**Capstone Project:** Predicting Carbon Stored (Biomass) in Forest Trees

**Team:** Paul Hicks (pdh2d), Jay Hombal (mh4ey), Francisco Estrada (fge8tj)

**Introduction and Project Goals.**

With the Paris Accord signed in 2015, many countries and companies have committed to carbon neutrality, necessitating incentives and innovations aimed at **carbon emissions reductions** or **through carbon capture**.

Several large companies including General Electric, Praxair, and Exxon Mobil[8], are positioning themselves to be leaders in the carbon capture market by establishing carbon offsetting projects such as tree plantings to sell their carbon capture capacity as carbon credits. The current carbon offset market is valued at approximately $0.6 billion, but it is estimated that the market size by the year 2050 will be valued at over $200 billion (Collective). If carbon neutrality is the goal, reducing the greenhouse gas emission needs to be the first priority.

In terms of the carbon capture side of neutrality efforts, forests play a significant role in absorbing carbon dioxide through photosynthesis and storing CO2 in their wood tissues as biomass as suggested in recent research in that "...targeted tree plantings on productive forestland have the potential to increase carbon storage, but it will also increase all of the benefits of trees, such as removing air pollution and purifying water..." [5]. Carbon credit markets aim to reduce greenhouse gas emissions, which will allow emitters to offset their emission by purchasing carbon credits. Progress for such markets has been slow, but there are signs of momentum including examples such as California, a U.S. state that relatively recently started a cap-and-trade program in 2013, which has been semi-successful and has tripled in size up to 2021 (California Air Resources Board).

Accurate quantification of how much carbon forests are capable of storing is essential to the success of the carbon credit markets, which is where the California program is struggling on a technical basis. More recently, climate change goals have taken on more prominence and they are a significant part of the new US administration’s (Biden) goals (as well as other countries), with several startup companies such as Pachama, Indigo, and ClimateCorp gaining financial backing to work in this area.

Our project goal is to build a model to accurately predict the Above Ground Biomass (AGB) stored in forest trees.

**Literature Review.**

"The Biomass estimation of the forest ecosystem enables us to estimate the amount of carbon dioxide that can be sequestered from the atmosphere by forests" [6]. The above-ground biomass can be estimated using either destructive or non-destructive methods. The destructive method is counterproductive, resource and time-intensive, and therefore not scalable.

The Non-destructive methods involve taking tree measurements either by ground[7] - or aerial[4] - data survey and acquisition processes. The standard data points collected via ground survey include tree diameters at breast height (1.5 meters above ground), and tree species, while aerial- survey teams typically gather data such as tree height and crown area of the individual trees in a forest. The aerial data is gathered larger than the ground teams can achieve. The ABG value is calculated using allometric equations from the tree stem diameter (from the ground survey dataset). Allometric equations generalize the carbon sink capacity of single species or multi-species forests from trees' previously mentioned physical parameters, such as stem diameter (diameter at breast height).

Previous above ground biomass estimation efforts to the AGB estimation efforts could not scale due to the time and workforce resource-intensive requirements of destructive and non-destructive ground data collection methods. However, the aerial surveys can not gather the same data as the ground survey, therefore unable to achieve highly accurate above-ground biomass predictions at scale.

Hence, there is a real need to design an approach that uses the ground and aerial survey data for predicting biomass. This is the area we researched heavily, and we will detail our methods below.

