Master's thesis

Master's Programme in Data Science

Estimating Above-Ground Biomass in Finnish Forests Using Remote Sensing Data

Bearjadat Valkama

June 14, 2022

Supervisor(s): Associate Professor Laura Ruotsalainen

Examiner(s): Associate Professor Laura Ruotsalainen

Associate Professor Arto Klami

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Bearjadat Valkama

Työn nimi — Arbetets titel — Title

Estimating Above-Ground Biomass in Finnish Forests Using Remote Sensing Data

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	June 14, 2022	72

Tiivistelmä — Referat — Abstract

Above-ground biomass (AGB) estimation is an important tool for predicting carbon flux and the effects of global warming. This study describes a novel application of remote-sensing based AGB estimation in the hemi-boreal vegetation zone of Finland, using Sentinel-1, Sentinel-2, ALOS-2 PALSAR-2, and the Multi-Source National Forest Inventory by Natural Resources Institute Finland as sources of data. A novel method of extracting data from the features of the surrounding observations is proposed, and the method's effectiveness was evaluated. The findings showed that the method showed promising results, with the model trained using the extracted features achieving the highest evaluation scores in the study. In addition, the viability of using free and highly available satellite datasets for AGB estimation in the hemi-boreal Finland was analyzed, with the results suggesting that the free Synthetic Aperture Radar (SAR) based products had a low performance. The features extracted from the optical data of Sentinel-2 produced well-performing models, although the accuracy might still be too low to be feasible.

ACM Computing Classification System (CCS):
Social and professional topics → Professional topics → Computing industry → Sustainability
Applied computing → Document management and text processing → Document management → Text editing

Avainsanat — Nyckelord — Keywords

layout, summary, list of references

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

Carbon has recently received a lot of attention as a major driving force behind global warming [78]. While it is not alone to blame for the changing climate, it is the major greenhouse gas and has a relatively long lifetime in the atmosphere. As much as 60% of the global warming effect has been attributed to carbon.

Carbon follows a natural cycle on earth. In the atmosphere, it mainly exists as the gas carbon dioxide [78]. Plants use carbon dioxide during photosynthesis and convert it into carbohydrate, while simultaneously releasing oxygen into the atmosphere. When the plants die and decay or are burnt, carbon is released back to the atmosphere.

Lately, the balance of the carbon cycle has been disrupted [78]. This is largely due to actions by the human population, such as widespread deforestation. As a result, the amount of carbon in the atmosphere has been steadily increasing. This in turn contributes to the greenhouse-effect, causing global climate warming. Consequently, preserving, monitoring, and analyzing the plant ecosystems — and the carbon stored within them — plays a crucial role in actions against climate change.

While carbon is stored in various different ways, the Intergovernmental Panel on Climate Change (IPCC) separates terrestrial carbon pools into five categories: above-ground biomass, below-ground biomass, litter, woody debris, and soil organic matter [78]. Out of these carbon pools, above-ground biomass (AGB) is generally considered as one of the most important. Above-ground biomass has been estimated to include 70 to 90% of total forest biomass [7], and around 30% of total terrestrial ecosystem carbon pool [42].

In addition, above-ground biomass is the most dynamic and manipulable of the various carbon pools [42]. While, for example, organic matter in soil contains two to three times the amount of carbon on global scale compared to above-ground biomass, the carbon in soil is much more protected and less prone to be perturbed and released back into the atmosphere. Above-ground biomass is in a state of constant change due to various different factors, such as logging, fires, storms, and changes in land use. Consequently, estimating above-ground biomass and its changes is crucial for modeling carbon flux, which allows for more accurate projections and estimates of climate change and its various impacts around the globe.

Above-ground biomass estimations are used in various initiatives [42], such as Reducing Emissions from Deforestation and Forest Degradation (REDD) and REDD+ [98]. Both are frameworks proposed by the United Nations Framework Convention on Climate Change which aim to mitigate climate change by various means, including various financing schemes and incentives to reduce deforestation and forest degradation. REDD and REDD+ require the participating nations to report — among other things — estimates of forest carbon stocks and the changes to them, and above-ground biomass estimation is an integral part of that.

Around the world, biomass is also used as a fuel [42]. It is a relatively popular source of energy, partly due to biomass being entirely renewable. Accurate estimation models are useful for the purposes of the energy industry, as the quality and location of the biomass greatly affect its feasibility as a source of energy.

While there is a great need for above-ground biomass estimation, the task itself has its difficulties [48]. Vast amounts of vegetation in both spatial and temporal sense needs to be measured in order to fulfill the variety of needs for AGB estimations. In addition, constant monitoring and measurements are required due to the dynamic nature of above-ground biomass. And the more accurate, timely, and accessible the estimations are, the more certain and useful the various resulting assessments — such as carbon stock modeling — can be.

Above-ground biomass estimation has seen a surge of new research in recent years [48]. This is largely due to the increased interest towards the results, but also partly because of advancements of computation methods and capacity. In particular, above-ground biomass estimation methodology relating to the domain of remote sensing has risen in popularity, which could partially be attributed to an increased availability of various remote sensing datasets [89]. The financial requirements of launching new satellites has been steadily going down, while various organizations — such as European Space Agency (ESA) — offer free and open access to relatively high accuracy remote sensing data.

This thesis describes a novel application of machine-learning based above-ground biomass estimation in the hemi-boreal vegetation zone of Finland. Part of the motivation behind this research was the lack of flexible, high accuracy, and low-cost applications of remote sensing based AGB estimation in the hemi-boreal vegetation zone. While biomass maps exist for the area [3], they are usually either outdated or have a low spatial resolution.

There is a similar problem with recent and upcoming biomass-focused satellites. NASA's Global Ecosystem Dynamics Investigation [94] (GEDI) uses narrow-swath LiDAR to get measurements, and the coverage of the highest-resolution product is very low. It is also designed to operate only up to 51.6° N latitude, meaning most of North-

ern Europe is out of its range. ESA's BIOMASS mission [93] (to be launched in 2023) uses a relatively low resolution of 200 m, rendering it unsuitable for estimating AGB accurately in small forest areas.

For Finnish forests, high-quality ground truth data is freely available [60, 51]. This is not the case for many other areas. While the quality and accessibility of the ground truth data limits the usefulness of the remote-sensing based AGB estimation in Finnish forest areas, the resulting methodology is hopefully applicable in other areas with similar vegetation, such as Estonia.

The aim of the research is to create a methodology that produces accurate AGB estimations, while also being flexible in both spatial and temporal sense. While research in the domain and public biomass products generally aim for large-scale estimation of lower accuracy, one of the goals of this research is to be able to create a methodology that can be used ad-hoc to up-to-date, accurate estimations for even small forest areas.

The methodology used for data processing, feature extraction, and feature selection is developed with the intention of being easily adaptable to areas apart from the sample region. While the trained models are likely to perform well only within the neighboring hemi-boreal zone, the data and feature processing methods should allow for prototyping and training of fitting models in any area of choice, as long as there is data available. The remote sensing data itself is sourced from open and free-to-use satellite products. The utilized satellite missions Sentinel-1 [99], Sentinel-2 [100], and ALOS-2 [90] are all ongoing and offer global coverage of recent data, although Sentinel-1 and Sentinel-2 are mostly focused on Europe.

This work also proposes a novel methodology for feature extraction. The data values of the encircling area (including vegetation indices and textural measures) for each of the observations are utilized as additional features. While there is research on using textural analysis to extract patterns from the surrounding observations for AGB estimation, to my knowledge there are no recent studies using the extracted features (e.g., spectral and polarization bands, vegetation indices, or textural metrics) of the neighboring data points as additional features.

As remote-sensing based AGB estimation tends to underestimate biomass values in areas with dense vegetation and high biomass, the hypothesis is that using the neighboring features might be able to counter this saturation effect. This is combined with multi-seasonal data-sourcing, as the characteristics of vegetation in hemi-boreal mixed forests, especially for the deciduous trees, depend largely on the period of the year. A goal of the study is to evaluate whether using the neighboring features, in conjunction with multi-seasonal data, might improve the AGB estimation accuracy of the trained models — especially for high-biomass areas.

This thesis is divided into two parts. The first part — primarily Chapter 2 —

contains a brief overview of the statistical and remote sensing -based methods used in estimating above-ground biomass. Forest-related methodology is the primary focus, as forests are the most wide-spread target of above-ground biomass-estimation, with a considerable impact on global climate [42, 48].

In the second part of the thesis — Chapters 3, 4, and 5 — the novel application of AGB estimation, and the methodology of extracting features from neighboring observations, are described and the results discussed. The research consists of multi-source remote sensing approach utilizing spectral bands, vegetation indices, and textural measures from both optical and SAR data.

In Chapter 3, the methodology of the research is described. The chapter contains a description of the algorithms and processes used for feature extraction and selection, including the novel method of adding features from the neighboring data points. In addition, there is a brief overview of the utilized machine learning model, various scoring formulas, and data preparation methods. The details of the data sources (Sentinel-1 [99], Sentinel-2 [100], ALOS-2 PALSAR-2 [90], and the Multi-Source National Forest Inventory of Finland [96]) are also outlined.

The results of the research are detailed in Chapter 4. A comparison of various combinations of features and data sources is provided, with the primary focus being on the performance of the models using neighboring features as additional data. The results of models are compared for general areas, and for high-biomass areas. Furthermore, the sample areas are shown.

In Chapter 5, the findings are discussed. The flaws and advantages of the methodology are analyzed, as well the implications of the results. In addition, possibilities for further research are briefly speculated on. And finally, in Chapter 6 is a brief summary of the thesis.

2. Forest Above Ground Biomass Estimation

Specifically, forest above-ground biomass consists of organic matter in living and dead plant materials [48]. This includes, for example, branches, leaves, and stem, and there are multiple approaches to estimating the amount of above-ground biomass. Broadly, these approaches can be divided into destructive and non-destructive methods [78]. In destructive methods, the trees are cut down and weighed, while non-destructive methods leave the forests intact. Destructive sampling is costly and invasive, and in certain ecosystems completely unviable due to threatened flora or fauna [42]. But, while having numerous downsides, destructive methods yield by far the most accurate measurements of biomass [6]. As such, they are useful for validating and developing non-destructive methods.

2.1 Allometric Equations

Allometric equations are one of the major ways of estimating biomass. Allometric equations are statistical models commonly based on easily collected biophysical properties of a tree [42, 6, 80]. These biophysical properties might include variables such Diameter at Breast Height (DBH), Commercial Bole Height (CBH), tree height, and wood density. The aim is to create equations that capture the scaling between a tree's form and biomass. Biomass in this context can either be an individual component — such as root, bark, bole, or needle biomass — or the total biomass.

An example of an approach for allometric AGB estimation is to first create linear models for each measured variable x for each of the n sampled trees [63]. A single model can be formulated as

$$y_i = bx_i + e_i \ , \tag{2.1}$$

where y_i is the biomass component y for the tree i, b is the vector of a fixed effect parameter for the biomass component, x_i is the vector of independent variables, and e_i is the residual error. The sampling can be improved by using multiple stands, and adding an error parameter u_k for the models, where the stand is denoted as k. The linear models can then be compiled into a multivariate model, which can be formulated as

$$\begin{aligned} y_{1ki} &= b_1 x_{2ki} + u_{1i} + e_{1ki} \\ y_{2ki} &= b_2 x_{2ki} + u_{2i} + e_{1ki} \\ &\vdots \\ y_{nki} &= b_n x_{nki} + u_{ni} + e_{nki} \ , \end{aligned} \tag{2.2}$$

where y_{1ki} to y_{nki} denotes the dependent variables of biomass for each component from 1 to n for tree i in stand k. Finally, the following is an example of an allometric equation for total AGB of scots pine derived from the multivariate model [63]:

$$ln(y_{ki}) = b_0 + b_1 \frac{d_{Ski}}{(d_{Ski} + 12)} + b_2 \frac{h_{ki}}{(h_{ki} + 20)} + u_k + e_{ki} \ . \tag{2.3}$$

To transform the equation into a linear form in addition to obtaining homoscedasticity of the variance, logarithmic transformation was used. The model uses mainly two variables, height h and stump diameter d_S. The stump diameter is defined as $d_S = 2 + 1.25d$, where d is the diameter at breast height.

Allometric equations have many sources of uncertainty. As biomass data is costly and invasive to collect via destructive methods, alternative sampling is often used [82]. This alternative sampling is generally based on either biomass estimation from volume, biomass estimation from predictor variables (such as DBH, CBH, or tree height), or a combination of both. The results from these sampling methods are rarely as accurate as actually felling and weighing a tree.

In addition, these allometric equations are often single tree models that assume that the sampled specimen are representative of the larger population [82]; Selecting the wrong individual trees as the sample set is going to skew the results of the model, as the specific properties (such as density or size) might be outliers or generally have large deviations. These models have limited use in heterogeneous forests with trees of various sizes, and are most accurate in environments with stands of similar age, such

as plantations or uniform forests.

This is further complicated by the fact that allometric equations are often applied to populations of trees that are different from the population the equations were developed for [80]. Even if the tree species and genetics are equal or very similar, growing in various conditions (e.g., climate, soil type, competition, tree age, availability of nutrients) affects the various properties of the trees. As a result, there might be significant amounts of error, in some cases even as much as 240% [82].

Allometric equations can also be used to estimate below ground biomass (BGB) [29]. While BGB is also an important variable for accurate modeling of carbon and total biomass, it is also considered much more difficult to estimate than AGB [29, 7]. As a consequence, various methods to get BGB estimations from above-ground characteristics exist. One of the more popular methods is the root-to-shoot ratio (R:S) [29], which is a relatively accurate way to estimate the amount of below-ground biomass from the number of above-ground tree shoots [81]. Above-ground biomass estimation, however, is both easier and usually more important than below-ground biomass estimation, and as such gets generally more focus in research [29].

Remote sensing biomass estimation methods often use allometric equations for validation or as training data. As this is the case, it is important to keep in mind that allometric equations themselves are very much predictive — with a high level of uncertainty — and not the hard truth. Partially due to this, the quality of results from many of the remote sensing estimation methods should be appraised with care.

2.2 Remote Sensing

Remote sensing is a category of methods where the physical characteristics of an area are measured at a distance — generally from above by an airborne or a spaceborne device — by the emitted and reflected radiation [89]. For the purposes of above-ground biomass estimation, the common remote sensing technologies can be divided into LiDAR, optical, and SAR [48]. All of these measure electromagnetic radiation, mainly visible light, ultraviolet, or low to medium infrared frequencies (see Fig. 2.1).

Electromagnetic radiation can interact with an object in three distinct ways: reflection, absorption, and transmission [89]. Objects generally interact with electromagnetic radiation in all three of these simultaneously, only the proportions vary. The deciding factor behind proportions is the wavelength of the radiation and the material and structure of the object. When it comes to above-ground biomass estimation, only some of the wavelengths are suitable. The chosen wavelengths need to interact with the vegetation and other objects in an environment in such a manner, that it is possible to differentiate — with sufficient accuracy — biological mass from the non-biological.

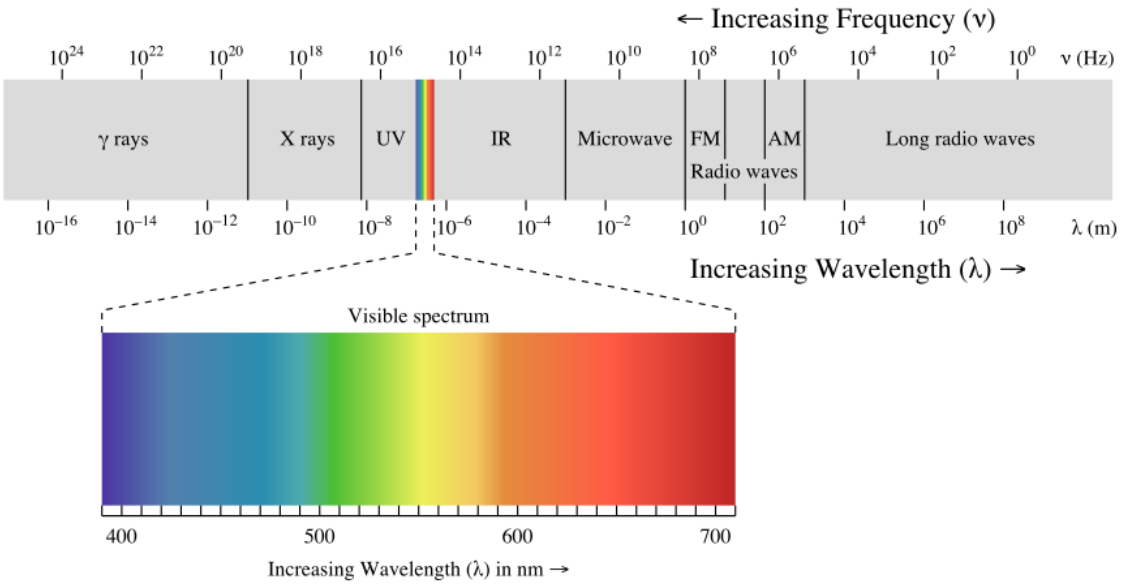

Figure 2.1: The electromagnetic spectrum [92].

Besides being able to capture varying objects, the wavelength of the radiation has multiple implications [89]. Atmospheric transmission between the sensor and the surface of earth is one of the main considerations. Different wavelengths are able to travel in the atmosphere with varying success; Some have a higher tendency to be absorbed or refracted by various airborne particles or to be contaminated by noise — of, for example, the solar background radiation. In addition, there are financial considerations: The technology for capturing or transmitting certain wavelengths in a large scale is sometimes too costly to be viable.

2.2.1 Light Detection and Ranging

Light Detection And Ranging (LiDAR) is an active remote sensing method, where the environment is scanned using laser [60, 59]. The scanning device — often mounted on an airborne vehicle, such as an airplane or an unmanned drone — sends pulses of laser energy, and then measures the amount of time for the pulse to reflect back to the sensor. Conventionally, the scanning results in a cloud of return pulses, that contains the intensity of the pulses as well as the coordinates of individual pulses as x, y, and z. These variables can then be used, for example, to map the environment as a 3D-model, or, for example, to infer biophysical properties of the vegetation in the scanned area.

LiDAR laser can be ultraviolet, visible, or near-infrared (NIR, ranging from 780 nm to 1550 nm) [89]. For the purposes of mapping vegetation, wavelengths of 905 nm, 1064 nm, and 1550 nm are commonly used, with 1550 nm being a costly but generally

the most suitable option. While it is possible to construct LiDAR systems that use multiple wavelengths simultaneously, airborne systems often use a single wavelength.

LiDAR is rapidly gaining popularity as a method of measuring forest data [60]. There are two main approaches of estimating the biophysical variables of vegetation using LiDAR: Tree-based and area-based. Both use the canopy height model, which is constructed from the differences of z coordinates between the lowest ground pulse and highest vegetation pulse. Whether a single pulse is returned from vegetation or ground can be predicted from the strength of the signal.

