AI4Good 2022: Team 3A

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

Presentation 3

Klim Troyan, Silviu Nastasescu, Dominic Wong

Agenda

- Current and General Issues
- Reproducing The Paper’s Results
- Outlier Detection
- Imputation
- Manual Labeling
- Literature Review: Matching Algorithm with HMM
- Work Plan Overview
- Questions

Current and General Issues

- **DeepForest**: tree crown detection algorithm
  → Problem: not good enough to detect the tree crowns
  → Solution: manually annotate the images (i.e., drawing the bboxes) and fine-tune the model on correct data

- **OneForest**: bbox to field (tabular) data matching algorithm
  → Problem: do not have the out-of-the-box model to run AND the model is not ideal
  → Solution: use our own from scratch or use the code they shared to obtain a working OneForest. Challenge being to deal with the high GPS noise.

- **Data/Measurements**: values obtained on the field manually
  Around 95% of the height values are missing. 230 trees have a height value.
  Around 44% of the diameter values are missing. 2631 trees have a diameter value.
  We do not have the true matching of an image of a tree and its group. Hence, we can only consider banana/non-banana.

Reproducing The Paper's Results

- Aim to redo the data creation pipeline (almost) completely.
  → We do NOT try to reproduce theirs
  → We follow a similar method, but try to improve it

- Using the given data, try to get similar results with a simple CNN model and then with the ResNet used.
  → We sample 100 data pairs from each site and per binary (i.e., banana / non-banana) group. Hence 1200 samples, which does not cause RAM Issues and allows for faster training as well as balanced dataset
  → We perform a train-val split of 70/30 and then a train-test split of 90/10
  → We did not get similar results for now with a simple CNN.

- **In any case, the current dataset arising from the paper is not meaningful and do not represent a useful distribution to learn**

Outlier detection

- Only the diameter values were considered
- Manually done, as most of the samples per class are in a specific range

Reminder: Imputation

Original:
Mean value

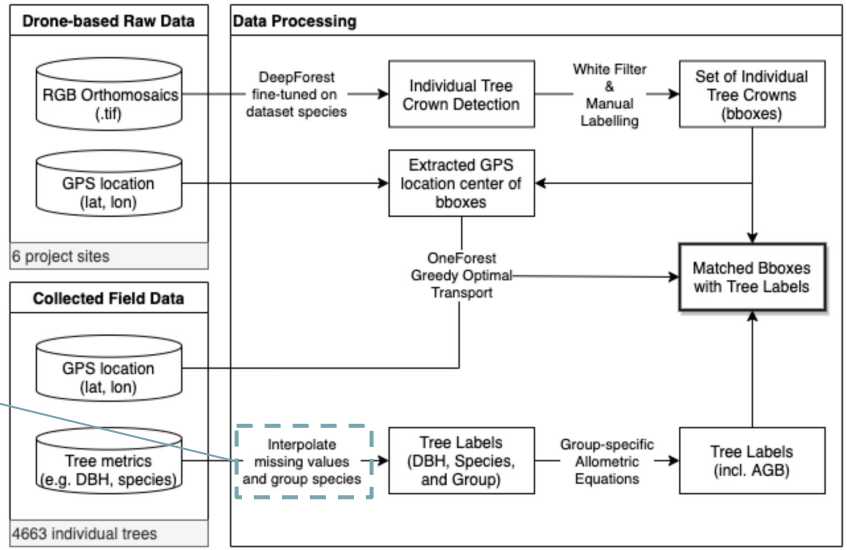


Figure 3: The raw data and data processing pipeline for the **ReforesTree** dataset, resulting in labels matched to bounding boxes per tree.

Reminder: Imputation

- Mean filter was used in the paper
- Implemented kNN Imputer
- Features used for imputation:
    - Species
    - Species group
    - Site
    - Year
    - Diameter
    - Height
- Future directions:
    - Current implementation of categorical features uses consecutive integers
    - Planning to implement one-hot encoding
    - Trying other imputation methods (MICE)
    - Choosing the best one