**Methodology.**

**Data preprocessing and feature engineering:**

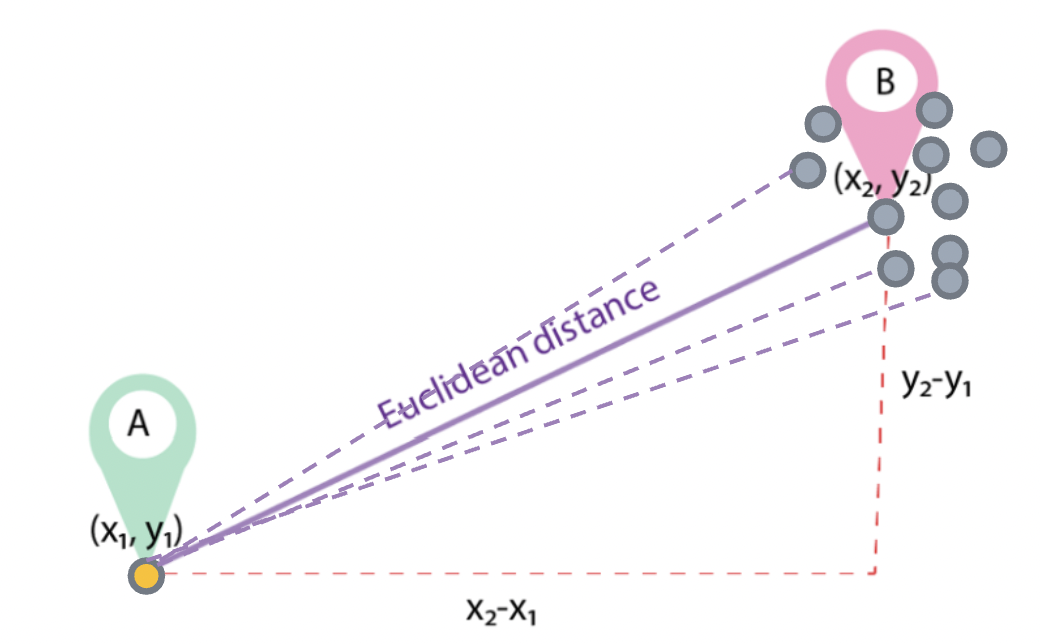
The ground survey dataset contains 53 species and 72555 trees, and the dataset has 2917297 trees for a much larger forest area, including the plots from the ground survey.

During the data cleaning and preprocessing phase, based on the data analysis, we cleaned the ground survey dataset by deleting the dead tree stems and records with missing data. Early in our research, we assumed that **the smaller trees measured in the ground survey dataset would not be visible in the aerial survey dataset, as large tree crown areas cover them**. So we filtered the smaller trees (< 20cm stem diameter) from the ground survey dataset. Similarly, we also filtered the aerial dataset to retain only trees corresponding to the ground survey dataset using geo-coordinates. We introduced a new column (**AGB**) by calculating the AGB from the Using allometric equations and the ground survey data. And ‘**estimated height**’ calculated using the stem diameter[9]. We also assumed **the stem\_diameter is correlated to the height of the trees.** To reduce the effects of minor observation errors, we introduced the bins based on tree heights in the aerial dataset and bins (**gdbin**) based on the stem diameter (dbh) in the ground dataset.

**Matching Algorithm:**

As previously mentioned, one area we stepped into that was novel included matching the ground survey data with the aerial data. These two data sets were acquired by different teams using different technologies and in the end, we needed information from both data sets so we could validate our model. We approached merging the two data sets using two different methods at different times as we were iteratively learning the most effective path forward through trial and error.

