

## 1 Introduction

This document details the processing workflow for the terrestrial LiDAR data collected during fieldwork in Bicuar National Park, Angola, and Kilwa District, Tanzania, with the aim of documenting a reproducible workflow.

The project aimed to understand the effects of tree species diversity and composition on tree canopy architecture and grassy biomass.

## 2 Field methods

Fieldwork was conducted at two sites, the first in Bicuar National Park, southwest Angola (S15.1°, E14.8°), and the second in Kilwa District, southeast Tanzania (S9.0°, E39.0°).

The two sites have similar climatic conditions.

Site	MAT (°C)	MAP (mm y <sup>-1</sup> )	Temp. range (°C)	CWD (mm y <sup>-1</sup> )
Bicuar	20.8 (0.70)	825.9 (52.01)	24.5 (0.90)	-844.8 (44.29)
Mtarure	25.7 (0.24)	958.4 (25.19)	12.0 (0.33)	-739.6 (8.06)

Table 1: Climatic data for each site, extracted from WorldClim at 2.5 minute resolution. Values are the mean and standard deviation (in brackets) of all pixels intersecting each protected area.

Fieldwork was conducted during the peak growth period of each site, in order to capture the highest leafy volume in the canopy and the largest grassy volume in the understorey.

At each site, 1 ha permanent plots were sampled. In Angola, 15 plots were sampled, while in Tanzania, only seven were sampled, following the curtailment of fieldwork due to COVID-19 travel restrictions.

Each permanent plot was further subdivided into nine 10 m diameter circular subplots arranged in a regular grid, with a buffer from the plot edge (Figure 1).

Within each subplot, a variable number of scans were recorded using a Leica HDS6100 phase-shift terrestrial laser scanner (TLS). The number and position of scans within a subplot was determined by the arrangement and density of canopy material in the subplot. Scan positions were arranged to minimise shadows within the canopy, and to maximise canopy penetration. Number of scans per subplot ranged between one and five in both Angola and Tanzania (Table 2).

Five Leica 6" planar tilt and turn targets were used at each subplot to align scans. To allow for registration of scans among subplots, the location of each target was registered using a Leica VIVA GS10 GNSS unit, set up in post-processed kinematic (PPK) configuration with a base-station located  $\sim 100$  m from the edge of each 1 ha plot. The location of each target was measured for at least 4 minutes.

We used SOME KIND OF GLOBAL SATELLITE MATCHING TO MAKE THE GNSS BETTER.

Table 2: Description of scan settings used for each scan.

Setting	Value
Scanner model	Leica HDS6100
Wavelength	650-690 nm
Spot size at exit	3 mm
Beam divergence	0.22 mrad
Range	79 m @90%; 50 m @18% albedo
Azimuth range	0-360°
Zenith range	0-155°
Increments	0.018°
Point spacing over 25 m	7.9 mm
Pixels per line	20000
Lines	10000
Compressed file size	~800 MB
Duration of scan	6 minutes 44 seconds

