Title: Terrestrial LiDAR processing workflow

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 (LAT-LONG), and the second in Kilwa District, southeast Tanzania (LATLOPNG).

The two sites have similar climatic conditions.

TABLE OF CLIMATE

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 used, while in Tanzania, only seven were used, 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).

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 1).

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 ~100 m from the edge of each 1 ha plot. The location of each target was measured for at least 4 minutes. MORE ABOUT DIALLING IN THE GNSS values.

Table 1: Description of scan settings used for each scan.

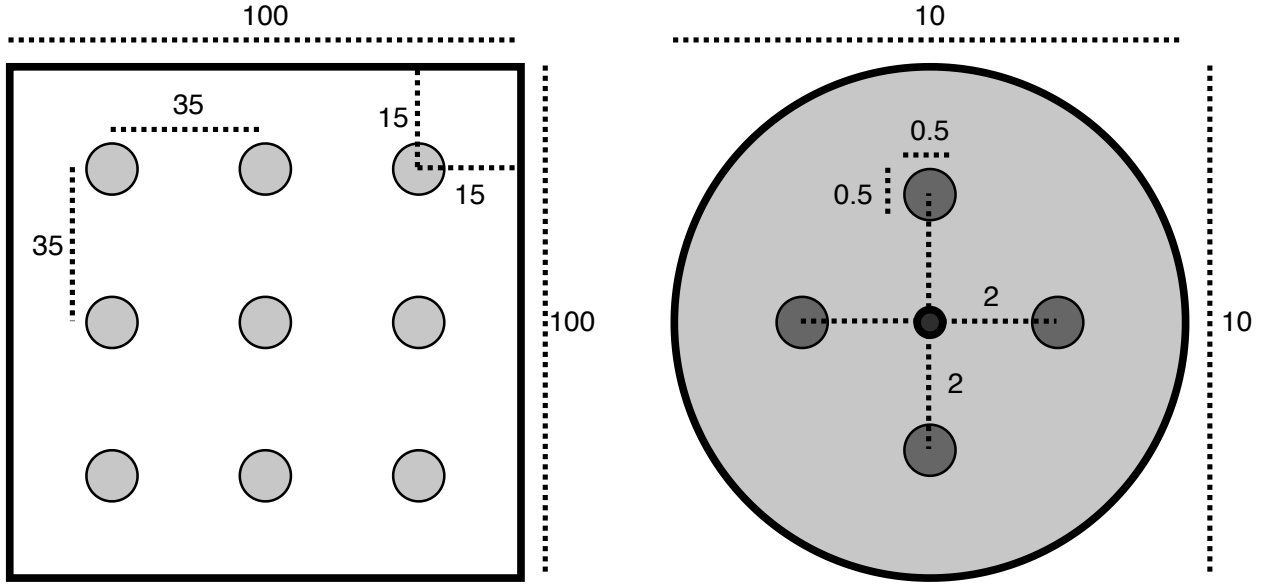| Setting	Value
Scanner model	Leica HDS6100
Wavelength	650-690 nm
Spot size at exit	3 mm
Beam divergence	0.22 mrad
Range	79 m @90%; 50 m @18% albedo
Azimuth range	0-360°
Zenith range	0-155°
Increments	0.018°
Point spacing over 25 m	7.9 mm
Pixels per line	20000
Lines	10000
Compressed file size	˜800 MB
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 XXXX. Targets from each scan were aligned using Cyclone's automatic target acquisition.

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

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

4   Voxelization

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, I used XXX cm^3 cubic voxels, while for subplot height profile estimation I used XXX cm^3 voxels, and for whole plot canopy rugosity I used XXX cm^3 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):

$$\overline{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mu_i \tag{1}$$

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

$$t = \mu + m\sigma \tag{3}$$

$$\text{outlier}_i = \begin{cases} \text{true}, & \text{if } \mu_i >= t \\ \text{false}, & \text{otherwise} \end{cases} \tag{4}$$

where N is the number of points in the scene, $\overline{\mu}$ is the mean distance to nearest neighbour points, and σ 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^3 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:

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

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

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

7 Canopy gap fraction

8 Grassy biomass estimation

9 Canopy rugosity

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.

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.

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.