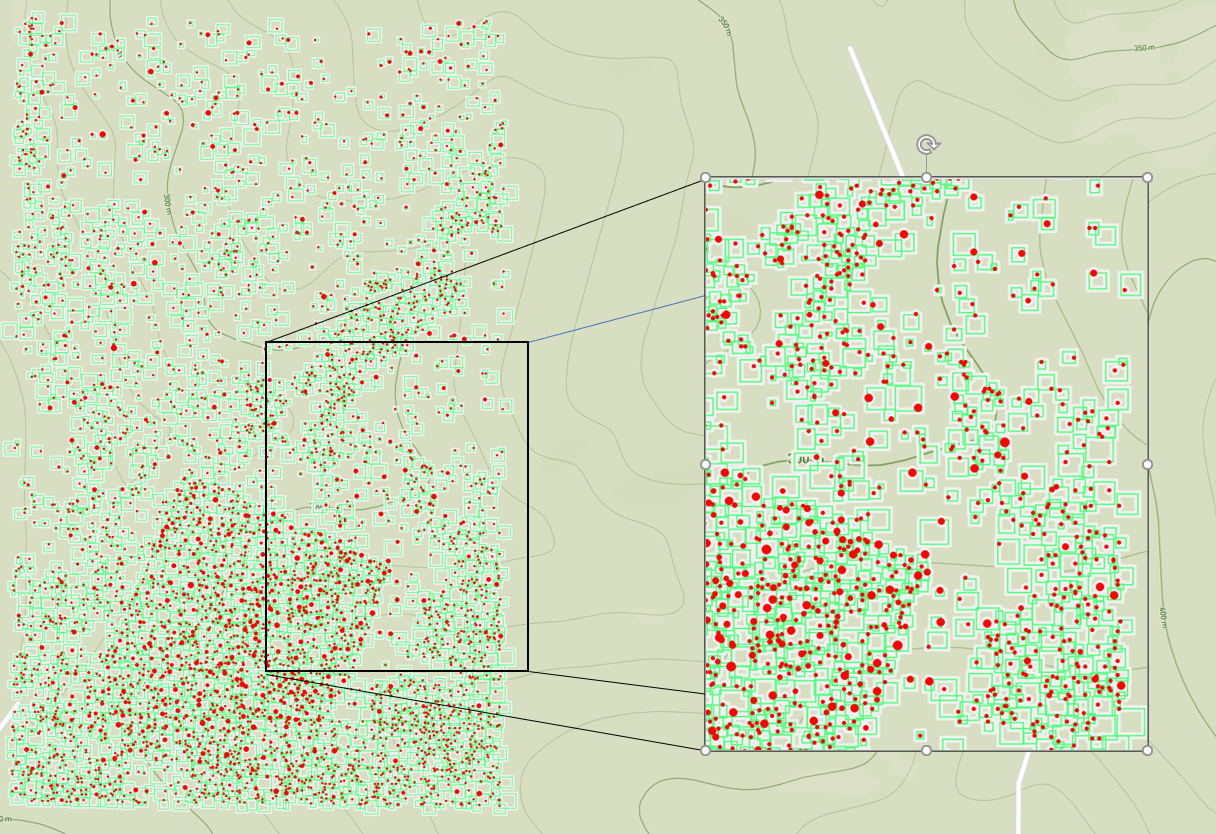
### **Approach 1: Matching Trees based on Euclidean distance calculated using geocoordinates:**



We calculated X, Y tree center geo-coordinates for the aerial tree dataset using the bounding box geo-coordinates, then computed the Euclidean distance between ground survey tree geo-coordinates and the newly calculated X, Y coordinates for a tree from the aerial survey dataset. To find an exact match from the candidate matches, we experimented by selecting a tree with max AGB and estimated height features. The tree from aerial survey data at the shortest distance from the selected ground survey tree was identified as the matched tree. The merged dataset included AGB, euclidean distance, stem diameter, height, tree crown area, and species as the features.