Imputation

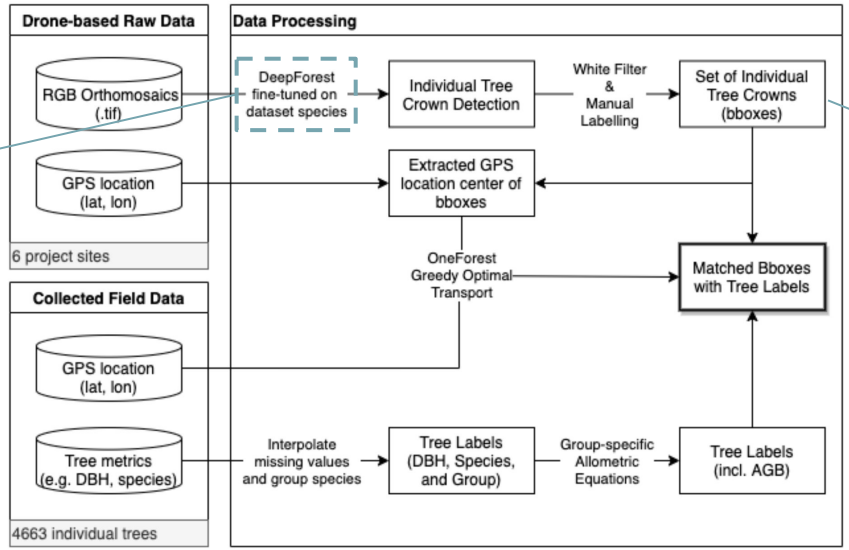
- Dropped the height column
- Categorical features were replaced by one-hot encoding
- Implemented MICE imputation
- kNN was preferred because of its simplicity and reproducibility

Manual labeling

- 200 trees were manually labeled with Roboflow
- Box format: x, y, width, height
- Only 2 labels were used for annotation:
  - Banana
  - Non-banana
- Useful for semi-supervised learning

Dataset: Processing

Retinanet-based object detection CNN



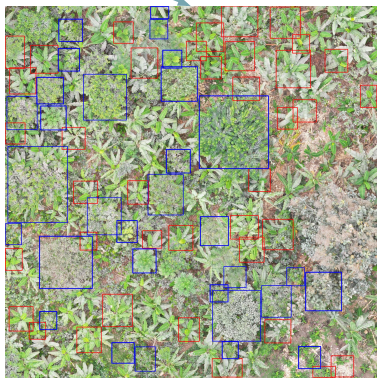
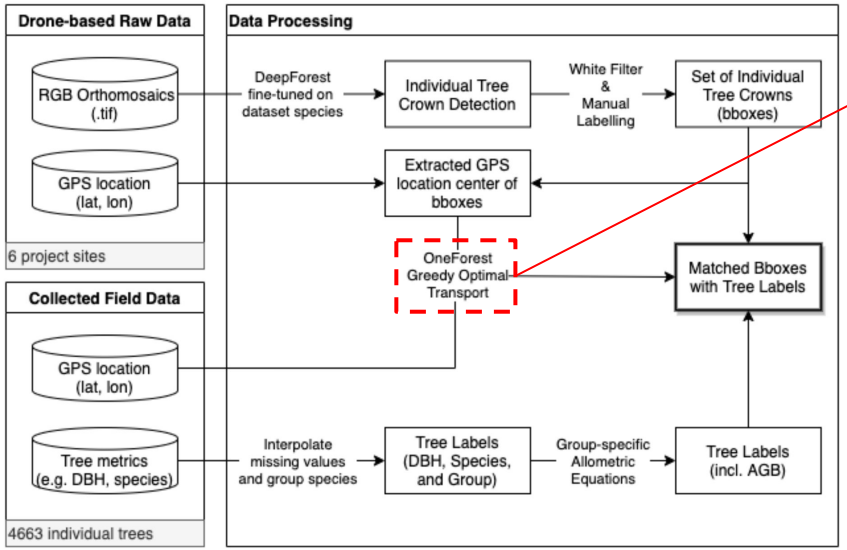


Figure 3: The raw data and data processing pipeline for the **ReforesTree** dataset, resulting in labels matched to bounding boxes per tree.

