documents/card/en/c/cb9360en

Fleischer, Walter E. 1934. "The Relationship Between Chlorophyll Content and Rate of Photosynthesis." *The Journal of General Physiology.* DOI: NA

Foody, Giles M., and Paul J. Curran. 1994. "Estimation of Tropical Forest Extent and Regenerative Stage Using Remotely Sensed Data." *Journal of Biogeography* 21 (3): 223-244. Web access: https://www.jstor.org/stable/2845527

Forkuor, Gerald *et al.* 2020. "Above-Ground Biomass Mapping in West African Dryland Forest Using Sentinel-1 and 2 Datasets: A Case Study." *Remote Sensing of Environment* (236): 1-15. DOI: https://doi.org/10.1016/j.rse.2019.111496

Genuer, R., Poggi J, Tuleau-Malot C. 2022. *VSURF: Variable Selection Using Random Forests.* R package version 1.2.0, <https://CRAN.R-project.org/package=VSURF>.

Gitelson, Anatoly A. *et al.* 2006. "Relationship Between Gross Primary Production and Chlorophyll Content in Crops: Implications for the Synoptic Monitoring of Vegetation Productivity." *Journal of Geophysical Research* 111: 1-13. DOI: https://doi.org/10.1029/2005JD006017

Global Wind Atlas 3.0. A free, web-based application developed, owned, and operated by the 565 Technical University of Denmark (DTU). The Global Wind Atlas 3.0 is released in partnership with the World Bank Group, utilizing data provided by Vortex, using funding provided by the Energy Sector Management Assistance Program (ESMAP). For additional information: https://globalwindatlas.info

GPP, NPP and Respiration. Government of Wales, United Kingdom. Website accessed 06 July 2023: http://resources.hwb.wales.gov.uk/VTC/env-sci/w23_id_resp_npp.htm

Guyot, G., F. Baret and S. Jacquemoud. 1992. "Imaging Spectroscopy for Vegetation Studies." In: *Imaging Spectroscopy: Fundamentals and Prospective Applications,* edited by: F. Toselli and J. Bodechtel: 145-165.

Hargraves, G.H. and Z.A. Samani. 1985. Reference Crop Evapotranspiration from Temperature, *Applied Engineering in Agriculture* 1 (2): 96-99.

Hennigar, C., A. Weiskittel, H. L. Allen, D. A. McLean. 2017. "Development and Evaluation of a Biomass Increment-Based Index for Site Productivity." *Canadian Journal of Forest Research* 47, 400-410. DOI: https://doi.org/10.1139/cjfr-2016-0330

Hijmans R. 2022. *terra: Spatial Data Analysis.* R package version 1.6-17, <https://CRAN.R-project.org/package=terra>.

Hirons, Andrew D., and Peter A. Thomas. 2018. *Applied Tree Biology.* Oxford: Wiley-Blackwell.

Ho, Tin Kam. 1995. "Random Decision Forests." *Proceedings of the Third International Conference on Document Analysis and Recognition* Montreal, QC, Canada, 14-16 August 1995: 278-282.

Hollister J., Shah T., Robitaille A., Beck M., and M. Johnson. 2021. elevatr: Access Elevation Data from Various APIs. doi:10.5281/zenodo.5809645, R package version 0.4.2, https://github.com/jhollist/elevatr/.

Hothorn, T.; Hornik, K; Zeileis, A. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15 (3), 651-674. DOI: https://doi.org/10.1198/106186006X133933

Hothorn, T.; Buehlmann, P.; Dudoit, S.; Molinaro, A.; Van Der Laan, M. 2006. "Survival Ensembles." *Biostatistics* 7 (3): 355-373.

Hothorn, T,, P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner. 2022. mboost: Model-Based Boosting, R package version 2.9-7, https://CRAN.R-project.org/package=mboost.

