# Africa Biomass Documentation

Author: ZLH

Subject: Biomass estimation of an area through satellite images.

Objective: Design a machine learning model to predict the biomass of an area.

### Input data

Only the three .h5 files in the training set and "TestFiles.zip" were used.

### Method

#### Outlier removal

Some agbd values from GEDI04 in the training set appear too large considering the landscape of the study area. The 99% percentile, about 301, was used as the upper bound to clip the agbd.

#### Feature engineering and model development

Biomass varies over space. As the footprint size of GEDI is about 25 m, a 3x3 pixel window at the center of each image was used to extract the data corresponding to the Lidar observations or the field measurements. Within the 3x3 windows, the features for further analysis include the 12 multispectral bands, the cloud score, the scene classification layer (SCL), the latitude, and the longitude. The geolocation coordinates might be useful to spatially interpolate the agbd. The features were flattened to form the input to the algorithm.

The test set requires further processing because the cloud cover is much higher than the training set. About 40% of the test samples were totally covered by cloud. As most of the bands have very limited capability of penetrating through clouds, the agbd of the cloud-covered samples are unlikely to be reflected by the image observations. Within the center 3x3 window in the test images, the pixels which look like "clear" observations were extracted. A pixel may be clear if the cloud score is less than 20 and the SCL value is not 1 (saturated or defective), 3 (cloud shadows), 8 (medium probability clouds), 9 (high probability clouds), or 10 (cirrus). If there is no clear pixel or the average red band (B04) reflectance is higher than 0.1 (pixel value > 1000), the sample is considered as cloud contaminated. All the samples were split into a "clear" group and a "cloud" group, and the latter was excluded when building the model.

The model is an ensemble of decision trees, whose hyperparameters were tuned by splitting the training set and visually investigating the metrics to strike a balance between overfitting and underfitting. For the "clear" group, the prediction is from the output of the model. For the "cloud" group, the samples were geographically categorized into 5 regions based on latitude and longitude, and in each region the agbd of the cloud samples were estimated using the predicted agbd of the clear samples in the same region. A linear relationship was assumed between the agbd and the canopy height for the estimation.

### Results

The method achieved RMSE = 51.52 on the public board and RMSE = 72.45 on the private board.