The tree-based approach interprets tree heights and locations from the canopy model [60]. Furthermore, these and other variables can be used to predict tree diameter, crown diameter, crown shape, and tree species. While the tree-based approach can yield accurate results, the main disadvantage is the tendency for larger individual trees to mask smaller trees in the tree height model, lowering the accuracy of the estimations.

The area-focused approach, on the other hand, is based on estimating the properties of a larger forest areas instead of individual trees [60]. The areas can be, for example, grids with specific dimensions. An area is scanned and then compared to a dataset of areas with known, already validated properties. The area from the dataset with the most similar LiDAR measurements is selected, and its properties are used to derive the properties of the scanned area. The advantage is that even non-uniform areas with substantial variety in tree species and sizes can be accurately estimated. However, a large dataset of validated measurements are needed, which might require considerable amounts of manual work.

The biophysical measurements from LiDAR can then be used to calculate biomass by using allometric equations [59]. Variables such as DBH or canopy height have been shown to correlate highly with above-ground biomass, and the methodology of deriving these variables from LiDAR data is relatively robust. In fact, there is strong evidence that with a small-footprint (<1 m) LiDAR system, measurements such as canopy height can be at least as accurate as ground measurements [2]. As a consequence, LiDAR is considered to be one of the most accurate and reliable methods of mapping biomass — and various other properties — in both small and large forest areas.

Airborne LiDAR scanning has its limitations, however [26]. It is expensive, time-consuming, and requires specific missions to get LiDAR vehicles over the areas where the measurements are wanted. As a consequence, it is difficult to get airborne LiDAR data, especially if it is needed for specific periods of time. This is opposed to many of the global satellite-based approaches, where measurements are constantly updated and stored, and where the data is often available to almost anyone at a low cost.

Satellite-based LiDAR exists as well [26]. Currently, however, the technology is greatly limited. Satellite-based LiDAR has high energy requirements, allowing only for

short bursts of scanning. Combined with the narrow scope of the scanning swath, we are far from a globally viable spaceborne LiDAR coverage.

As of June 2022, there are six different satellites equipped with LiDAR-based technology, with varying purposes [26]. For example, some were built to scan ice-caps, while some measure clouds and aerosols. At the current level of technology, It has been estimated that 12 satellites working together would be needed to produce a global LiDAR map every 5 years at 30 m resolution, or 418 satellites for 5 m resolution. Even a mission of 12 satellites would be a considerably expensive and time-consuming project, and the resolution of 30 m might not be sufficient for many applications.

In the domain of remote sensing, LiDAR-based AGB estimation methods are usually the most accurate and reliable, especially when it comes to dense vegetation [48]. However, due to the cost and difficulty of obtaining up-to-date measurements, other methods have been widely explored.

2.2.2 Optical Imagery

Optical imaging is one of the most common remote sensing methods [89]. Many airborne vehicles from airplanes to drones are fitted with optical sensors. In addition, tens of satellites with optical imaging capabilities are orbiting Earth, the first of which — Landsat 1 — was launched in 1972. While the satellites are controlled by varying organizations and nations, much of the data is easily accessible to anyone.

Optical imaging sensors can utilize visible, NIR, and short-wave infrared (SWIR) from the electromagnetic spectrum [89]. Typically, the systems produce panchromatic, multispectral, or hyperspectral images. Panchromatic images are either grayscale or black and white, captured by systems with monospectral channel detectors that are sensitive to a broad range of radiation.

Multispectral images are generated by multichannel detector sensors that are able to measure several spectral bands [89]. Each band is recorded to its own layer of data, which includes both brightness and spectral (color) information. Often bands are divided into specific, meaningful ranges, such as red, green, and blue of the visible light. A multispectral image often contains 2-15 separate bands.

Hyperspectral images are similar to multispectral images, except that they can contain hundreds of separate bands [89]. Each band is generally very narrow, with ranges as low as a fraction of a micrometer. The large number of bands allows for accurate discrimination of reflectance and detection of various objects at different wavelengths. In multispectral images, the bands can be too broad for fine-grained analysis.

While hyperspectral images store a large amount of information, processing the images is costly and time intensive. In addition, remote sensing devices with hyper-

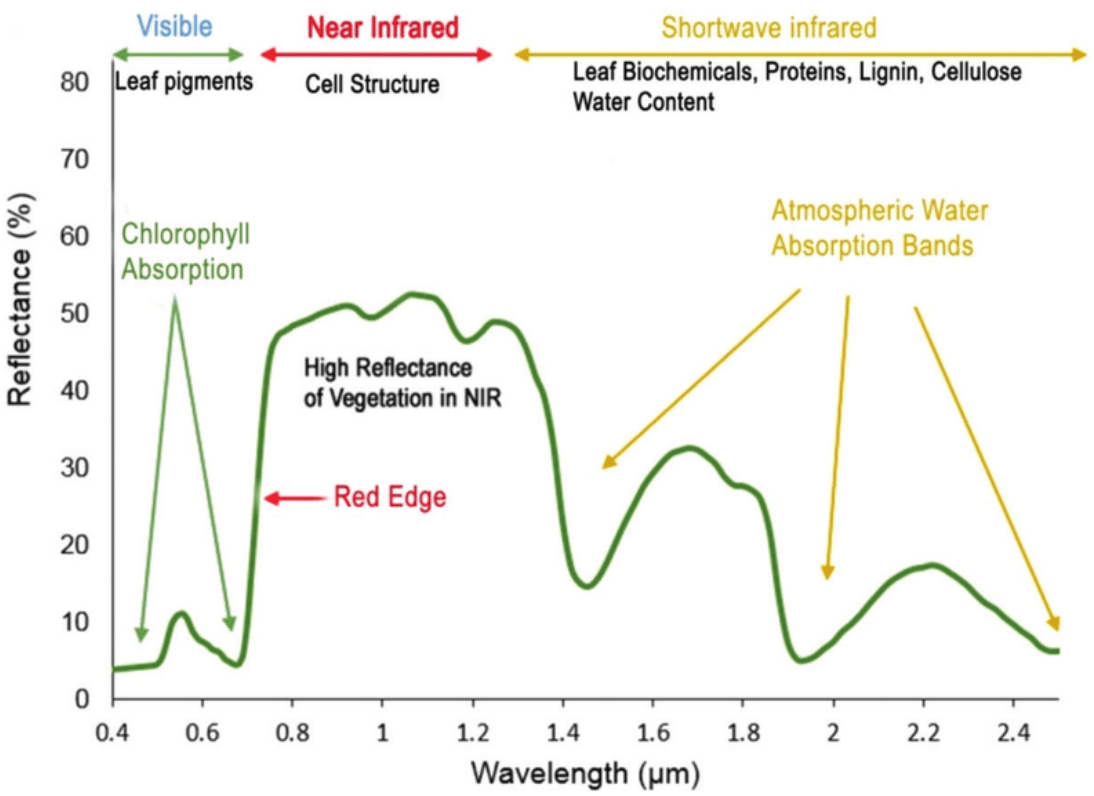

Figure 2.2: The reflectance of light from green vegetation [65].

spectral sensors are rarer than their multispectral counterparts, and the vast amount of information might be redundant for many purposes. Consequently, for AGB estimation using optical imagery, most research is centered around multispectral images.

Various transformations can be used to improve the cost and efficiency of processing imagery, and to normalize values or introduce new features [8]. Transformations are often used for both multispectral and hyperspectral images. Common transformations include Principal Component Analysis (PCA) and Tasseled Cap Transformation (TCT).

TCT is a conversion that is used especially for vegetation mapping [39]. It uses spectral bands to produce several distinct component bands. Originally, six bands were proposed, out of which three are most commonly used: greenness, wetness, and brightness. For example, the equation for the component band greenness — a measure of amount of vegetation — is defined as

$$v_{GREENNESS} = c_1(B1) + c_2(B2) + c_3(B3) + c_4(B4) + c_5(B5) + c_6(B7) \tag{2.4}$$

where the coefficients $c_1, \ldots, c_6$ are values ranging from -1 to 1, and the constants $B1$,

$B2$, $B3$, $B4$, $B5$, and $B7$ are the spectral bands. Note that this specific equation is for the sensors of the Landsat-satellite only. For other sensors the coefficients, bands, and number of bands used needs to be determined, and the process can be complicated [56].

In short, TCT uses the weighted sum of spectral bands to produce the transformed components, with the goal of extracting the most relevant information and capturing the variation of the spectral bands. The components can then be directly used to estimate the amount of vegetation and biomass in an area.

PCA is one of the most common methods of reducing dimensionality, and is often used in remote sensing [8]. Similar to TCT, PCA is a linear transformation with the aim of capturing the most relevant information of data. For a more detailed explanation of the algorithm, see subsection 3.3.1.

In addition, the spectral bands can be directly used for above-ground biomass estimation. Vegetation reflects various bands to different degrees [65]. In Fig. 2.2 the percentage of reflected radiation for green vegetation in various bands can be observed. Chlorophyll in plants mostly absorb blue and red bands from the visible spectrum, reflecting back mainly radiation within the green band. Near infrared is reflected heavily by the mesophyll cells of vegetation. Leaf biochemicals, proteins, lignin, and cellulose water content reflect shortwave infrared (SWIR). The atmosphere, however, more readily absorbs certain ranges of the radiation, which needs to be considered.

The differences between the reflectance of certain bands has led to a large amount of research and experimentation [70]. The high contrast between visible red and NIR (see Fig. 2.2) is known as the Red Edge (RE), and it is especially useful for vegetation estimation. The exact position and features of the red edge is affected by a multitude of factors, including the combination of plant and tree species, and the number of layers of leaves [30]. Furthermore, these factors can, for example, correlate with canopy thickness and biomass.

Similar to other spectral transformations, multiple bands can be used to calculate a component band with a better representation of the various characteristics of vegetation [70]. These component bands are called Vegetation Indices (VI), and they are considered to be among the most optimal methods of detecting AGB via spectral features.

Two of the most common vegetation indices in AGB estimation is the Simple Ratio (SR) and the Normalized Difference Vegetation Index (NDVI) [48]. SR is defined as NIR divided by the visible red band. It is among the most straightforward methods of depicting the red edge. While SR allows the separation of vegetation from the environment, the resulting measurement scale is not linear, often complicating analysis. In addition, SR is susceptible to division by zero errors.

Vegetation Index		Equation	Ref.
Difference Vegetation Index	DVI	$NIR - RED$	[76]
Simple Ratio	SR	NIR/RED	[37]
Modified Simple Ratio	MSR	$\frac{NIR/(RED-1)}{\sqrt{NIR/RED+1}}$	[11]
Normalized Difference Vegetation Index	NDVI	$\frac{NIR-RED}{NIR+RED}$	[68]
Enhanced Vegetation Index	EVI	$G\frac{NIR-RED}{NIR-C_1\cdot RED-C_2\cdot BLUE+L}$	[32]
Specific Leaf Area Vegetation Index	SLAVI	$\frac{NIR}{RED+SWIR}$	[50]
Normalized Difference Water Index	NDWI	$\frac{NIR-SWIR}{NIR+SWIR}$	[28]
Wide Dynamic Range Vegetation Index	WDRVI	$\frac{a\cdot NIR-RED}{a\cdot NIR+RED}$	[22]
Triangular Vegetation Index	TVI	$0.5[120(NIR-GREEN)-200(RED-GREEN)]$	[16]
Transformed Triangular Vegetation Index	TTVI	$\sqrt{ABS(\frac{NIR-RED}{NIR+RED}+0.5)}$	[74]
Corrected Transformed Vegetation Index	CTVI	$\frac{NDVI+0.5}{ABS(NDVI+0.5)}\sqrt{ABS(NDVI+0.5)}$	[57]
Renormalized Difference Vegetation Index	RDVI	$\frac{NIR-RED}{\sqrt{NIR+RED}}$	[67]
Ratio Vegetation Index	RVI	$\frac{RED}{NIR}$	[64]
Normalized Ratio Vegetation Index	NRVI	$\frac{RVI-1}{RVI+1}$	[5]
Soil Adjusted Vegetation Index	SAVI	$(1+L)\frac{NIR-RED}{NIR+RED}$	[33]
Normalized Difference Index using bands 4 and 5 (Sentinel 2)	NDI45	$\frac{RE_1-RED}{RE_1+RED}$	[17]
Inverted Red-Edge Chlorophyll Index (Sentinel-2)	IRECI	$\frac{RE_3-RED}{RE_1/RE_2}$	[20]
Sentinel-2 Red-Edge Position	S2REP	$705+35\frac{(NIR+RED/2)-RE_1}{RE_2-RE_1}$	[20]
Optimized Soil-Adjusted Vegetation Index	OSAVI	$\frac{NIR-RED}{NIR+RED+Y}$	[66]
Green Normalized Difference Vegetation Index	GNDVI	$\frac{NIR-GREEN}{NIR+GREEN}$	[23]
Transformed Normalized Difference Vegetation Index	TNDVI	$\sqrt{\frac{NIR-RED}{NIR+RED}+0.5}$	[87]
Green Ratio Vegetation Index	GRVI	$\frac{GREEN-RED}{GREEN+RED}$	[18]
Moisture Stress Index	MSI	$\frac{NIR}{SWIR}$	[38]
Infrared Percentage Vegetation Index	IPVI	$\frac{NIR}{NIR+RED}$	[14]

Table 2.1: Various vegetation indices used in AGB estimation.

NDVI is the difference between NIR and the visible red band, divided by the sum of NIR and red [70]. Similarly to SR, NDVI is based on the high contrast of reflectance between the visible red band and NIR, and shows the separation of vegetation from soil and other topography. It also produces a linear measurement scale and is less likely to contain division by zero errors.

Multiple other vegetation indices use the red and NIR bands, or are directly inspired by NDVI or SR. Some of these include the Transformed Vegetation Index (TVI), Transformed Normalized Difference Vegetation Index (TNDVI), and Infrared Ratio Vegetation Index (IPVI). Many bands also use other wavelengths, such as Green Normalized Vegetation Index (GNDVI) or Enhanced Vegetation Index (EVI). See Table 2.1 for various vegetation indices used in AGB estimation.

There is a large amount of research about vegetation indices, with a variety of equations proposed [48]. The vegetation indices have their individual strengths and weaknesses, and are often specifically suitable for certain types of vegetation, topography, imagery, or environmental factors. When estimating AGB via spectral features, multiple bands and vegetation indices are commonly used in combination.

Vegetation indices and separate bands are well suited as input variables in machine learning [48]. Regression models are one of the most common methods of optical imagery based AGB estimation, with algorithms such as Random Forest (RF), Artifical Neural Network (ANN), and Support Vector Machine (SVM) generally showing good results. Using data derived from manual sampling or LiDAR measurements as the ground truth data is a common approach. The problem can be formulated as

$$f(x_1, x_2, \cdots, x_n) = E[Y|X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] \,, \tag{2.5}$$

where $X_1, \cdots, X_n$ are the chosen bands and vegetation indices as input variables, and Y is the amount of AGB as derived from the ground truth data.

While using the correlating bands and vegetation indices to detect biomass seems straightforward, the subject is complicated by the reflectance of other materials in the environment [48]. In Fig. 2.2 can be seen the reflectance of dry soil, snow, litter, and water alongside vegetation. The choice of bands and VIs used is heavily influenced by a variety of factors, including soil type, amount of water and moisture, the current season, and the presence of human-made objects, such as structures and litter. In addition, different vegetation reflects radiation in different proportions. As such, choosing a reliable set of bands and vegetation indices to estimate biomass is often a difficult task. This is, however, made easier by the fact that for most forest areas the composition

of biomass is dominated by trees, and many forests only contain a few main species of trees.

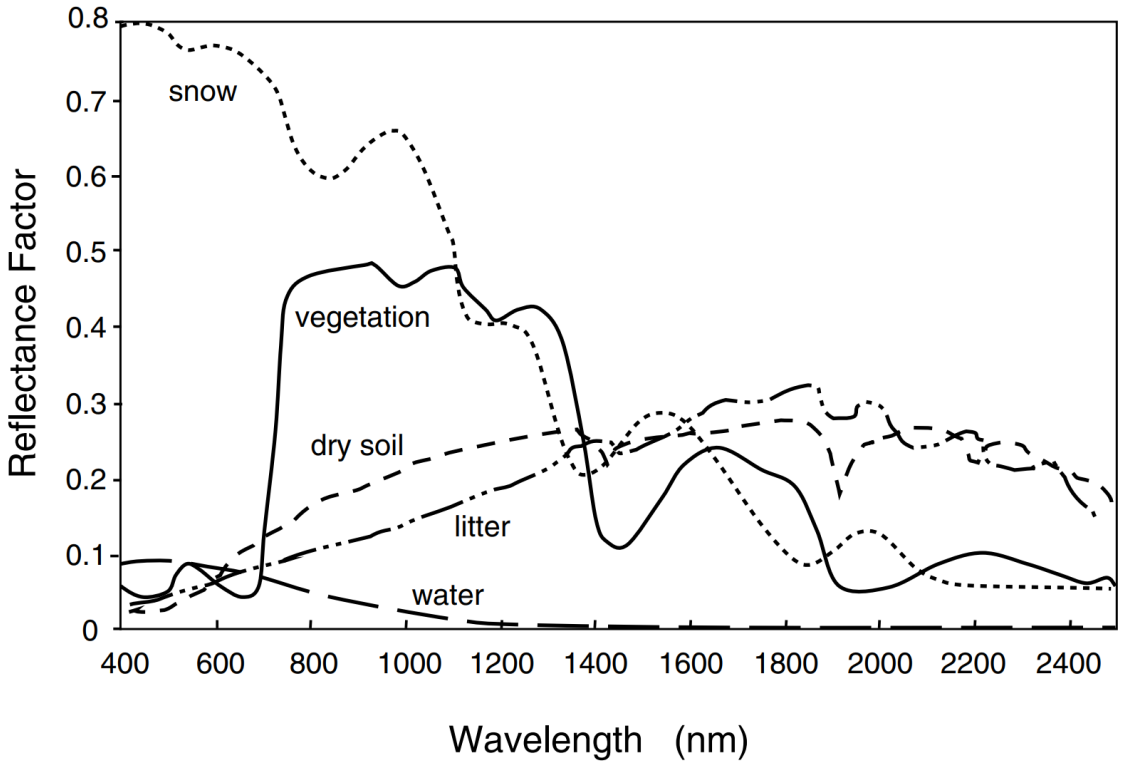

Figure 2.3: The reflectance of various materials in the environment [34].

Besides spectral analysis, spatial analysis can also be used for AGB estimation [45]. This is often achieved by using textural features. Texture is a small-scale arrangement of spectral values that repeats as a pattern in a larger spatial scale. These various arrangements can be identified as specific elements, such as vegetation or soil.

There are three main classes of methods for extracting textural features: configurative methods, statistical methods, and spectrum decomposition methods. Out of these, statistical methods are most common in AGB estimation, especially Grey Level Co-occurrence Matrix (GLCM). See subsection 3.2.2 for an overview of the GLCM algorithm.