Matching algorithm

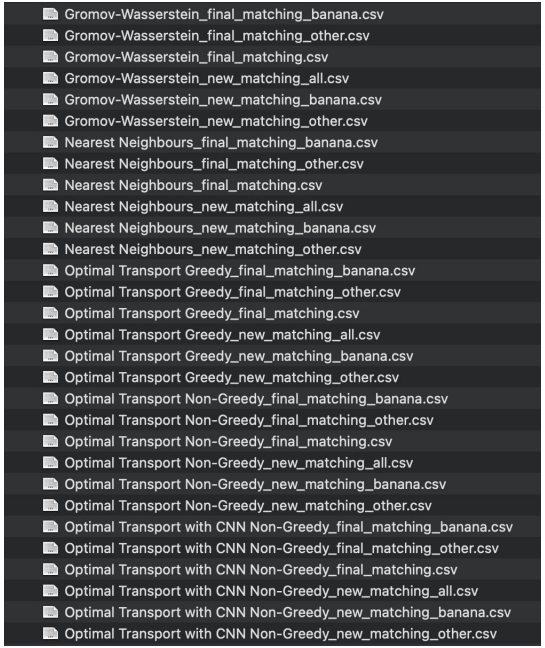


Can we improve on this?

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

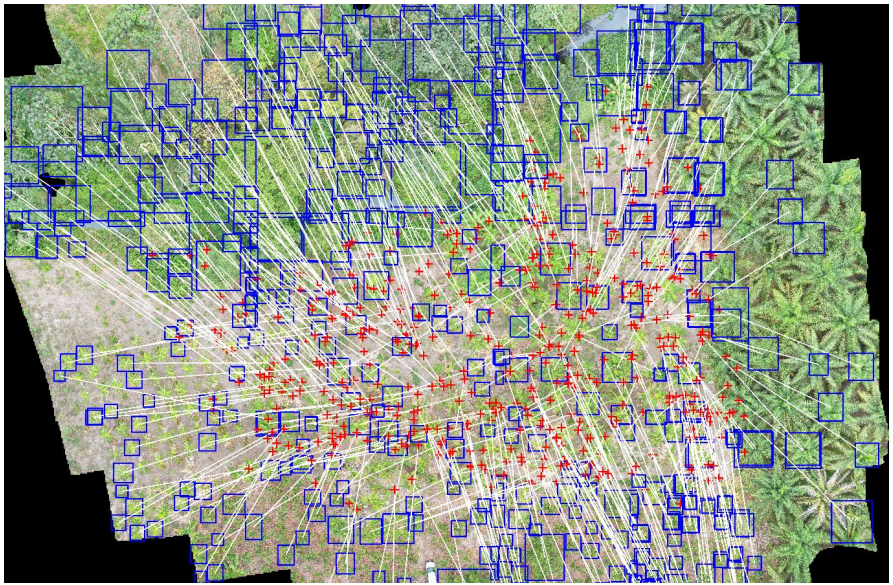
Optimal transport

- Original authors tried a variety of optimal transport algorithms
- Uses OneForest
- Inspect results of OneForest
  - Why was “Optimal Transport Greedy” chosen?
  - How do the other mappings look like?
  - What are the constraints?

