Making things right:
Cropping Drone Imagery and
Benchmarking AGB Maps again for ReforesTree

Data Science Project Report

Autumn Nguyen, Sulagna Saha
Mentors: David Dao, Donát Köller, Molontay Roland

1. Introduction

ReforesTree is a valuable new dataset for the community of Machine Learning and data scientists who want to apply their skills to help the environment, especially through forestry applications. The dataset comprises of the field data of over 4,600 individual trees in six agro-forestry sites, and high-resolution RGB drone images of these sites. The information about each tree in the field data is matched with the corresponding point in the drone image. Using these data, we can train models to estimate the aboveground biomass (AGB), hence carbon stock, of a forest area through aerial imagery, and we can also benchmark existing models. There have been studies in this area, but mainly for temperate forests. For a tropical agro-forestry area, ReforesTree was the first publicly available dataset with both aerial imagery and ground truth field measurements.

There are shortcomings in the original pipeline and dataset of ReforesTree, however. One of the core problems, discovered by Barenne et al, was that the areas captured by the drone images were bigger than areas that the ground truth field measurements were taken on. As a result, the comparison of the AGB values calculated from field measurements versus AGB values estimated from aerial imagery was not a fair

comparison. Therefore, the conclusion in the original ReforesTree paper that the satellite-based maps overestimated forest AGB couldn't be proven right. To know whether those maps overestimated AGB or not, we would have to crop the drone images so that they capture exactly the same field measured area and recalculate the AGB estimation for comparison. This was what Barenne et al didn't do, and we set out to do this.

The learning curve for us when working on this project was very steep. There were quite a few tutorials, many of which are incomplete; the open-source code had few helpful comments but many errors; and no one in our college has domain knowledge about this topic to help us. We were very lucky to receive help from our mentors, but we have determined to make it less difficult for people to get started with ReforesTree in particular, and in this intersection of data science and forestry applications in general. Therefore, we wrote this report not only as a deliverable for our Data Science course, but also as a documentation to help future students and researchers. Specifically, we hope that this report will:

- **demonstrate what we have learned** through this project, especially about data science processes, remote sensing data, and forestry domain knowledge.
- **showcase how we have exercised our data science skills** through this project
- **convey the results of our work to the ReforesTree authors**, so that they can make corrections in their dataset and additions to their documentation.
- **be clear and thorough documentation** that will help anyone who will be working with ReforesTree in the future to have a more efficient time understanding ReforesTree than we did.

To relate the sections of this report to the components of a data science process that we learned in our AIT Data Science course: Related work is in the Introduction; Data Understanding is in section 2.1, 2.2, 2.3; Data Preparation and Data Analysis are in section 2.4, 2.5; Machine Learning model is in section 3.3.

2. Methods

2.1. Understanding the dataset

The ReforesTree dataset consists of data from six agroforestry sites: Carlos Vera Arteaga, Carlos Vera Guevara, Flora Pluas, Leonor Aspiazu, Manuel Macias, Nestor Macias. It has four components:

i) Raw drone imageries

data/wwf_ecuador/RGB Orthomosaics directory has the drone images in 3 different formats (tif, kml, tfw). We mainly use the tif file which includes raster graphics and image information.

ii) Hand measured tree parameters (diameter at breast height, species, biomass, and location) of every tree

data/field_data.csv has the main information about each of the trees that people have collected from Ecuador.

iii) Set of bounding boxes of trees for each site cleaned by hand and labeled as banana or not banana

data/annotations/cleaned/clean_annotations.csv provides the bounding box information(xmax, xmin, ymax, ymin) discovered by the DeepForest model that are manually labeled.

iv) Mappings of these bounding boxes with tree labels based on GPS location

data/mapping/final_dataset.csv gives the mappings for each of the field trees to corresponding bounding boxes in the drone images.