Hu, Yang *et al.* 2020. "Estimating Forest Stock Volume in Hunan Province, China, by Integrating In Situ Plot Data, Sentinel-2 Images, and Linear and Machine Learning Regression Models." *Remote Sensing* 12 (186): 1-23. DOI: https://doi.org/10.3390/rs12010186

Imran, A. B., *et al.* 2020. "Narrow Band Based and Broadband Derived Vegetation Indices using Sentinel-2 Imagery to Estimate Vegetation Biomass." *Global Journal of Environmental Science and Management* 6 (1): 97-108. DOI: https://doi.org/10.22034/gjesm.2020.01.08

Jones, H.G. and Vaughan, R.A., 2010. Remote Sensing of Vegetation: Principles, Techniques, and Applications. Oxford University Press, USA.

Kanke, Yumiko *et al.* 2012. "Red Edge as a Potential Index for Detecting Differences in Plant Nitrogen Status in Winter Wheat." *Journal of Plant Nutrition* 35 (10): 1526-1541. DOI: https://doi.org/10.1080/01904167.2012.689912

Kuhn M. 2022. *caret: Classification and Regression Training.* R package version 6.0-93, <https://CRAN.R-project.org/package=caret>.

Liaw, A. and M. Wiener. 2002. Classification and Regression by randomForest. R News 2(3), 18-22.< https://cran.r-project.org/web/packages/randomForest/index.html>

Lin, Shangrong *et al.* 2019. "Evaluating the Effectiveness of Using Vegetation Indices Based on Red-Edge Reflectance from Sentinel-2 to Estimate Gross Primary Productivity." *Remote Sensing* 11: 1-25. DOI: https://doi.org/10.3390/rs11111303

Monteith, J. L. 1972. "Solar Radiation and Productivity in Tropical Ecosystems. *The Journal of Applied Ecology* 9: 747-766. DOI: https://doi.org/10.2307/2401901

MODIS GPP/NPP PROJECT (MOD17). University of Montana. Website accessed 06 July 2023: https://www.umt.edu/numerical-terradynamic-simulation-group/project/modis/mod17.php

Moradi, Fardin *et al.* 2022. "Estimating Aboveground Biomass in Dense Hyrcanian Forests by the Use of Sentinel-2 Data." *Forests* 13 (1): 1-18. DOI: https://doi.org/10.3390/f13010104

Natural Resource Conservation Service (NRCS) Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys. 2nd ed. 1999. United States Department of Agriculture. Agriculture Handbook 436. Washington DC.

Obata, Shingo, *et al.* 2021. "Random Forest Regression Model for Estimation of the Growing Stock Volumes in Georgia, USA, Using Dense Landsat Time Series and FIA Dataset." *Remote Sensing* 13 (218): 7-24. DOI: https://doi.org/10.3390/rs13020218

Peng, Fang. 2021. "Aboveground Biomass Mapping of Crops Supported by Improved CASA Model and Sentinel-2 Multispectral Imagery." *Remote Sensing* 13 (2755): 1-26. DOI: https://doi.org/10.3390/rs13142755

Prior, Elizabeth M., Valerie A. Thomas, and Randolph H. Wynne. 2022. "Estimation of Mean Dominant Height Using NAIP Digital Aerial Photogrammetry and LiDAR Over Mixed Deciduous Forest in the Southeastern USA. *International Journal of Applied Earth Observation and Geoinformation* 110: 1-12. DOI: https://doi.org/10.1016/j.jag.2022.102813

PRISM Climate Group, Oregon State University. https://prism.oregonstate.edu, data created 16 March 2013, accessed 15 April 2023.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahimzadeh-Bajgiran, Parinaz, Masashi Munehiro, and Kenji Omasa. 2012. "Relationships Between the Photochemical Reflectance Index (PRI) and Chlorophyll Fluorescence Parameters and Plant Pigment Indices at Different Leaf Growth Stages." *Photosynthesis Research* 113: 261-271. DOI: https://doi.org/10.1007/s11120-012-9747-4

