AI4Good 2022: Team 3A

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

Presentation 2

Klim Troyan, Silviu Nastasescu, Dominic Wong

Agenda

- More information from the ReforesTree authors
- Reproducing the Baseline: Data and ResNet CNN
- From previous presentation: Dataset Shortcomings
- EDA
- Missing value imputer
- Next Ideas to Explore
- First Task: the Dataset
- Work Plan Overview
- Questions

More information from the authors

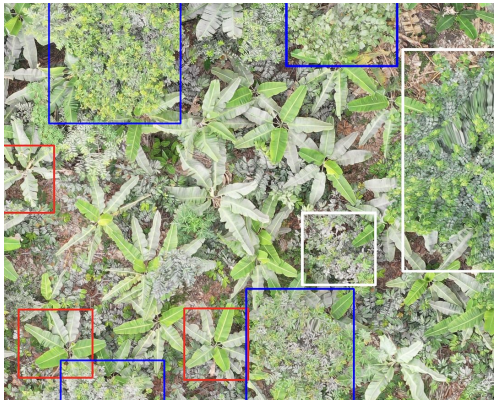
- Received the Master’s thesis on which the paper was based
- Waiting to receive the OneForest repo and associated Master’s thesis
- Waiting for the manually labeled trees in the images
- Waiting for the complete notebook used to create the dataset
- Further information useful to reproduce the results

Reproducing the baseline: Data and ResNet

- Following the pipeline presented in paper, we:
  - Extract the individual trees from each tile image based on the bboxes coordinates
  - Add zero-padding to reach a dimension of 800x800
  - Resize to 224x224 to match the ResNet CNN input dimension
  - Train the ResNet CNN and predict the AGB on the holdout test set

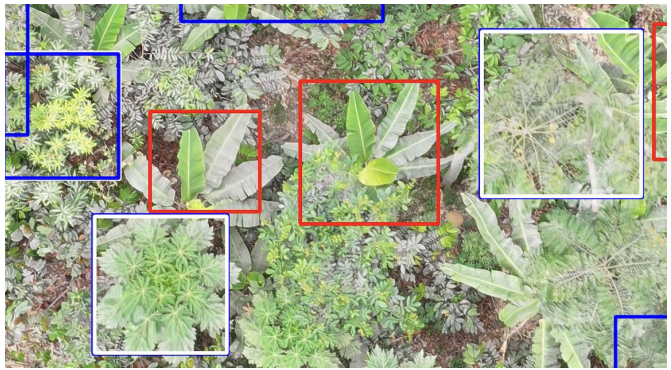
From previous presentation: Dataset Shortcomings

- 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



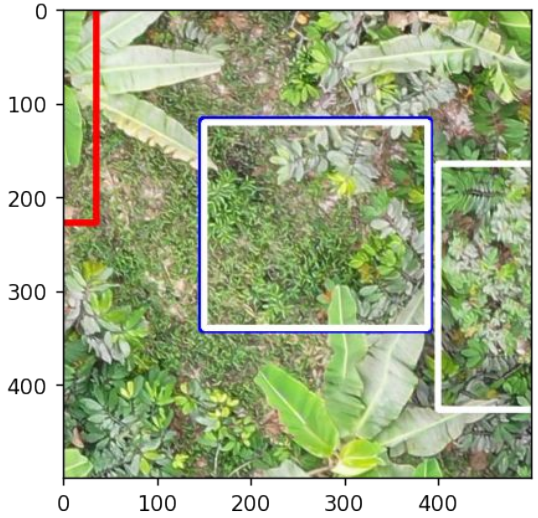
From previous presentation: Dataset Shortcomings

- 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**!



EDA: Overlapping bounding boxes

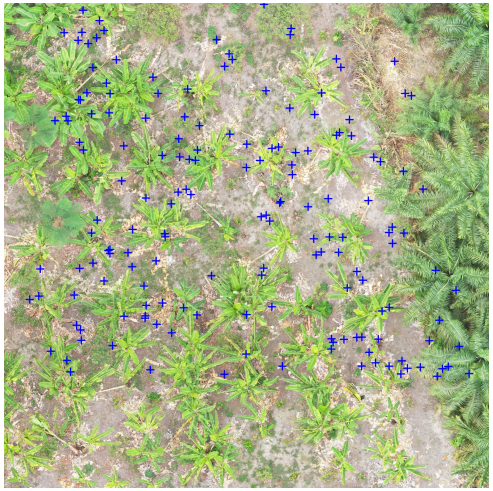
- Number of overlapping bounding boxes: 1572
  - At least 786 trees have wrong bounding box
  - 16.9% of labels
  - `banana:  160 (10.64%)`
  - `cacao:  816 (39.96%)`
  - `citrus:  29 (42.65%)`
  - `fruit:  433 (57.66%)`
  - `other:  67 (41.61%)`
  - `timber:  67 (48.91%)`
- Only two trees