2.2. Understanding the original ReforesTree pipeline

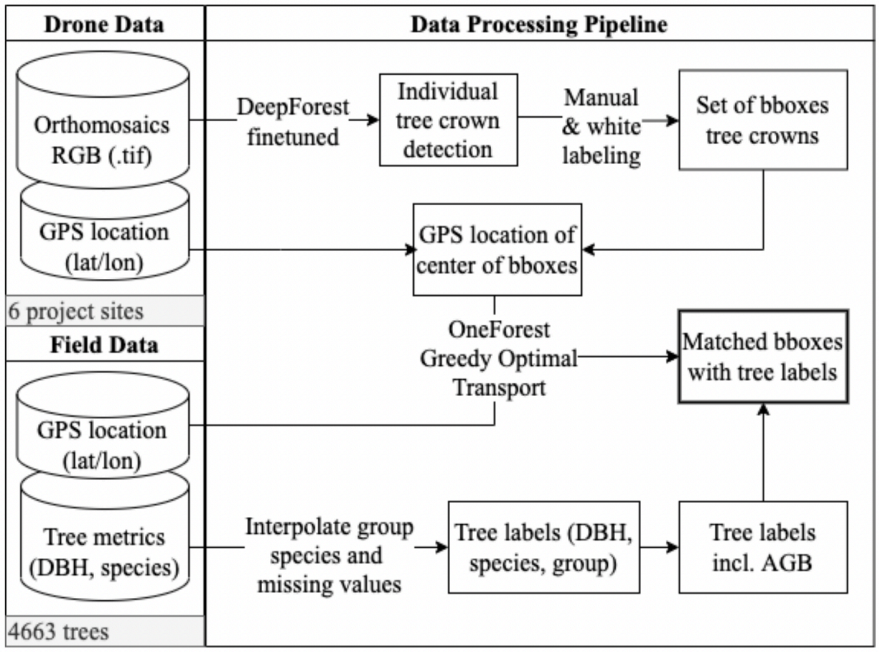

Figure 1. Original figure in the ReforesTree paper (Reiersen et al, 2022): The raw data and data processing pipeline for the ReforesTree dataset, resulting in labels matched to bounding boxes per tree.