Rahimzadeh-Bajgiran, Parinaz, Chris Hennigar, Aaron Weiskittel and Sean Lamb. 2020. "Forest Potential Productivity Mapping by Linking Remote-Sensing-Derived Metrics to Site Variables." *Remote Sensing* 12 (12): 1-17. DOI: https://doi.org/10.3390/rs12122056

Rees, Gareth W. *et al.* 2021. Estimation of Boreal Forest Growing Stock Volume in Russia from Sentinel-2 MSI and Land Cover Classification." *Remote Sensing* 13 (4483): 1-17. DOI: https://doi.org/10.3390/rs13214483

Rouse, J. *et al.* 1974. "Monitoring Vegetation Systems in the Great Plains with ERTS." *NASA Special Publication.*

Rowell, Roger M. *et al.* 2012. *Handbook of Wood Chemistry and Wood Composites.* Routledge: Princeton NJ.

Running, Steven W. *et al.* 2004. "A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production." *Bioscience* 54 (6): 547-560. DOI: NA

Schroeder, Todd A. *et al.* 2022. "Evaluating Statewide NAIP Photogrammetric Point Clouds for Operational Improvement of National Forest Inventory Estimates in Mixed Hardwood Forests of the Southeastern US." *Remote Sensing* 14 (17): 1-29. DOI: https://doi.org/10.3390/rs14174386

Schumacher, Johannes *et al.* 2019. "Combination of Multi-Temporal Sentinel 2 Images and Aerial Image Based Canopy Height Models for Timber Volume Modelling." *Forests* 10 (746): 1-19. DOI: https://doi.org/10.3390/f10090746

Sims, Daniel A., *et al.* 2008. "A New Model of Gross Primary Productivity for North American Ecosystems Based Solely on the Enhanced Vegetation Index and Land Surface Temperature from MODIS." *Remote Sensing of Environment* 112: 1633-1646. DOI: https://doi.org/10.1016/j.rse.2007.08.004

Sharma, Lakesh K, *et al.* 2015. "Active-Optical Sensors Using Red NDVI Compared to Red Edge NDVI for Prediction of Corn Grain Yield in North Dakota, U.S.A." *Sensors* 15: 27832-27853. DOI: https://doi.org/10.3390/s151127832

Smith, P., and Metcalfe, P. 2018. dynatop: an implementation of the dynamic TOPMODEL Hydrologic Model in R. Available from https://cran.r-project.org/web/packages/dynatop/index.html.

Strobl, C.; Boulesteix, A., Zeileis, A., Hothorn, T. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8 (25). DOI: https://doi.org/10.1186/1471-2105-8-25

Strobl, C.; Boulesteix, A.; Kneib, T; Augustin, T; Zeileis, A. 2008. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics* 9 (307). DOI: https://doi.org/10.1186/1471-2105-9-307

Venables, W. N. and Ripley, B. D. 2002. *Modern Applied Statistics with S.* Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1-29. URL: https://www.jstatsoft.org/v40/i01/

Wu, Chaoyang *et al.* 2009. "Remote Estimation of Gross Primary Production in Wheat Using Chlorophyll-Related Vegetation Indices." *Agricultural and Forest Meteorology* 149, 1015-1021. DOI: https://doi.org/10.1016/j.agrformet.2008.12.007

Zeileis, A.; Hothorn, T.; Hornik, K. 2008. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics* 17 (2): 492-514. DOI: https://doi.org/10.1198/106186008X319331

Zhao, Maosheng *et al.* 2005. "Improvements of the MODIS Terrestrial Gross and Net Primary Production Global Data Set." *Remote Sensing of Environment* 95 (2): 164-176. DOI: http://dx.doi.org/10.1016/j.rse.2004.12.011

BIOGRAPHY OF THE AUTHOR

Peter G. Larson is from Hampden, Maine. He began his MS in Ecology and Environmental Science in 2020. In 2022 he began working with Professor Parinaz Rahimzadeh-Bajgiran. He also served as Teaching Assistant for GIS and participated in field work. Peter is a candidate for the Master of Sciences from the University of Maine in December 2023.