A variety of different spatial resolutions of optical imagery have been used in AGB estimation [48], from low (1000 m) to very high (under 1 m). The data from higher resolution sensors have an advantage in prediction accuracy, but are costly to acquire and process. The lower resolution images are generally easier and financially more viable to use for large areas, and are more readily available for different areas and time periods. Consequently, low to medium resolution images are often used for monitoring AGB change over time, or to complement other measurements — for example, from a forest inventory collected using other sampling methods.

In addition to the widespread availability of the optical imaging data — especially at lower spatial resolutions — its main advantages are intuitiveness and ease of use [48]. The flat 2D images, mostly captured directly from above, are comparable to traditional photography. This is contrasted by SAR or LiDAR imagery, where geometry is depicted from an odd angle and the data is generally less intuitive.

The main problem of optical imagery is the low penetration depth of the radiation [48]. Generally, only the reflectance from surface objects and vegetation is retrieved, rendering much of the underlying biomass invisible in the optical imagery. Especially in thick vegetation, the data tends to saturate and the volume of AGB gets underestimated, meaning that using optical imagery alone is not always viable. In addition, images can also be partially or completely obscured by clouds, rendering certain images or areas of images unsuitable for analysis.

2.2.3 Synthetic Aperture Radar

The Synthetic Aperture Radar (SAR) is an increasingly popular technology in the field of remote sensing [19]. It is frequently used on both airborne and spaceborne platforms, with many of both the current and the upcoming satellites carrying a SAR sensor. The use cases of SAR are numerous, including airport surveillance, urban monitoring, global mapping, maritime navigation, and remote archaeology.

SAR is also a major remote sensing method for AGB estimation [86]. Similarly to LiDAR, SAR is an active data collection system that sends its own electromagnetic radiation and then measures the backscatter radiation after it has reflected from the objects. Instead of using laser, SAR emits and measures microwaves, mostly working at wavelengths in the range of 0.8 - 100cm.

As with optical data, the wavelength ranges are divided into specific bands [54]. The most common bands are denoted as Ka, K, Ku, X, C, S, L, and P (see Table 2.2). For the task of AGB estimation, L and P are considered the most feasible bands [86]. Generally, when the particles are on the scale of the radiation wavelength, most backscattering occurs. This is due to these longer wavelengths (15-30cm for L and 30-100cm for P) being able to efficiently penetrate even thick canopy covers and reflect from tree trunks and branches instead of surface leaves. This allows for better modelling of the height and 3D structure of forests. Higher frequency bands, such as C and X, have still been successfully used for modelling biomass in regions with sparse vegetation, such as savannas, grasslands, or various crops [83].

SAR transmits its own radiation, which allows precise control over the signal. This has an advantage over optical sensors using reflected solar radiation, which has a wide range of wavelengths and random phase. Phase is a term for a point in time on a

Band	Wavelength (cm)	Typical Application
Ka	0.8 - 1.1	Airport surveillance (rarely used).
K	1.1 - 1.7	H_2O absorption (rarely used).
Ku	1.7 - 2.4	Satellite altimetry (rarely used).
X	2.4 - 3.8	Urban monitoring, ice and snow.
C	3.8 - 7.5	Global mapping, mapping areas with moderate vegetation, change detection. (Commonly used)
S	7.5 - 15	Earth observation, agriculture monitoring.
L	15 - 30	Geophysical monitoring, biomass and vegetation mapping.
P	30 - 100	Biomass mapping.

Table 2.2: Various SAR bands and their typical applications [19].

waveform cycle, and measuring allows SAR devices to estimate the distance from the sensor to the reflected surface with accuracy up to the centimeter level. In addition to phase, amplitude (signal intensity) can be measured precisely and the polarization of the wave can be controlled.

Polarization is an important aspect of SAR data [54]. The transmitted electromagnetic wave oscillates on a plane, and the orientation of the plane is referred to as polarization. Generally, the wavelengths transmitted by SAR devices are linearly polarized, meaning they are either aligned horizontally (H) or vertically (V) (see Fig. 2.4).

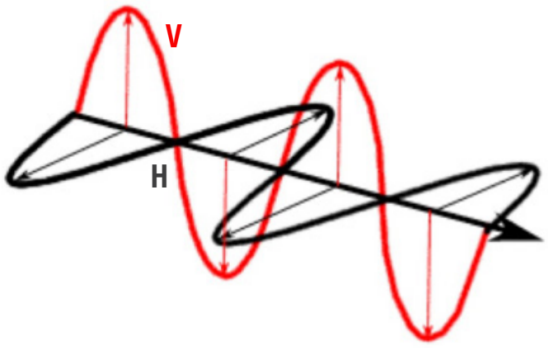

Figure 2.4: Waves with vertical (V) and horizontal (H) polarizations [19].

SAR sensors can control the polarization both when transmitting and receiving [19]. For example, a signal can be sent as H, and received as V. This would be denoted as HV, horizontal vertical. There are four different polarization operation modes

for SAR sensors: single-polarization, cross-polarization, dual-polarization, and quad-polarization. In single-polarization mode, the radar system receives wavelengths in the same orientation as it was sent (VV or HH). In cross-polarization, different polarization is used for transmit and reception (VH or HV). Dual-polarization mode allows the radar to transmit in a single orientation, while receiving in both polarizations simultaneously (VV and VH, or HH and HV). Finally, in quad-polarization mode signals are transmitted in alternating polarizations and received simultaneously in both, allowing the collection of all the four different polarization types (VV, VH, HH, HV).

Each combination of polarizations reveals information of the imaged surface in a distinct way, giving additional control for the requirements of various specific tasks [19]. For example, a vertically oriented standing tree reflects the radiation back in a different pattern, depending on the polarization (V or H). The scatter is also different between the upright trunk of the tree and its horizontally or diagonally oriented branches and leaves.

The reception polarization is also affected by the surface types [19]. For example, strong HH scattering indicates a pattern of double-bounce scattering, which is often due to man-made structures of stemmy vegetation. VV relates to scattering from rough surfaces, such as water or bare ground. Cross-polarized data (VH or HV) is most sensitive to volume scattering, which can be an indicative of leaves and branches of trees.

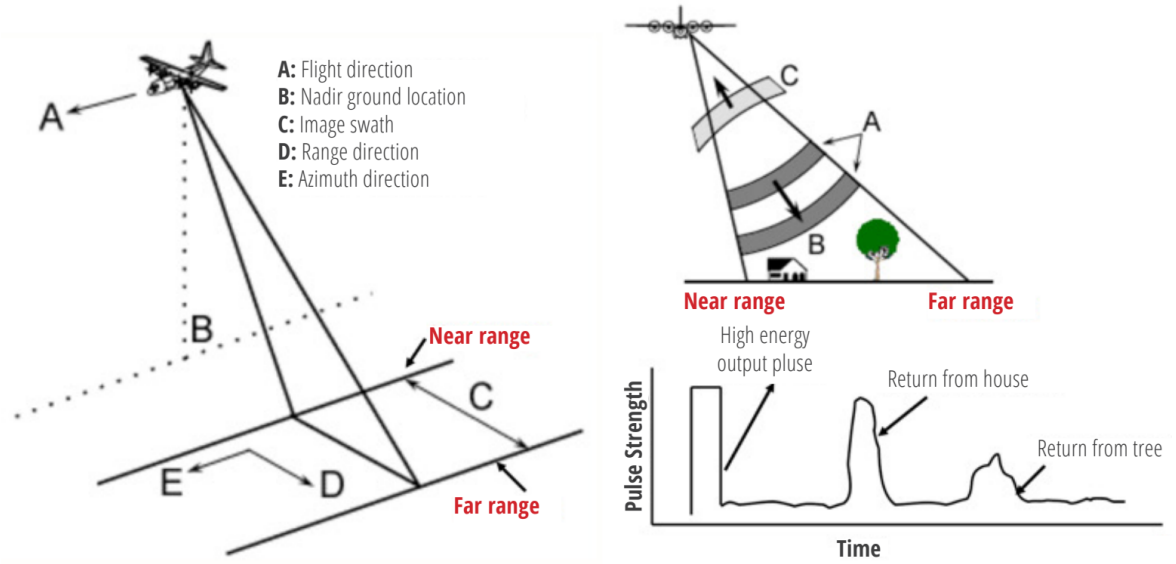

Figure 2.5: The geometry of typical SAR observations. Note the off-nadir angle of the signal and how it allows the differentiation between objects that are near (the house) and far (the tree) by temporal analysis [19].

Using SAR data reliably for AGB estimation is often difficult [19]. There are multiple factors of ambiguity resulting from measuring difficulties, which can be divided

into two categories: geometrical and environmental. Geometrical factors arise from the combination of the off-nadir angle of observations and the geometry of the imaged objects. Unlike with typical optical sensors or many LiDAR arrangements, SAR uses a side-looking swathe in order to estimate the distance of objects and their proportional relations within the scanned area (see Fig. 2.5). The observation geometry results in measurement differences between the near and the far range of the image swath.

In addition, the angle of observation and direction of flight can vary for each pass over [19]. Furthermore, for spaceborne SAR, it is important to consider whether the orbit is descending or ascending. All of these factors can cause variances in the observation data.

The geometry of the targets will also have an effect on the observations [19]. In cases of trees and other vegetation, this includes the orientation of branches and leaves. The geometry of branches and leaves is constantly changing, resulting from growth and from environmental factors, such as wind.

Other environmental factors include soil moisture, surface inundation, intercepted water, and water content of vegetation [19]. Soil moisture can change backscatter values, especially when using HH or VV polarization. Surface inundation means the temporary, full or partial submergence of the imaged object, for example by a seasonal flood. This increases the power of backscattering significantly, especially for HH and VV.

Intercepted water can be, for example, from rain or dew [19]. Depending on the wavelength and droplet size, it can either increase backscattering power (X and C bands) or lower it (L and P bands). Water content of trees and leaves can also affect backscattering. The water content changes for various reasons, which include annual seasons, temperature, time of day, weather, and stress.

Due to all the geometrical and environmental factors, it is difficult to get reliable, comparable observations on SAR [19]. Generally, intensive preparation and processing is required to get SAR data into an analysis-ready form. One of the typically undesirable characteristics of SAR images is speckling. The speckle effect is the result of the interference from the multitude of radiation echoes reflecting and scattering within the area of a single imaged pixel. Speckling is generally most pronounced in the areas of the image with the highest amount of radiation reflected back, i.e., the brightest regions.

The speckle effect is absolutely multiplicative. It can be modelled as

$$I(t) = R(t)n(t) \ , \tag{2.6}$$

where $I(t)$ and $R(t)$ are the speckled and speckle-free backscatter of a SAR measurement, respectively [88]. Noise factor is noted as $n(t)$, and it is independent with $R(t)$.

The speckling can result in a grainy, noise-like texture in SAR images, introducing differences in the captured backscatter values even in cases where the imaged objects are identical [19]. Filters can be used to reduce the effect. There are a variety of commonly used speckle filtering algorithms, such as the Lee filter [44]. The Lee filter utilizes minimum mean squared error (MMSE) filtering, and it can be formulated as

$$R(t) = \bar{I}(t) + W(t)I(t) - \bar{I}(t) \ , \tag{2.7}$$

where $\hat{R}(i,j)$ is the speckle-filtered backscatter, and also the estimated value of $R(t)$ [88]. I is the speckled backscatter, and $\bar{(}I)$ is its mean value. The $W(t)$ represent a weighing function, defined as

$$W(t) = 1 - \frac{C_v}{C_I} \ , \tag{2.8}$$

where C_v is the variance coefficient $\sigma_v/\bar{v}$ of the speckled backscatter, while C_I is the variance coefficient $\sigma_I/\bar{I}$ of the speckle-free backscatter.

As a standard deviation based (sigma) filter, the Lee filter is able to suppress noise while preserving the quality and sharpness of the image [19]. However, the Lee filter assumes that speckling is uniform throughout the whole image. If there are large deviations among different parts of the image — such as when imaging largely varied areas — the Lee filter does not effectively retain all the details of the image.

Other examples of commonly used speckle filtering algorithms are Quegan [61], Enhanced Lee [46], and Frost [21]. The filters, similarly to the Lee filter, also have their advantages and disadvantages. Among other factors, the choice of de-speckling methodology depends largely on the properties of the SAR device, the imaged region, and the intended purpose of the data.

Besides speckle filtering, SAR imagery often needs radiometric and geometrics corrections [19]. As the observations are captured using a side-looking geometry, topographic shading is compounded on the sensor-facing surfaces. If a SAR is imaging a hill, for example, the slope positioned towards the sensor gets radiometrically overexposed. Additionally, pixels are often mislocated when imaging topographically inclined areas, leading to geometric distortion. Both of these effects reduce the viability of the

imagery for many applications, as the true structure of the targeted areas is masked. Additionally, the effects depend on the angle and position of the sensor, complicating a comparison of multiple observations.

A single pixel value σ in calibrated SAR data is defined as

$$\sigma = \frac{I_r}{I_i} 4\pi R^2 \ , \tag{2.9}$$

where I_r is the received signal intensity and i_i the signal intensity at the incident location [19]. R is the distance to the target, in meters. The σ is also termed Radar Cross-Section (RCS). RCS, however, neglects the effect of incidence angle on the measurement. For the purposes of angle-correction, RCS can be defined as

$$\sigma = \sigma^0(\theta_i) A_\sigma(\theta_i) \ , \tag{2.10}$$

where A_σ is the surface covered by the pixel, and θ_i is the local incidence angle. Sigma-nought (σ^0) is the normalized, incidence angle -dependent RCS. To estimate sigma-nought, the effect of $A_\sigma(\theta_i)$ is corrected for in a process called Radiometric Terrain Correction (RTC).

Radiometric terrain correction is able to compensate for both radiometric and geometric distortion [19]. To correct geometric distortion, geometric terrain correction (geocoding) is employed. For the removal of radiometric distortion, Radiometric Terrain Normalization (RTN) is utilized in a pixel-by-pixel basis. Both RTN and geocoding require additional data in the form of a Digital Elevation Map (DEM), which is a topographic map containing elevation information.

While there are many methods of RTC, one of the most common methods was introduced by Small [2011]. Notably, the Small methodology produces backscatter values in a gamma-nought (γ^0) normalized form, which are used by multiple SAR-related algorithms [19].

Once SAR data has been calibrated and processed into an analysis-ready form, above-ground biomass can be estimated using similar methods as with optical imagery. [47] Statistical models — such as random forest — can be used with features extracted from the SAR data. The polarization bands can be used alone as predictors of biomass, and textural algorithms — such as GLCM — can be used to depict patterns in the imagery.

In addition, the SAR polarizations can be used to calculate vegetation indices [40].

These SAR-derived vegetation indices are relatively uncommon in research, especially in biomass estimation. This is partly because many of the common SAR remote sensing devices only produce — at most — dual-polarized observations (i.e., VV and VH, or HH and HV) at a given time, while many of the indices require both H and V -polarization transmitted bands [19]. For example, the Radar Vegetation Index (RVI) is formulated as

$$RVI = \frac{8\gamma^0_{HV}}{\gamma^0_{HH} + \gamma^0_{VV} + 2\gamma^0_{HV}} , \tag{2.11}$$

utilizing the polarizations HH, HV, and VV [40]. The γ^0_{XX} denotes that backscatter values of polarization XX are in the normalized gamma-nought form.

The major advantage of SAR is the ability to penetrate thick vegetation and canopy cover, especially when using the long wavelength L and P -bands. Additionally, as an active sensor, SAR functions without external radiation, meaning it can image unlit areas. SAR is also more resistant to obstruction by clouds, as the microwaves can mostly travel through cloud cover without refracting notably.

2.2.4 Combining AGB Estimation Methods

Selecting the most viable approach for an AGB estimation problem is often difficult. While there are multiple methods of estimating AGB, all of them have their advantages and disadvantages.

Allometric equations have among the highest prediction accuracy out of the statistical approaches. Manual sampling, however, is expensive and time-consuming. Models using allometric equations with LiDAR-based data can have — in some cases – an even higher accuracy than models using ground measurements [2]. As a downside, LiDAR measurements are expensive, the data is limited, and spaceborne measurement devices with high spatial resolution and coverage do not exist yet. A large-scale airborne LiDAR scanning is often a massive undertaking that might take multiple years.

Satellite-based imagery is abundant and often inexpensive or free to use. Optical and SAR data models to cover large areas can be easily built without the need for a large financial investment to acquire measurement data. The methods utilizing optical and SAR data, however, do not generally reach the accuracy of the methods based on LiDAR or ground measurements.

Processing optical imagery is usually straightforward. Simple statistical AGB estimation models can be built by using the raw spectral bands. Further, the complexity and accuracy of the models can be improved by using vegetation indices. In addition,

textural measures have been shown to produce high accuracy models when used in conjunction with spectral band and vegetation indices. Textural measures, however, lower the spatial resolution of the data. To counteract the loss of information, very high-resolution data sources are used for best effect.

In general, much of the state-of-the-art research focuses on combining multiple data sources and feature extraction methods [48]. Combining SAR backscatter with optical vegetation indices and spectral bands is a common approach [77, 43]. Another recent approach is to combine features based on textural measures, vegetation indices, and spectral bands derived from multiple high resolution optical data sources [47].

The intuition is clear, especially when combining optical and SAR sources. Optical data produces high-quality results when used alone, but tends to saturate in high biomass areas, as the imaged wavelengths are unable to penetrate canopy cover and vegetation. While high accuracy estimation models utilizing solely SAR data are rare [48], the SAR wavelengths are able to penetrate surface vegetation, thus resulting in better approximations of the underlying geometry of the targeted forest areas.

Using both sources of data in conjunction can be thought to negate the disadvantages of each dataset. SAR, however, also suffers from saturation problems. Depending on the used wavelength, the radiation might not be able to penetrate very far into the canopy. Highest wavelength sensors that are most resistant to saturation (P-band) still do not exist on satellite platforms. Even the upcoming ESA's BIOMASS satellite carrying a P-band sensor will have a low spatial resolution, rendering it unsuitable for many biomass estimation purposes.

Compared to optical data, SAR data is more difficult to analyze and process. The geometry of SAR measurements the physics of backscattering radiation require complicated processing steps to prepare the data into a cohesive form. The diversity of the data is usually also lower: Whereas an optical image might have tens or hundreds of different bands, SAR images generally only contain 1 to 4 polarizations. As a result, SAR based textural measures are less flexible, and vegetation indices fewer and less useful. An advantage of satellite based SAR compared to optical imagery, however, is that it is much more resistant to obstruction by clouds and can take measurements in the dark.

A major problem of remote sensing -based estimation is the resolution of imagery [48]. A single unit of data (a pixel) might correspond to a 10 x 10 m area in a relatively high-resolution image. This pixel has to represent everything in the area, which might include multiple trees of different species, other vegetation, soil, and human made objects. Selecting data of sufficient resolution is important, while also having to balance the downsides of high-resolution data: financial and processing cost, lower availability, and the increased difficulty of analysis and development. If the estimation model needs

to be applicable outside the area used for training, a diverse and large enough dataset is required.