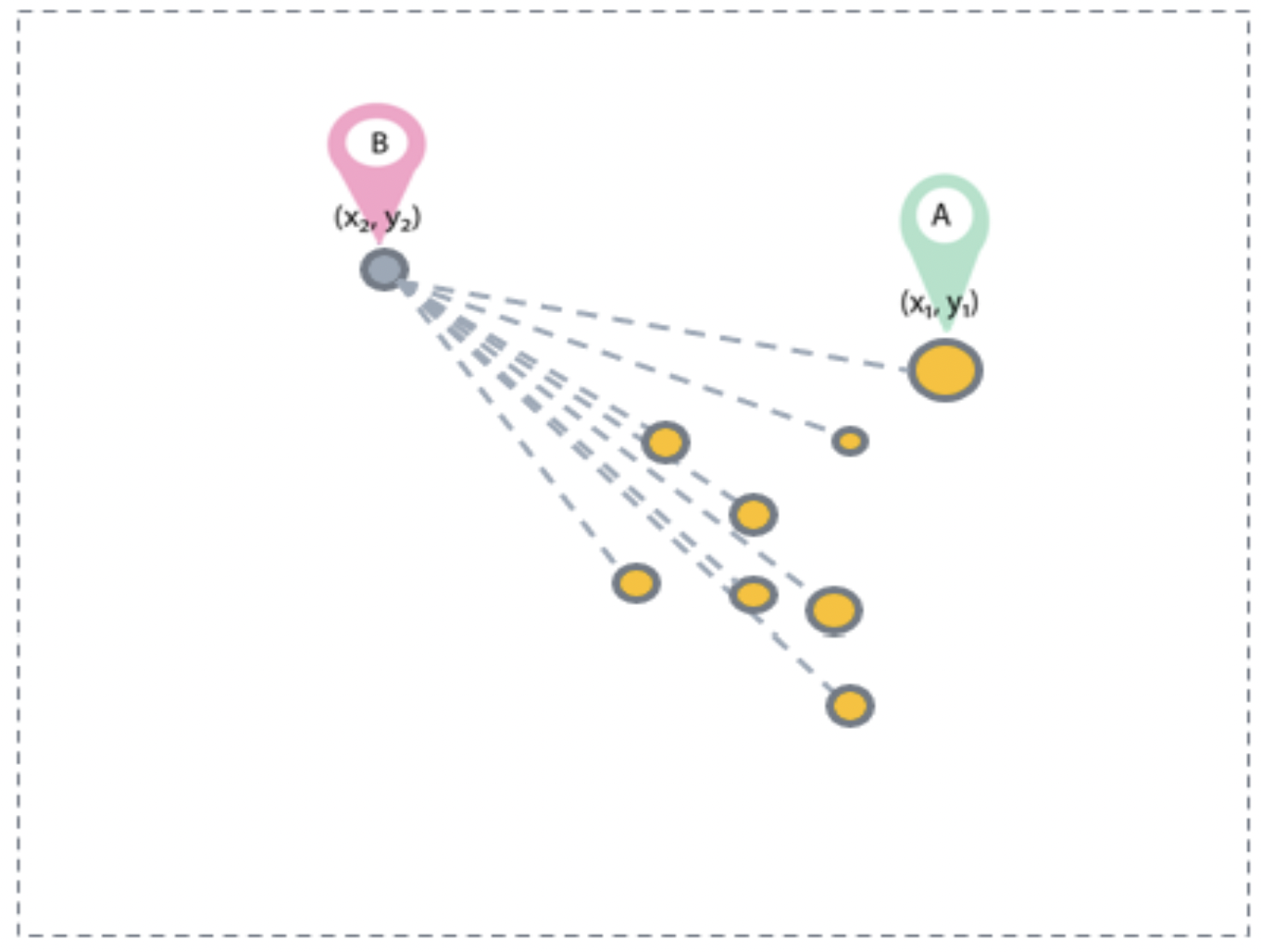
The matching based on AGB and estimated\_height introduced significant noise in the combined dataset, so our models. This led to a lack of correlation between our key target variable of above-ground biomass from the ground data and the extracted features of height and area from the LiDAR data. This initial method leveraged the geocoordinate data we had in both data sets, where we had the geolocation of each individual tree from the ground survey team and we had a bounding box of geo-coordinates from the aerial survey team. Initial filtering criteria incorporated using euclidean’s distance to understand which ground trees might match the aerial bounding .

Despite numerous attempts at modeling the AGB, using linear regression, random forest, and/or XGBoost, we were unable to gain an adjusted R-squared above 11%. This complication led to a re-evaluation of the matching algorithm to try and find a better use of physical principles based on domain expertise with forests in order to enable a more coherent matching algorithm formulation.