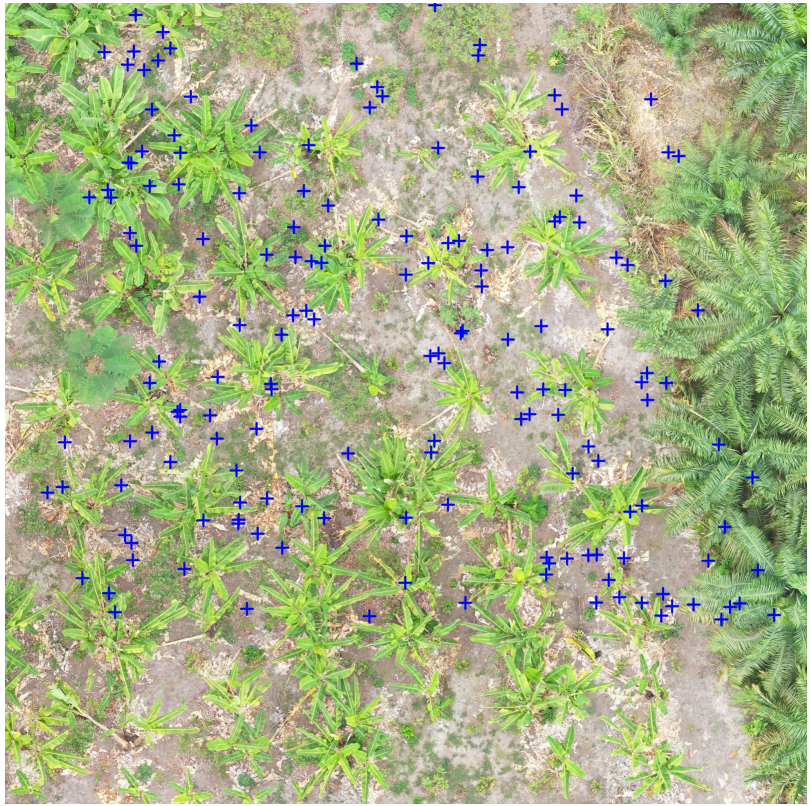


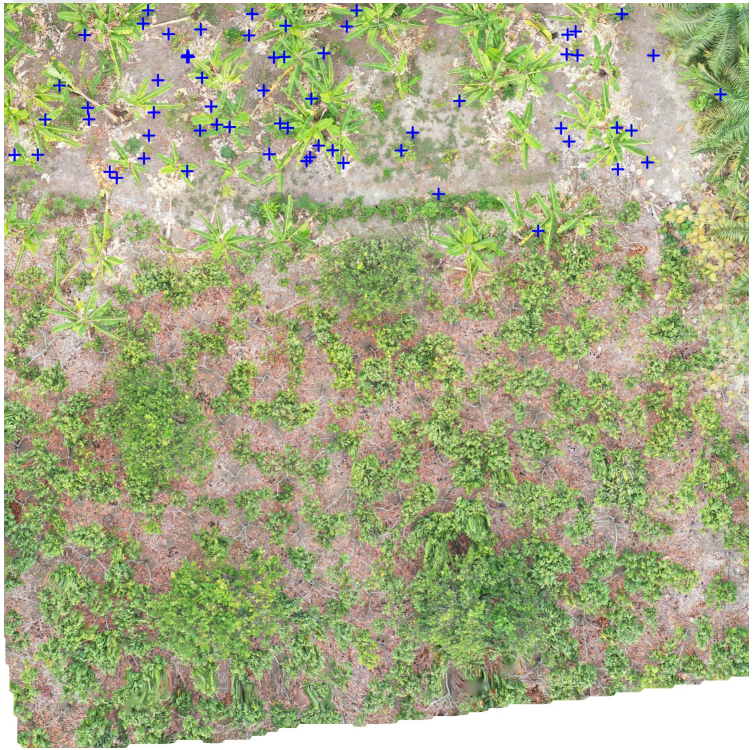
EDA: GPS location of measurements

> “The biggest challenge with the dataset is the high noise in the GPS data for the field data.”