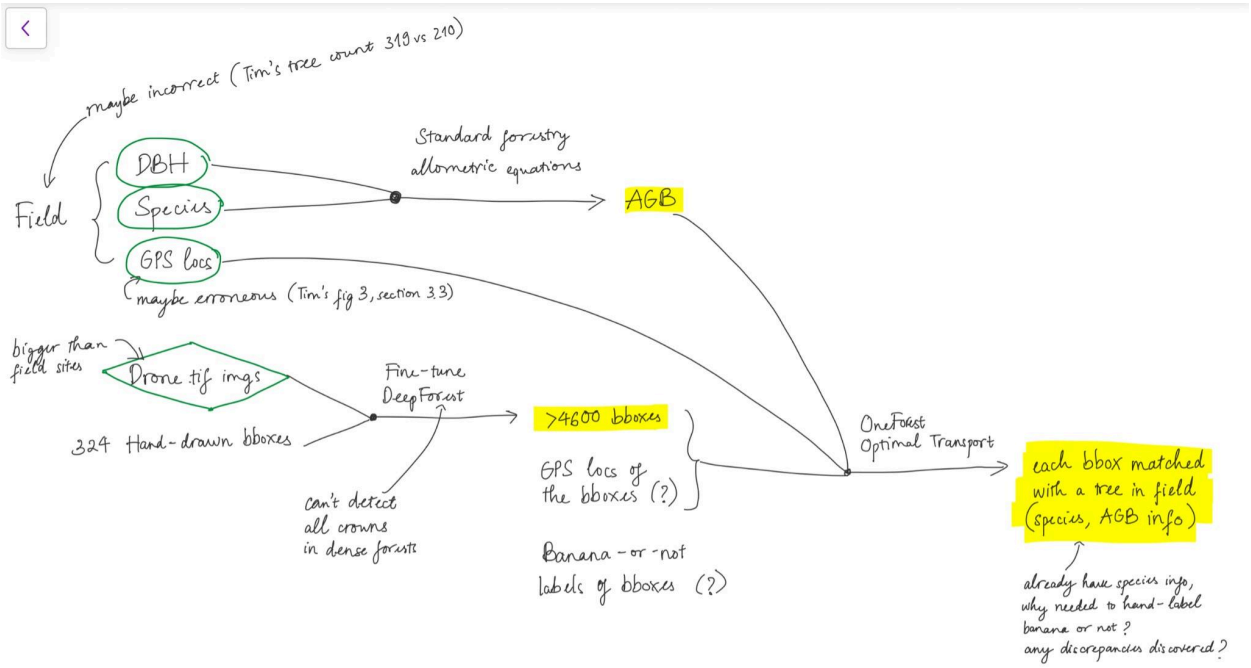

Figure A new pipeline figure that we drew to showcase more clearly the process, the problems, and the ambiguities of the pipeline (to-be-improved).

2.3. Understanding the problems

Figure 2. GPS locations of trees from field data (red dots) are plotted on drone images of the 6 sites. Coordinate system is EPSG:4326; the axes units are pixels.

In the whole pipeline for the main paper described earlier, the whole area of the drone images have been assumed to correctly fit into the area covered by the field measurements. As we can see in the pictures, that's far from the truth. As well as this error, there's also errors in the GPS locations from the field measurements; Moreover, there's potential mistakes in the number of trees from the field measurements. Our aim is to figure out, even with the mistakes, how this dataset can still help the machine learning community to learn and get some tangible results.

2.4. Tackle problem 1: area captured in drone images bigger than area got field measured.

The main ReforesTree pipeline works on an incorrect assumption that field data boundaries and the drone image boundaries are exactly the same. To fix that we aim to fix the drone imageries for the six agroforestry sites. Here are the steps taken:

i) Get the field data in a GeoDataframe

Initially, we create GeoDataframes (from geopandas library) using the longitude and latitude information of each field data point. It creates point geometries from numeric data for us to work with them easily and visualize the data.

ii) Get the boundary of the collected field data points using alphashape

We used alphashape[1] library to find the convex hull[2] that encloses a set of points. The bigger the alpha value is, the more border points the convex hull will fit around, resulting in more tight and complex hulls. We chose an alpha value of 15000 as it was a reasonable value also chosen by Barenne et al 2021.

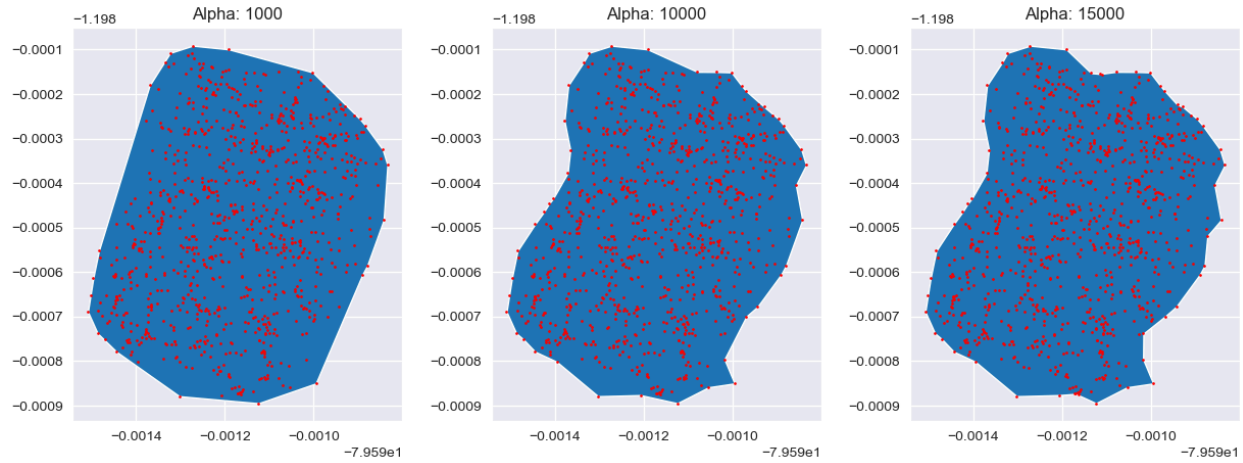

Figure 3. Convex hulls over field data points with different alpha values.

