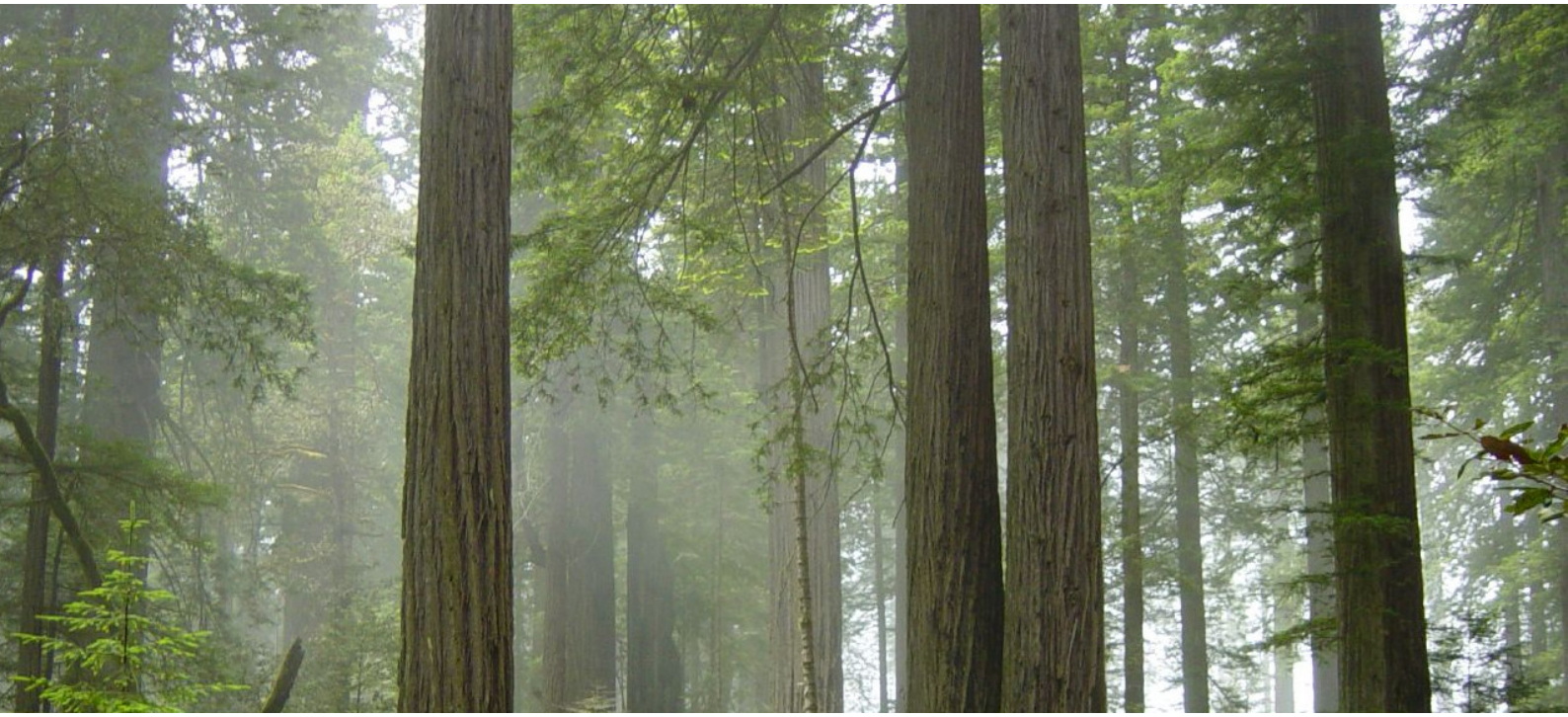

Preparation of a California Canopy Height Model

Proof of concept and methods report prepared for Vibrant Planet

Authors:
Fabian Döweler
Ian Reese

DRAGONFLY
Data Science

PO Box 27535, Wellington 6141
New Zealand
dragonfly.co.nz

Cover Notes

To be cited as:

this (**this**). **this**, 35 pages. Proof of concept and methods report prepared for Vibrant Planet.

Cover image:

https://www.flickr.com/photos/schweppe/21556590/

CONTENTS

EXECUTIVE SUMMARY

The expansion of Vibrant Planet's platform for natural resource management requires high-quality data to guide forestry management in the state of California. Best practice decision-making is currently limited to areas where LiDAR datasets are available. Within this model we provide a proof-of-concept how we aim extrapolate point cloud derived canopy height information to the whole state of California.

The prototype ingests publicly-available data, Sentinel-2 spectral bands, vegetation classification layers, to provide an approximation of relevant information without the deployment of costly LiDAR collection procedures.

This documents gives an overview of the state of this project, a brief breakdown of the methodological framework for the prototype and a literature research - revealing the extent of research considered to be deployed. A detailed documentation for geospatial processes, machine learning concepts and data ingestions via AWS can be found on the linked repositories.

TODO: A summary of the models developed, and performance of them

1. INTRODUCTION

This section provides an overview of the data sources considered and associated methods from the remote sensing and forestry research community. The canopy height model (CHM) derived from raw LiDAR point clouds constitutes the 'heart' of the model and will be used to train a machine learning model that can upscale satellite pixels (Sentinel-2) information to a state-wide layer.

While the initial proof-of-concept proposed within this document uses a CHM provided by the OpenTopography database, we already developed workflows to approximate this output and operate with known and scalable datasets which we intend to refine in alignment with Vibrant Planet's quality standards.

Sentinel-2 data

Sentinel data offers complete coverage of the area with a 5 day revisit time. Sentinel-2 derived vegetation indices can be used to to correlate with above-ground vegetation and offers higher resolution (10 - 20 m) than traditionally used Landsat data (30 m). Imagery has been post-processed by the European Space Agency (ESA) and cleaned uploads are available since 2018. Working with state-wide Sentinel-2 imagery requires reprojection into a uniform coordinate system and workflows to implement our own gridded tiling system. Once these workflows are in place we can use the whole database via S3 and query filter imagery with low cloud-obstruction to derive spectral indices which will aid us to upscale CHM information. We will also explore options for time-series analysis to monitor vegetation disturbance, recovery and improve existing vegetation classification.

LiDAR

LiDAR measurements are the state of the art technology to calculate above-ground biomass based on tree height estimates in forest ecosystems. We receive all currently available LiDAR point clouds from the opentopography database[1]. The datasets date back to 2010 and come in varying degrees of resolution and have been sampled with different techniques. Current workflows at Vibrant Planet rely on vendor provided digital terrain models (DTM) and use the Fusion software to derive a digital surface model (DSM) and a resulting canopy height model (CHM) as the difference between both layers. Since vendor provided CHM are not universally available for all datasets and the workflows to derive these are unknown, we are currently working on our own process to generate a DTM and further, a CHM using pdal[2]. This process will enable us to work with the full point cloud database and guarantee replicability with future LiDAR datasets. To verify the accuracy of our CHM the output is compared against current 'best practice' approaches within Fusion (vendor provided DTM, software calculated CHM) and freely available CHM's on the opentopo database. Before deploying these steps the procedure will be signed of by Vibrant Planet. Generating CHM's with pdal will enable us to scale data processing more readily than relying on Windows based software (Fusion).

For now we will focus our approach we use an area-based rather than a tree-centric approach for CHM's which is the current standard due to LiDAR restriction to detect below canopy/merged trees (**coomes_area-based_2017**). Further, to be able to delineate individual trees CHM resolutions < 1m are required (Van Kane, April 29 meeting). One

[1] Available here: https://portal.opentopography.org/datasets
[2] pdal documentation: https://pdal.io/

possibility to bridge the gap between area-based and tree-centric approaches could be the creation of tree approximate objects (TAO). These objects can be generated within Fusion using a watershed analysis and the area processor. We will work together with Colton to generate these outputs and test the viability of training satellite data with TAO information or ecological objects (Ecobject[3]).

Field data

The field plots are required in order to indicate the accuracy of our predictions in areas outside of LiDAR avaiability and to verify the quality of the generated CHM. Plot locations reveal information about tree species, DBH, categorized size classes (e.g. 8 - 16 m) and traditionaly consist of 4 radial plots (FIA & LiDAR plots). A known issue is the limited availability exact plot locations to ground-truth the output. It is unlikely that we will receive plot locations for a non-university project (Van Kane, April 29 meeting). Vibrant Planet provided us with LiDAR verification plots for the lake Tahoe area, but they do not contain individually geo-referenced trees and accurate height measurements are limited (e.g. 5 out of 20 trees for a sample plot). A current workflow as agreed with one of Cibrant Planet's forest ecologist to identify dominant trees and verify them against our CHM output or generate CHM's near Scott Conway's home to have an additional ground-truth reference. Alternatively we could artificially create gaps in in the LiDAR plots to verify our output against.