The most common approach of creating a satellite-data based AGB estimation model is to use high-quality ground truth data — either derived from manual sampling or LiDAR — in conjunction with satellite imagery. To reach highly accurate models, the quality of both ground truth data and the satellite data needs to be high. Extensive coverage by LiDAR or manual sampling is financially costly, often limiting research to areas with pre-existing datasets. High-quality satellite data is easier to acquire, but the expenses still limit the covered area, especially when combining multiple data sources.

3. Data and Methods

In this work, a variety of statistical methods were applied to estimate above-ground biomass for selected forest areas in Finland. The goal was to fit regression-based Random Forest models that are able to predict the amount of biomass in an area using satellite-based data. A major goal of the research was to evaluate the efficacy of the novel method of using neighboring features for estimation. This method is described in subsection 3.2.3.

Pre-existing biomass data was used as the target data for training the models, and the openly available forest inventory dataset from the Finnish Forest Center [97] and the forest inventory and biomass dataset from Natural Resources Institute Finland [96] are partly behind the inspiration for this research. While the high-quality forest inventory and biomass data renders the results of this work impractical in Finland, the methods are hopefully applicable to areas with similar vegetation outside of Finland. For this reason, the hemiboreal zone of Finland — residing mainly in the coastal south-west — was chosen as the target of analysis. The vegetation zone (see Fig. 3.1) extends to Norway, Sweden, Estonia, Latvia, Lithuania, Ukraine, and Russia. Many of these areas might not have extensive, high-quality forest inventory datasets or biomass maps readily available.

The satellites Sentinel-1 [99], Sentinel-2 [100], and ALOS-2 [90] were chosen as the sources of data. The selection was motivated by two major factors: the availability and the quality of the data. Products from Sentinel-1, Sentinel-2, and ALOS-2 are freely available and of relatively high quality. The satellites are still functioning and producing new data, and recent data exists for the target areas of south-western Finland. Additionally, these data sources are also popular in recent research, which provides a firm basis for the methods used in this work and the option to compare the results.

The work has been divided into four parts: The preparation of data, the extraction of training features from the data, feature selection, and the training of the models.

Figure 3.1: The European hemi-boreal vegetation zone, in orange highlight [52]. The figure has been edited for clarity.

3.1 Data Preparation

As the data is from distinct satellite sources and imaging equipment, different methods need to be employed to make the data analysis-ready. The processing of Multi-Source National Forest Inventory of Finland, Sentinel-1, Sentinel-2 and ALOS-2 PALSAR-2 data is described separately.

3.1.1 The Multi-Source National Forest Inventory of Finland Data

The biomass data from The Multi-Source National Forest Inventory of Finland (MS-NFI) [96] was used as the ground truth data for the statistical models. The data is provided by the National Resources Institute Finland (LUKE). 2019

The methodology behind MS-NFI for year 2015 (MS-NFI-2015) has been openly published [51]. However, the exact methods for the most recent data of year 2019

(MS-NFI-2019) have not been disclosed in detail as of yet. The MS-NFI-2019 data was used in this work, and it should be noted the true accuracy of the dataset is difficult to evaluate. As a consequence, estimating the error stemming from the inaccuracies and the unknown methodology of MS-NFI-2019 was left out of the scope of this research.

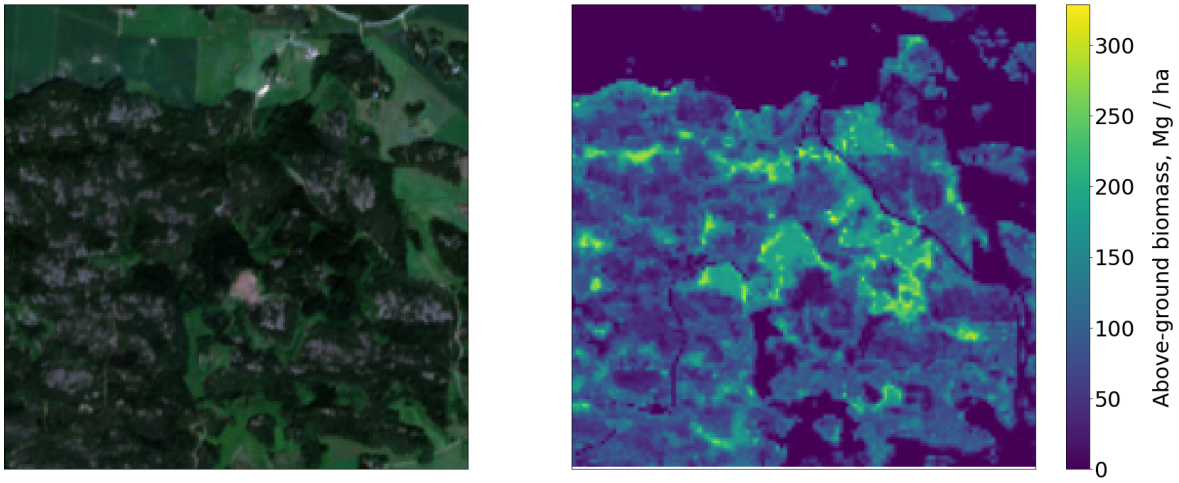

Figure 3.2: An RGB composite of Sentinel-2 optical imagery (on the left) and the MS-NFI-2019 total above-ground biomass data for the same area.

Both MS-NFI-2015 and MS-NFI-2019 offer comparable products for biomass and other forest parameters [96]. MS-NFI-2015 uses multiple sources of data to calculate the variables, with field data from 11th and 12th Finnish National Forest Inventories forming the basis of the dataset [51]. The inventory contains field data from 78312 systemically located, manually sampled plots across Finland. Both stand level characteristics and tree level measurements have been recorded for the plots.

In addition, satellite data, digital map data, and other geo-referenced data were used for MS-NFI-2015 [51]. As the field data plots cover only a small portion of the dataset area, the additional data sources were used to extrapolate the data and extend the coverage. Notably, optical satellite imagery from Sentinel-2 and Landsat-8, and a Digital Elevation Model (DEM) were employed among other sources of data.

Biomass was derived for sample plots using allometric equations [51]. The tree-level measurements in the field data were used in single-tree allometric models, including the models from Repola [2008] and Repola [2009]. For example, the equation for biomass of living branches for birch is

$$\ln(y_{ki}) = b_0 + b_1 \frac{d_k}{(d_k + 16)} + b_2 \frac{h}{(h + 10)} + u_{3k} + e_{3ki} \,, \tag{3.1}$$

where y_{ki} is the living branch biomass (in kilograms) and d_{ki} the diameter at breast height, both for the tree i in stand k. Height is denoted by h, while u_{3k} and e_{3ki} are the random stand parameter and the residual error, respectively.

MS-NFI-2019 contains 6 distinct component biomass products: living branches, stem residual, roots, stump, dead branches, stem and bark, and foliage [96]. The products are available individually for three categories of trees: pine, spruce, and broad-leaved trees. To get the total biomass, the component biomasses from each of tree classes were summed together. As the goal of this work is to estimate above-ground biomass, the root component was left out of the calculations.

The biomass products are available in grids with pixels 16 x 16 meter size [96]. The maximum resolution of other datasets in this work is 10 x 10 meters (Sentinel-1 and Sentinel-2), so the MS-NFI-2019 data was resampled to a matching 10 x 10 meter resolution. The value for each new pixel was calculated by the average of neighboring pixels, i.e., by using bilinear interpolation. A visualization of the processed data can be seen in Fig. 3.2

3.1.2 Sentinel-2

The optical data used in this work is from the Sentinel-2 mission by the European Space Agency (ESA) [100]. The mission currently consists of two satellites, Sentinel-2A and Sentinel-2B, the first of which was launched in 2015. The satellites focus on imaging land and coastal water with an orbital swath width of 290 km.

Each satellite carries a Multi-Spectral Instrument (MSI) [100]. The MSI captures 13 distinct wavelength bands, with resolutions of either 10 m, 20 m, or 60 m (see Table 3.1). The Red, Green, Blue, and NIR bands have the highest spatial resolution of 10 x 10 meters, while the Red Edge 1-3, Narrow NIR, and SWIR 1-2 have resolutions of 20 x 20 m. The three lowest 60 meter resolution bands are used for detecting coastal aerosols, water vapor, clouds, and other airborne particles, partly in order to estimate the amount of obstruction between the sensor and the imaged surface on the ground.

Sentinel-2 data products are publicly available in two formats: 1C and 2A [100]. Category 1C data consists of radiometrically and geometrically corrected tiles, in cartographic format. Each of the tiles contains top-of-atmosphere (TOA) reflectance for a 100 x 100 km area. As the Sentinel-2 satellites orbit outside the atmosphere, the reflectance of objects on Earth's surface is interfered and refracted by the particles in the atmosphere. To alleviate these effects, the 2A data product has been atmospherically corrected to transform the TOA reflectance into bottom-of-atmosphere (BOA) reflectance.

Band	Band Description	Resolution	Central Wavelength (S2A / S2B)
B1	Coastal Aerosol	60 m	442.7 nm / 442.2 nm
B2	Blue	10 m	492.4 nm / 492.1 nm
B3	Green	10 m	559.8 nm / 559.0 nm
B4	Red	10 m	664.6 nm / 664.9 nm
B5	Red Edge 1	20 m	832.8 nm / 833.0 nm
B6	Red Edge 2	20 m	704.1 nm / 703.8 nm
B7	Red Edge 3	20 m	740.5 nm / 739.1 nm
B8	NIR	10 m	782.8 nm / 779.7 nm
B8A	Narrow NIR	20 m	864.7 nm / 864.0 nm
B9	Water Vapor	60 m	945.1 nm / 943.2 nm
B10	SWIR - Cirrus	60 m	1373.5 nm / 1376.9 nm
B11	SWIR 1	20 m	1613.7 nm / 1610.4 nm
B12	SWIR 2	20 m	2202.4 nm / 2185.7 nm

Table 3.1: The captured spectral bands, resolutions, and central wavelengths by Sentinel-2A (S2A) and Sentinel-2B (S2B) satellites.

Additionally, the 2A product has been processed using cloud and cloud shadow detection algorithms, and a cloud masking dataset is provided [100]. In this work, the 2A imagery for sample areas was cloud masked using the provided dataset, and a multi-temporal mosaic was generated. Several images captured within a time-span were selected. The clouds and cloud shadows within the images were removed, and then the mean value of each pixel over the collection of images was selected, resulting in a single, cloud-free image.

3.1.3 Sentinel-1

Akin to Sentinel-2, the Sentinel-1 is a mission launched by ESA, consisting of two satellites [99]. Both satellites carry a C-band SAR sensor, with the ability to produce measurements in VV, VH, HH, and HV polarizations. The satellites have four different operational modes: Strip Map (SM), Interferometric Wide Swath (IW), Extra Wide Swath (EW), and Wave (WV). The SM mode features an 80 km swath with 5 x 5 meter resolution, the highest available on the Sentinel-1. However, it is mainly reserved for emergencies, and the high-resolution data is not generally available.

The IW is the main operational mode for imaging land [99]. It has a 250 km swath and a 5 x 20 meter resolution. The data is available in multiple different processing levels, out of which the Level 1 Ground Range Detected (GRD) was used exclusively in this work. The GRD data has been calibrated and processed in a multitude of ways, including speckle reduction, and radiometric and geometric correction. The end product has been projected to ground range and is in a cartographic format. The GRD data has a resolution of 20 x 22 m, with a 10 x 10 m pixel spacing.

To make the GRD data more viable for analysis, further processing can be applied [55]. In the present work, a collection of images was first acquired. Next, as the GRD images are prone to backscatter noise along the image edges, border noise correction was applied to the images. The methodology used is as per Stasolla and Neyt [2018], and resulted in noisy pixels being masked out.

Further, the Refined Lee Filter (RLF) [85] was used to reduce speckling. The RLF in an alteration of the original Lee algorithm (see Eq. 2.8) [44], with the addition of selecting an adjustable maximum number of neighboring pixels (k) to consider in the calculations. The advantage of RLF is that the optimal parameters can be chosen for each task, with lower values of k resulting in sharper images at the cost of overall cohesiveness. In the present work, mainly lower values of $k \in \{5, 7, 9\}$ were experimented with to retain image features and sharpness for textural analysis.

In the present study, RFL was used with the multi-temporal filtering framework proposed by Quegan and Yu [2001]. The framework allows the application of mono-

temporal speckle filters — such as RFL — over a collection of images. The methodology of Quegan and Yu has the goal of preserving image radiometry, while optimally reducing speckle by utilizing the additional information available from the multi-temporal dataset. The filtering framework also serves to even out the highest variations from the selected image. In the present work, the multi-temporal filtering conceivably resulted in a more robust statistical model, as data from different time-spans with differing characteristics were used, depending on the location of the sample areas.

Further, radiometric terrain normalization is applied to the images to mitigate the effect of topography on the backscatter values. The methodology used is as per Vollrath et al. [2020]. As RTN requires an additional topographical elevation map, the Global Digital Elevation Model (GDEM) [1], produced from data by the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) was used. The post-RTN imagery is in a gamma-nought format, with shadow areas and pixels falling under an active layover being masked.

The final SAR pre-processing step is selecting a single image from the series and correcting the pixels masked both by border noise correction and RTN. The additional advantage of the Quegan and Yu multi-temporal filtering framework is the increased cohesion between images, which allows the noise-masked pixels in an image to be feasibly replaced by the values in other images of the collection. For analysis, an image from the middle of the time-series was selected, with unmasked pixels derived from temporally the nearest images, where applicable. An example of the processed imagery can be seen in Fig. 3.3.

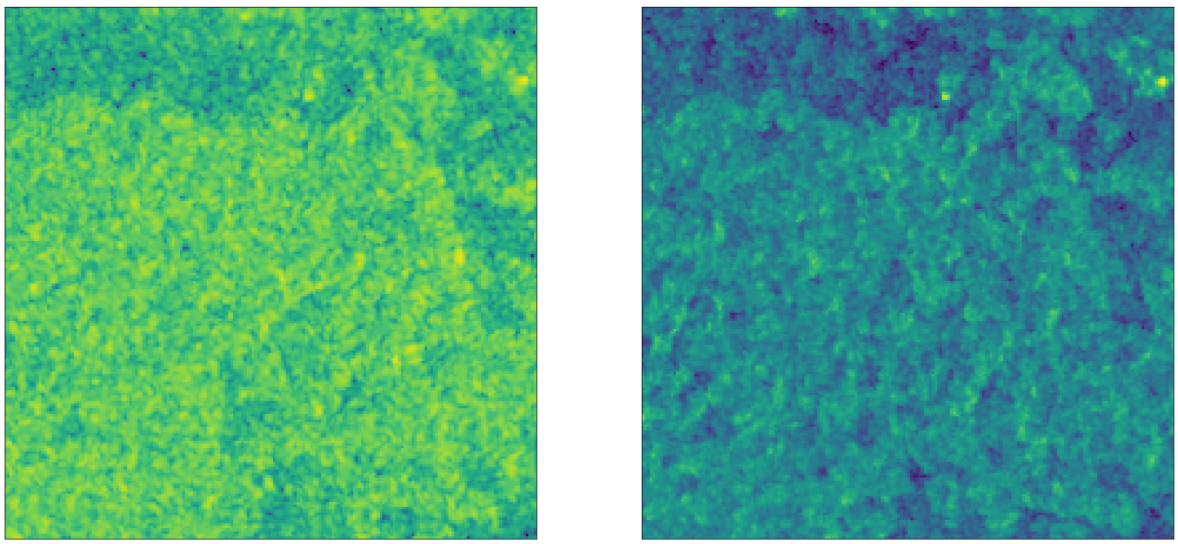

Figure 3.3: A visualization of the VH polarization band of Sentinel-1 Ground Range Detected imagery, before and after processing (on the right). The sample area is the same as in Fig. 3.2.

3.1.4 ALOS-2 PALSAR-2

The Advanced Land Observing Satellite-2 (ALOS-2) is a satellite launched by the Japan Aerospace Exploration Agency, launched in 2014 [90]. It carries the PALSAR-2 L-band SAR sensor, which is able to produce measurements with resolutions as high as 3 x 1 m. In the present research, however, the 25 x 25 m resolution imagery is used. The data is available as a global mosaic composed of manually screened images, with a single high-quality observation for each area per year [91].

The imagery in the dataset is orthorectified and slope corrected, meaning the effects of planar tilt and terrain relief on the imagery have been compensated for [91]. A 90 m resolution DEM, constructed from the data provided by the Shuttle Radar Topography Mission (SRTM) was used in the process. In addition, as neighboring images in the mosaic can be from differing parts of the year, a destriping process [69] was applied to even out the differences resulting from seasonal and environmental factors.

In this work, the global mosaic data was further processed by applying a refined Lee speckle filter. In addition, the imagery was resampled to 10 x 10 m resolution using bilinear interpolation.

3.2 Feature Extraction Methods

A variety of feature extraction methods were utilized in this research. These are roughly divided into three categories: vegetation indices, textural analysis, and deriving features from the neighboring pixels.

3.2.1 Vegetation Indices

Various vegetation indices are calculated from Sentinel-2 data. Interestingly, Sentinel-2 offers multiple wavelength bands that are close together, notably the red edge bands and the NIR bands B8 and B8A (see Table 3.1) [100]. This variety of relatively interchangeable bands is especially useful in vegetation detection and estimation, increasing the flexibility of building various vegetation indices. For example, NDVI can be formulated as

$$NDVI = \frac{B_8 - B_4}{B_8 + B_4}\,, \tag{3.2}$$

where B_8 is NIR and B_4 the red band. But, the NIR band can be substituted with the

Narrow NIR band (B_{8A}), leading to a formulation of

$$NDVI2 = \frac{B_{8A} - B_4}{B_{8A} + B_4}\,, \tag{3.3}$$

which, in practice, possibly performs differently compared to the standard NDVI.

As the specifics of the captured spectral bands — and especially the features of the red edge — depend on the environment and the imaged species of vegetation, a multitude of vegetation indices with several band variations were calculated tested in the present work. All the used formulations consisted of straightforward arithmetic operations, such as in Eq. 3.2 and Eq. 3.4. For a full list of the 64 distinct Sentinel-2 vegetation indices used in this research, including the equations, see Appendix A.

In addition to Sentinel-2 vegetation indices, the HH and HV bands from ALOS-2 PALSAR-2 were used to calculate the Radar Forest Degradation Index (RFDI) [19]. Although RFDI is mainly used for evaluating forest and loss and recovery, it can also be useful in biomass estimation. RFDI is defined as

$$RFDI = \frac{\gamma^0_{HH} - \gamma^0_{HV}}{\gamma^0_{HH} + \gamma^0_{HV}}\,, \tag{3.4}$$

where the HH and HV bands are gamma-nought (γ^0) normalized.