**Figure 1.** Forest level overview (left). Detailed view (right) revealing canopy area and tree circumference mismatch.

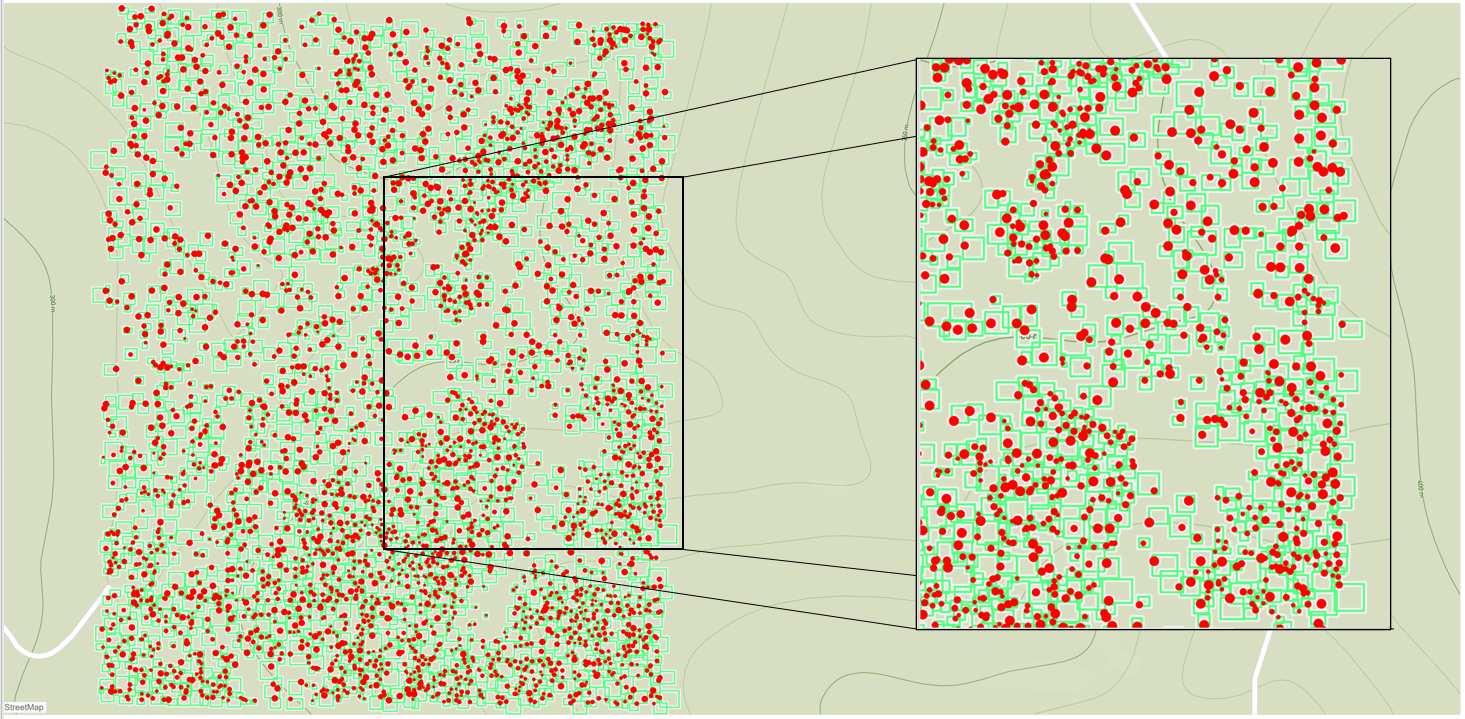
### **Approach 2: Bin Matching Tree from aerial dataset to a ground dataset:**

****

Using geolocation data from the ground and the aerial dataset was essential for matching trees accurately. In the second approach, we used the binning or bucketing data pre-processing technique to reduce the effects of minor observation errors and improve the filtering of trees using physical principles. We designed a bin-matching aspect to capitalize on the fact that we understand taller trees with larger canopy areas should have larger AGB values.

Using data binning, we bucketed the ground data into five bins based on stem diameter and the aerial dataset into five bins based on tree height. Next, we found candidate matches from the ground dataset that fall within the bounding box of a tree from the aerial dataset. Finally, we choose the tree with the highest stem diameter as the matched tree among these matches. To accommodate for the uncertainty, we allowed for matching trees from ground survey data with +1/-1 bin, corresponding the bin # from the aerial dataset or exact matching results (bin number must be the same in each data set). We saw an improvement in the matching algorithm observed below. The exact matching results resulted in over 1700 trees remaining for modeling, but the results were as pure and coherent as possible. The merged dataset included AGB, stem diameter, height, tree crown area, and species as the features.

Given our confidence in the matching algorithm’s modifications, we achieved an adjusted R-square of 87% using the exact matching algorithm, which was a significant milestone for our efforts.



**Figure 2.** Forest level overview (left). Detailed view (right) revealing canopy area and tree circumference improved matching results.

**Modeling Results.**

In the modeling phase,we placed emphasis on using predictive modeling to evaluate how well our merging methods were performing. We performed several foundational linear and non-linear models, including XGBoost, linear regression and random forest models. XGBoost (XGB) is a decision tree based ensemble machine learning algorithm that utilizes weak learning prediction models, where errors are minimized by a gradient descent algorithm, that can be used to solve regression, classification, ranking and user defined prediction problems. Our top key results for our second merging approach are shown below in Table 1. We continued exploring additional modelings that can be seen in Table 2. These results represent a significant improvement (several orders of magnitude improvement) in predictive capacity of our data merging methods and combined with the modeling results serve to motivate further research and feature extraction (discussed below).