Vegetation classification

For the initial proof of concept we use a downscaled version of the extensive 'Classification and Assessment with Landsat of Visible Ecological Groupings database' (CALVEG) which summarizes a total of 32 vegetation classes for the entire state. The layer is a patchwork from 20+ years of vegetation mapping with careful literature research and field verification with the aim to deliver repeated measurements in a 15 year timeframe. The 'fire return interval departure dataset' (FRID[4]) groups CALVEG classification into subclasses and will be used to discriminate forested areas and non-vegetated surfaces (urban, lakes) in this proof of concept. In a later stage of the model the CALVEG database will serve as a more detailed predictor. We have already developed workflows to reproject, merge and rasterize the CALVEG datasets. We will retain the FRID dataset for future use, since it contains valuable information for historic fires in the area.

Upscaling

The upscaling process to translate locally available canopy height information to a state-wide layer involves the training of individual Sentinel-2 pixels. The protoype attached to this report incorporates a linear regression machine learning process to demonstrate the feasibility of this approach with no refined expectations towards it's accuracy. In the future the machine learning algorithm will be trained with a vast range of vegetation indices, LiDAR metrics and vegetation layers. While a random forest has been widely used in this context to approximate above-ground biomass estimations (**ghosh_aboveground_2018**, **matasci_large-area_2018**, **pham_monitoring_2017**, **huang_integration_2019**) we will explore more sophisticated options to optimize this

[3]EcObject documentation: https://buttecounty.opennrm.org/assets/e106ca2a359a122e74e33ef183a0fb4a/application/pdf/EcObjectProductGuide_Final_V1.pdf

[4]Documentation here: https://data.sacriver.org/assets/5ca70e8ae658de1d1e0331ff747c3f76/application/pdf/California_FRID_GIS_metadata_11-10-20117.pdf

process.

1.1 Sentinel 2

Table availability in header as footnote[5]

Compared to Landsat multispectral bands, the Sentinel-2 multispectral bands provide two red edge bands and one NIR band with improved spatial resolution. The Sentinel 2 bands can be used to derive various vegetation indices useful for discriminating between species and estimating forest biomass Table 1

SNAP is an open source common architecture for ESA Toolboxes ideal for the exploitation of Earth Observation data [6]. SNAP provides toolboxes to derive information (e.g. vegetation indices) from Sentinel data.

[5] Available here: https://www.arcgis.com/home/item.html?id=fd61b9e0c69c4e14bebd50a9a968348c
[6] Available here: https://step.esa.int/main/toolboxes/snap/

1.1.1 Vegetation Indices

Vegetation indices including near-infrared wavelength have weaker relationships with biomass than those including shortwave infrared wavelength, especially for forest sites with complex stand structures. The results of image transformations such as the first principal component from the PCA showed stronger relationships with biomass than individual spectral bands, somehow independent of different biophysical conditions. However, in a study area with poor soil conditions and relatively simple forest stand structure, near-infrared band or relevant vegetation indices had a strong relationship with biomass (**lu_survey_2016**).

Table. 2: Vegetation Indices derived from Sentinel 2 information (adapted from **pandit_estimating_2018**, **wang_estimating_2020-1**).

Vegetation.Indices	Equations	References
NDVI	(NIR - Red)/(NIR + Red)	(Tucker 1979)
RGR (Red Green Ratio)	Red / Green	(Sims \& Gamon 2002)
EVI (Enhanced Vegetation Index)	2.5*((NIR - Red)/ (NIR + 1 + 6 * Red + 7.5 * Blue))	(A. Huete et al. 2002)
SR (Simple ratio)	NIR/RED	(Jordan 1969, Wicaksono et al. 2016)
SAVI (Soil-Adjusted Vegetation Index)	(NIR - Red)/(NIR + Red + L)* (1 + L)	(Huete 1988)
DVI	NIR - Red	(Zhu et al. 2017)
FDI (Forest Discrimination Index)	NIR - (Green + Red)	(Kamal et al. 2015)
TNDVI (Transformed Normal Difference vegetation index)	sqrt[(NIR - Red)/(NIR + Red) + 0.5]	(Castillo et al. 2017)
Red edge based		
PSRI (Plant Senescence Reflectance Index)	(Red - Green)/RE2	(Hill 2013; Merzlyak et al. 1999)
RE NDVI	(NIR- RE2)/(NIR + RE2)	(Chen et al. 2007)
NDVI1	(NIR - RE1)/(NIR + RE1)	(Kross et al. 2015)
NDVI2	(NIR - RE2)/(NIR + RE2)	(Gitelson \& Merzlyak 1994; Kross et al. 2015)
NDVI3	(NIR - RE3)/(NIR + RE3)	(Sharma et al. 2015)
NDVI4	(NIR - RE4)/(NIR + RE4)	(Kross et al. 2015)
Clg-RE1*	RE1 / Green - 1	(Gitelson et al. 2003)
Clg-RE2*	RE2 / Green - 1	(Gitelson et al. 2003)
Clg-RE3*	RE3 / Green - 1	(Gitelson et al. 2003)
IRECI (Inverted Red-Edge Chlorophyll Index)**	(RE3 - Red)/(RE1/RE2)	(Castillo et al. 2017)
MTCI (terrestrial chlorophyll index)	RE2 - RE1)/(RE1 - Red)	(Dash \& Curran 2004)
MCARI (modified chlorophyll absorption in reflectance index)	[(RE1 - Red) - 0.2 * (RE1 - Green)]×(RE1/Red)	(Daughtry et al. 2000)
MSRren (Modified Simple Ratio red-edge narrow)	[(RE4 / RE1) - 1] / sqrt[(RE4 / RE1) + 1]	(Fernandez-Manso et al. 2016)
PSSRa (pigment specific simple ratio)	RE3 / Red	(Blackburn, 1998)
S2REP (Sentinel-2-red-edge position)	RE1 + 35*(0.5*(RE3 + Red)/2) - RE1)/(RE2 -RE1)]	(Frampton et al. 2013)
RECI (Inverted Red-Edge Chlorophyll Index)	NIR - Red/(RE1/RE2)	(Frampton et al. 2013)
Shortwave infrared based***		
NDII (Normalized Difference Infrared Index)	(NIR - SWIR1)/(NIR + SWIR1)	(Hardisky et al. 1983)
MNDWI (Normalised difference water index)****	(Green-SWIR1)/(Green + SWIR1)	(Ji et al. 2009)

References to include from the table: (**tucker_red_1979**, **sims_relationships_2002**, **huete_overview_2002**, **jordan_derivation_1969**, **wicaksono_mangrove_2016**, **huete_soil-adjusted_1988**, **zhu_exploring_2017**, **kamal_object-based_2015**, **castillo_estimation_2017**, **hill_vegetation_2013**, **merzlyak_non-destructive_1999**, **kross_assessment_2015**, **gitelson_spectral_1994**, **sharma_active-optical_2015**, **gitelson_relationships_2003**, **dash_meris_2004**, **daughtry_estimating_2000**, **fernandez-manso_sentinel-2a_2016**, **blackburn_spectral_1998**, **frampton_evaluating_2013**, **hardisky_influence_1983**, **ji_analysis_2009**)

1* Subtraction of NIR reciprocal reflance from from reciprocal reflectance (520 to 550 nm and 695 to 705) relates closely to total chlorophyll content (**gitelson_relationships_2003**)

2* IRECI had highest correlation to biomass (r = 0.8) in **castillo_estimation_2017** (Mangrove Forest Philippines) and further improved with added elevation

3* SWIR bands are used as important indicators in forest regeneration studies (**boonprong_random_2018**). Important for vegetation mapping as a standalone (**immitzer_first_2016**). Large changes for seasonal reflectance in European forest species (**grabska_forest_2019**)

4* Delineating water bodies if required (**ji_analysis_2009**)

Papers:

Estimating AGB in sub-tropical Nepal using Sentinel 2 (**pandit_estimating_2018**). Field-based AGB as a dependent variable, as well as spectral band values and spectral-derived vegetation indices as independent variables in the Random Forest (**breiman_random_2001**). In this algorithm, decision trees are generated to the maximum extent without pruning using a randomly-selected two thirds of the samples as training data with bootstrapping (re-sampling the data many times with replacement), which strengthens the flexibility by aggregating the prediction across individual trees to make a final prediction. The paper ranks importance of spectral band data and vegetation indices from above **??** as a typical output of a random forest to estimate AGB (R2 = 0.81 and RMSE = 25.57 t ha-1; **pandit_estimating_2018**.

Estimating AGB in in tropical forest using Sentinel-1 (SAR) and vegetation indices from Sentinel-2 (**ghosh_aboveground_2018**). Sentinel-1 data does not work as a good predictor variable for the study area, as C-band SAR backscatter saturates at a biomass density of 100 Mg/ha. R package, CARET was used to implement RF regression to rank important predictor variables. Stochastic gradient boosting modelling (SGB, same R package) combines both regression tree and boosted algorithms to predict the response variable. SGB models were fitted, with varying values for the number of regression trees (50–5000), tree complexity of 1, 3 and 5. The value for learning rate was fixed at 0.01. Combination of satellite sensors shows the best result for S. robusta forest, with a coefficient of determination value of 0.71 and an RMSE value of 105.027 t/ha (**ghosh_aboveground_2018**).