3.2.2 Gray-Level Co-occurrence Matrix

Gray-Level Co-occurrence Matrix (GLCM) was chosen as a method of extracting textural features. GLCM is based on the distribution of co-occurring values at a given offset [27]. A grayscale image is used for analysis, e.g. a single band satellite image.

Given an image I with L gray levels, a GLCM G is a square matrix with dimensions of N x N, where $N = L$. Each element $g_{i,j}$ in G is the number of times a pixel with the value i occurs adjacent to a pixel with the value j in I [27]. In Fig. 3.4 a 4 x 5 image with 8 gray levels (left) can be observed, along with a corresponding GLCM (right). In the GLCM, $g_{1,1} = 1$, as the gray level value 1 occurs exactly once next to an 1 in the original image. Similarly, $g_{1,2} = 2$, since the value 2 occurs twice next to pixels with the value 1.

In Fig. 3.4, the GLCM is constructed by considering the pixel value directly to the right of the original pixel. However, the analysis can be computed using any direction horizontally, vertically, or diagonally. Multiple pixel positions can be calculated for a

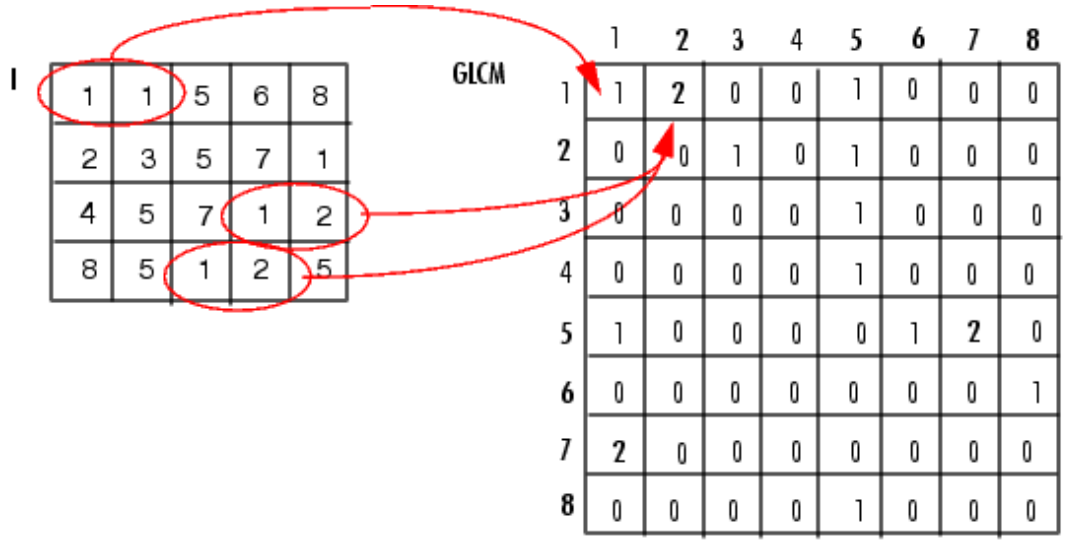

Figure 3.4: A GLCM (on the right) and the original image I (on the left) [95]. The GLCM is created using horizontal direction, considering the value of a pixel directly to the right of each original pixel.

single image, resulting in multiple co-occurrence matrices. These matrices can then be combined into one, e.g. by averaging the directional occurrences for each value.

The matrix G can be made symmetric with

$$G^S = G + G^T \,, \tag{3.5}$$

where G^T is the transposition of G. This addition ensures that each relationship i to j is the equal to the relationship j to i. Consequently, the sum of all the elements i, j in G^S is now twice the element-wise sum of the original matrix G. This can be corrected by normalization

$$P = \frac{G^S}{\sum_{i=1}^{N} \sum_{j=1}^{N} g^s_{i,j}} \,, \tag{3.6}$$

where each element i, j in G^S is divided by the sum of all the elements in G^S. This results in the matrix P, where each element i, j is the probability of the relationship i, j or j, i occurring in the original image I.

Finally, texture measures can then be calculated from P [27]. There are multiple measures, although for remote sensing and AGB estimation there are 8 commonly used ones [49, 24]: angular second moment, inverse difference moment, variance, correlation, dissimilarity, contrast, and entropy. The equations can be seen in Table 3.2. As an

example, the formulation for Angular Second Moment (AMS) is

$$\sum_i \sum_j (P_{i,j})^2 \ , \tag{3.7}$$

where P is a normalized GLCM, and $P_{i,j}$ is the probability of value j occurring in relation to value i.

Angular second moment, Entropy (ENT), and Difference Entropy (DEN) are measurements of textural uniformity [27]. ASM is used to calculate how regular the pixel value differences are within the image; Higher ASM values correlate with more regular differences, meaning the textures are more uniform. Entropy functions in the opposite way: The higher the entropy value, the less regular the pixel differences are, and the less texturally uniform the image is. Disorder entropy is similar to entropy, except that it uses the differences of the gray levels, instead of the gray levels directly.

Contrast (CON), Dissimilarity (DIS), and Inverse Difference Moment (IDM) are measures of homogeneity [27]. Contrast measures the amount of local variations in the image. The more similar the nearby pixels are in relation to each other, the less contrast the image has. Dissimilarity also measures the amount of local variances, except that it scales linearly, instead of exponentially, like contrast. IDM is strongly inversely correlated with contrast and dissimilarity: When all the elements in the image are the same, IDM is at its maximum value.

Variance (VAR) and Correlation (COR) are part of the descriptive statistical texture measures [27]. Variance (VAR) is similar to contrast and dissimilarity, and can be thought of as a measure of heterogeneity. While CON and DIS measure the difference between highest and lowest values, VAR measures the dispersion around the mean. Correlation, on the other hand, measures the linear dependency of neighboring pixel values. Higher correlation equates with a more predictable linear relationship. And lastly, the Sum Average (SAV) measures the relationship between occurrences of pairs with high values and the pairs with low values.

In applications of AGB estimation, a moving window W is often used instead of the full image I [10, 12, 49]. A GLCM is calculated for the subset of M within W, and then the window is moved to the next position. The process is repeated until the image (or a specified region of the image) has been analyzed. The window can be of any size, although sizes ranging from 3 x 3 to 15 x 15 are commonly used in AGB estimation research.

A common practice is to start with a small window size and increase the dimensions until sufficient textural patterns emerge [48]. In general, large window sizes result

Texture measure		Formulation
Angular Second Moment	ASM	$\sum_i \sum_j (P_{i,j})^2$
Entropy	ENT	$\sum_i P \log(P_{i,j})$
Difference Entropy	DEN	$\sum_{i=0}^{N-1} P_{x-y}(i) \log\{P_{x-y}(i)\}$
Contrast	CON	$\sum_i \sum_j (i-j)^2 P_{i,j}$
Dissimilarity	DIS	$\sum_i \sum_j P_{i,j} \lvert i-j \rvert$
Inverse Difference Moment	IDM	$\sum_i \sum_j \frac{1}{1+(i-j)^2} P_{i,j}$
Variance	VAR	$\sum_i \sum_j (i-\mu)^2 P_{i,j}$
Correlation	COR	$\sum_i \sum_j P_{i,j} [\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}}]$
Sum Average	SAV	$\sum_{i=2}^{2N} i P_{x+y}(i)$

Table 3.2: GLCM texture measures most commonly found in above-ground biomass estimation research [27].

in lower spatial resolution and increase the estimation errors near separate spatial instances, while small window sizes result in noisier textural features. The range of viable window sizes greatly depend on the resolution of the source image, with high-resolution images allowing higher window sizes to extract textural patterns. For AGB estimation, textural features are usually less useful when derived from low-resolution imagery.

3.2.3 Utilizing the Neighboring Values

As a novel method of both compensating for the saturation effect of AGB estimation and improving the prediction accuracy, the surrounding feature values for each pixel were also included in the set of features. After extracting the desired features, we had a three-dimensional N x M x J matrix, with each cell $C_{n,m}$ — where $n \in \{1, 2, \ldots, N\}$ and $n \in \{1, 2, \ldots, M\}$ — containing values $x_{m,n,j}$, where $j \in \{1, 2, \ldots, J\}$. Furthermore, $x_{m,n,j} \in X_j$, where X is a set containing J extracted features.

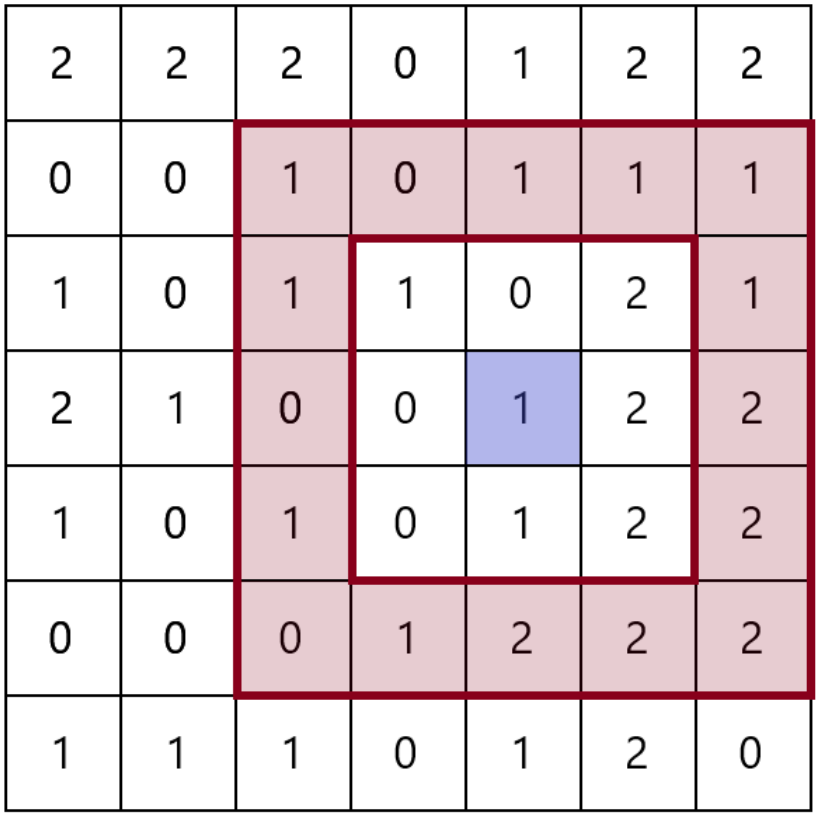

Figure 3.5: A figure illustrating the geometry of a 5 x 5 kernel used in the analysis of neighboring feature values. Only the cells highlighted with red are used for the particular calculations. The resulting values are added as additional feature values to the center cell, which is highlighted with blue.

For each of the cells $c_{n,m}$, additional features were created by considering the neighboring values. This was done by creating a sliding kernel with the dimensions of

d x d, which was moved over the N and M dimensions of the matrix. The kernel was shaped like a hollow square, and it only considered the cells on the edges of the shape. As such, the kernel can be intuited as a cell wide band that encircles the center. See Fig. 3.5 for clarification.

The feature values for each of the k cells falling within the kernel were summed together and divided by k. The value of k, by default, was $d^2 - (d-2)^2$. If a cell contained no data and was masked, or the kernel was out of the bounds of the matrix, k was subtracted from. The process can be formulated as

$$x^B_{n,m,j} = \frac{1}{k} \sum_{i=1}^{k} x_{i,j} \tag{3.8}$$

where, if the cell $c_{n,m}$ is at the center of the kernel, $x^B_{n,m,j}$ is the new feature value in the set X^B for feature X^B_j. x_i, and is a feature value belonging to a cell within the kernel. The process was carried out for each feature in $\{X_1, X_2, \ldots, X_J\}$, resulting in the new feature set X^B having J features.

In short, a duplicate feature set containing the encircling feature values of each cell, within a certain range was created. In this work, kernel size d was either 3, 5, or 7. Depending on the resolution and characteristics of the source data, as well as the details of the application, higher values can be used as well.

Before the applying the process, the values of the features used were standardized by removing the mean and scaling to variance. This is formulated as:

$$z = \frac{x - \mu}{\sigma} \tag{3.9}$$

where x is the sample value, μ is the mean, and σ is the standard deviation. As the features were of different distributions and scales, the standardization was performed in order to make the neighbor processing feasible for as many features as possible. In addition, as PCA was applied to the new features, the standardization was useful.

3.3 Statistical Methods

3.3.1 Principal Component Analysis

To reduce the dimensionality of the data, Principal Component Analysis (PCA) was used. Specifically, PCA was mostly applied to the new feature sets derived from the

neighboring values of data points. The primary purpose of the neighboring feature processing was to have a representation of the thickness of the surround vegetation for each observation. As such, minimizing its possible negative impact on the trained models — by adding unneeded complexity — was done by extracting principal components from the features.

In PCA, a linear transformation is applied to maximize the variance of the data [84]. The bands are first standardized, and then a series of components are produced. The components are linear combinations of original bands, with the intention of representing the variation within the original data while correlating minimally with each other. The resulting component bands are ordered by the amount of scene variance, in descending order.

PCA is a flexible process that can be used for a wide variety of purposes, ranging from visualization to unsupervised learning [84]. In addition, as a dimensionality reduction tool, it is applicable to datasets with various characteristics. It is, for example, resistant to missing data and imprecise measurements. While some accuracy is lost as PCA removes redundant data, the simpler data is faster and more efficient to explore, analyze, and process.

As a downside, PCA is a linear method, and does not consider complex relationships. However, for the purposes of this research — especially as a tool of reducing the dimensionality of neighbor-derived feature sets — capturing simple relationships is likely sufficient.

3.3.2 Metrics

To rank and evaluate the performance of the statistical models, the adjusted coefficient of determination (R_a^2), the Root Mean Squared Error (RMSE), and the Mean Percentage Error (MPE). R^2 is a measure of the proportion of the variance in the dependent variable that is explainable by the independent variables. It is used to measure how well a regression model fits the data. The equation for R^2 is

$$R^2 = 1 - \frac{\sum_i^n e_i^2}{\sum_i^n (y_i - \bar{y})} , \tag{3.10}$$

where e_i^2 is the square of the residual error for the prediction i, $\bar{y}$ is the mean of the dependent variable y, and n is the number of observations. In essence, R^2 is the ratio between the sum of squared residuals and the total sum of squares of the variable y. Generally, the value of R^2 ranges from 1 (all the variance is explainable, and the model fits perfectly) to 0.

While R^2 is often used for regression tasks, it has disadvantages when used with multiple input variables; Adding new variables to the model will often increase R^2, even if the variable was redundant and the quality of the model did not increase. In fact, the actual goodness of the fit might decrease due to additional complexity introduced by the redundant variable.

For the purposes of this work, the adjusted R^2 — denoted as R_a^2 — was used instead of R^2. Using the original equation of R^2, R_a^2 can be formulated as

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p} \tag{3.11}$$

where n is the number of observations, and p is the total number of variables in the model. As such, R_a^S is a method of effectively considering the number of variables and adjusting the score accordingly.

In addition to R_a^2, RMSE was used as well. RMSE is a measure of how close the predicted values are to the observed values. It is defined as

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n} e_i^2}\ . \tag{3.12}$$

Lower RMSE generally means better fit of the model. As the RMSE and the outcome variable are measured in the same unit, the error metric is often intuitive to interpret. In addition, the percentage value of RMSE — denoted as RMSE(%) — is used to make the results more comparable. It is calculated by dividing RMSE with the mean value of observations.

For feature evaluation purposes, Spearman's rank correlation coefficient (R_S) and mutual information (I) were used. Spearman's correlation is a measure of monotonic relationship between two variables. Its value can range from -1 to 1. If the nature of the monotonic relationship is such that the value of one variable decreases as the other's increases (i.e., the correlation is negative), the value of R_S is negative. If the variables have a positive correlation — meaning an increase in the value of one variable equals an increase in the value of the other one — the R_S is positive. The further away from zero the value of R_S is, the higher the negative or positive correlation between the variables.

The Spearman's coefficient is similar to other correlation statistics, such as the Pearson's correlation coefficient (r). Unlike r, (R_S) uses the rank of the variable values, instead of the actual variables. This is advantageous when evaluating two variables with

different distributions; For statistics such r, the data generally needs to be normalized, which can lead to loss of information.

The Spearman's rank correlation coefficient is formulated as

$$R_S = \frac{\mathrm{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} , \tag{3.13}$$

where $R(X)$ and $R(Y)$ are the rank variable of X and Y, function cov notates covariance between two variables, and σ is the standard deviation.

Mutual information (I) measures mutual dependence between two random variables [75]. Its value ranges from 0 to 1, where 0 means there is no relationship between the variables. Generally, the metric is based on the entropy values of the variables and can be formulated as

$$I(X,Y) = \iint dxdy\mu(x,y) \log \frac{\mu(x,y)}{mu_x(x)\mu_y(y)} , \tag{3.14}$$

where X and Y are random variables, and $\mu_x(x) = \int dy(x,y)$ and $\mu_y(y) = \int dx(x,y)$ the marginal densities of X and Y.

The problem with this variation of I is that the densities μ, μ_x, and mu_y need to be either known or estimated, which is not always feasible [41]. To calculate I using only the set X, Y, a k-Nearest Neighbor (k-NN) based method can be used. While there are multiple k-NN based mutual information algorithms, the approach utilized in this work is defined as

$$I(X,Y) = \psi(k) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(N) , \tag{3.15}$$

where N is the number of realizations (x_i, y_i), $i = 1, \ldots, N$ of the random variable (X, Y), and $\psi(x)$ is the digamma function $\psi(x) = \Gamma(y)^{-1} d\Gamma(x)/dx$. Here, $C = 0.5772156$ is the Euler-Maschoni constant, and the digamma function satisfies the recursion $\psi(x+1) = \psi(x) + 1/x$ and $\psi(1) = -C$.

Intuitively, the simplified version of the idea is to consider the distribution of X, Y as a space where the distance to neighboring values can be measured. The I is then estimated from the average k-nearest neighbor, normalized by an averaging function.

The averaging function $\langle \ldots \rangle$ in Eq. 3.15 is defined as

$$\langle \ldots \rangle = N^{-1} \sum_{i=1}^{N} \mathbb{E}_{1_n}^{N}[1, \ldots, N] \; . \tag{3.16}$$

If we use $\mathbb{E}_x(i)$ to denote the distance from x_i to its k-th neighbor, then $\mathbb{E}(i) = \max\{\mathbb{E}_x(i), \mathbb{E}_y(i)\}$. In short, $\langle \ldots \rangle$ averages over a distance measure of all the observations (x_i, y_i), $i = 1, \ldots, N$.

The advantage of mutual information is its ability to find almost any sort of relationship between two datasets. While R_S measures only the monotonic relationship between two variables, I can measure more complex relationships. As such, it is suited for selecting performant variables in complicated datasets where relationships are not immediately obvious. Consequently, I is also sensitive to noise and is prone to detecting ineffective relationships. In addition, to reach the same level of reliability as more straightforward metrics, mutual information generally requires larger datasets.