**Table 1. Modeling Results.**

| **Model** | **Performance**  **(R Squared)** | **Error**  **(RMSE)** |
| --- | --- | --- |
| Random Forest (log AGB ~ area + height) | 0.879 | 0.421 |
| Polynomial Linear Regression (log AGB ~ area(4) + height) | 0.867 | 0.444 |
| Linear Regression (log AGB ~ area + height) | 0.502 | 0.859 |

**Table 2. Other Experimental Modeling Results**

| **Model** | **Performance**  **(R Squared)** | **Error**  **(RMSE)** |
| --- | --- | --- |
| XGBoost (log AGB ~ area + height) | 0.995 | 6.177 |
| Random Forest (log AGB ~ area + height + RGB values) | 0.895 | 0.395 |
| Polynomial Linear Regression (log AGB ~ area(5) + height) | 0.867 | 0.444 |
| Linear Regression (log AGB ~ area + height) | 0.502 | 0.859 |

**Challenges and Future Research**

We understand that the ground data survey teams are a major resource constraint and that relying on their data will limit AGB predictive modeling to surfaces of the Earth in areas of wealth and scientific resource abundance. We also know that areas that will be most impacted by climate change are communities with low resources and a need for financial assistance. With these two contradictions in mind, we believe that an area ripe for research includes how to scale these predictive models in such a way that one can use only aerial or satellite imagery data in order to make AGB predictions, which will unlock more surfaces of the earth and affected communities to the carbon credit market participation at a time when global resources are being gathered for these and other climate change interventions coming out of COP26 (Czapla, 2021). Such a method will unlock scaling into tough to reach corners of the globe as well as under-resourced communities. To do so, we have initially looked at incorporating information from the visible light spectrum (RGB) from high resolution imagery. One can see modest improvements in performance in Table 2 in terms of experimental preliminary results. However, this area was not heavily researched for this project and we consider feature extraction of additional spectral data as a subject ripe for further research. Spectral data likely holds significant potential information dealing with tree species biochemistry and therefore represents an enticing path forward, particularly for overhead data that can glean multispectral and hyperspectral data that extends predictive power and may hold the key to scaling and reducing uncertainty in AGB modeling results.

**Conclusion**.

We believe our research has proven a path forward for merging important information from ground and aerial survey teams using physical principles and guided by modeling performance. This type of research is important for the ability to develop predictive models for underpinning carbon credit markets globally such that benefits are scientifically grounded in physical principles. This principle will play into adoption of such models globally but it will also serve to level the playing field, particularly if low-cost methods for data acquisition, such as gaining data from overhead imagery can further develop and become even more ubiquitous. To do so, more investigation into spectral data acquisition and validation against species data will need to move forward to prove themselves in a community and field that has not only technical challenges but geopolitical ones as well.

# **References**

1. California Air Resources Board. *Cap-and-Trade Program | California Air Resources Board*. n.d. Website. 2 November 2021.
2. Collective, Countercurrents. *https://countercurrents.org/2021/03/big-companies-find-profit-in-carbon-offsets-market/*. n.d. 31 October 2021.
3. Czapla, Ewelina. *American Action Forum*. 17 November 2021. https://www.americanactionforum.org/insight/the-results-of-cop26/. 21 November 2021.
4. Goulden, Tristan. *National Science Foundation National Ecological Observatory Network*. 8 April 2021. https://www.neonscience.org/resources/learning-hub/tutorials/calc-biomass-py. 2021.
5. Hodgins, Jane. *Growing Trees and Capturing Carbon*. 31 August 2021. Website. 30 October 2021.
6. Jayakumar, Kuimi T. Vashum and S. "Methods to Estimate Above-Ground Biomass and Carbon Stock in Natural Forests - A Review." *Journal of Ecosystem & Ecography* (2012): 2-9.
7. Smithsonian. *Forest Global Earth Observatory Network*. n.d. https://forestgeo.si.edu/. 2021.
8. "Top 20 Companies in Carbon Capture and Storage 2019." Market Research. 2019. 31 October 2021.
9. Comparison of lidar- and allometry-derived canopy height models in an eastern deciduous forest https://harvardforest1.fas.harvard.edu/sites/harvardforest.fas.harvard.edu/files/publications/pdfs/Sullivan\_ForestEcoMgt\_2017.pdf