1.1.2 Normalised Burn Ratio

The Normalized Burn Ratio (NBR) is an index designed to highlight burnt areas in large fire zones and can be derived from landsat or sentinel imagery. The formula is similar to NDVI Equation 1, except that the formula combines the use of both near infrared (NIR) and shortwave infrared (SWIR) wavelengths Figure 1. It can also be used to monitor forest disturbanches such as logging activities (**shimizu_using_2016**).The formula is similar to NDVI, NBR uses the ratio between NIR and SWIR bands. A high NBR value indicates healthy vegetation while a low value indicates bare ground and recently burnt areas. Non-burnt areas are normally attributed to values close to zero.

$$NBR = (NIR - SWIR)/(NIR + SWIR) \tag{1}$$

For any burn related damage/regeneration calculations we can verify fire activity locations via the Fire Information for Resource Management System (FIRMS[7]). The data come from the Collection 6 Near Real Time (NRT), extracted from the standard MCD14ML fire product produced at the MODIS Fire Science Computing Facility (**roteta_development_2019**).

Sentinel 2 data to create fire database in sub-Saharan Africa (Roteta et al., 2019). Sentinel-2 MSI reflectance measurements in the short and near infrared wavebands plus the active fires detected by Terra and Aqua MODIS sensor. They were able to detect smaller fires than with common MODIS approach, but Sentinel-2 based products have lower temporal resolution and consequently are more affected by cloud/cloud shadows. visually created using BAMS (Burned Area MappingSoftware) methodology (**roteta_development_2019**), which consists in a trained classification to detect burned areas between images from two dates[8].

Fire on Madeira, Spain (**navarro_evaluation_2017**). Sentinel-2 data (5 days, 10 m resolution) for pre- and post-fire image assessments (sometimes just two). The framework can be used for the assessment of many other burnt areas globally. Enabling an extremely unprecedented perspective with a unique set of accurate, robust, timely and easily accessible information. No real measurement of accuracy provided.

Classification of burn severity for wildfire in Spain using Sentinel-2 (**fernandez-manso_sentinel-2a_2016**) (with NBR). Superiority of red-edge spectral indices (particularly, Modified Simple Ratio Red-edge, Chlorophyll Index Red-edge, Normalized Difference Vegetation Index Red-edge) over conventional spectral indices. Fisher's Least Significant Difference test confirmed that Sentinel-2A MSI red-edge spectral indices are adequate to discriminate four burn severity levels.

[7] Available here: https://firms.modaps.eosdis.nasa.gov/download/
[8] Available here: https://climate.esa.int/en/projects/fire/data/

Figure 1: Comparison of the spectral response of healthy vegetation and burned areas (USFS)

1.1.3 Problems with Sentinel 2 Data

General problem with spatial resolution (30 m). The round FIA plots overlap with multiple pixels and sometimes the forest edge. Relatively new, so does not have the high temporal value as landsat. Will probably fail to distinguish tree phenology (e.g. growth stages) and can reach a spectral saturation point if biomass is too high.

Papers: Bamboo forest in China: Seasonality, different growth phenomena in different years. Correlation between spectral bands and biomass varies within a period of months. Biomass calculation based on DBH and age (**chen_exploring_2019**). Also used random forest to evaluate key variables.

Sentinel-2 cleaning

The Sen2Cor atmospheric correlation processor can be used to conduct atmospheric correction. The Sen2Cor processor was developed for formatting and generating Sentinel-2 Level-2A products. The processor is freely available from the European Space Agency website[9].

Sentinel-2 estimating AGB

Tested ability of Sentinel imagery (Sentinel-1 (SAR) + Sentinel-2) for the retrieval and predictive mapping of above-ground biomass of mangroves and their replacement land uses (Mangrove Forest, Philippines) **castillo_estimation_2017**. Developed biomass prediction models through the conventional linear regression and novel Machine Learning algorithms (SAR raw polarisation backscatter, multispectral bands, vegetation indices, canopy biophysical variables). Model based on biophysical variable Leaf Area Index (LAI) derived from Sentinel-2 was more accurate in predicting the overall above-ground biomass (RMSE 27.8–28.5 Mg ha-1). Among the Sentinel-2 multispectral bands, the red and red edge bands (bands 4, 5 and 7), combined with elevation data, were the best variable set combination for biomass prediction. The red edge-based Inverted RedEdge Chlorophyll Index had the highest prediction accuracy among the vegetation indices.

Estimation of aboveground biomass for mangrove forests in the Philippines **castillo_estimation_2017**. All modelling tasks were implemented using IBM SPSS Statistics version 23 (IBM, USA) and the Waikato Environment for Knowledge Analysis (WEKA, version 3.8.0, The University of Waikato, NZ). The WEKA software is a collection of machine learning algorithms (**hall_weka_2009**). To assess if the correlation with biomass and prediction error from the linear models can still be improved, the set of predictors from the linear models with the highest r and lowest RMSE for each part was further subjected to 17 machine learning algorithms available in the WEKA machine learning software. The model/algorithm with highest r and lowest RMSE was selected for use in predictive mapping of biomass which was implemented in ArcGIS (version 10.3.1, ESRI, USA). Four biomass predictive maps were produced which were derived from Sentinel-1 SAR channels, Sentinel-2 bands, Sentinel-2 vegetation index, and Sentinel-2 vegetation biophysical variable **castillo_estimation_2017**.

[9] Available here: http://step.esa.int/main/third-party-plugins-2/sen2cor/

1.2 LiDAR

LiDAR data is the new standard when it comes to estimating AGB. Common procedure to calculate AGB via vegetation/canopy height and measure accuracy by comparing it to AGB derived from forest plots. Everything > 2m is typically classified as canopy (Jonathan Kane, 29 April meeting). Identification of individual trees only possible if resolution matches the size of the individual canopy (< 1m, Van Kane, 29 April meeting). LiDAR is useful to estimate vegetation metrics (basal area and stem density with r2 values of 0.86–0.95 across multiple studies, **zald_influence_2014**)

Recent paper with Jonathan and Van Kane:

Used a Carbon Monitoring System (CMS) to produce annual estimates of above-ground biomass using Random Forests (RF) for a regional and landscape approach. Field plots (mostly FIA) with self calculated AGB as a response variable to predict AGB from LiDAR derived canopy height and density information (R2 = 0.8, RMSE = 115 Mg ha-1, Bias = 2 Mg ha-1). A stratified random sample of AGB pixels from landscape-level AGB maps then served as training data for predicting AGB regionally from Landsat image time series variables processed through LandTrendr. Climate metrics calculated from downscaled 30 year climate normals were used as predictors for both models (landscape and regional), as were topographic metrics calculated from elevation data; these environmental predictors allowed AGB estimation over the full range of observations with the regional model (R2 = 0.8, RMSE = 152 Mg ha-1, Bias = 9 Mg ha-1), including higher AGB values (>400 Mg ha-1) where spectral predictors alone saturate (**hudak_carbon_2020**).

Combination of LiDAR and other remotely sensed information for biomass estimation

A) Combination of LiDAR and QuickBird image did not improve AGB estimation in mixed coniferous forests in California; LiDAR data alone provided a better performance (**hyde_mapping_2006**).

B) LiDAR and hyperspectral combination has lower accuracy than LiDAR alone in tropical forest Costa Rica (**clark_estimation_2011**).

C) LiDAR and Synthetic Aperture Radar (SAR). The paper provides an overview table for all combinations and their output accuracies (Figure 2, **kaasalainen_combining_2015**). Upscaling refers to the extrapolation of LiDAR to areas where LiDAR is not available. Generally increases accuracy.

D) LiDAR and Landsat

matasci_large-area_2018 provide an overview of combination of LiDAR with spaceborne imagery (mainly MODIS & LandSat, Figure 3)

1.2.1 LiDAR remove noise

A) Lastool in QGIS (LASnoise)

Figure 2: Overview of combinations of LiDAR and radar in research (Kaasalainen et al., 2015)

B) LiDAR360 3.1 software (GreenValley, Being, China) to remove the noise points floating between the flight altitude and the mangroves. (**wang_estimating_2020**).

C) Triangulated irregular network (TIN) densification as a filtering algorithm to deal with complex forest landscapes (**zhao_improved_2016**). Improved progressive TIN densification filtering algorithm (IPTD) performs better than other general filtering algorithms (Figure 4). The strength of 537 the IPTD lies in its ability to retain hilltops and handle break lines and steep slopes. Four parameters are used in the IPTD, i.e., k (neighboring number), r (threshold), 0 (iterative angle), and s (iterative distance). Densifies ground points and accounts for ground surface structure with (upward/downward densification) (**zhao_improved_2016**).

C) The Fusion[10] software (Windows) used by the USFS introduces filters and smoothing algorithms to create digital surface model. The workflow necessary to create a clean digital terrain model still have to be assessed. While an outlier filter could be used to remove below-ground points the scalability of the approach over larger areas and difficicult topography is questionable. A common procedure is to rely on vendor provided DTM's and create a DSM and a CHM in Fusion.