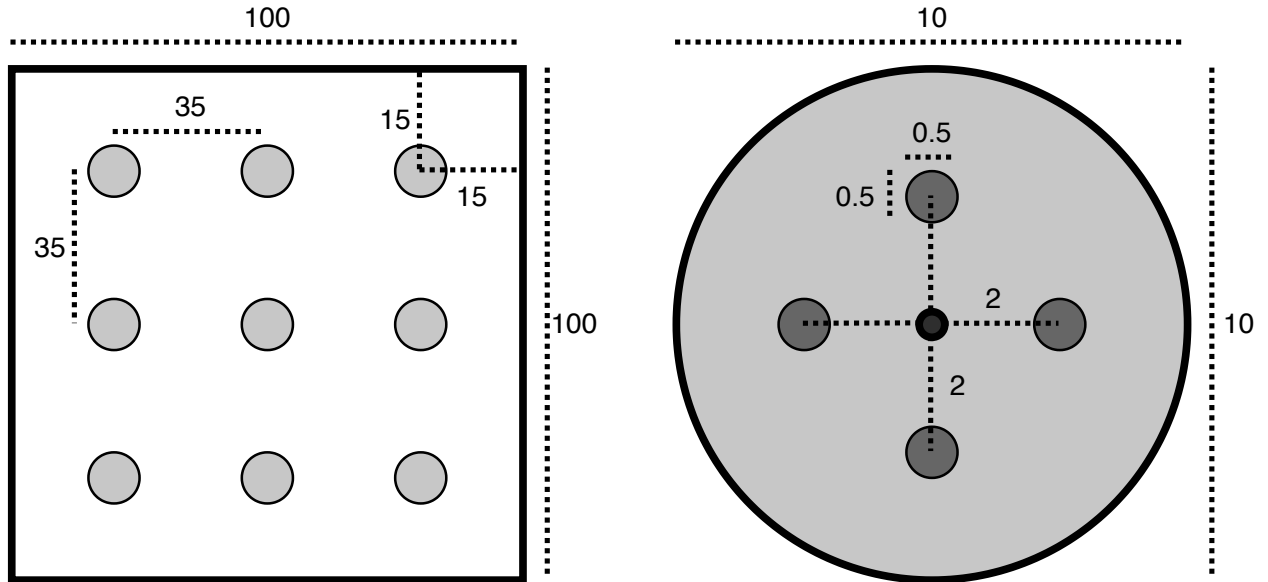


Figure 1: The layout of 10 m diameter subplots with each larger 1 ha square plot. Each subplot is situated inside a 15 m buffer from the plot edge, with 35 m between subplot centres. Subplots are arranged in a 3x3 grid. Disc-pasture measurements and biomass samples are located in cardinal directions 2 m from the centre of the subplot. All distances are in metres.

### 3 Registration

Scan registration for each subplot was conducted in Leica Cyclone (version 9.1). Targets from each scan were aligned using Cyclone’s automatic target acquisition.

After registration, scan scenes were exported from Cyclone as PTX files, one per subplot.

### 4 Voxelisation

PTX files were converted to compressed LAZ files using PDAL (). The exact code used to extract and apply the PTX rotation matrix to each point in the PTX file can be found IN THIS APPENDIX HERE.

LAZ files were voxelised to different voxel sizes depending on the application of the data. For grassy biomass estimation, we used 2 cm<sup>3</sup> cubic voxels, while for subplot height profile estimation we used 5 cm<sup>3</sup> voxels, and for whole plot canopy rugosity we used 10 cm<sup>3</sup> voxels. WHY THO

### 5 Noise reduction

Outlier detection and noise reduction was conducted in PDAL using the `filters.outlier` filter, using the “statistical method” (sensu Rusu et al. 2008), with  $k = 8$  (mean number of neighbours), and  $m = 1.96$  (standard deviation threshold, approximating a 95% confidence interval):

$$\bar{\mu} = \frac{1}{N} \sum_{i=1}^N \mu_i \quad (1)$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\mu_i - \bar{\mu})^2} \quad (2)$$

$$t = \mu + m\sigma \quad (3)$$

$$\text{outlier}_i = \begin{cases} \text{true,} & \text{if } \mu_i \geq t \\ \text{false,} & \text{otherwise} \end{cases} \quad (4)$$

where  $N$  is the number of points in the scene,  $\bar{\mu}$  is the mean distance to nearest neighbour points, and  $\sigma$  is the standard deviation of these distances.  $t$  is the threshold distance used to define an outlier.

### 6 Foliage density profiles

To estimate subplot foliage density profiles, first the point cloud was cropped to a 10 m diameter cylinder of infinite height. Then the `filters.pmf` (Progressive Morphological Filter - PMF) PDAL function was used to identify ground points (sensu Zhang et al. 2003). The `filters.hag_nn` (Nearest Neighbour) PDAL function was used to generate height above ground of each point within

the cylinder. Points below ground level were then discarded. Height profile points were exported to a XYZ file then imported into R for further processing.

We excluded points above the 99.9th percentile of height, under the assumption that these often constituted noise that had not been adequately removed by PDAL.

In R, within each 1 cm width vertical layer, we calculated the gap fraction as the proportion of unfilled 1 cm<sup>3</sup> voxels. We filtered the point cloud data to the tree canopy, excluding grass. We identified the breakpoint between the grassy understorey and the tree canopy as the first local minima above 1.3 m from the ground.

We extracted statistics from the foliage density profile for use in statistical analysis. We first smoothed the density profile using a loess model with a span of 0.1. We then calculated the number of local maxima and minima along the profile. We defined local maxima and minima as points regions where the gap fraction of the surrounding 50 cm of 1 cm bins was lower or higher, respectively.

We calculated the effective number of layers (ENL), using the Shannon diversity index (sensu (Ehbrecht et al., 2016)).

We calculated the area under the curve of foliage density using trapezoid estimation.