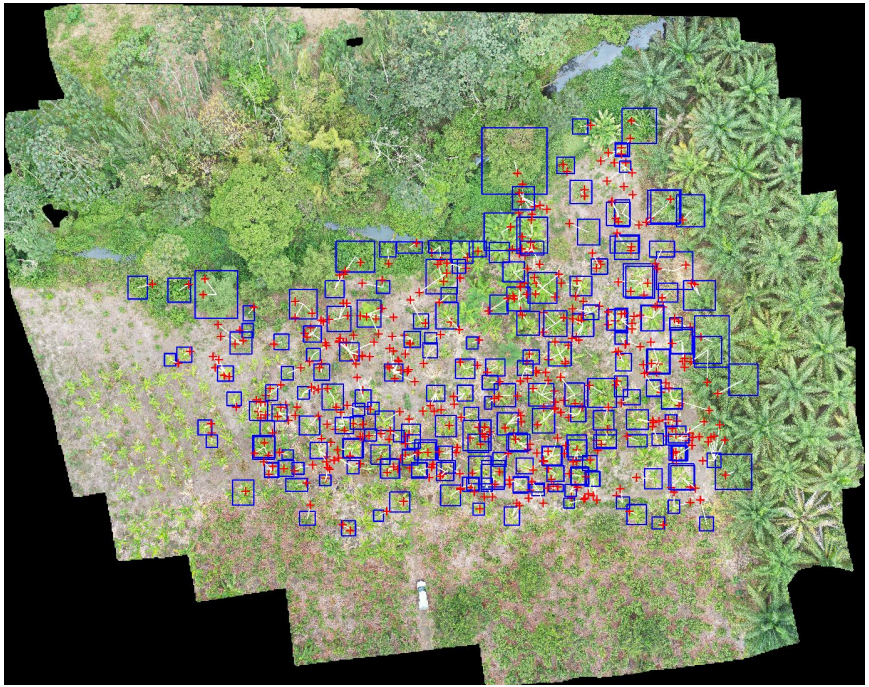


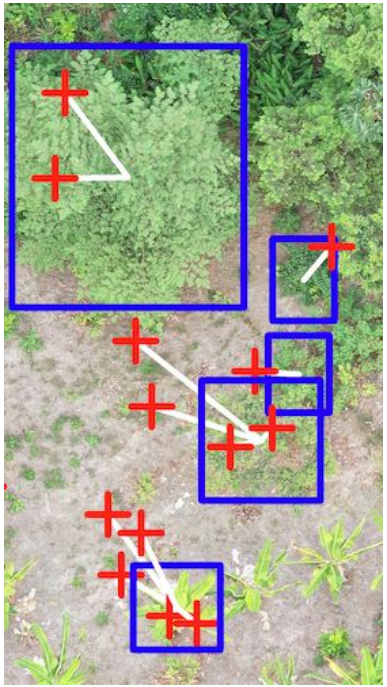
Optimal transport

- Distribution of measurements != distribution of detected trees

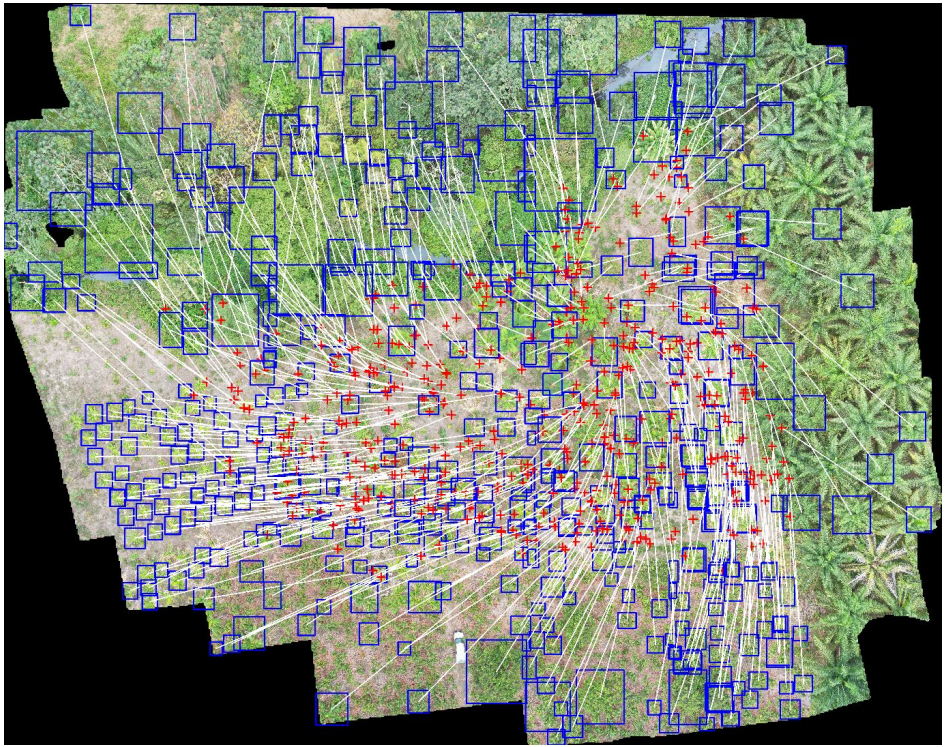


Nearest Neighbours





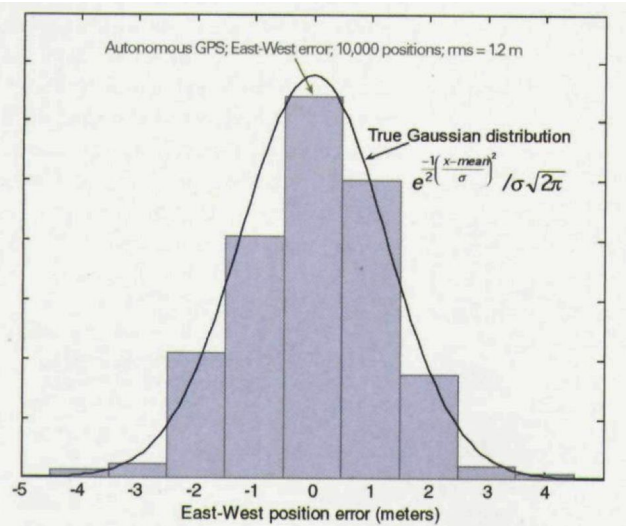
Optimal transport: Gromov-Wasserstein



Optimal transport - constraints

- Trees in admissible region
- Constraint by group
  - Tree classification and detection
- Bijective mapping

GPS noise is (kinda) Gaussian!