D) pdal[11] offers a wide and specialised library to work with pooint cloud information. Filtering, classification and creation of terrain models can be implemented into a comphrensive script and assures scalability of the approach. Accuracy of the output will be tested against canopy height models provided by the vendors and created in Fusion (currently 'best practice').

1.2.2 Identifying individual trees vs canopy area approach

Tree-centric approaches are not accurate enough yet in order to justify a much more hardware intensive calculation. Tree-centric modelling is appealing because it is based on summing the biomass of individual trees, but until algorithms can detect understory trees reliably and estimate biomass from crown dimensions precisely, areas-based modelling will remain the method of choice (**coomes_area-based_2017**).

Canopy height estimates from LiDAR (methods)

A) Quantifying australian wheat (**walter_estimating_2019**). 90 points/m2 by using sensor on farming vehicle coupled with multispectral to identify vegetation. Canopy height was extracted through percentile algorithm in R. Two step approach: (1) Identifying the 98th percentile of maximum returned height in each scan line and (2) taking the 86th percentile of these values to provide an estimate of overall canopy height. 86th percentile was selected through optimization of Pearson's correlation coefficient and RMSE between LiDAR derived canopy height and measured canopy height for all sample times. More details in supplementary (**walter_estimating_2019**).

[10] Available here: http://forsys.cfr.washington.edu/fusion/fusionlatest.html
[11] pdal documentation: https://pdal.io/

Figure 3: Overview of combinations of LiDAR (air - or spaceborne), and optical imagery to map forest structural attributes (Matasci et al. 2018)

B) Honkong forest with thousands of individual tree measurements (dbh, height; **chan_estimating_2021**).

LiDAR processed to TIN by extracting only the ground returns. Te extraction and processing were performed in ArcMap with LAStools[12] extension and FUSION[13]. The canopy surface was defined using all non-ground returns with height above 2.0 m. The reason of choosing 2.0 m as the threshold was to avoid canopy returns to be mixed with ground returns. Entire 1 ha study area was divided into circular plots with three plot sizes (10, 5 and 2.5 m radius; **chan_estimating_2021**).

Canopy Surface - Ground TIN = Normalized height of canopy points -> derive LiDAR metrics

Circular plots were considered more favorable than rectangular or square plots, since the periphery-to-area ratio was the smallest and thus minimized the number of edge trees (**kohl_sampling_2006**).

Relevant plot metrics (LiDAR, climate) were derived by the 'cloudmetrics' function in FUSION version 3.7. The LiDAR metrics, and its log-transformed metrics were input into stepwise regression model as independent predictors of AGB. Significant predictors were selected ($F < 0.5$ were entered and $F > 1.0$ were removed) into the regression model. Tree sets of regression models were generated and the allometric models tended to be linear and normal after logarithmic transformation (old ecology ref available if necessary). Logarithmic transformation increased accuary. Model predicted three parameters : (1) canopy cover by first returns, (2) 95th Height Percentile and (3) median of absolute deviation from mode (MADmode). Accuracy of predication increased with plot size (**chan_estimating_2021**).

[12] Available here: https://rapidlasso.com/lastools/
[13] Available here: http://forsys.cfr.washington.edu/fusion/fusionlatest.html

Figure 4: Comparison of kappa values (total error, TE) for LiDAR filtering algorithms for 15 sites tested. Lowest value in bold (Zhao et al., 2016)

1.2.3 Common LiDAR point metrics

There is a wealth of LiDAR plot metrics derived from the point clouds **??** and they can even be used do discriminate forest tree species (with and without leaves; **shi_important_2018**). They are created using height profiles and point distancing (for detailed table with references see **dong_selection_2017**) and can be grouped in four categories:

1) height (canopy height distributions within the plot) estimator for carbon and AGB (**lim_estimation_2004**, **patenaude_quantifying_2004**)

2) intensity (similar to the height metrics except that they are statistics of the intensity value rather than the height value of the point clouds, only finds application in high resolution LiDAR datasets)

3) density (canopy return density)

4) canopy volume (portray canopy morphology for canopy cover index or leaf area density)

1.2.4 Upscaling LiDAR with Sentinel-2

LiDAR upscaling by integrating field plots (point), UAV-LiDAR strip (line) data and Sentinel-2 imagery (polygon) based on a point-line-polygon framework for AGB estimation in a mangrove forest (**wang_estimating_2020**). Field plots were linked to UAV-LiDAR plots with a random forest model and then extrapolated to all UAV-LiDAR grid cells. UAV-LiDAR serves as a linear bridge between forest plots and sentinel data (polygon). For method comparison, a traditional model was also constructed using only field plots and Sentinel-2 imagery. Since the validation plots were gradually added to the G-S2 (without LiDAR) model, there were not enough observations to validate the model performance with an independent subset. Therefore, a 10-fold cross-validation method was employed and repeated 10 times as suggested by Ghosh and Behera (2018). The same 10-fold cross-validation process that was iterated 10-times was also applied to the G-LiDAR-S2 model to obtain cross-validation accuracy. A variable selection process was also conducted via the backward feature elimination method prior to the final estimation due to the large number of Sentinel-2 (32) and UAV-LiDAR variables (53) input. As usual, a model constructed from a small number of variables is more interpretable, and eliminating irrelevant and highly correlated variables may improve the predictive power (**gregorutti_correlation_2017**). The backward feature elimination approach is based on the random forest algorithm and compares the cross-validated prediction results of model as a proportion of the predictors is eliminated. This method was implemented by the rfcv function in the random Forest package and replicated 100 times with 5-fold cross-validation to get the optimal variables (**pham_monitoring_2017**).

AGB estimates were assessed using in-dependent validation samples by comparing the predicted to observed values using the coefficient of determination R2, root mean square error (RMSE) and RMSE expressed as a percentage of the observed mean (RMSE

RF has been widely used in biomass upscaling approaches (**matasci_large-area_2018**, **pham_monitoring_2017**, **huang_integration_2019**, **ghosh_aboveground_2018**).

1.2.5 Upscaling LiDAR with GeoEye-1

GeoEye-1 image and LiDAR data were segmented using region growing approach to delineate individual tree crowns; and the segmented crowns of tree were further used to establish a relationship with field measured carbon and total tree height (**wangda_species_2019**)

Upscaling of biomass and carbon estimates by tree-centric approaches are less accurate (**coomes_area-based_2017**), but studies on the upscaling carbon estimates by species stratification using VHR and LiDAR in smaller areas to a landscape level are limited (**latifi_stratified_2015**)

Field plots (DBH (> 10 cm), tree height, species and crown diameter), calculated biomass with mangrove-specific equation and converted to carbon stock using IPCC conversion factor. GeoEye-1 data to delineate trees and canopy height model to estimate individual height. eCognition for object based image analysis and resulting species separation with tree height. Training dataset comprising 70

1.3 Field data

Application of LiDAR for forest inventory requires field plot data to ground-truth the information. Field plots need to incorporate relevant parameters (volume, basal area, biomass) (**hudak_carbon_2020**). FIA plot data provide an unbiased, systematic sample of forest conditions in space and time and are more applicable over larger areas and for successive effects such as climate change and forest regeneration rates (**tinkham_applications_2018**)

Shana sampling plots (LiDAR verification plots):

Sampling plot data[14] from the Forest Inventory and Analysis National Program (FIA).

- Allometric: DBH, size classes (e.g. saplings) > height groups, largest tree measured, height to live crown, crown ratio, crown width, age data sparse and based on 1-2 cores - Biomass derived from LiDAR datasets - plot center GPS using Javad - plot date and size - Ancillary data: Slope, ground cover, vegetation cover by type (e.g. shrub, forb, etc), modal vegetation height by different types of vegetation, fuel models, fuels data, seedlings, site history (e.g. plantation, if there was a fire, etc)

Forest plots are triangle with 4 circular subplots (Figure 5), just assume they are representative of 1 ha of forest (Van Kane, 29 April meeting).Spatial mismatch between the 7.3 m radius, round configuration of an FIA subplot and 30 m × 30 m square Landsat pixels (**tinkham_applications_2018**). Inevitably, the four subplots will intersect a different number of pixels and in varying proportions.

Criteria to match LiDAR with FIA plots (**hudak_carbon_2020**):

- fixed-area plots

- geo-referenced with a GNSS capable of differential correction

- established within ±3 years of an overlapping LiDAR collection

- not disturbed in the time between field and LiDAR data collections

1.3.1 Biomass estimation in the field

Collection of a large number of biomass reference data at the plot level is time-consuming and labor-intensive. It is only suitable for a small area and cannot provide the spatial distribution. However, this kind of data is a prerequisite for developing biomass estimation models (**lu_survey_2016**). Allometric models most common, but require data about soil, land use history and climate influence (**clark_tropical_2012**).