We extracted the height of the maximum foliage density peak, and calculated the difference between the highest and lowest local maxima. We also extracted the maximum canopy height within the subplot.

We calculated the coefficient of variation of the point cloud height distribution.

To describe the uniformity of the foliage density distribution we used Ripley's L function, which is more commonly used in describing spatial variation across a 2 dimensional surface. Ripley's L is an adjustment to Ripley's K, defined as:

$$\hat{K}(t) = \lambda^{-1} \sum_{i \neq j} \frac{I(d_{ij} < t)}{n} \quad (5)$$

$$\hat{L}(t) = \left( \frac{\hat{K}(t)}{\pi} \right)^{1/2} \quad (6)$$

Finally, we calculated the Shannon diversity index on the gap fraction of 50 cm bins:

$$H' = - \sum_{i=1}^N p_i \ln p_i \quad (7)$$

Where  $N$  is the number of 50 cm bins in the height profile, and  $p_i$  is the proportion of filled voxels in layer  $i$  (gap fraction).

## 7 Canopy gap fraction

Simple canopy gap fraction as seen from the ground was measured using two methods: 1) hemispherical photography and 2) terrestrial LiDAR.

## 7.1 Hemispherical photography

At the centre of each subplot a single photograph was taken with a Nikon D750 full-frame DSLR camera, with a circular fisheye lens. The lens had an equisolid (equal area) projection, which avoids image distortion. The projection function is given as:

$$R = 2f \sin(\theta/2) \quad (8)$$

Where  $R$  is the radial position of a point on the image on the sensor,  $f$  is the focal length of the lens, and  $\theta$  is the angle in radians of the desired angular radius of the cropped image.

The photo was taken facing directly to zenith, with the top of the camera facing to magnetic north, at a height of 1.3 m or above understorey vegetation, whichever was higher. Table 3 shows describes the camera settings for each hemispherical photo:

Table 3: Description of camera settings used for each hemispherical photo. Note that the values of shutter speed and ISO are deliberately variable within sensible thresholds to adapt to light conditions.

Setting	Value
Camera model	Nikon D750
Lens model	Sigma 8 mm f/3.5 EX DG Circular Fisheye
Pixel pitch	5.95 $\mu\text{m}$
Sensor resolution	24.3 MP
Shutter speed	>1/60s
Aperture	5-7
ISO	100-200
Exposure compensation	-0.7
Focus	$\infty$
Image size	Large Fine JPEG 6016x4016 px
Orientation	Landscape

Photos were captured under uniform light conditions as much as possible, either under overcast skies or early in the day before direct sunlight could be seen on the photo.

ImageJ (Fiji, version XXX) was used to binarize hemispherical photos (). We first split each image into red, green and blue channels. We used the Huang algorithm to automatically threshold images to separate sky from plant material, using the blue channel only, under the assumption that plant material reflects little blue light, while the sky reflects much more (). Images were saved as PNG at the original pixel resolution.

## 7.2 Terrestrial LiDAR

Due to terrestrial LiDAR measurement locations being spread over the subplot to avoid occlusion of canopy material, we simulated a scan position at the centre of the subplot using the point cloud data from all scans per subplot. Similar to the processing chain for the foliage density profiles, PDAL was used to crop the point cloud to a 20 m cylinder around the subplot centre, then used `filters.hag_nn` to classify ground points and recalculate height above ground. We cropped the point cloud to points above 1.3 m, with a 50 cm exclusion sphere around the scan position at 1.3 m above the ground. The point cloud was converted to a POV-Ray object, where each point was transformed to a 1 cm<sup>3</sup> cube. POV-Ray was then used to produce a ray-traced image. As with the hemispherical photos, we used a fisheye lens with an equisolid projection and a view angle of 180°, located at the subplot centre, at the same height as the hemispherical photo, with the top of the camera facing magnetic north and the camera facing straight up. Each cube was set as a non-reflective object, and the sky had an equal gamma of 1.0. POV-Ray produced an image of 4016x4016 px, identical to the cropped circular dimensions of the image produced by the hemispherical photo.

## 7.3 Gap fraction estimation

Hemiphot () was used to estimate gap fraction from both the hemispherical photos and the TLS POV-Ray simulation. Hemiphot calculates canopy gap fractions in 90 evenly sized concentric rings. To obtain the total gap fraction of an image:

$$G_{\text{tot}} = \sum_{\alpha=0.5}^{\alpha=89.5} (G_{\alpha} A_{\alpha} / A_{\text{tot}}) \quad (9)$$

Where  $G_{\alpha}$  is the fraction of unfilled pixels in ring  $\alpha$ ,  $A_{\alpha}$  is the sky area of the ring segment, and  $A_{\text{tot}}$  is the total sky area of the hemisphere.