▲ **FIGURE 3** Experimental and theoretical GPS error distribution for a receiver operating in autonomous mode

Source: F. van Diggelen, "GNSS Accuracy, Lies, Damn Lies, and Statistics," GPS World, Vol. 18, No. 1, 2007, pp. 26-32

Literature Review: Map matching

Hidden Markov Map Matching Through Noise and Sparseness

Paul Newson and John Krumm
Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA
+1 425 705 4507, +1 425 703 8283
{pnewson, jckrumm}@microsoft.com

isprs International Journal of Geo-Information

MDPI

Article

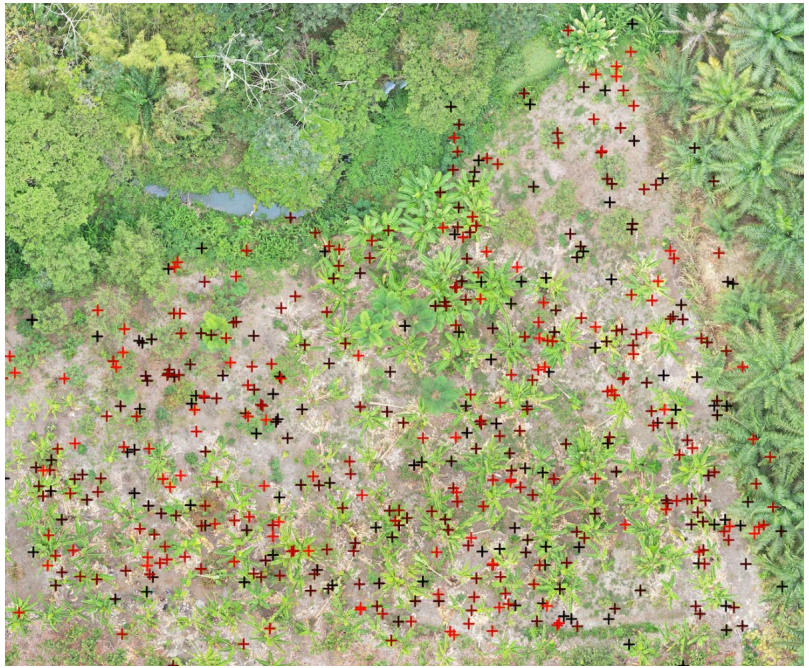
Enhanced Map-Matching Algorithm with a Hidden Markov Model for Mobile Phone Positioning

An Luo [1], Shenghua Chen [2,*] and Bin Xv [2]

Figure 1: Map matching consists of matching measured locations (black dots) to the road network in order to infer the vehicle's actual path (light gray curve). Merely matching to the nearest road is prone to mistakes.

Matching with HMM

- Initial idea:
  - Can we get some temporal information from order of measurements
- Colour of cross = position in CSV file



Matching with HMM

- Assumptions
  - GPS noise is “kinda” Gaussian
  - Distance between two measurements corresponds to distance between two trees
- Measurement likelihood: how likely is a measurement made at a given tree
  - Gaussian w.r.t. distance between measurement and a given tree
- Transition probability
  - Function of: (distance between two trees) - (distance between measurements)

Work Plan Overview

Until now

- Experimentation with imputation
- Discovery of unit error
- Investigate original repository
- Literature review into map matching with HMM

From now (see next presentation!)

- Further investigate matching algorithm
- Creation of a new model or improvement of the ResNet CNN (baseline)

Questions ?

Contributions

- All group members contributed equally to the following
  - Read the ReforesTree paper in details and studies the associated Master thesis
  - Performed further data analysis and reflected on the found issues
  - Worked on the reproduction of the paper's results
  - Performed the outlier detection
  - Implemented the imputation
  - Manually labeled a small part of the dataset
  - Went through the repo of the master thesis and of OneForest