Two methods:

1) Use data of national forest inventories. Calculate with volume expansion factor (VEF),

[14]Sampling techniques used: https://www.fia.fs.fed.us/program-features/index.php

Figure 5: FIA sampling plot layout. a) generated canopy height model from LiDAR and b) landsat 30m imagery overlay (taken from Hudak et al., 2020)

average wood density (WD), biomass expansion factor (BEF) (**brown_biomass_1989**, **lehtonen_biomass_2004**, **wang_uncertainties_2011**):

$$AGB(kg/ha) = volume(m3/ha) * VEF * WD * BEF + e \quad (2)$$

Calculate AGB from field plots using equations from Fire and Fuels Extension (FFE) of the Forest Vegetation Simulator (FVS)[15]. These equations are based on a series of regional volume and biomass equations. The above-ground portion of the live and standing dead trees were summed to a single, plot-level AGB value; the belowground portion of the trees and non-tree species were excluded from the AGB estimates. These plot-level AGB estimates were the response variable for a landscape model (**hudak_carbon_2020**).

Problems with field data

Taken from **lu_survey_2016**:

(1) tree variables, including sampling, measurement, recording and grouping errors when tree variables such as DBH and height are measured

(2) conversion coefficients and models including variation of conversion factors from volume to biomass and then to carbon, inappropriate selection and usage of allometric models for relationship of tree volume and DBH and height, and incorrect regression models relating forest biomass/carbon to spectral variables

(3) uncertainties of spectral values due to unbalanced platforms, scanner motions, poor atmospheric conditions, and slope; inappropriate spatial interpolation methods for geometrical and radiometric corrections, and incorrect methods for image enhancement and analysis

(4) sample plot locations, including global positing system (GPS) coordinates used to locate the sample plots, geometric correction and the uncertainties due to mismatch of sample plots with spatial resolutions of remotely sensed data

(5) differences in sizes of sample plots and image pixels, disagreement between remotely sensed data and plot observations when portions of trees on boundaries are outside plots although both sample plots and pixels have the same spatial resolutions

(6) temporal differences between field plot measurements and remotely sensed data

Depending on the protocol and the quality of the Global Navigation Satellite System (GNSS) receiver used in the field; the three peripheral subplots, while systematically laid out from the center subplot by consistent distances (36.6 m) and bearings (120◦, 240◦, 360◦) are usually not georeferenced, making them more subject to locational inaccuracy due to additive errors in accounting for horizontal distance on slopes and for magnetic declination on compass azimuths (**zald_influence_2014**).

1.3.2 Forest classification

Van Kane (meeting Apr 29): Forest service cannot reveal plot locations, only if project output is open access (ideally uni projects). Suggestion: Use Gradient nearest neighbours

[15] FVS documentation: https://www.fs.fed.us/fmsc/ftp/fvs/docs/gtr/EssentialFVS.pdf

(GNN) so nearest plot imputes tree list

NLCDB

The National Land Cover Database (NLCD) is updated every five years and stands as the definitive land cover database for the United States[16].. The latest iteration is NCLD 2016. It contains 28 different land cover products characterizing land cover and land cover change across 7 epochs from 2001-2016, urban imperviousness and urban imperviousness change across 4 epochs from 2001-2016, tree canopy and tree canopy change across 2 epochs from 2011-2016 and western U.S. shrub and grassland areas for 2016. NLCD 2019 is scheduled to be released in mid-2021.

CALVEG

Vegetation mapping has been completed for all of the National Forest lands of Pacific Southwest Region. Updates and map improvements are scheduled as part of the Monitoring program (aim is to update layers in 15 year intervals). Through cooperative mapping projects, all ownerships within North Interior, Northern Sierras, North Coast and Montane, & South Coast and Montane CALVEG Zones have been completed. Classification processes include detailed field studies and literature reviews[17]. A map status of the CALVEG Ecological zones has been provided by the USFS (**??**). The dataset can be downloaded as polygon shapefiles for the different zones.

FRID

The fire return interval departure (FRID) analysis quantifies the difference between current and presettlement fire frequencies and downscales the CALVEG classification into 32 dominant vegetation groups. It receives an annual update (just released the 2020 version two weeks ago). This polygon layer consists of information compiled about fire return intervals for major vegetation types on the 18 National Forests in California and adjacent land jurisdictions. Comparisons are made between pre-Euroamerican settlement and contemporary fire return intervals (FRI). Current departures from the pre-Euroamerican settlement FRIs are calculated based on mean, median, minimum, and maximum FRI valuesFRID documentation: https://www.fs.fed.us/r5/rsl/projects/gis/data/FRID/FRID_Metadata.html.

[16]NLCDB legend: https://www.mrlc.gov/data/legends/national-land-cover-database-2016-nlcd2016-legend

[17]CALVEG documentation: https://www.fs.fed.us/emc/rig/documents/protocols/vegClassMapInv/EVTG_v2-0_June2015.pdf

Figure 6: Current status of the CALVEG vegetation classification program for California（USFS）

1.4 Landsat

Landsat data (mainly thematic mapper (TM)

The deployment of Landsat imagery to refine the machine learning process has been discussed, but any form of implementation still requires testing. While the Landsat data comes at a coarser resolution than the Sentinel-2 imagery it offers a much higher temporal resolution reaching back to 1984. Landsat imagery became publicly available in 2008 and resulted in great interest by researchers. Pixel based image interpolations have been developed recently and are best pratice in generating annual a gap-free image tiles for the period of interest (e.g. growing season). We aim to replicate this procedure using the best available pixel approach (BAP, **white_pixel-based_2014**) in combination with breakpoint detection procedure (Composite2Change, **hermosilla_integrated_2015**) to fill gaps within the Landsat imagery. Normalised Burn Ratio (NBR, **key_landscape_2005**) derived from shortwave infrared bands (SWIR) are currently serving as the best reference indices to assess forest disturbance (**kennedy_detecting_2010**). Based on the gap-free tiles we will be able to estimate forest recovery after logging and wildfire events using the normalised burn ratio (NBR, **white_nationwide_2017**). The output will aid us in predicting biomass regeneration of burned areas (30 m). Further, it will allow us to label forest types (from CALVEG) with their expected timespan to return to a pre-fire state. This will be an important indicator to estimate the severity of the fire damage and aid us in estimating above-ground biomass where post-fire LiDAR information is unavailable.

Available Landsat Data:

- Landsat 8 Operational Land Imager (OLI): April 2013 to present

- Landsat 7 Enhanced Thematic Mapper Plus (ETM+): July 1999 to present

- Landsat 5 Thematic Mapper (TM): March 1984 to May 2012

- Landsat 4 Thematic Mapper (TM): July 1982 to December 1993

- API for the Landsat data available[18]

- USGS Landsat archive: L1T products are systematically corrected for radiometric, geometric, and terrain distortions

- Preference for Landsat-5 over Landsat-7 for best available pixel analysis (details below, **white_pixel-based_2014**

eDART, an Ecosystem Disturbance and Recovery Tracker system for monitoring landscape disturbances has been deleveoped for Landsat imagery (**koltunov_edart_2020**), but they do not provide a tool they publicly share (Shana, 28 May).

LandTrendr is set of spectral-temporal segmentation algorithms that are useful for change detection in a time series of moderate resolution satellite imagery (primarily Landsat) and for generating trajectory-based spectral time series data largely absent of inter-annual signal noise. Can be used to assess forest regeneration over larger timescales (Fig. 3). They provide a github repository evaluating how to extract the data from e.g. Google Earth(FVS)[19].

[18] Access here: https://espa.cr.usgs.gov/static/docs/api-readme.html
[19] Access here: https://emapr.github.io/LT-GEE/landtrendr.html

1.4.1 Normalised Burn Ratio (NBR) for LandSat timeseries analysis

NBR recommended spectral indices to perform land cover change analysis (**key_landscape_2005**). NBR2 as an alternative to NBR. NBR2 modifies the Normalized Burn Ratio (NBR) to highlight water sensitivity in vegetation and may be useful in post-fire recovery studies. NBR2 is calculated as a ratio between the SWIR values, substituting the SWIR1 band for the NIR band used in NBR Equation 3. Rasters with calculated NBR2 (LC8NBR2, LE7NBR2, LT5NBR2, or LT4NBR2) may be readily available, but otherwise the procedure can be done starting from Sentinel 4-5 data (1984).

$$NBR2 = (SWIR1 - SWIR2)/(SWIR1 + SWIR2) \tag{3}$$

1.4.2 Pre-processing of Landsat data

A) Cloud & Shadow:

Fmask is an object-based algorithm designed to identify clouds and cloud shadows, as well as clear land and water pixels, snow, and areas of no data (Zhu and Woodcock 2012). Fmask is run on 30 m aggregated TOA reflectance[20].

B) Surface reflectance correction:

The Landsat Ecosystem Disturbance Adaptive System (LEDAPS) is used to generate surface reflectance values. LEDAPS produces top-of-atmosphere (TOA) reflectance from Landsat TM and ETM+ digital numbers (DN) and applies atmospheric corrections to generate a surface reflectance product. LEDAPS applies the dark dense vegetation method of Kaufman et al. (1997) to estimate aerosol optical thickness (AOT) directly from the imagery and the AOT is one of the inputs used in the 6S radiative transfer model. The LEDAPS surface reflectance output includes an AOT map derived from the Landsat TM or ETM+ blue band commonly referred to as opacity. New standard LaSCR instead of LEDAPS? LaSRC is based on the 6S radiative transfer model and a heritage from the MODIS MCD09 products as well as the earlier LEDAPS algorithm implemented for Landsat-5 and Landsat-7.