## 8 Grassy biomass estimation

Grassy volume and biomass within each subplot was estimated from four sample points located 2 m from the subplot centre in cardinal directions (Figure 1). At each point, a disc-pasture measurement was taken with a 45.8 cm radius disc weighing exactly 1.5 kg. Small woody stems were removed from disc-pasture sample points before the disc-pasture measurement was taken. Within each 1 ha plot, biomass harvesting was conducted at nine randomly allocated disc-pasture sample points. Tree leaf litter was removed from biomass samples. Grass samples in Angola were dried until the mass remained constant ( $\pm 5$  g) for more than 48 hours, then weighed to ascertain the grassy biomass. Grass samples from Tanzania could not be processed due to curtailment of field-work due to COVID-19 travel restrictions.

An allometric model was developed to estimate grassy biomass at every disc-pasture sample point using the grassy biomass sample masses. This model was only developed for Angola where grassy biomass samples were weighed. The model consisted of a linear mixed effects regression testing

the relationship between disc-pasture height (independent) and grassy biomass (dependent), with a random slope term for each 1 ha plot.

Grassy volume was measured from TLS point cloud data following the methodology of. First the point cloud was cropped to points below 2 m. The point cloud was then aggregated to cubic voxels of 2 cm<sup>3</sup>. Within each vertical 2 cm<sup>2</sup> column, the mean height of points was calculated, then the volume below the mean was assumed to be entirely filled with grassy material.

Figure 2: Point cloud with mean heights for each 2 cm<sup>2</sup> column labelled and the estimated grassy volume below.

## 9 Canopy rugosity

The canopy rugosity of each 1 ha plot was estimated. All scans from each plot were merged to a single point cloud, and noise reduction was performed as described above and the cloud was voxelised to 10 cm<sup>3</sup> cubic voxels. The point cloud was cropped to the plot boundaries, which were located with dGPS similar to the LiDAR targets.

A canopy height model was produced to describe the upper canopy surface. The 99th percentile of height in each 10 cm<sup>2</sup> vertical column was extracted. The maximum height was not used as this occasionally constituted a severe outlier which skewed further canopy height model smoothing. We used the pit-filling algorithm described in **Khosravipour2014** to smooth the canopy height profile by removing gaps within trees caused by incomplete penetration of the LiDAR beam.

From the canopy height profile we extracted a number of statistics for use in statistical modelling. We calculated the mean and coefficient of variation of canopy height across the plot (canopy rugosity), following (**Parker2004**). We calculated the Topographic Ruggedness Index (TRI) as the mean of absolute differences between the heights of each column and the height of its eight surrounding cells (Wilson et al., 2007). From this we estimated the plot level mean TRI and coefficient of variation.

We also calculated a second measure of canopy rugosity ( $R_c$ ) following Hardiman et al. (2011), using all point cloud data rather than just the top surface:

$$R_c = \sigma(\sigma G_z)_x \quad (10)$$

Where  $G_z$  is the vertical height axis  $z$ ,  $x$  is the horizontal axis, and  $\sigma$  is the standard deviation.

## References

Ehbrecht, Martin et al. (2016). “Effective number of layers: A new measure for quantifying three-dimensional stand structure based on sampling with terrestrial LiDAR”. In: *Forest Ecology and Management* 380, pp. 212–223. DOI: 10.1016/j.foreco.2016.09.003.

- Hardiman, Brady S et al. (2011). “The role of canopy structural complexity in wood net primary production of a maturing northern deciduous forest”. In: *Ecology* 92.9, pp. 1818–1827. DOI: 10.1890/10-2192.1.
- Rusu, Radu Bogdan et al. (2008). “Towards 3D Point cloud based object maps for household environments”. In: *Robotics and Autonomous Systems* 56.11, pp. 927–941. DOI: 10.1016/j.robot.2008.08.005.
- Wilson, Margaret F. J. et al. (2007). “Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope”. In: *Marine Geodesy* 30.1-2, pp. 3–35. DOI: 10.1080/01490410701295962.
- Zhang, Keqi et al. (2003). “A progressive morphological filter for removing nonground measurements from airborne LIDAR data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 41.4, pp. 872–882. DOI: 10.1109/tgrs.2003.810682.