# LiDAR for Tree Canopy Analysis in MapInfo Pro Advanced

Sam Roberts August 2020

MapInfo Pro Advanced 2019 Patch 3, due to be released in October 2020, introduces a new tool for tree canopy analysis. This tool uses airborne acquired LiDAR survey data to measure tree canopy coverage, tree canopy density and tree canopy height. On other GIS platforms, calculating these canopy parameters will involve multiple steps and intermediate rasters. Not in MapInfo Pro. These new capabilities are fast, simple, accurate and can process huge surveys. In this document I describe what is going on “under the hood” in detail so that users can have confidence in the methodology.

**Input**

We accept LiDAR data as input stored in either LAS or LAZ format, but there is no limit to the number of files you can use as input to a single analysis operation. There is no theoretical limit on the number of points that can be processed, and we have tested the software with billions of input points. However, please have reasonable expectations if you want to process billions of points or more. You may require significant hardware resources to do so efficiently.

The software will attempt to detect the coordinate system of the data from the LAS files, which is expected to be in a projection like UTM. We do not expect the data to be in any kind of geodetic coordinate system and we do not support this. You can enter the coordinate system manually if required. Note that the output raster will be in the same coordinate system as the input points and no reprojection of point data is supported in these operations.

The software will automatically select the “Z” channel (i.e. elevation). However, this is only to keep the software happy at this stage of the processing chain - the tree canopy analysis operations do not depend on this channel selection in any way.

We do not support any windowing or clipping of the input data points and we do not provide any data conditioning of input points. Also, there is no direct access to the LiDAR filtering properties which will be applied to select returns from the LiDAR data stream.



**Output**

The output, in all cases, is a raster in MRR format. You cannot output in any other raster format. If you require the output to be in a different format, then you will have to run a “Convert” operation at completion to convert from MRR to the desired output raster format.

You can set the compression codec and compression level as well as choose to use predictive encoding if desired. Only lossless codecs are offered so that data quality is never compromised. We do not recommend using predictive encoding in these operations. The MRR will be fully finalised with overviews and base level statistics and ready to consume at the end of the processing operation.



**Methodology**

All three operations depend on identifying and counting LiDAR returns that come from the ground (hereafter called GND) and LiDAR returns that come from vegetation (hereafter called VEG). It follows that the LiDAR returns must be classified and that the best results will be obtained when ground and vegetation returns are clearly identified and recorded with appropriate classification codes.

The user must specify the classification codes that represent ground and the classification codes that represent vegetation. You can select one or more codes in either case and, if you are using non-standard classification codes, you can enter those code numbers manually. If the LiDAR data contains extended classification codes, then you can direct the operation to use the extended classification channel rather than the standard classification channel.

The “Class Rules” option gives you a fallback option if your LiDAR data is not appropriately classified. Normally you would choose “Ground and Vegetation” to specify classification codes for GND and VEG. If you choose “Ground” then you only specify the ground classification codes. You might use this option if your LiDAR data has only been partially classified and only ground returns have been identified (which is quite common). In this case, the operations will count ground returns and then consider all returns that are not ground to be vegetation. Clearly, this is not likely to be the case, and the result will have to be interpreted carefully. You might expect that the coverage and density computed in this case is overestimated because there will be more “vegetation” returns than otherwise. Similarly, if you choose “Vegetation” then you only specify vegetation classification codes and all returns that are not identified as vegetation will be assumed to be ground. In this case you might expect the coverage and density to be underestimated.

 

**Filtering**

When we input the LiDAR data, we select returns based on a LiDAR filter. In fact, the only difference between canopy Coverage and Density operations is the filter used to select returns for processing. The rules for this filter are hard-coded.

Firstly, some returns have special designations which are Synthetic, Key Point, Withheld and Overlap. We reject Synthetic returns in Coverage and Density operations and accept them in the Height operation. We accept Key Point and Overlap returns in all operations. We reject Withheld returns in all operations.