1.4.3 Timeseries analysis for Landsat imagery

The process is a multi-method approach, introducing recent developments (pixel based compositing) to achieve the most up to date analysis of landsat information mainly developed around a team of Canada based scientists (Hermosilla, White, Wulder).

A) Best available pixel analysis (BAP; **white_pixel-based_2014**

Four scores were calculated for each pixel. Sensor and Day-of-the-year score (DOY) were calculated at the image level (i.e., all pixels within the image receive the same score), whilst the cloud/cloud shadow and opacity scores were unique to each pixel. Results for

[20]Detailed method: https://www.l3harrisgeospatial.com/docs/calculatecloudmaskusingfmask.html

Figure 7: Pixel change over updated intervals for the Normalised Burn Ratio (NBR)

BAP analysis in Canada show 83

B) Filling gaps in BAP with breakpoint analysis (Composite2Change; **hermosilla_integrated_2015**)

Characterisation of trends in pixel series. A breakpoint represents a change in the temporal development of a pixel's values through time. Robust identification of breakpoints enablesthe assignment offinal proxy values to pixels with missing observations, as well as enabling change events, and the year in which theevent occurred, to be correctly identified. The breakpoint detection pro-cess is performed over the Normalized Burn Ratio (NBR; **key_landscape_2005**) pixel series, which has been demonstrated as sensitive and con-sistent for the retrieval of disturbance events over forest environments (**kennedy_detecting_2010**). Once breakpoints (temporal) have been identified us contextual analysis for spatial correlation. The pre-infill process of data gaps may result in spatially discordant disturbance events labeled with an incorrect year for change. If we consider discrete events, such as a wildfire, the information need is to ensure that these change events are allocated to the correct year (**hermosilla_integrated_2015**).

Analysis based on BAP and Composte2Change (**hermosilla_regional_2015**)

Set of spectral metrics are computed for selected bandsand indices: average of the spectral values of the objects before thechange event, average and standard deviation after the change event,and range, average and standard deviation of the values of the pixelseries for the spectral bands 3, 4, 5, and 7, and for the indices NBR, and Brightness, Greenness and Wetness components from the TasseledCap. Most important predictors within the trend analysis metrics (e.g. average change magnitude, standard deviation post change).

Method for entire Canada (**hermosilla_mass_2016**)

The annual image composites were created by considering as candidate images all available Landsat Thematic Mapper (TM) and Enhanced Thematic Mapper Plus (ETM+) images (more than 81,000) from the 1285 scenes (path/rows) of the Landsat Worldwide Referencing System (WRS2) in the USGS archive that covered at least part of the terrestrial area of Canada, had less than 70

Nationwide forest regeneration study after logging and wildfire (Canada; **white_nationwide_2017**)

Identified national trends in forest disturbance and recovery by disturbance type (1985–2010). 57.5 Mha of Canada's forests were disturbed by wildfire (71

Papers:

Landsat time series predictors and information of disturbance eventsfrom 1984 to 2016 were obtained from gap-free Landsat composites,created using the Composite2Change (C2C). Specifically, summer image composites (August1 ± 30 days) were created through a best available pixel (BAP) selection process which accounted for distance to clouds,

Figure 8: Schematic of nbr recovery metrics: deltaNBR - regrowth, Recovery Indicator (RI), and year of recovery (Y2R) (taken from White et al., 2017)

atmospheric con-tamination, as well as acquisition sensor. Spectral indices (bold performed best): Tasseled Cap Brightness (TCB), Greenness (TCG),Wetness (TCW), and the NBR (**bolton_optimizing_2020**).

Annual time series of gap-free, surface reflectance composites and the detection, delineation and characterization of annual forest changes. Detected forest changes were attributed to a change type (i.e. fire, harvest, road and non-stand-replacing disturbance) following a random forest classification model based on their spectral, geometrical and temporal characteristics. To avoid potential noisy information at boundaries, samples are 45 m away from polygon boundaries.This figure appears in colour in the online version of Forestry. In other words: samples only taken further within the polygon (**wulder_satellite-based_2020**).

Jonathan Kane paper (**kolden_mapped_2012**). Characterized unburned area within fire perimeters by fire size and severity, characterized distance to an unburned area across the burned portion of the fire, and investigated patch dynamics of unburned patches within the fire perimeter. From 1984 through 2009, the total area within the fire perimeters that was classified as unburned from dNBR was 37

The complex biophysical environments and vegetation characteristics, e.g. phenology, species composition, growth phase, and health – will affect vegetation spectral signatures; thus, biomass estimation models based on optical spectral features cannot be directly transferred to different study areas for biomass mapping (**foody_predictive_2003**, **lu_aboveground_2005**).

Combination of Landsat and land cover to estimate AGB in Uganda (**avitabile_capabilities_2012**). A regression tree-based model (Random Forest) produces good results (cross-validated R^2 0.81, RMSE 13 T/ha) when trained with a sufficient number of field plots representative of the vegetation variability at national scale. Specific limitations are mainly related to saturation of the optical signal at high biomass density and cloud cover, which hinders the compilation of a radiometrically consistent multi-temporal dataset. Land cover data increases the model performance because it provides information on vegetation phenology.

1.5 Other relevant datasets

1.5.1 Synthetic aperture radar (SAR)

Most radar-based biomass estimation studies use L-band SAR data, especially the ALOS PALSAR L-band data. The SAR C-band data have not been extensively used because of the C-band's inability to capture forest biomass features. In summary, it is difficult to use radar data for distinguishing vegetation types (**lu_aboveground_2012**) because radar data reflect the roughness of land cover surfaces instead of the difference between the vegetation types, thus resulting in difficulty of biomass estimation. The speckle in radar data is another problem affecting its applications. Properly employing filtering methods to reduce noise and outliers in InSAR data is needed to improve the vegetation height estimation performance (**lu_survey_2016**).

1.5.2 NASA-GEDI LiDAR

GEDI has the highest resolution and densest sampling of any LiDAR ever put in orbit (25 m to 1 km resolution). It has been used in an upscaling study of forest biomass in China (**chen_improved_2021**). However, satellite was launched in 2018 and only designed to operate for 2 years, will be wrapped up in 2021[21].

Papers:

Comparison of biomass estimates for 3 satellites (GEDI, ICESat-2 and NISAR) over Sonoma county in California (**duncanson_biomass_2020**)

GEDI and ICESat-2 were simulated from airborne lidar point clouds, while UAVSAR's L-band backscatter was used as a proxy for NISAR. To estimate biomass for the lidar missions we used GEDI's footprint-level biomass algorithms, and also adapted these for application to ICESat-2. For UAVSAR, they developed a locally trained biomass model, calibrated against the ALS reference map. Each mission simulation was evaluated in comparison to the local reference map at its native product resolution (25 m, 100 m transect, and 1 ha) yielding RMSEs of 57

ICESat-2 underestimates biomass on average, and this underestimation increases with canopy height, canopy cover, and to a small degree with slope. The comparison between simulated GEDI and ICESat-2 height metrics confirms that at least with respect to GEDI simulations, ICESat-2 underestimates height, and these underestimations are more pronounced in the higher height metrics. A comparison of ICESat-2 simulations from the higher signal photon return rate shows that the signal rate drives much of this error, so in areas where ICESat-2 has little atmospheric attenuation ICESat-2 will likely perform well for forest structure (**duncanson_biomass_2020**).

As expected, the highest accuracies were from GEDI's power beam, while accuracies for ICESat-2 depended heavily on the signal photon rate, and NISAR accuracies will depend on the availability of high-quality biomass training data, and be limited to lower biomass forests over relatively flat areas. Nonetheless, we find all three sets of mission simulations promising for biomass (**duncanson_biomass_2020**).

[21] Project website: https://gedi.umd.edu/

1.5.3 NOAA-AVHRR, MODIS, ASTER, Quickbird, IKONOS

The use of coarse resolution satellite images such as NOAA-AVHRR, MODIS, etc. for biomass estimation is limited due to the occurrence of mixed pixels and inconsistent accuracy at regional or local scale (**wangda_species_2019**).

The use of moderate resolution satellite images (e.g. Landsat, ASTER) for AGB estimation is also confronted with the problem of mixed pixels and data saturation in complex biophysical environments (**wangda_species_2019**).

Quickbird and IKONOS have very high resolution, but are frequently obstructed by cloud cover in tropics (**wangda_species_2019**), might be more viable in California.

1.5.4 RapidEye

ESA is offering access to the full RapidEye archive for scientific research and application development. Access is only available to submitted proposals that are accepted. 5 m pixel size (RGB, RE and NIR). Overall, RapidEye data are not suitable for AGB estimation, but when AGB falls within 50–150 Mg/ha, support vector regression based on stratification of vegetation types provided good results. Problem for biomass estimates can be fixed incorporating tree height if LiDAR data or stereo image is available (Brasilian Rainforest, **feng_examining_2017**.

