AI4Good 2022: Team 3A

ReforesTree: A Dataset for Estimating Tropical Forest Carbon Stock with Deep Learning and Aerial Imagery

Presentation 1

Klim Troyan, Silviu Nastasescu, Dominic Wong

Agenda

- Problems to Tackle
- Goals
- Dataset
  - Creation
  - Processing
  - Initial Exploration
  - Shortcomings
- Work Plan Overview
- Questions

Problems to Tackle

- Tropical forests are important **source of biodiversity** and a **climate regulator** (i.e., carbon offsetting)

- Currently, forest carbon stock inventory is done **manually**
  → non-scalable
  → high-cost
  → labor intensive
  → error prone (e.g., systematical overestimations of forestry carbon offsetting)

- Lack of transparency and accountability of the monitoring, verification and reporting (MVR) → decrease of incentives for forest owners and buyers to enter the market

- Difficulties in designing an accurate, reliable and effective end-to-end ML pipeline (e.g., no good dataset available, high variance of tree species, occlusion of tree crowns)

Goals

Generally

- Leverage advancements in Machine Learning for high quality remote sensing technologies in order to replace the current forest stock protocols for certifications.
- Proving individual tree detection from low cost RGB drone imagery is enough to accurately estimate forest carbon stock within official carbon offsetting certification standards.

Concretely

- Estimate reliably and at scale the tropical biomass (and hence carbon stock) in tropical forests.
- Improve the bounding boxes used for individual tree detection.
- Improve the dataset creation pipeline.

Dataset: Creation

- It consists of 6 different types of agroforestry sites with planted trees, namely:
  - Carlos Vera Arteaga
  - Carlos Vera Guevara
  - Flora Plua
  - Leonor Aspiazu
  - Manuel Macias
  - Nestor Macias

- For each site, the following is given:
  - Raw drone RGB images.
  - Tree features: name/species, location, latitude, longitude, diameter at breast height (DBH), height, year, group (species), biomass (AGB), carbon stock)

Dataset: Processing

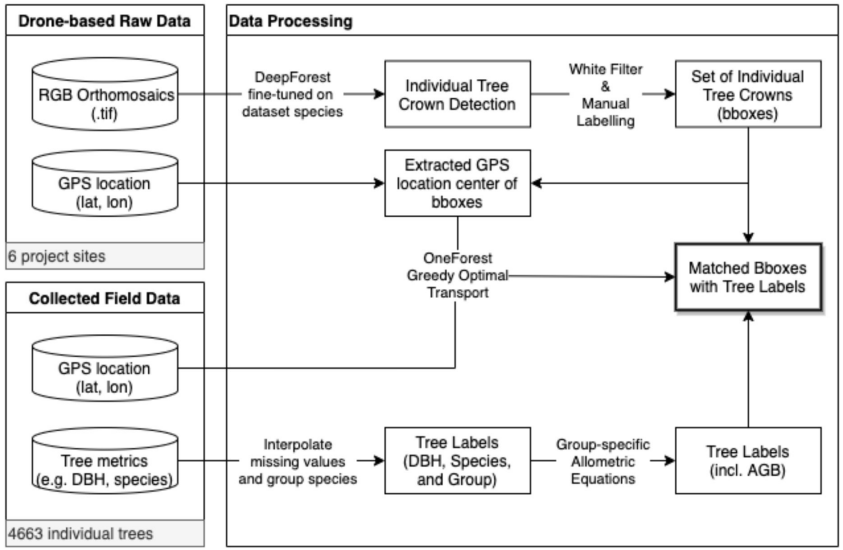
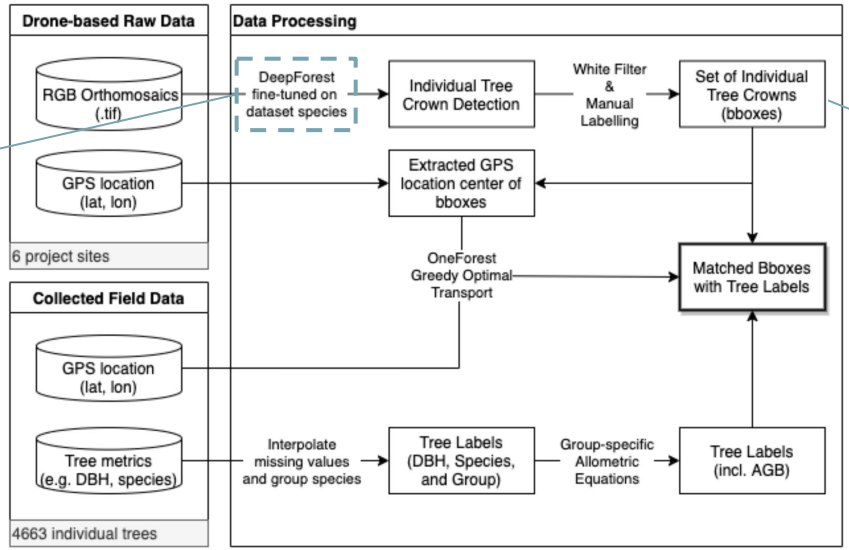