Secondly, we can apply filters to accept returns that have either Scan Angle, Intensity or Z values within certain ranges. We do not use any of these filters, but we do not override them. If you execute these operations programmatically then you can enable these filters, but you cannot do so via the MapInfo Pro user interface.

If you supply ground and vegetation classes, then we filter returns by Classification and then by Return. This means we reject all returns that have a classification outside of the supplied set. If you only supply ground or vegetation classes, then we consider all returns regardless of classification and only filter by Return.

Each shot from a LiDAR instrument will generate one or more returns. In other words, the laser may penetrate the first object it encounters and go on to hit another object. Both hits will generate a measurable return pulse and will be recorded in the LAS file. Often, the laser will hit and penetrate vegetation and continue to the ground.

To measure canopy Coverage we select the first return of any shot to identify the highest object in the path of the laser. Often, that will be vegetation. If the first return is classified as ground, it means we are looking at bare earth at that location.

To measure canopy Density we select all returns in the shot. This means we count the number of returns that hit vegetation and the number of returns that make it to the ground. The ratio of vegetation to ground returns is a proxy for the density of the vegetation within the area of interest.

To measure canopy Height we select all returns in the shot. All ground returns are used to build a model of the ground surface but vegetation returns are screened so that the maximum difference between the vegetation height and the ground surface is retained.

In forested areas you may not have any other “hard surface” classifications apart from ground. In urban areas, you might consider objects like buildings and roads as ground as well.

When we read a shot, we firstly match the classification. Then, we make sure the return is not rejected by the special designations or by any of the value filtering rules. When we look for the ‘first’ return we look for the first return that passes the classification and other tests. In other words, it may not be the first return in the shot, but it is the first return that matches our other criteria in the shot.

**Tree Canopy Coverage**

Tree Canopy Coverage measures the percentage of the survey area, as seen from above, over which tree canopy obscures the ground. Each raster cell will be populated with a value from 0 to 100 representing the percentage of canopy coverage over the area of that cell. At the base resolution level, the percentage is rounded to the nearest integer value.

The raster stores 32 bit floating point values so the percentage reported in overview levels may be fractional rather than integer. Overview resolution levels are generated using the “Discrete” band value type which ensures that, even when invalid cells are present with no value, the average cell value stored in overviews is spatially accurate. In effect, invalid cells are assigned a zero value for the purpose of averaging. This means that a statistical analysis of coverage on an overview level will return the same result as the base level cells that were used to populate that overview level.

The coverage percentage is computed as (100 \* VEG) / (GND + VEG) where VEG is the number of vegetation returns and GND is the number of ground returns. Because we have only retained the first returns, this returns an approximation of the vegetation to bare earth ratio.

Coverage can be computed by considering returns that are spatially limited to the boundary of each cell, or the algorithm can consider returns within a radius of the centre of each cell. This is set using the “Integration” option which can either be “Over cell” or “Over radius”. In the second case you must specify a radius in the horizontal units of the coordinate system (usually meters or feet). You can produce a smoother result by turning on the “Smooth Kernel” option. In this case, a Quartic weighting (see Heat Mapping) factor is applied which tapers the contribution of samples further from the cell centre.



If there are no suitable returns in a cell, or within the designated radius of the centre of the cell, then by default the cell value will be recorded as invalid and empty. If you prefer you can turn on the “Make empty zero” option and populate these cells with zero indicating no coverage or no vegetation density or zero vegetation height.

**Tree Canopy Density**

Tree Canopy Density measures the density of the vegetation. Each raster cell will be populated with a value from 0 to 100 representing the percentage of vegetation density over the area of that cell. At the base resolution level, the percentage is rounded to the nearest integer value.