[1] Inspired from https://github.com/TimEngelmann/ai4est/blob/main/exploration/data_ground_truth.ipynb

[2] A convex hull can be thought of as a rubber band stretched around the outermost points of a set.

iii) Overlap the alphashape on the tif image and cropping it

We chose to use the coordinate of the tif images and plot the field data points in the same system. We used a mask from the rasterio library, to overlap the alphashape on the tif images. We cropped out the unnecessary parts outside of the boundary replacing them with white pixels.

iv) Fixing the white pixels

After cropping the tif images, we found out the bounds of non-white pixels of the images and we made sure it is fitted around the square shape correctly to be used for the AGBench library later.

2.5. Tackle problem 2: getting inaccurate AGB estimation from satellite-based maps

AGBench is a Python library that benchmarks satellite-based AGB maps by filtering and overlapping them with ReforesTree's drone imagery and comparing with the AGB estimations from ReforesTree's field data. We followed their AGBench[3] tutorial, and made changes to solve issues we encountered along the way, to benchmark the satellite-based maps again using the correctly cropped drone images.

3. Results

3.1. Cropped drone images

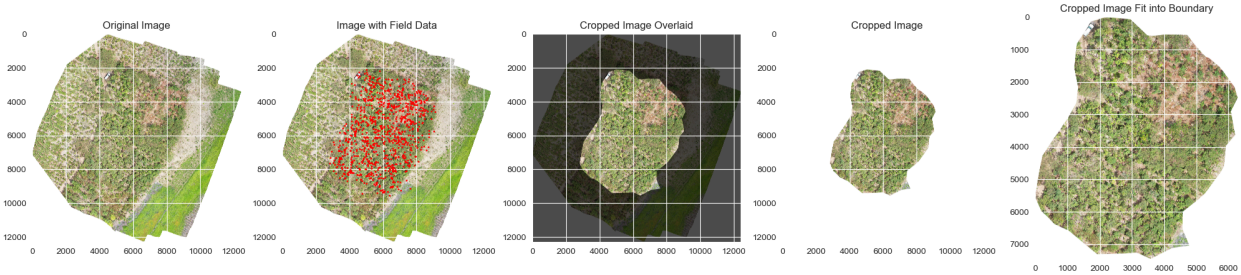

Figure 4. Process of cropping drone images to the correct field boundary.

[3] https://github.com/gyrrei/AGBench/blob/master/test/AGBench_tutorial.ipynb

Final result shown in the fifth plot after the steps taken above

3.2. Correct satellite-based AGB

[To be done] Getting the correct AGB numbers out. Comparing the field AGB numbers to see how much bigger/smaller they are.

3.3. A ML model to classify banana trees

For the easily recognizable features of banana trees, there's manual labels given for each of the bounding boxes of the drone images as banana or not-banana. We have worked on preparing the data from final_dataset.csv to train and test a simple classification model to observe how well it can recognise the species. We have cut out the image part using the bounding box information and visualized the corresponding is_banana label. As each of the bounding boxes are of different size, we have resized them to 64*64 pixels and normalized to pass into the ML model. We have 4664 images overall in the final dataset. Currently we have a (4664, 64, 64, 3) shaped array to pass into the classification model.

Figure 5: Bounding box cutouts from area Carlos Vera Guevara and Manuel Macias

We have trained a simple convolutional machine learning model using 3730 entries as training and 933 entries for testing. We used two convolutional layers(32 and 64 filters), a Maxpooling layer with (2,2) pool size and dense layer with 64 nodes and 'relu' activation function. We used categorical_crossentropy as loss function and adam as optimizer for training 10 epochs. Our training accuracy is 0.9875, validation accuracy is 0.9068 and test accuracy is 0.9067.