3.3.3 Feature Selection

In this work, hundreds of features were extracted from the data, and only the most relevant features needed to be selected for training the models. Denoting Y as the outcome variable and X as the input feature set of n feature variables $\{X_1, X_2, \ldots X_n\}$, the following automatic process was employed:

(i) Spearman's rank correlation coefficient $R_S(X_i, Y)$ and mutual information $I(X_i, Y)$ was calculated for all variable pairs (X_i, Y) from i to n.

(ii) $R_S(X_i, X_j)$ was calculated for each variable pair (X_i, X_j), $i, j \in \{1, 2, \ldots, n\}$ where $i \neq j$. If $|R_S(X_i, X_j)|$ — the absolute value of $R_S(X_i, X_j)$ — was higher than the constant c, either X_i or X_j was removed from the set of features. To decide which, the sum of the two scoring metrics

$$\begin{aligned} M(X_i) &= |R_S(X_i, Y)| + I(X_i, Y) \;\; \text{and} \\ M(X_j) &= |R_S(X_j, Y)| + I(X_j, Y) \end{aligned} \tag{3.17}$$

were compared. If $M(X_i) \geq M(X_j)$, then X_j was removed; Else, X_i was removed. Denoting the number of removed features as r, the set X now contained $m = n - r$ features.

(iii) The features in the reduced set $\{X_1, X_2, \ldots, X_m\}$ were ranked according to the scores $\{M(X_1), M(X_2), \ldots, M(X_m)\}$. Finally, top k features from the ranked set were selected.

The motivation behind this specific process was to evaluate both monotonic and more complex relationships between the input features and the outcome variable. The method allows the selection of variables with complex relationships, while still having reliably performing variables with high monotonic correlation. Furthermore, in item (ii), the $R^S(X_i, X_j)$ was used to measure the monotonic correlation between two feature variables, in order to not have redundant, highly similar features in the final dataset. The constant s can be a value between 0 and 1, although in this work, generally values from 0.8 to 0.95 were experimented with.

However, while both $|R_S|$ and I range from 0 to 1, they are not directly comparable. As such, it is simplistic to sum together the two values, as was done in items (ii) and (iii). In addition, using the straightforward statistical metrics is unlikely to predict the performance of the features in complex machine learning models. As such, the method is only fit for crude feature selection, in order to exclude tens or hundreds of features quickly.

3.3.4 Random Forest

In this work, Random Forest (RF) was used as a statistical model to predict the amount of above-ground biomass. Random forest is a popular, well performing algorithm for a wide variety of tasks. While RF can also be used for classification, in the present research it is used for regression.

Random forest is an ensemble model, meaning it consists of several sub-models. For RF, the sub-model is a decision tree. A decision tree is a tree-like structure that is created by stratifying the predictor space into small sections. Intuitively, the tree is defined by its branching construction, where each junction of the branches — known as a node — is an evaluation of an observation. The tree is traveled from the trunk onward, eventually leading to an end node, which results in a prediction produced by the tree.

Formally, the predictor space $x_1, x_2, \ldots, x_n$ of a regression decision tree is divided into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$. Then, for every observation that belongs into a region R_j, same prediction is made. The prediction is usually the mean value of responses for all the training observations in R_j.

A decision tree is a highly flexible model prone to overfitting. To make the predictions more generalizable, a random forest can be used. An RF is built by splitting the

training data into B subsets and training a decision tree for each one. The prediction given by a random forest is the mean prediction of the B decision trees.

Generally, the training data is split using bootstrapping. In bootstrapping, n observations are randomly selected from the data. Some observations can be sampled multiple times, while others not at all. This method is used to estimate variability, without having to sample independent data from the full population.

Random forests generally only consider a subset of the predictor variables at each node. Combined with the variability from bootstrapping, the effect of strong predictors is mitigated. This allows for more variation in the trees and reduces the correlation between them, ultimately lowering variance.

As a non-parametric model, random forest is well suited for AGB estimation. RF does not require the training data to follow any specific distribution, or that the predictor values are normalized. This is useful, since the ranges and distributions of values from the various feature extraction methods can differ greatly — as is the case in the present work.

4. Results

The training data was sampled from 15 plots within the hemiboreal zone of Finland (see Fig. 4.1). Each area was a 2 x 2 km square, split into 10 x 10 m observation cells. A single sample area contained 40 000 individual observations, totaling a combined 600 000 observations over the 15 plots. From the total, 169 601 were discarded due to missing data, primarily either from the MS-NFI biomass dataset or Sentinel-2 imagery. Ultimately, the final training outcome data Y contained $n = 430\,399$ individual observations. In addition, observations from five separate 2 x 2 km square areas were used for evaluating the performance of the trained models. The test sample areas can be seen in Fig. 4.2.

In the training outcome set $Y = \{y_1, y_2, \ldots, y_n\}$, each outcome y_k equaled the combined amount of above-ground biomass (in 10 kg / ha) for scots pine, spruce, and broad-leaved trees for the corresponding 10 x 10 m area. The testing outcome set Y^{test} was similar to Y, except that it was sampled from separate testing areas, resulting in a total of 200 000 samples. In addition, a separate, high biomass testing outcome set was created from the observations of Y^{test}. The observations had biomass volume higher than 200 Mg / ha, although this requirement limited the sample size to 27 302. All the measurements were from the year 2019.

The feature set X of the training dataset $D = \{(X, Y)\}$ consisted of multiple subsets. Each of the subsets was a collection of features, and various combinations of these subsets were evaluated in the research. The primary subsets were the features extracted from Sentinel-2 data (X^{S2}), Sentinel-1 data (X^{S1}), and ALOS-2 PALSAR-2 data (X^{A}). For each of the subsets, a large set of features were evaluated using several metrics, and the highest performing ones selected.

In addition, for Sentinel-1 and Sentinel-2 each there were two separate datasets, for a total of four: S1-HVG, S1-LVG, S2-HVG, S2-LVG. The S1-HVG and S2-HVG datasets — with HVG denoting High Vegetation Growth — were from imagery captured during summer, in order to get observations of the highest possible amount of grown vegetation. The datasets were from the period of time from June 1st, 2019 to August 15th, 2019. To supplement the high-growth dataset, the low-growth S1-LVG and S2-LVG datasets were from spring — before the major vegetation growth of early

Figure 4.1: The sample areas (marked in red), residing in the South-Western coastal area of Finland.

Figure 4.2: The separate testing areas (marked in red), used to evaluate the performance of the trained models.

summer, while having low snow cover. The time-span for the *LG*-datasets was from March 1st, 2019, to May 15th, 2019.

A total of 75 distinct features were initially chosen for Sentinel-2, including of 11 spectral bands and of 64 vegetation indices calculated using the bands (see A for the full list of vegetation indices). All 75 features were sampled from both high-vegetation (S2-HVG) and low-vegetation (S2-LVG) datasets, resulting in the feature sets of X^{S2HVG} and X^{S2LVG}. To select only the most relevant features, the procedure outlined in subsection 3.3.3 was applied to X^{S2HVG} and X^{S2LVG} separately. In the final step (iii), however, the two sets were joined together, and for each feature the summed score from

	HVG (Jun - Aug)		LVG (March - May)	
Feature	R_S	I	R_S	I
B2	0.67	0.39	0.62	0.44
EVI2	0.66	0.60	0.86	0.76
SARVI2-C	0.64	0.59	0.67	0.54
B1	0.59	0.30	0.51	0.33
MSI-B	0.56	0.36	0.71	0.41
B3	0.54	0.48	0.54	0.49
MARI	0.53	0.32	0.79	0.51
B4	0.45	0.47	0.45	0.53
MSI	0.40	0.28	0.55	0.27
MCARI2-B	0.39	0.40	0.53	0.57

Table 4.1: The Sentinel-2 bands and vegetation indices used for training the models. The respective R^2 and I scores are shown separately for high vegetation and low vegetation datasets.

	HVG (Jun - Aug)		LVG (March - May)	
Feature	R_S	I	R_S	I
VH-D-SAV-5	0.61	0.41	0.61	0.41
VH-A-SAV-7	0.60	0.39	0.60	0.37
VH-A-SAV-5	0.60	0.37	0.60	0.37
VH-A-SAV-3	0.59	0.39	0.58	0.36
VH-D-SAV-3	0.59	0.38	0.61	0.40
VV-A-SAV-7	0.59	0.36	0.50	0.31
VH-D-SAV-3	0.59	0.38	0.60	0.40
VH-D-DEN-7	0.57	0.28	0.56	0.29
VV-D-SAV-7	0.56	0.35	0.55	0.33
VV-A-SAV-3	0.56	0.35	0.48	0.29

Table 4.2: The Sentinel-1 features used for training the models. The respective R^2 and I scores are shown separately for high vegetation and low vegetation datasets.

both X^{S2LVG} and X^{S2LVG} determined the ranks to get the top k features.

Values $s = 0.95$ and $k = 10$ were used to select the final set of Sentinel-2 derived features. The features and the corresponding R_S and I values can be seen in Table 4.1.

For selecting features from Sentinel-1, the process was identical to Sentinel-2 data. The set X^S1, however, consisted of 110 features: nine GLCM texture metrics (as seen in Table 3.2), for the two polarization bands used (VV and VH), for both ascending and descending orbits, for three separate window sizes (3 x 3, 5 x 5, 7 x 7), separately — in addition to the original two bands.

Ultimately, 10 features were selected for the final training set X^{S1} (see Table 4.2). In the order of the highest sum of R_S and I scores, the features were: VH-D-SAV-5, VH-A-SAV-7, VH-A-SAV-5, VH-A-SAV-3, VH-D-SAV-3, VV-A-SAV-7, VH-D-SAV-3, VH-

Feature	R_S	I
HV-SAV-3	0.75	0.52
HH-SAV-3	0.67	0.44
RFDI	0.44	0.19
HH-VAR-7	0.32	0.10
HH-COR-7	0.31	0.09
HV-CON-7	0.30	0.10
HV-VAR-7	0.30	0.12
HH-VAR-5	0.29	0.08
HH-CON-5	0.29	0.07
HH-ENT-7	0.28	0.08

Table 4.3: The ALOS-2 PALSAR-2 features used for training the models, along with the respective R^2 and I score.

D-DEN-7, VV-D-SAV-7, and VV-A-SAV-3. The names are derived from polarization (VV or VH), orbit (ascending or descending), GLCM texture measure (refer to Table 3.2), and the window size of the GLCM algorithm.

As the ALOS-2 dataset only offered a single image per year for an area, a seasonal division could not be made. Otherwise, the feature extraction process for X^A was similar to X^{S1} and X^{S2}. The original feature set consisted of the HH and HV polarization bands, the ratio of HH and HV, the radar forest degradation index, and nine GLCM textural measures calculated for both polarizations, using window sizes of 3 x 3, 5 x 5, and 7 x 7.

Before selection, X^A consisted of 58 features, out of which 10 were chosen (see Table 4.3). Ordered by the sum of R_S and I scores, the features were: HV-SAV-3, HH-SAV-3, RFDI, HH-VAR-7, HH-COR-7, HV-CON-7, HV-VAR-7, HH-VAR-5, HH-CON-5, and HH-ENT-7. Similarly to Sentinel-1 features, the names are composed of polarization (HH or HV), GLCM textural measure, and the GLCM window size. Orbital variations were not available for the ALOS-2 dataset.

In addition, the neighboring feature sets for each data source were created as per subsection 3.2.3, with LVG and HVG datasets being processed separately. The created subsets are denoted as X^{S2N_k}, X^{S1N_k}, and X^{ALOSN_k}, where k is the kernel size used for the algorithm. Values of 3, 5, and 7 were used, and datasets created for each. A separate selection process was used to get the most viable features for the sets, with the addition of using PCA to reduce the features to three components for each dataset. The PCA components accounted for 98.9%, 98.2%, and 99.8% of variance of the ALOS2, Sentinel-1, and Sentinel-2 features, respectively.

Several models using various combinations of the described sets of features were trained. The models were evaluated using observations from the test areas, with R_a^2, RMSE, RMSE(%) scores calculated. In addition, the models were separately tested for their performance in estimating high-volume biomass (more than 200 Mg / ha), although the observation set was very low, with a population of 27302. The results can be seen in Tables 4.4 and 4.5.

The terms S2 and S1 without HVG or LVG suffix denote full Sentinel-1 and Sentinel-2 datasets, using both HVG and LVG data. In addition, control datasets (denoted as CTRL) were trained. For Sentinel-1 the CTRL-set contained the VV, VH, VV-A, and VH-V features. For Sentinel-2, the selected features were all the available spectral bands (B1 to B11), in addition to NDVI. For ALOS, the CTRL-set contained HH and HV polarization bands. In addition, for Sentinel-2 a model using all of the 74 features was evaluated. This model is denoted as S2-ALL.

	All areas			High biomass areas		
Model	R_a^2	RMSE	RMSE(%)	R_a^2	RMSE	RMSE(%)
S2-HVG-ALL	0.71	34.18	37.0%	0.46	36.91	35.9%
S2-HVG-CTRL	0.77	32.21	35.9%	0.50	35.92	33.8%
S2-HVG	0.79	31.68	35.6%	0.52	34.88	33.7%
S2-HVG-N3-NPCA	0.80	30.41	34.9%	0.51	34.62	33.5%
S2-HVG-N3	0.81	30.36	34.8%	0.53	34.68	33.5%
S2-HVG-N3-N5	0.81	30.33	34.8%	0.52	34.62	33.5%
S2-HVG-N3-N5-N7	0.80	30.38	34.8%	0.51	34.70	33.5%
S2	0.80	30.82	35.1%	0.51	34.79	33.6%
S2-N3	0.81	30.20	34.7%	0.52	34.41	33.2%
S2-N3-N5	0.80	30.41	34.9%	0.51	34.66	33.5%
S2-N3-N5-N7	0.79	30.44	34.9%	0.51	34.67	33.5%

Table 4.4: The trained models and their respective R_a^2, RMSE, and RMSE(%) scores. The results are shown for both unrestricted observations and for high-biomass observations.

	All areas			High biomass areas		
Model	R_a^2	RMSE	RMSE(%)	R_a^2	RMSE	RMSE(%)
S1-HVG	0.37	52.22	47.6%	-0.29	57.80	58.8%
S1-HVG-CTRL	0.38	52.29	47.6%	-0.30	57.69	58.7%
S1-HVG-N3	0.36	52.18	47.5%	-0.29	57.73	58.7%
S1	0.36	52.11	57.5%	-0.29	57.90	58.9%
ALOS	0.44	50.31	46.4%	-0.08	52.41	52.9%
ALOS-CTRL	0.45	50.28	46.4%	-0.07	51.41	51.8%
ALOS-N3	0.45	49.31	45.9%	-0.09	52.46	53.0%
S2	0.80	30.62	35.0%	0.50	34.49	33.3%
S2 + S1-HVG	0.81	30.49	34.9%	0.51	34.47	33.3%
S2 + S1-HVG + ALOS	0.79	31.26	35.3%	0.50	34.55	33.4%
S2-N3 + ALOS	0.80	31.24	35.3%	0.51	34.52	33.3%
S2-N3 + S1-HVG	0.78	31.90	35.7%	0.50	36.52	35.5%
S2-N3 + ALOS + S1-HVG	0.77	32.90	36.3%	0.49	36.62	35.6%

Table 4.5: The trained models and their respective R_a^2, RMSE, and RMSE(%) scores. The results are shown for both unrestricted observations and for high-biomass observations.

5. Discussion

Somewhat surprisingly, the NIR bands B8 and B8A were removed from the Sentinel-2 feature set by the selection process (see 4.1). However, most of the vegetation indices within the final feature set included B8 and B8A in their formulations. According to the model metrics, the feature selection process for Sentinel-2 showed some improvements.

For Sentinel-1 feature set (Table 4.2), by far the most correlating textural measure was the Summed Average, with the only other feature being the Difference Entropy. The summed average metrics of various GLCM window sizes performed better than either of VV or VH bands, which might be indicative of the noisiness of the Sentinel-1 imagery, even after de-speckling and radiometric corrections.

For ALOS-2, the features were more varied (4.3). Interestingly, the vegetation index RFDI was a high-performing feature. Similarly to Sentinel-1, neither of the original polarization bands were included in the final set.

Considering the performance of the Sentinel-1 and ALOS-2 models, the feature selection for the two SAR datasets meant little. While the final models performed slightly better than the control models using the most basic features, the difference was negligible. In addition, the Sentinel-1 and ALOS-2 models using the feature values of neighboring pixels showed very little improvement. Both of these factors can be explained by the low spatial resolution of the imagery, and the lack of original bands in the SAR data — all the subsequent textural features were derived from two bands. In addition, the specific images of the ALOS-2 mosaic might be unviable for vegetation measurements, or from vastly different time periods (e.g., some samples being taken during winter and some during summer), making it difficult for the statistical models to fit to the data.

The feature selection of Sentinel-2 seemed to have some, but little relevance in the results. Compared to the control model (S2-HVG-CTR) and the model utilizing all 74 features (S2-HVG-ALL), the S2-HVG with the selected 10 features performed slightly better. In addition, the neighbor selection for Sentinel-2 seemed to perform quite well, with the RMSE scores of the S2-N3 model (30.20) surpassing the scores of models utilizing both Sentinel-1 and Sentinel-2 data (S2-N3 + ALOS with RMSE of 31.13, S2-N3 + S1-HVG with RMSE of 31.90, and S2-N3 + ALOS + S1-HVG with

RMSE of 32.90). While this is indicative of the usefulness of the neighbor-feature extraction, the results are also partly explained by the lack of performance from the features derived from Sentinel-1 and ALOS-2. As such, using these free sources of data for AGB estimation in the hemiboreal zone might not be feasible.

However, the results of the SAR features can also be explained by the preprocessing methods. For example, the speckle filters, resampling methods, and radiometric corrections can be experimented with. As such, more testing is needed to ascertain the conclusion that the data is unsuitable for AGB estimation in hemi-boreal Finland.

In addition, the Spearman rank correlation coefficient and mutual information -based feature extraction process might not have been optimal for the SAR datasets. In fact, the method used was very simplistic, and further feature selection would be needed to increase the performance of the models.

The lowest RMSE of the trained models (30.20) was achieved by the S2-N3 model, which was simply using high vegetation and low vegetation features, along with the extracted neighboring features. This is indicative of the efficacy of using the neighboring features as a method of enriching Sentinel-2 optical data for biomass purposes. Further research, however, is needed, as the sample sizes, especially for high biomass test set, were small. As such, the effectiveness of counter-acting the saturation effect could not be evaluated. As a note, all the models performed worse in the task of estimating the high-biomass samples, which shows the saturation effect taking place even for the relatively low biomass volumes (200 - 300 Mg / ha) used in the sample set.

While the results were on par with comparable research [77, 58, 73], the lowest prediction RMSE of 30.20, with RMSE(%) of 34.7%, is still high for many AGB estimation purposes. In addition, evaluation of the data and features using models beside Random Forest would be a useful addition in future research.