Figure 3: The raw data and data processing pipeline for the **ReforesTree** dataset, resulting in labels matched to bounding boxes per tree.

Dataset: Processing

Retinanet-based object detection CNN



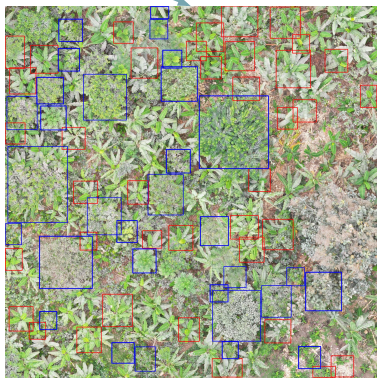


Figure 3: The raw data and data processing pipeline for the **ReforesTree** dataset, resulting in labels matched to bounding boxes per tree.

Dataset: Processing

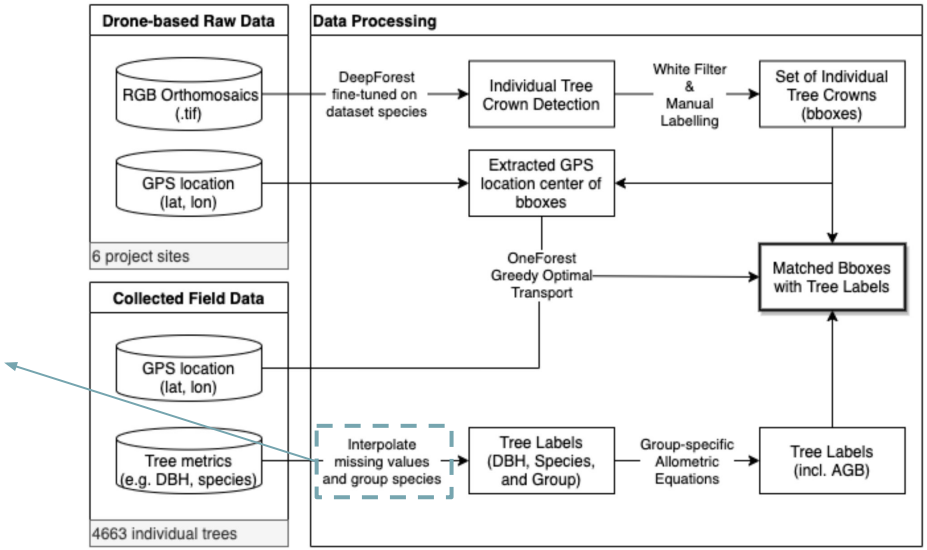
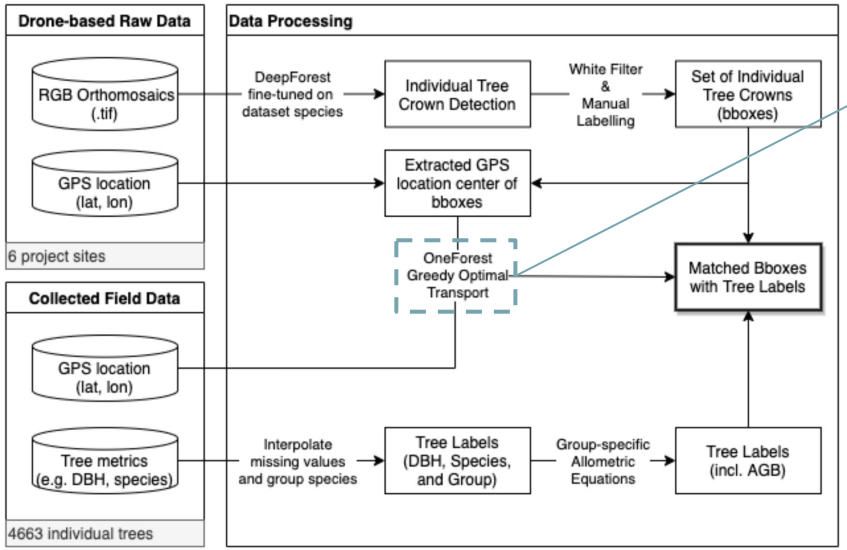