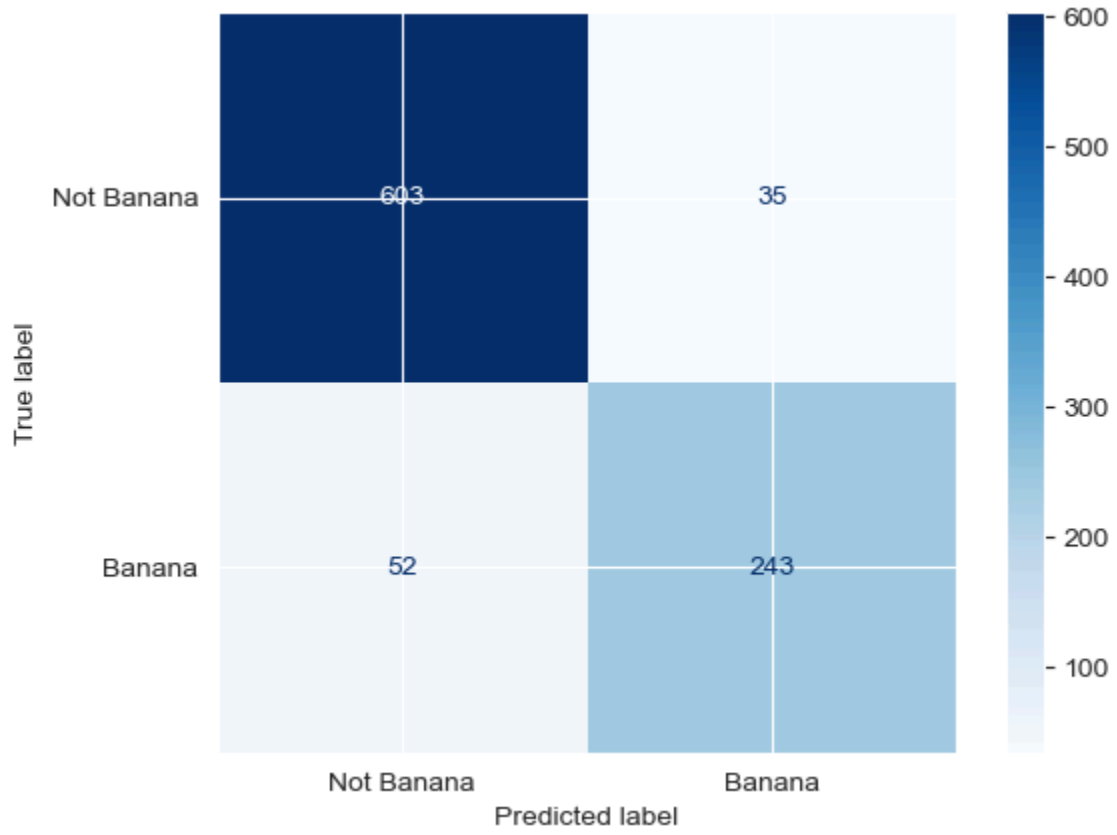

Figure 6. Confusion matrix of our baseline CNN model's result.

Bibliography

Barenne, Victoria, Jan Philipp Bohl, and Dimitrios Dekas. "Tropical Forest Carbon Stock Estimation Using RGB Drone Imagery," n.d.

Reiersen, Gyri, David Dao, Björn Lütjens, Konstantin Klemmer, Kenza Amara, Attila Steinegger, Ce Zhang, and Xiaoxiang Zhu. "ReforesTree: A Dataset for

Estimating Tropical Forest Carbon Stock with Deep Learning and Aerial Imagery." *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 11 (June 28, 2022): 12119–25. https://doi.org/10.1609/aaai.v36i11.21471.