The raster stores 32 bit floating point values so the percentage reported in overview levels may be fractional rather than integer. Overview resolution levels are generated using the “Discrete” band value type which ensures that, even when invalid cells are present with no value, the average cell value stored in overviews is spatially accurate. In effect, invalid cells are assigned a zero value for the purpose of averaging. This means that a statistical analysis of coverage on an overview level will return the same result as the base level cells that were used to populate that overview level.

The measurement of how difficult it is for the LiDAR Laser shots to penetrate through the canopy to the ground is a proxy for the canopy density.

The density percentage is computed as (100 \* VEG) / (GND + VEG). Because we have retained all returns, this ratio is a proxy for the vegetation density.

**Tree Canopy Height**

Tree Canopy Height measures the height of the vegetation above the ground surface. To compute this the algorithm must build a model of the ground surface and then find, in each raster cell, the maximum height of any vegetation returns above the ground surface.

All ground returns are used to build a triangulated ground surface model. Once this has been done, the vegetation returns in each raster cell are examined to find the maximum height of the vegetation above the ground surface, computed from the triangle mesh by linear interpolation at the centre of the cell. The height is clipped so it is always greater than or equal to zero.

Unlike the other operations, height is always computed from returns that are located within the spatial boundary of each raster cell.

**Processing workflow**

Understanding the processing workflow can assist if you are trying to process large amounts of data with limited computing resources.

Firstly, the LiDAR data is imported from the supplied LAS or LAZ files. This import process caches the point data and spatially sorts it. The cache is a kind of raster and, if possible, it is retained in memory. But if the raster tile cache fills up the system will begin to write the point data out to a temporary file. To maximise performance, it is advantageous to have as much RAM as possible to ensure the point cache remains in memory. If this is not possible, it is advantageous to open the temporary cache file on a fast SSD that has plenty of free space. This will generally be the drive that your operating system is stored on. Note that you need to make sure that MapInfo Pro allocates memory to the tile cache. You do this via the Advanced preferences dialog from the backstage.

Secondly, the operation is executed. All operations are multi-threaded and will benefit from a multi-core processor. However, at the present time no operation will execute more than 16 threads so there is little advantage in having more than 16 CPU cores available. As the output raster is generated it will be stored in the tile cache, competing for room with the point cache. If there is no room in the cache, you may see it written to the output MRR progressively. If you want to clip the output raster to a polygon, then you can do this as part of the operation. It is more efficient to perform this clipping as a part of the processing operation than to clip the raster afterwards with a separate Clip operation.

When you run canopy coverage or density you have the option to use an integration radius or integrate over the cell. For highest performance, integrate over the cell and use a cell size large enough to reduce noise to an acceptable level.

When you run the canopy height operation, there is an option called “Patch Size”. This can be set from 1 – 5. On 1 it uses less memory but takes longer to run. On 5 it uses more memory but runs faster. Unless you are running the processing on a Commodore 64, leave it on 5 all the time.



Finally, we calculate statistics for the output raster and generate overviews. Throughout this process the output raster will also get committed to the output file. Both these operations require the entire base resolution level of the raster to be read. It is advantageous if the raster fits in memory. If not, it is better if the output raster is being written to a fast SSD. Performance in this stage largely depends on the compression codec in use. If you are using LZMA and a compression level of 5 or higher, you can expect much slower performance.

**Other considerations**

The integration area might be a radius in the case of Coverage and Density, or it may be the raster cell extent. If the integration area is too small, then you will see larger variations in the calculations. In effect, the output raster will become noisy. If the integration area is too large, then the raster will be smooth and you will lose detail. It is important to choose both the cell size and the integration radius carefully to optimise the processing time and the computed result.

When you turn on “Make empty zero” then you will see a variable amount of padding around the edges of the output raster. An MRR is always a collection of tiles, and these tiles will be getting filled with zero values. To prevent this, you may have to use a clipping polygon.

Whilst the tool is designed for tree canopy analysis, you can turn it towards other tasks by changing the classification selections for ground and vegetation. For example, by selecting buildings instead of vegetation you can use the tool to determine building height rather than tree height.