Figure 3: The raw data and data processing pipeline for the **ReforesTree** dataset, resulting in labels matched to bounding boxes per tree.

Dataset: Processing



Fusing tree data with drone imagery

Figure 3: The raw data and data processing pipeline for the **ReforesTree** dataset, resulting in labels matched to bounding boxes per tree.

Dataset: Initial Exploration (Drone Imagery)

- DeepForest model was used to create bounding boxes for each tree

- OneForest algorithm was used to choose the trees positions with the distribution closest to the distribution of the GPS values in the annotations

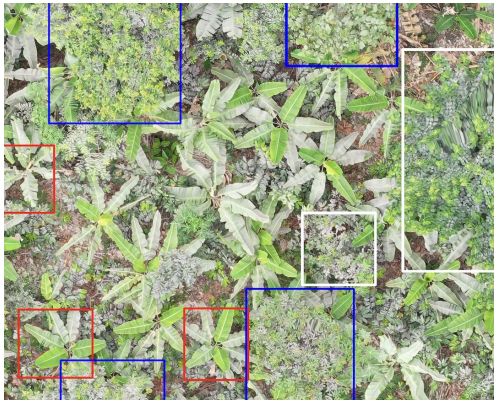
Dataset: Initial Exploration (Annotations)

- Number of tree labels: 4663
  - Banana: 1504
  - Cacao: 2042
  - Citrus: 68
  - Fruit: 751
  - Timber: 137
  - Other: 161

- Number of “NaN” for diameter measurement: 2042 (44%)
  - Replaced with mean value

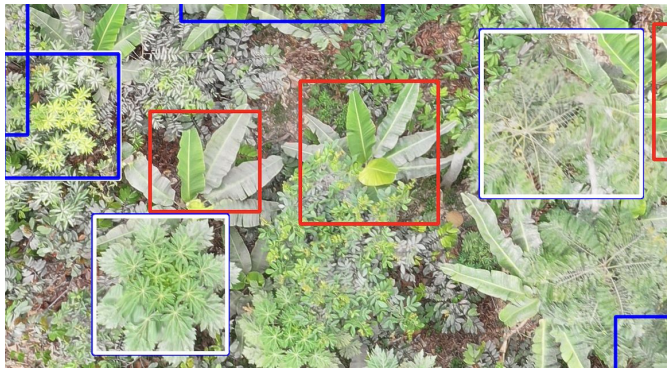
Dataset: Shortcomings

- The dataset in unbalanced: 43% cacao, 32% banana, 16% fruit, 3% timber, 2% citrus, 4% other
- The number of detected trees by DeepForest is almost twice the number of annotated trees (8520 vs 4663)
- Even after data cleaning, the number remains very large (7969 vs 4663)
- Many trees do not have bounding boxes, while (apparently) many boxes have no tree inside
- Bad bboxes → bad positions → OneForest will try to match the GPS locations to a wrong distribution



Dataset: Shortcomings

- Large number of datapoints (trees) without a measured diameter
  - 44% of entire dataset does not have a measure diameter
  - Cacao: 1806 out of 2042 trees do not have a diameter
- Multiple tree labels assigned to a single tree
- All these problems would lead us to fit a model to a **meaningless distribution**!



Work Plan Overview

Until now

- Read (critically) the ReforesTree paper
- Clearly defined the goals
- Explored the data provided as well as the processing pipeline
- Contacted the authors to obtain further info (e.g., info missing for reproducibility)
- Worked on the reproducibility of the dataset AND of the results (waiting for missing information)

From now (see next presentation!)

- One-to-one reproduction of the paper results
- Creation of a new model or improvement of the ResNet CNN (baseline)
- Modification of the dataset creation pipeline to seek improvements (e.g., imputation of missing values, OneForest for the bbox-labels matching, etc.)

Questions ?