1.5.5 ICESat-2

Ice, Cloud and Land Evalutation Satellite, 91-day repeat, began its 3-year mission in September 2018. While ATLAS onboard ICESat-2 was primarily designed to determine changes in ice sheet elevation and mass, it will provide information about vegetation that may be used to estimate AGB. ATLAS is a photon counting system, operating in the visible wavelengths, at 532 nm. It generates three pairs of tracks, with each pair approximately 3.3 km apart and each track within a pair separated by 90 m. LiDAR footprints are produced every 70 cm in the along-track direction and measure approximately 14 m in diameter. Given the unprecedented coverage and spatial detail from ICESat-2, translating ICESat-2 measurements to AGB estimates would allow for large-scale AGB and forest carbon assessments.

Paper:

ICESat-2 for mapping forest biomass with deep learning (**narine_synergy_2019**). A first set of models were developed using vegetation indices calculated from single-date Landsat imagery, canopy cover, and land cover, and a second set of models were generated using metrics from one year of Landsat imagery with canopy cover and land cover maps. With the extended dataset containing metrics calculated from Landsat images acquired on different dates, substantial improvements in model performance for all data scenarios were noted. The R2 values increased to 0.64, 0.66, and 0.67. Comparisons with Random forest (RF) prediction models highlighted similar results, with the same R2 and root mean square error (RMSE) range (15–16 Mg/ha) for daytime and nighttime scenarios. ICESat-2 profiles, especially with the nighttime scenario (R2 = 0.66), highlight the potential for generating a wall-to-wall AGB product with ICESat-2 by adopting a synergistic approach with Landsat optical imagery, canopy cover, and land

cover (**narine_synergy_2019**).

ICESat-2 in combination with land cover data for mapping AGB in Texas (Figure 9, **narine_using_2020**). Extrapolation of ICESat-2-derived AGB estimates from segments along the ICESat-2 transect, to 30 m pixels across the study site was carried out with RF, using Landsat and NLCD products. Results are indicative of the utility of ICESat-2 data for characterizing AGB. As more ICESat-2 tracks become available and with the availability of corresponding reference data, an improved understanding of the data and their ability to characterize vegetation structure will be gleaned. Nonetheless, possibilities with the data, including synergistic use with freely available data from other space-based missions, including GEDI, are exciting. Future work will involve the application of the methodology for regional-scale mapping with corresponding uncertainty analyses and an investigation of deep-learning approaches with multi-source data (**narine_using_2020**).

1.5.6 LISS-3 (ISRO)

Satellites provide multispectral images at 24 m (LISS-3) and 56 m (AWiFs) meter resolution (Table 2). IRS data products are free of charge to all data users including the general public, scientific and commercial users[22]. EarthExplorer and USGS Global Visualization Viewer (GloVis) can be used to search, preview, and download ISRO data. In EarthExplorer the IRS AWiFS and IRS LISS-3 datasets are located under the ISRO Resourcesat category.

1.5.7 Climate data

Paper:

For climate, data were used from the North American Regional Reanalysis (NARR) (**mesinger_north_2006**), the National Center for Environmental Prediction's high-resolution combined model and assimilated dataset available eight-times daily 1979-2000 at 32 km resolution. Ecosystem demography model trained with hourly land surface weather data to calculate carbon assimilation rate and transpiration (2 m air temperature, dew point, downward solar radiation, precipitation, soil temperature; **hurtt_beyond_2019**).

Three parameters to account for climatic variations: Temperature seasonality (TS), Long-term Maximum Climatological Water Defcit, and Precipitation Seasonality (Hongkong Forest; **chan_estimating_2021**).

[22]USGS EROS Archive: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-products-overview

Figure 9: Regression models for estimating AGB with ICESat-2 data from the strong beam and weak beam (Narine et al., 2020)

Table 1: Overview Sentinel 2 bands

Band	Description	Wavelength (um)	Resolution (m)
1	Coastal Aerosol	0.433 - 0.453	60
2	Blue	0.458 - 0.523	10
3	Green	0.543 - 0.578	10
4	Red	0.650 - 0.680	10
5	RE1	0.698 - 0.713	20
6	RE2	0.733 - 0.748	20
7	RE3	0.773 - 0.793	20
8	NIR	0.785 - 0.900	10
8a	Narrow NIR	0.855 - 0.875	20
9	Water vapour	0.935 - 0.955	60
10	SWIR-Cirrus	1.365 - 1.385	60
11	SWIR-1	1.565 - 1.655	20
12	SWIR-2	2.100 - 2.280	20

Table 2: Resourcesat Sensor Specifications

Sensor	AWiFS	LISS-3
Number of Bands	4	4
Spectral Band 2 (μ)	0.52 – 0.59 (green)	0.52 – 0.59 (green)
Spectral Band 3 (μ)	0.62 – 0.68 (red)	0.62 – 0.68 (red)
Spectral Band 4 (μ)	0.77 – 0.86 (NIR)	0.77 – 0.86 (NIR)
Spectral Band 5 (μ)	1.55 – 1.70 (SWIR)	1.55 – 1.70 (SWIR)
Resolution (m)	56	24
Swath (km)	740	140
Revisit Period (days)	5	24

2. METHODS

The method section describes the selected Geospatial Tools and documents the workflows to develop the accuracy assessment (Lemon Canyon) and upscaling-prototype of our test region (Lake Tahoe basin). Currently GDAL[23] is used as the primary source of geospatial processing and all relevant geospatial tools created to convert, reproject and align the datasets are reported below. A detailed documentation for the geospatial processing deployed for our test region can be found on the Github repository[24]. While this initial proof-of-concept is not considered to meet the required level of accuracy, it is designed to demonstrate the overall feasibility of the data collation, exercising all required pathways from data acquisition to output generation.

3. G

eospatial Tools

3.1 Geospatial Data Abstraction Library (GDAL) Usage

The Geospatial Data Abstraction Library or GDAL, is an open source library specifically developed for working with raster and vector geospatial data. For raster processing, GDAL libraries are implemented in the backend across the majority of geospatial software tools used today including ArcGIS, FME, and QGIS. The power in GDAL is its command line capability for geospatial data processing. GDAL is is relatively easy to script/scale, is well doucmented, and works with virtually all major geospatial data types.

For the purposes of our project, we are leveraging GDAL 3.2.2 for all our important raster processing:

- Reprojections (`gdalwarp`)
- Resampling (`gdal_translate/gdalwarp`)
- Masking (`gdal_translate/gdalwarp`)
- Data type conversions (`gdal_translate`)
- Metadata queries (`gdalinfo`)
- Raster calculations (`gdal_calc.py`)
- Rasterization (`gdal_rasterize`)

Additionally, we are implementing GDAL's capabilities to create:

- Hillshade
- Slope

[23] GDAL documentation: https://gdal.org/
[24] GitHub repository: https://github.com/Vibrant-Planet/vp-csm

- Aspect
- Roughness

3.2 PostGIS/PostgreSQL Usage

PostgreSQL (13.3) used to manage vector data as tables. PostGIS is an extention for PostgreSQL providing capabilites to manage these tables as geospatial layers. Vector data is uploaded to PostgreSQL using ogr2ogr. This method allows us to directly load FDGB file types to our database without the need to covert to shapefile first. Conversion to shapefile truncates the coloumn names; therefore, this is very useful tool helping keep our data intact. Utilizing ogr2ogr as our upload also allows us to force all geometires to multiploygon, further cleaning and generalizing the data.

Both the Calveg and FRID (vegetation classification layers) require cleaning and correction of invalid geometries. These are large complex data sets containing millions of verticies and complex geometries. PostGIS (2.4) is an excellent tool for handling these types of large operations. Cleaning is a necessary step for upstream processes like rasterization, but also necessary to work with the data properly. Once clean, we export the tables using ogr2ogr. Using this command as an export provides us the opportunity to build a single unified shapefile. Working with a single shapefile, although large, permits easy clips of the data using a bounding box. In the future, it is more likley we will skip the export back to shapefile and query regions directly from the PostgreSQL table instead, levergaing the speed and power of properly indexed data. For now, our process is tested and functioning.

3.3 Vector Rasterization

Rasterization of vectors is necessary for our machine learning models. If we are doing pixel to pixel comparisons, the vector data must be in the same format. Rather than rasterizing the entire vector data set, we clip the data in the shapefile as needed and perfrom the rasterization on the fly. At the time of the conversion to raster we pass the necessary extents and resolution for the output. We've found this method to be quick since the read and storage of vector data is more efficient.

3.4 Point Cloud Data Abstraction Library (PDAL) Usage

PDAL (2.3.0) is an open library written in C++ for managing and processing point cloud data. This tool is similar to LASTools in the Windows envrionment and offers similar capabilities. The advantages for us in using PDAL for this project are many. One, PDAL is an easily scripted tool allowing for scaling. At its core, PDAL utilizes JSON configurations called pipelines and allows us to string multiple processes into single documents. This method for developing individual complex pipelines for point clouds gives us the control needed to run these processes across myriad cores. Two, PDAL is developed to handle varied point cloud data types outside of the traditional ASPRS LAS. This may be important as we begin to generate surfaces from varied sources. Three, being an open library, we can if needed build additional capabilities in to the tool. One example being, adding additonal libraries for reading LAS and LAZ formats.