6. Conclusion

In this study, a novel application of above-ground biomass estimation in the hemi-boreal zone of Finland was described. In addition, a novel method of extracting features from neighboring observations was proposed and evaluated. The findings show that the method is useful, having the highest RMSE score among the model tested. However, more research is needed. In addition, the effectiveness of the method as a tool to compensate for the saturation effect of high-biomass areas is yet to be ascertained. This is partly due to the low biomass density of hemi-boreal Finland, in general.

Additionally, the viability of using free and highly available satellite datasets for AGB estimation in sample area was analyzed, with the results suggesting that the free SAR based products had a lacking performance, possibly due to low spatial resolution or incorrect preprocessing adjustments. The features extracted from the optical data of Sentinel-2 produced well-performing models, although the prediction accuracy might still be too low to be usable.

Bibliography

[1] Michael Abrams, Robert Crippen, and Hiroyuki Fujisada. ASTER global digital elevation model (GDEM) and ASTER global water body dataset (ASTWBD). *Remote Sensing*, 12(7):1156, April 2020. doi: 10.3390/rs12071156. URL https://doi.org/10.3390/rs12071156.

[2] Gregory P. Asner, George V. N. Powell, Joseph Mascaro, David E. Knapp, John K. Clark, James Jacobson, Ty Kennedy-Bowdoin, Aravindh Balaji, Guayana Paez-Acosta, Eloy Victoria, Laura Secada, Michael Valqui, and R. Flint Hughes. High-resolution forest carbon stocks and emissions in the amazon. *Proceedings of the National Academy of Sciences*, 107(38):16738–16742, September 2010. doi: 10.1073/pnas.1004875107. URL https://doi.org/10.1073/pnas.1004875107.

[3] Valerio Avitabile and Andrea Camia. An assessment of forest biomass maps in europe using harmonized national statistics and inventory plots. *Forest Ecology and Management*, 409:489–498, February 2018. doi: 10.1016/j.foreco.2017.11.047. URL https://doi.org/10.1016/j.foreco.2017.11.047.

[4] A. Bannari, D. Morin, F. Bonn, and A. R. Huete. A review of vegetation indices. *Remote Sensing Reviews*, 13(1-2):95–120, August 1995. doi: 10.1080/02757259509532298. URL https://doi.org/10.1080/02757259509532298.

[5] F. Baret and G. Guyot. Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sensing of Environment*, 35(2-3):161–173, February 1991. doi: 10.1016/0034-4257(91)90009-u. URL https://doi.org/10.1016/0034-4257(91)90009-u.

[6] T.M. Basuki, P.E. van Laake, A.K. Skidmore, and Y.A. Hussin. Allometric equations for estimating the above-ground biomass in tropical lowland dipterocarp forests. *Forest Ecology and Management*, 257(8):1684–1694, March 2009. doi: 10.1016/j.foreco.2009.01.027. URL https://doi.org/10.1016/j.foreco.2009.01.027.

[7] Michael A. Cairns, Sandra Brown, Eileen H. Helmer, and Greg A. Baumgardner. Root biomass allocation in the world's upland forests. *Oecologia*, 111(1): 1–11, June 1997. doi: 10.1007/s004420050201. URL https://doi.org/10.1007/s004420050201.

[8] Jeannine Cavender-Bares, John A. Gamon, and Philip A. Townsend, editors. *Remote Sensing of Plant Biodiversity*. Springer International Publishing, 2020. doi: 10.1007/978-3-030-33157-3. URL https://doi.org/10.1007/978-3-030-33157-3.

[9] Pietro Ceccato, Stéphane Flasse, Stefano Tarantola, Stéphane Jacquemoud, and Jean-Marie Grégoire. Detecting vegetation leaf water content using reflectance in the optical domain. *Remote Sensing of Environment*, 77(1):22–33, July 2001. doi: 10.1016/s0034-4257(01)00191-2. URL https://doi.org/10.1016/s0034-4257(01)00191-2.

[10] D. Chen, D. A. Stow, and P. Gong. Examining the effect of spatial resolution and texture window size on classification accuracy: an urban environment case. *International Journal of Remote Sensing*, 25(11):2177–2192, January 2004. doi: 10.1080/01431160310001618464. URL https://doi.org/10.1080/01431160310001618464.

[11] Jing M. Chen and Josef Cihlar. Retrieving leaf area index of boreal conifer forests using landsat TM images. *Remote Sensing of Environment*, 55(2):153–162, February 1996. doi: 10.1016/0034-4257(95)00195-6. URL https://doi.org/10.1016/0034-4257(95)00195-6.

[12] Lin Chen, Yeqiao Wang, Chunying Ren, Bai Zhang, and Zongming Wang. Optimal combination of predictors and algorithms for forest above-ground biomass mapping from sentinel and SRTM data. *Remote Sensing*, 11(4):414, February 2019. doi: 10.3390/rs11040414. URL https://doi.org/10.3390/rs11040414.

[13] E. A. CLOUTIS, D. R. CONNERY, D. J. MAJOR, and F. J. DOVER. Airborne multi-spectral monitoring of agricultural crop status: effect of time of year, crop type and crop condition parameter. *International Journal of Remote Sensing*, 17 (13):2579–2601, September 1996. doi: 10.1080/01431169608949094. URL https://doi.org/10.1080/01431169608949094.

[14] R CRIPPEN. Calculating the vegetation index faster. *Remote Sensing of Environment*, 34(1):71–73, October 1990. doi: 10.1016/0034-4257(90)90085-z. URL https://doi.org/10.1016/0034-4257(90)90085-z.

[15] C Daughtry. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sensing of Environment*, 74(2):229–239, November 2000. doi: 10.1016/s0034-4257(00)00113-9. URL https://doi.org/10.1016/s0034-4257(00)00113-9.

[16] DW Deering. Measuring" forage production" of grazing units from landsat mss data. In *Proceedings of the Tenth International Symposium of Remote Sensing of the Envrionment*, pages 1169–1198, 1975.

[17] Jesús Delegido, Jochem Verrelst, Luis Alonso, and José Moreno. Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. *Sensors*, 11(7):7063–7081, July 2011. doi: 10.3390/s110707063. URL https://doi.org/10.3390/s110707063.

[18] Michael J. Falkowski, Paul E. Gessler, Penelope Morgan, Andrew T. Hudak, and Alistair M.S. Smith. Characterizing and mapping forest fire fuels using ASTER imagery and gradient modeling. *Forest Ecology and Management*, 217 (2-3):129–146, October 2005. doi: 10.1016/j.foreco.2005.06.013. URL https://doi.org/10.1016/j.foreco.2005.06.013.

[19] Africa Flores, K. Herndon, Rajesh Thapa, and Emil Cherrington. Synthetic aperture radar (sar) handbook: Comprehensive methodologies for forest monitoring and biomass estimation. 2019. doi: 10.25966/NR2C-S697. URL https://gis1.servirglobal.net/TrainingMaterials/SAR/SARHB_FullRes.pdf.

[20] William James Frampton, Jadunandan Dash, Gary Watmough, and Edward James Milton. Evaluating the capabilities of sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 82:83–92, August 2013. doi: 10.1016/j.isprsjprs.2013.04.007. URL https://doi.org/10.1016/j.isprsjprs.2013.04.007.

[21] Victor S. Frost, Josephine Abbott Stiles, K. S. Shanmugan, and Julian C. Holtzman. A model for radar images and its application to adaptive digital filtering of multiplicative noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(2):157–166, March 1982. doi: 10.1109/tpami.1982.4767223. URL https://doi.org/10.1109/tpami.1982.4767223.

[22] Anatoly A. Gitelson. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. *Journal of Plant Physiology*, 161(2):165–173, January 2004. doi: 10.1078/0176-1617-01176. URL https://doi.org/10.1078/0176-1617-01176.

[23] Anatoly A. Gitelson, Yoram J. Kaufman, and Mark N. Merzlyak. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, 58(3):289–298, December 1996. doi: 10.1016/s0034-4257(96)00072-7. URL https://doi.org/10.1016/s0034-4257(96)00072-7.

[24] Sérgio Godinho, Nuno Guiomar, and Artur Gil. Estimating tree canopy cover percentage in a mediterranean silvopastoral systems using sentinel-2a imagery and the stochastic gradient boosting algorithm. *International Journal of Remote Sensing*, 39(14):4640–4662, November 2017. doi: 10.1080/01431161.2017.1399480. URL https://doi.org/10.1080/01431161.2017.1399480.

[25] D Haboudane. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, 90(3):337–352, April 2004. doi: 10.1016/j.rse.2003.12.013. URL https://doi.org/10.1016/j.rse.2003.12.013.

[26] Steven Hancock, Ciara McGrath, Christopher Lowe, Ian Davenport, and Iain Woodhouse. Requirements for a global lidar system: spaceborne lidar with wall-to-wall coverage. *Royal Society Open Science*, 8(12), December 2021. doi: 10.1098/rsos.211166. URL https://doi.org/10.1098/rsos.211166.

[27] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, November 1973. doi: 10.1109/tsmc.1973.4309314. URL https://doi.org/10.1109/tsmc.1973.4309314.

[28] Michael Hardisky, Victor Klemas, and and Smart. The influence of soil salinity, growth form, and leaf moisture on the spectral radiance of spartina alterniflora canopies. *Photogrammetric Engineering and Remote Sensing*, 48:77–84, 01 1983.

[29] Huaijiang He, Chunyu Zhang, Xiuhai Zhao, Folega Fousseni, Jinsong Wang, Haijun Dai, Song Yang, and Qiang Zuo. Allometric biomass equations for 12 tree species in coniferous and broadleaved mixed forests, northeastern china. *PLOS ONE*, 13(1):e0186226, January 2018. doi: 10.1371/journal.pone.0186226. URL https://doi.org/10.1371/journal.pone.0186226.

[30] D.N.H. Horler, M. Dockray, J. Barber, and A.R. Barringer. Red edge measurements for remotely sensing plant chlorophyll content. *Advances in Space Research*, 3(2):273–277, January 1983. doi: 10.1016/0273-1177(83)90130-8. URL https://doi.org/10.1016/0273-1177(83)90130-8.

[31] A Huete. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sensing of Environment*, 59(3):440–451, March 1997. doi: 10.1016/s0034-4257(96)00112-5. URL https://doi.org/10.1016/s0034-4257(96)00112-5.

[32] A Huete, K Didan, T Miura, E.P Rodriguez, X Gao, and L.G Ferreira. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1-2):195–213, November 2002. doi: 10.1016/s0034-4257(02)00096-2. URL https://doi.org/10.1016/s0034-4257(02)00096-2.

[33] A.R Huete. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25(3):295–309, August 1988. doi: 10.1016/0034-4257(88)90106-x. URL https://doi.org/10.1016/0034-4257(88)90106-x.

[34] A.R. HUETE. REMOTE SENSING FOR ENVIRONMENTAL MONITORING. In *Environmental Monitoring and Characterization*, pages 183–206. Elsevier, 2004. doi: 10.1016/b978-012064477-3/50013-8. URL https://doi.org/10.1016/b978-012064477-3/50013-8.

[35] E. Raymond Hunt, C. S. T. Daughtry, Jan U. H. Eitel, and Dan S. Long. Remote sensing leaf chlorophyll content using a visible band index. *Agronomy Journal*, 103(4):1090–1099, July 2011. doi: 10.2134/agronj2010.0395. URL https://doi.org/10.2134/agronj2010.0395.

[36] Z JIANG, A HUETE, K DIDAN, and T MIURA. Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, 112(10):3833–3845, October 2008. doi: 10.1016/j.rse.2008.06.006. URL https://doi.org/10.1016/j.rse.2008.06.006.

[37] Carl F. Jordan. Derivation of leaf-area index from quality of light on the forest floor. *Ecology*, 50(4):663–666, July 1969. doi: 10.2307/1936256. URL https://doi.org/10.2307/1936256.

[38] E.R Hunt Jr and B Rock. Detection of changes in leaf water content using near- and middle-infrared reflectances73. *Remote Sensing of Environment*, 30(1):43–54, October 1989. doi: 10.1016/0034-4257(89)90046-1. URL https://doi.org/10.1016/0034-4257(89)90046-1.

[39] R. J. Kauth and G. S. Thomas. The tasseled cap – a graphic description of the spectral-temporal development of agricultural crops as seen by landsat. *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*, 1976.

[40] Yunjin Kim and J.J. van Zyl. A time-series approach to estimate soil moisture using polarimetric radar data. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8):2519–2527, August 2009. doi: 10.1109/tgrs.2009.2014944. URL https://doi.org/10.1109/tgrs.2009.2014944.

[41] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6), June 2004. doi: 10.1103/physreve.69.066138. URL https://doi.org/10.1103/physreve.69.066138.

[42] Lalit Kumar and Onisimo Mutanga. Remote sensing of above-ground biomass. *Remote Sensing*, 9(9):935, September 2017. doi: 10.3390/rs9090935. URL https://doi.org/10.3390/rs9090935.

[43] Gaia Vaglio Laurin, Johannes Balling, Piermaria Corona, Walter Mattioli, Dario Papale, Nicola Puletti, Maria Rizzo, John Truckenbrodt, and Marcel Urban. Above-ground biomass prediction by sentinel-1 multitemporal data in central italy with integration of ALOS2 and sentinel-2 data. *Journal of Applied Remote Sensing*, 12(01):1, January 2018. doi: 10.1117/1.jrs.12.016008. URL https://doi.org/10.1117/1.jrs.12.016008.

[44] Jong-Sen Lee. Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(2):165–168, March 1980. doi: 10.1109/tpami.1980.4766994. URL https://doi.org/10.1109/tpami.1980.4766994.

[45] Mingshi Li, Ying Tan, Jie Pan, and Shikui Peng. Modeling forest aboveground biomass by combining spectrum, textures and topographic features. *Frontiers of Forestry in China*, 3(1):10–15, March 2008. doi: 10.1007/s11461-008-0013-z. URL https://doi.org/10.1007/s11461-008-0013-z.

[46] A. Lopes, R. Touzi, and E. Nezry. Adaptive speckle filters and scene heterogeneity. *IEEE Transactions on Geoscience and Remote Sensing*, 28(6):992–1000, 1990. doi: 10.1109/36.62623. URL https://doi.org/10.1109/36.62623.

[47] Patrícia Lourenço, Sérgio Godinho, Adélia Sousa, and Ana Cristina Gonçalves. Estimating tree aboveground biomass using multispectral satellite-based data in mediterranean agroforestry system using random forest algorithm. *Remote Sensing Applications: Society and Environment*, 23:100560, August 2021. doi: 10.1016/j.rsase.2021.100560. URL https://doi.org/10.1016/j.rsase.2021.100560.

[48] PatrÃcia LourenÃ§o. Biomass estimation using satellite-based data. In Ana Cristina GonÃ§alves, AdÃ©lia Sousa, and Isabel Malico, editors, *Forest Biomass*, chapter 3. IntechOpen, Rijeka, 2021. doi: 10.5772/intechopen.93603. URL https://doi.org/10.5772/intechopen.93603.

[49] D. Lu. Aboveground biomass estimation using landsat TM data in the brazilian amazon. *International Journal of Remote Sensing*, 26(12):2509–2525, June 2005. doi: 10.1080/01431160500142145. URL https://doi.org/10.1080/01431160500142145.

[50] Leo Lymburner, Paul Beggs, and Carol Jacobson. Estimation of canopy-average surface-specific leaf area using landsat tm data. *Photogrammetric Engineering and Remote Sensing*, 66:183–191, 02 2000.

[51] Kai Mäkisara, Matti Katila, and Jouni PerÃ€saari. The multi-source national forest inventory of finland â methods and results 2015. 03 2019.

[52] Michael Manton, Charles Ruffner, Gintautas Kibirkštis, Gediminas Brazaitis, Vitas Marozas, Rūtilė Pukienė, Ekaterina Makrickiene, and Per Angelstam. Fire occurrence in hemi-boreal forests: Exploring natural and cultural scots pine fire regimes using dendrochronology in lithuania. *Land*, 11(2):260, February 2022. doi: 10.3390/land11020260. URL https://doi.org/10.3390/land11020260.

[53] Mark N. Merzlyak, Anatoly A. Gitelson, Olga B. Chivkunova, and Victor YU. Rakitin. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia Plantarum*, 106(1):135–141, May 1999. doi: 10.1034/j.1399-3054.1999.106119.x. URL https://doi.org/10.1034/j.1399-3054.1999.106119.x.

[54] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1(1):6–43, March 2013. doi: 10.1109/mgrs.2013.2248301. URL https://doi.org/10.1109/mgrs.2013.2248301.

[55] Adugna Mullissa, Andreas Vollrath, Christelle Odongo-Braun, Bart Slagter, Johannes Balling, Yaqing Gou, Noel Gorelick, and Johannes Reiche. Sentinel-1 SAR backscatter analysis ready data preparation in google earth engine. *Remote Sensing*, 13(10):1954, May 2021. doi: 10.3390/rs13101954. URL https://doi.org/10.3390/rs13101954.

[56] R. Nedkov. Orthogonal transformation of segmented images from the satellite sentinel-2. *Comptes rendus de l'AcadÃ©mie bulgare des sciences: sciences mathÃ©matiques et naturelles*, 70:687–692, 05 2017.

[57] Charles R. Perry and Lyle F. Lautenschlager. Functional equivalence of spectral vegetation indices. *Remote Sensing of Environment*, 14(1-3):169–182, January 1984. doi: 10.1016/0034-4257(84)90013-0. URL https://doi.org/10.1016/0034-4257(84)90013-0.

[58] Henrik J. Persson, Jonas Jonzén, and Mats Nilsson. Combining TanDEM-x and sentinel-2 for large-area species-wise prediction of forest biomass and volume. *International Journal of Applied Earth Observation and Geoinformation*, 96:102275, April 2021. doi: 10.1016/j.jag.2020.102275. URL https://doi.org/10.1016/j.jag.2020.102275.

[59] Sorin C. Popescu. Estimating biomass of individual pine trees using airborne lidar. *Biomass and Bioenergy*, 31(9):646–655, September 2007. doi: 10.1016/j.biombioe.2007.06.022. URL https://doi.org/10.1016/j.biombioe.2007.06.022.

[60] Timo Pukkala. Delineating forest stands from grid data. *Forest Ecosystems*, 7 (1), March 2020. doi: 10.1186/s40663-020-00221-8. URL https://doi.org/10.1186/s40663-020-00221-8.

[61] S. Quegan and Jiong Jiong Yu. Filtering of multichannel SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 39(11):2373–2379, 2001. doi: 10.1109/36.964973. URL https://doi.org/10.1109/36.964973.

[62] Jaakko Repola. Biomass equations for birch in finland. *Silva Fennica*, 42(4): 605–624, 2008. doi: https://doi.org/10.14214/sf.236.

[63] Jaakko Repola. Biomass equations for scots pine and norway spruce in finland. *Silva Fennica*, 43(4):625–647, 2009. doi: https://doi.org/10.14214/sf.184.

[64] Arthur J Richardson and CL Wiegand. Distinguishing vegetation from soil background information. *Photogrammetric engineering and remote sensing*, 43(12): 1541–1552, 1977.