- Noise in GPS location of measurements
- Can we quantify this?
- Reproduce mapping from GPS location to image coordinates







EDA: AGB calculation

- Carbon values in the field/final data files are wrong
    - Should be around 64% of AGB, instead of around 40%
- Fruit trees allometric equation from paper is different from the one in the Master thesis; we believe the master thesis because their referenced paper uses the same equation.

$$AGB_{fruit} = 0.1466 * DBH^{2.223}$$

$$AGB_{fruit} = 0.0776 * DBH^{2.64}$$

$$AGB_{other} = 0.1466 * DBH^{2.223}$$

Missing Value Imputer

Original:
Mean value

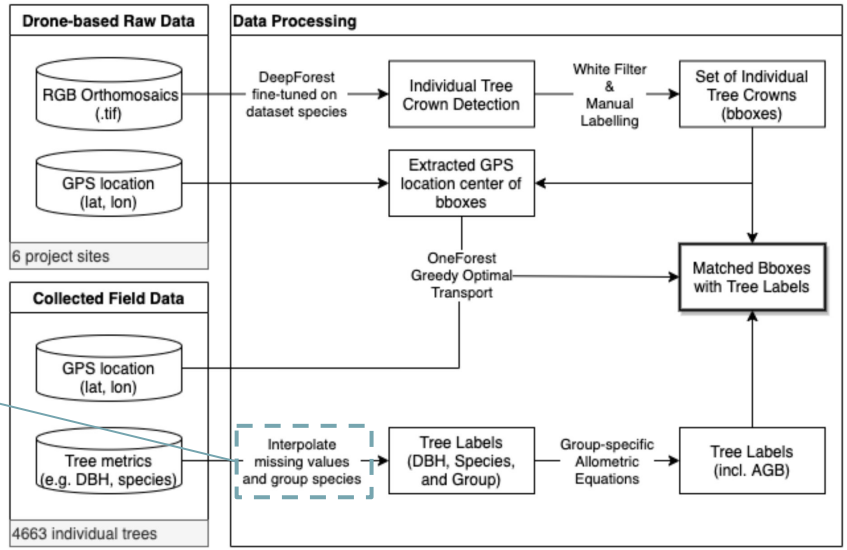


Figure 3: The raw data and data processing pipeline for the **ReforesTree** dataset, resulting in labels matched to bounding boxes per tree.

Missing Value Imputer

- Mean filter was used in the paper
- Implemented more complex imputation methods:
    - kNN Imputer
    - MICE
- Features used for imputation:
    - Species
    - Species group
    - Site
    - Year
    - Diameter
    - Height

Next Ideas to Explore

- For imputation:
  - Current implementation of categorical features uses consecutive integers
  - Planning to implement one-hot encoding
  - Trying other imputation methods
  - Choosing the best one
- Alleviate noise in GPS location of measurements
  - Quantify noise/error
  - Reproduce OneForest matching algorithm
- Removing outliers

First task: The Dataset

- It does **not** make sense to work on a model to fit a distribution considerably different from its use cases
- Trained model might not learn the actual desired behaviour
- However, the dataset has many merits
  - Importance
  - Field measurements
  - Availability of all raw data
- Idea: improve on what we have

Work Plan Overview

Until now

- Explored the data provided as well as the processing pipeline
- Modification of the dataset creation pipeline to seek improvements (e.g., imputation of missing values)
- Worked on the reproducibility of the dataset AND of the results (waiting for missing information)
- One-to-one reproduction of the paper results (not 100%)

From now (see next presentation!)

- Still waiting for authors' manually labeled data
- For imputation, the categorical data (site, species, species group) was numerically represented (site 1 → 1, site 2 → 2, ...); We plan to implement a one-hot encoding version of this (is_banana, is_cacao, ...)
- Reproduce matching procedure with OneForest
- Creation of a new model or improvement of the ResNet CNN (baseline)

Questions ?

Group members' contribution

- All group members contributed equally to the following:
  - EDA
  - Imputation
  - Reproducing results