In our project, we are implementing PDAL for point cloud filtering plus, generation of Digital Terrain Models (DTM) and Digital Surface Models (DSM). In a basic PDAL process we will:

1. Refine point clouds into ground and non-ground points
2. Generate Digital Terrain Model (DTM) interpolation from ground points
3. Generate Digital Surface Model (DSM) interpolation from highest points

Results from the generation of the DTM and DSM are then used to calculate the Canopy Height Model (CHM).

PDAL is not developed specifically for the generation of surface grids (rasters); however, it does provide this capability when generating output in geoTIFF format. For the generation of surface rasters, an interpolation method is necessary to create a uniform surface. PDAL integrates an additional library called, points2grid to accomplish this. Points2grid utilizes the Inverse Distance Weighted (IDW) method for its interpolations. This method is powerful in creating accurate interpolations, but can have limitations. Noteably, the IDW method does not interpolate over large areas where data is not present. Specifically, in our surface generations, this results in regions of 'nodata' where there are insufficiant ground points for interpolation. This is a known outcome and currently addressed by increasing the window size for the IDW to search for neighboring points. Currently, we are confident in this solution for addressing 'nodata' regions of the surface model; however, we are continuing to research other solutions (e.g. Traingulated Irregular Networks).

3.5 Sentinel 2 Handling

Handling and manipulation of Sentinel 2 data is performed using GDAL. Raw Sentinel data comes to us in JP2 format and is projected in its corresponding UTM zone. Our first operation is to bring all the Sentinel into a singlular projection, California Albers (EPSG:3310). For the purposes of geopatial oeprations, we also convert images into GeoTiff format. If needed, Sentiel images are resampled into 10m resolution. Our resampling method is to use gdal_tranlate and to split the pixel, e.g a 20m resolution pixel is split into four new pixels to achieve 10m resolution.

3.6 Pixel Alignment

In order to have clean pixel sampling, we need to ensure all our data aligns to be the best of our abilities. This is ensured through several processes. First, projection. All data for our project is ensure to be in the same projectio wih the smae origin. Second, resolution. Each data set is always forced into a 10m resolution when GDAl commands allow. Third, extents. extents are always captured the the beginning of a process and reapplied to the outputs. Forth, scripting. Scripting allows us to build repeatable processes capturing all the perameters needed along the way.

3.7 Canopy Height Model (CHM) Calculation

Calculation of the CHM Equation 4 is an important aspect of our project, but a rather easy process.

$$CHM = DSM - DTM \tag{4}$$

We are using `gdal_calc.py` for this operation. Gdal_calc.py is a commandline tool offered by the GDAL library. Importantly, this tool provides us control over extents and data types of the outputs. Setting and swtiching between data types, like UInt16 and Float32, is important for keeping all our data in the same format.

4. ACCURACY ASSESSMENT CANOPY HEIGHT MODELS

To demonstrate our confidence in our CHM model the prototype development is preceded by an accuracy assessment for the generation of CHM's from different sources. Our test region is located in the Lemon Canyon (Californa) approximately 50 km North from Lake Tahoe. We chose this area since one of the LiDAR verification plots falls within the test boundaries and the area

To demonstrate our confidence in our CHM model the prototype development is preceded by an accuracy assessment for the generation of CHM's from different sources. Our test region is located in the Lemon Canyon (Californa) approximately 50 km North from the Lake Tahoe test region. We selected this area due to it's proximity to our prototype, valley structure with a range of distinct topographic features (e.g. slopes, aspects) and one of the available LiDAR verification plots located within the test boundaries (1.6 km2). For this plot we downloaded a subset of the 2014 LiDAR dataset from the national center for airborne laser mapping (NCALM) with 5 - 35 cm (8.93 pts/m2) accuracy from a publicly available source (OpenTopography, 2014). The dataset provides pre-calculated digital terrain model (DTM), digital surface model (DSM) and derived canopy height model (CHM).

In the accuracy assessment we will use the OpenTopography provided model (1) as a reference against the Fusion (2) and PDAL (3) generated outputs. While the OpenTopography workflows to achieve these layers are not documented, the products have been tested in regions with distinctive ridgelines and steep (e.g. grand canyon[25]) . We provided Vibrant Planet with the DTM and point cloud from Opentopography to calculate a CHM within Fusion. For the third CHM model we used the raw point cloud to generate our own CHM model using PDAL and GDAL pipelines:

1. CHM provided by Opentopography using their own DTM and DSM (all 1 m2)
2. CHM generated in Fusion using a DTM provided by Opentopography
3. CHM generated with PDAL and GDAL using the raw point cloud

CHM (2), Fusion workflow:

[25] Opentopography CHM: https://opentopography.org/news/opentopography-releases-canopy-height-model-tool

The Fusion workflow to create a CHM uses the DTM provided by Opentopography and has been produced by Vibrant Planet using the following steps:

1. Conversion of *.tif file into *.asc using ArcGIS
2. Conversion of *.asc into *.dtm to ingest into Fusion using unbuilt 'ASCII2DTM' function
3. Producing a CHM using the converted DTM and raw point cloud information
4. Conversion of the CHM from *.dtm to *.tif using Fusion 'DTM2TIF' function

The resulting CHM (2) will be used in the consecutive accuracy assessment to compare it against the Opentopography generated CHM (1).

CHM (3), PDAL workflow:

Before we generated our final CHM (3) with PDAL we produced a range of preliminary outputs and verified alignment, projection and pixel values with the existing datasets. Particularly the production of a gap-free DTM is a crucial process where a prior groundpoint filter has to be implemented to extract true last-return information from points which have been reflected by other surface features and aboveground vegetation. It is inevitable that the ground filter produces gaps within the point cloud for large trees with dense canopies. Therefore it is necessary to carefully select and adjust the ground filter as well as an interpolation process which does not exaggerate the height within the DTM. Successful generation of a DTM is verified by creating a hillshade output to check that the interpolation process does not accidentally create 'surface bumps' as a result of of incorporating mis-identified ground points into the process. The consecutive production of a gap-free DSM is an easier process since it does not require an intital classification of ground points. For this process all points can be used and the interpolation 'drapes' a surface on top of the point cloud. Once DTM and DSM are generated the CHM can be calculated using a raster calculation in GDAL by subtracting the ground information from the surface information (e.g. canopy). The result is a CHM model for the Lemon Canyon extending 1.6 km2, 2.6 Million pixels at a resolution of 1m2.

In order to assess the difference between the three different CHM's the outputs (2) and (3) were subtracted from the Opentopography CHM (1). The resulting layer can be analysed within QGIS raster statistics to estimate prediction differences.

5. PROTOTYPE DEVELOPMENT

The proof of concept demonstrates the capability of our model to upscale canopy height from areas with available LiDAR point cloud information to areas where information has not been collected. The prototype uses CHM information to train state-wide Sentinel-2 spectral information layers. We used a canopy height model generated with our PDAL workflows from the raw point cloud of the Lake Tahoe basin area. The dataset has been collected from the USGS in a 2010 survey[26] and has a resolution of 13.20 pts / m2. The

[26]USFS Lake Tahoe point cloud: https://portal.opentopography.org/datasetMetadata?otCollectionID=OT.032011.26910.1

LiDAR dataset provides pre-classified ground points at a resolution of 2.26 pts / m2. Within our workflow we will re-classify the points to improve the accuracy in the DTM.

The test region used in the final proof-of-concept is located within the city of Incline Village (California) located at the Northern tip of the Lake Tahoe. We selected the area due to Vibrant Planet's high confidence in the quality of the LiDAR data collection and the proximity of one of VP's members home. This process provides us with an additional ground-truth component in our CHM due to the limited availability of field plots.We also incorporated a rasterized version of the fire returnal interval departure (FRID 2019) dataset which contains 34 dominant vegetation classes from the CALVEG. All the datasets are projected into California Albers (EPSG:3310) and raster alignment is verified. We then calculate the normalised difference vegetation index (NVDI) from Sentinel-2 bands (NIR & Red, 10 m). The 1 m resolution canopy height model and the FRID layer are used to train the 10x10 m Sentinel-2 pixels using a linear correlation between the mean canopy height and the NDVI values calculated from the spectral bands. The upscaling process predicts mean canopy height for areas where traditional LiDAR is unavailable.

add info from Kiarie and Bogdan here

for results/discussion:

5.1 Additional Refinement for CHM Generation Methods

As noted previously, we are confident with our first generations of a CHM. Through our testing, we are able to attain similar results to CHM's produced by OpenTopography and Vibrant Planet. We are aware; however, these results can be improved. We plan to continue to explore better methods of interpolation via SAGA and GRASS GIS. We are also likely to explore LAS to surface generation using GRASS GIS.

Generating TIN models is an excellent method for surface creation of sparsely populated datasets. We have not ruled out the usage of these models; however, due to the point density for our test regions being high, we landed on IDW as our preferred method for the intial surface generation.