[65] Anamaria Roman and Tudor Ursu. *Multispectral satellite imagery and airborne laser scanning techniques for the detection of archaeological vegetation marks*, pages 141–152. 12 2016. ISBN 978-606-543-787-6.

[66] Geneviève Rondeaux, Michael Steven, and Frédéric Baret. Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, 55(2):95–107, February 1996. doi: 10.1016/0034-4257(95)00186-7. URL https://doi.org/10.1016/0034-4257(95)00186-7.

[67] Jean-Louis Roujean and François-Marie Breon. Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, 51(3):375–384, March 1995. doi: 10.1016/0034-4257(94)00114-3. URL https://doi.org/10.1016/0034-4257(94)00114-3.

[68] John Wilson Rouse, Robert H. Haas, John A. Schell, and D. W. Deering. Monitoring vegetation systems in the great plains with erts. volume 351, pages 309–317, 1973.

[69] M. Shimada and O. Isoguchi. JERS-1 SAR mosaics of southeast asia using calibrated path images. *International Journal of Remote Sensing*, 23(7):1507–1526, January 2002. doi: 10.1080/01431160110092678. URL https://doi.org/10.1080/01431160110092678.

[70] Nikolaos G. Silleos, Thomas K. Alexandridis, Ioannis Z. Gitas, and Konstantinos Perakis. Vegetation indices: Advances made in biomass estimation and vegetation monitoring in the last 30 years. *Geocarto International*, 21(4):21–28, December 2006. doi: 10.1080/10106040608542399. URL https://doi.org/10.1080/10106040608542399.

[71] David Small. Flattening gamma: Radiometric terrain correction for SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 49(8):3081–3093, August 2011. doi: 10.1109/tgrs.2011.2120616. URL https://doi.org/10.1109/tgrs.2011.2120616.

[72] Mattia Stasolla and Xavier Neyt. An operational tool for the automatic detection and removal of border noise in sentinel-1 GRD products. *Sensors*, 18(10):3454, October 2018. doi: 10.3390/s18103454. URL https://doi.org/10.3390/s18103454.

[73] Nikos Theofanous, Irene Chrysafis, Giorgos Mallinis, Christos Domakinis, Natalia Verde, and Sofia Siahalou. Aboveground biomass estimation in short rotation forest plantations in northern greece using ESA's sentinel medium-high resolution multispectral and radar imaging missions. *Forests*, 12(7):902, July 2021. doi: 10.3390/f12070902. URL https://doi.org/10.3390/f12070902.

[74] Amadou Khoudiedji Thiam. *Geographic information systems and remote sensing methods for assessing and monitoring land degradation in the Sahel region: the case of southern Mauritania*. Clark University, 1998.

[75] MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

[76] Compton J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, May 1979. doi: 10.1016/0034-4257(79)90013-0. URL https://doi.org/10.1016/0034-4257(79)90013-0.

[77] Sasan Vafaei, Javad Soosani, Kamran Adeli, Hadi Fadaei, Hamed Naghavi, Tien Pham, and Dieu Tien Bui. Improving accuracy estimation of forest aboveground biomass based on incorporation of ALOS-2 PALSAR-2 and sentinel-2a imagery and machine learning: A case study of the hyrcanian forest area (iran). *Remote Sensing*, 10(2):172, January 2018. doi: 10.3390/rs10020172. URL https://doi.org/10.3390/rs10020172.

[78] Kuimi T. Vashum. Methods to estimate above-ground biomass and carbon stock in natural forests - a review. *Journal of Ecosystem &amp Ecography*, 02(04), 2012. doi: 10.4172/2157-7625.1000116. URL https://doi.org/10.4172/2157-7625.1000116.

[79] Andreas Vollrath, Adugna Mullissa, and Johannes Reiche. Angular-based radiometric slope correction for sentinel-1 on google earth engine. *Remote Sensing*, 12(11):1867, June 2020. doi: 10.3390/rs12111867. URL https://doi.org/10.3390/rs12111867.

[80] Anthony G. Vorster, Paul H. Evangelista, Atticus E. L. Stovall, and Seth Ex. Variability and uncertainty in forest biomass estimates from the tree to landscape scale: the role of allometric equations. *Carbon Balance and Management*, 15(1), May 2020. doi: 10.1186/s13021-020-00143-6. URL https://doi.org/10.1186/s13021-020-00143-6.

[81] Xiangping Wang, Jingyun Fang, and Biao Zhu. Forest biomass and root–shoot allocation in northeast china. *Forest Ecology and Management*, 255(12):4007–4020, June 2008. doi: 10.1016/j.foreco.2008.03.055. URL https://doi.org/10.1016/j.foreco.2008.03.055.

[82] Aaron R. Weiskittel, David W. MacFarlane, Philip J. Radtke, David L.R. Affleck, Hailemariam Temesgen, Christopher W. Woodall, James A. Westfall, and

John W. Coulston. A call to improve methods for estimating tree biomass for regional and national assessments. *Journal of Forestry*, 113(4):414–424, July 2015. doi: 10.5849/jof.14-091. URL https://doi.org/10.5849/jof.14-091.

[83] Jean-Pierre Wigneron, Albert Olioso, Jean-Christophe Calvet, and Patrick Bertuzzi. Estimating root zone soil moisture from surface soil moisture data and soil-vegetation-atmosphere transfer modeling. *Water Resources Research*, 35(12):3735–3745, December 1999. doi: 10.1029/1999wr900258. URL https://doi.org/10.1029/1999wr900258.

[84] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, August 1987. doi: 10.1016/0169-7439(87)80084-9. URL https://doi.org/10.1016/0169-7439(87)80084-9.

[85] Aiyeola Sikiru Yommy, Rongke Liu, , and Shuang Wu. SAR image despeckling using refined lee filter. In *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*. IEEE, August 2015. doi: 10.1109/ihmsc.2015.236. URL https://doi.org/10.1109/ihmsc.2015.236.

[86] Yifan Yu and Sassan Saatchi. Sensitivity of l-band SAR backscatter to aboveground biomass of global forests. *Remote Sensing*, 8(6):522, June 2016. doi: 10.3390/rs8060522. URL https://doi.org/10.3390/rs8060522.

[87] Xiaonong Zhou, Lin Dandan, Yang Huiming, Chen Honggen, Sun Leping, Yang Guojing, Hong Qingbiao, Leslie Brown, and J.B Malone. Use of landsat TM satellite surveillance data to measure the impact of the 1998 flood on snail intermediate host dispersal in the lower yangtze river basin. *Acta Tropica*, 82(2):199–205, May 2002. doi: 10.1016/s0001-706x(02)00011-6. URL https://doi.org/10.1016/s0001-706x(02)00011-6.

[88] Junling Zhu, Jianguo Wen, and Yafeng Zhang. A new algorithm for SAR image despeckling using an enhanced lee filter and median filter. In *2013 6th International Congress on Image and Signal Processing (CISP)*. IEEE, December 2013. doi: 10.1109/cisp.2013.6743991. URL https://doi.org/10.1109/cisp.2013.6743991.

[89] Lingli Zhu, Juha Suomalainen, Jingbin Liu, Juha Hyyppä, Harri Kaartinen, and Henrik Haggren. A review: Remote sensing sensors. In *Multi-purposeful Application of Geospatial Data*. InTech, May 2018. doi: 10.5772/intechopen.71049. URL https://doi.org/10.5772/intechopen.71049.

[90] Japan Aerospace Exploration Agency – ALOS-2 Project / PALSAR-2, . Accessed May 27, 2022. https://www.eorc.jaxa.jp/ALOS-2/en/about/palsar2.htm.

[91] Earth Engine Data Catalog – Global PALSAR-2/PALSAR Yearly Mosaic, . Accessed June 01, 2022 https://developers.google.com/earth-engine/datasets/catalog/JAXA_ALOS_PALSAR_YEARLY_SAR.

[92] Chemistry Libretexts – Electromagnetic Spectrum. Accessed May 10, 2022. https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Map:_Introductory_Chemistry_(Tro)/09:_Electrons_in_Atoms_and_the_Periodic_Table/9.03:_The_Electromagnetic_Spectrum.

[93] European Space Agency – BIOMASS. Accessed May 19, 2022. https://www.esa.int/Applications/Observing_the_Earth/FutureEO/Biomass.

[94] GEDI Ecosystem Lidar. Accessed May 19, 2022. https://gedi.umd.edu/.

[95] Image processing toolbox user's guide. Accessed May 24, 2022. http://http://matlab.izmiran.ru/help/toolbox/images/enhanc15.html.

[96] National Resources Institute Finland – File Service. Accessed May 27, 2022. https://kartta.luke.fi/index-en.html.

[97] Forest Centre – Spatial Datasets. Accessed May 27, 2022. https://www.metsakeskus.fi/en/node/946.

[98] Redd+ web platform. Accessed Jun 07, 2022. https://redd.unfccc.int/.

[99] European Space Agency – Sentinel-1. Accessed May 27, 2022. https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1.

[100] European Space Agency – Sentinel-2. Accessed May 27, 2022. https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2.

Appendix A. Features Used

The following tables contain the formulations and references for the Sentinel-2 based vegetation indices used in this work. The band variables ($B1, B2, B3$, etc.) denote the Sentinel-2 spectral band used. The alphabetically notated indices (e.g., EVI-B, S2REP-C) are variations of the original equations, where some of the bands are substituted by similar bands.

Vegetation Index		Equation	Ref.
Difference Vegetation Index	DVI	$B8 - B4$	[76]
Difference Vegetation Index B	DVI-B	$B8A - B4$	[76]
Enhanced Vegetation Index	EVI	$2.5\frac{B8A - B5}{B8A - 6 \cdot B5 - 7.5 \cdot B2 + 1}$	[32]
Enhanced Vegetation Index B	EVI-B	$2.5\frac{B8 - B4}{B8 - 6 \cdot B4 - 7.5 \cdot B2 + 1}$	[32]
Enhanced Vegetation Index C	EVI-C	$2.5\frac{B8A - B4}{B8A - 6 \cdot B4 - 7.5 \cdot B2 + 1}$	[32]
Enhanced Vegetation Index D	EVI-D	$2.5\frac{B7 - B4}{B7 - 6 \cdot B4 - 7.5 \cdot B2 + 1}$	[32]
2-band Enhanced Vegetation Index	EVI2	$2.5\frac{B8A - B5}{B8A + \cdot B5 + 1}$	[36]
2-band Enhanced Vegetation Index 2	EVI2-2	$2.5\frac{B8A - B5}{B8A + 2.4 \cdot B5 + 1}$	
Green Normalized Difference Vegetation Index	GNDVI	$\frac{B8 - B3}{B8 + B3}$	[23]
Green Normalized Difference Vegetation Index B	GNDVI-B	$\frac{B7 - B3}{B7 + B3}$	[23]
Green Ratio Vegetation Index	GRVI	$\frac{B3 - B4}{B3 + B4}$	[18]
Infrared Percentage Vegetation Index	IPVI	$\frac{B8}{B8 + B4}$	[14]
Infrared Percentage Vegetation Index B	IPVI-B	$\frac{B8A}{B8A + B4}$	[14]
Inverted Red-Edge Chlorophyll Index	IRECI	$\frac{B7 - B4}{B5/B6}$	[20]
Modified Anthocyanin Reflectance Index	MARI	$(B3^{-1} - B5^{-1}) \cdot B8A$	[9]
Modified Chlorophyll Absorption in Reflectance Index	MCARI	$[(B5 - B4) - 0.2(B5 - B3)](\frac{B5}{B4})$	[15]
Modified Chlorophyll Absorption in Reflectance Index 1	MCARI1	$1.2(2.5(B8 - B4) - 1.3(B8 - B3))(\frac{B5}{B4})$	[25]
Modified Chlorophyll Absorption in Reflectance Index 1 B	MCARI1-B	$1.2(2.5(B8A - B4) - 1.3(B8A - B3))(\frac{B5}{B4})$	[25]
Modified Chlorophyll Absorption in Reflectance Index 2	MCARI2	$1.5\frac{2.5(B8 - B4) - 1.3(B8 - B3)}{\sqrt{(2 \cdot B8 + 1)^2 - (6 \cdot B8 - 5\sqrt{B4}) - 0.5}}$	[25]
Modified Chlorophyll Absorption in Reflectance Index 2 B	MCARI2-B	$1.5\frac{2.5(B8A - B4) - 1.3(B8A - B3)}{\sqrt{(2 \cdot B8A + 1)^2 - (6 \cdot B8A - 5\sqrt{B4}) - 0.5}}$	[25]
Modified Normalized Vegetation Index	MNDVI	$\frac{B8 - B4}{B8 + B4 + 2 \cdot B1}$	[31]
Modified Normalized Vegetation B Index	MNDVI-B	$\frac{B8A - B4}{B8A + B4 + 2 \cdot B1}$	[31]

Table A.1: The Sentinel-2 vegetation indices used in this work

Vegetation Index		Equation	Ref.
Modified Soil Adjusted Vegetation Index	MSAVI	$\frac{2 \cdot B8A + 1 - \sqrt{(2 \cdot B8A + 1)^2 - 8(B8A - B5)}}{2}$	[17]
Modified Soil Adjusted Vegetation Index Hyper	MSAVIH	$0.5[(2 \cdot B8 + 1) - \sqrt{(2 \cdot B8 + 1) - 8(B8 - B4)}]$	[17]
Modified Soil Adjusted Vegetation Index Hyper B	MSAVIH-B	$0.5[(2 \cdot B8A + 1) - \sqrt{(2 \cdot B8A + 1) - 8(B8A - B4)}]$	[17]
Moisture Stress Index	MSI	$\frac{B4}{B12}$	[38]
Moisture Stress Index B	MSI-B	$\frac{B4}{B11}$	[38]
Modified Simple Ratio	MSR	$\frac{B8/(B4-1)}{\sqrt{B8/B4 + 1}}$	[11]
Modified Simple Ratio B	MSR-B	$\frac{B8A/(B4-1)}{\sqrt{B8A/B4 + 1}}$	[11]
Modified Simple Ratio C	MSR-C	$\frac{B7/(B4-1)}{\sqrt{B7/B4 + 1}}$	[11]
Normalized Difference Index using bands 4 and 5	NDI45	$\frac{B5 - B4}{B5 + B4}$	[17]
Normalized Difference Vegetation Index	NDVI	$\frac{B8 - B4}{B8 + B4}$	[68]
Normalized Difference Vegetation Index B	NDVI-B	$\frac{B8A - B4}{B8A + B4}$	[68]
Normalized Difference Water Index	NDWI	$\frac{B8 - B10}{B8 + B10}$	[28]
Normalized Difference Water Index B	NDWI	$\frac{B8A - B10}{B8A + B10}$	[28]
Normalized Ratio Vegetation Index	NRVI	$\frac{B8/B4 - 1}{B8/B4 + 1}$	[5]
Normalized Ratio Vegetation Index B	NRVI-B	$\frac{B8A/B4 - 1}{B8A/B4 + 1}$	[5]
Optimized Soil-Adjusted Vegetation Index	OSAVI	$\frac{B8 - B4}{B8 + B4 + 0.16}$	[66]
Optimized Soil-Adjusted Vegetation Index B	OSAVI-B	$\frac{B8A - B4}{B8A + B4 + 0.16}$	[66]
Plant Senescence Reflectance Index	PSRI	$\frac{B4 - B2}{B6}$	[53]
Renormalized Difference Vegetation Index	RDVI	$\frac{B8 - B4}{\sqrt{B8 + B4}}$	[67]
Renormalized Difference Vegetation Index B	RDVI-B	$\frac{B8A - B4}{\sqrt{B8A + B4}}$	[67]

Table A.2: Some vegetation indices used in AGB estimation.

Vegetation Index		Equation	Ref.
Red Edge 1	RE1	$\frac{B5}{B4}$	[13]
Red Edge 2	RE2	$\frac{B5 - B4}{B5 + B4}$	[13]
Sentinel-2 Red-Edge Position	S2REP	$705 + 35\frac{(B7+B4/2) - B5}{B6 - B5}$	[20]
Sentinel-2 Red-Edge Position B	S2REP-B	$705 + 35\frac{(B8+B4/2) - RE1}{RE2 - RE1}$	[20]
Sentinel-2 Red-Edge Position C	S2REP-C	$705 + 35\frac{(B8A+B4/2) - RE1}{RE2 - RE1}$	[20]
Soil and Atmospherically Resistant Vegetation Index 2	SARVI2	$2.5\frac{B8 - B5}{1 + B8 + 6 \cdot B5 - 7.5 \cdot B1}$	[31]
Soil and Atmospherically Resistant Vegetation Index 2 B	SARVI2-B	$2.5\frac{B8A - B5}{1 + B8A + 6 \cdot B5 - 7.5 \cdot B1}$	[31]
Soil and Atmospherically Resistant Vegetation Index 2 C	SARVI2-C	$2.5\frac{B9 - B5}{1 + B9 + 6 \cdot B5 - 7.5 \cdot B1}$	[31]
Simple Ratio	SR	$B8/B4$	[37]
Simple Ratio B	SR-B	$B8A/B4$	[37]
Triangular Chlorophyll Index	TCI	$1.2(B5 - B3) - 1.5(B4 - B3) \cdot \sqrt{B5/B4}$	[35]
Transformed Normalized Difference Vegetation Index	TNDVI	$\sqrt{\frac{B8 - B4}{B8 + B4} + 0.5}$	[87]
Transformed Normalized Difference Vegetation Index	TNDVI	$\sqrt{\frac{B8 - B4}{B8 + B4} + 0.5}$	[87]
Transformed Normalized Difference Vegetation Index B	TNDVI-B	$\sqrt{\frac{B8A - B4}{B8A + B4} + 0.5}$	[87]
Transformed Normalized Difference Vegetation Index C	TNDVI-C	$\sqrt{\frac{B8A - B5}{B8A + B5} + 0.5}$	[87]
Transformed Triangular Vegetation Index	TTVI	$\sqrt{ABS(\frac{B8 - B4}{B8 + B4} + 0.5)}$	[74]
Transformed Triangular Vegetation Index B	TTVI-B	$\sqrt{ABS(\frac{B8A - B4}{B8A + B4} + 0.5)}$	[74]
Triangular Vegetation Index	TVI	$0.5[120(B4 - B3) - 200(B4 - B3)]$	[16]
Triangular Vegetation Index B	TVI-B	$0.5[120(B8A - B3) - 200(B4 - B3)]$	[16]
Transformed Vegetation Index	TraVI	$\sqrt{\frac{5 - 4}{5 + 4} + 0.5}$	[4]
Wide Dynamic Range Vegetation Index	WDRVI	$\frac{0.1 \cdot B8 - B4}{0.1 \cdot B8 + B4}$	[22]
Wide Dynamic Range Vegetation Index B	WDRVI-B	$\frac{0.1 \cdot B8A - B4}{0.1 \cdot B8A + B4}$	[22]

Table A.3: Some vegetation indices used in AGB estimation.