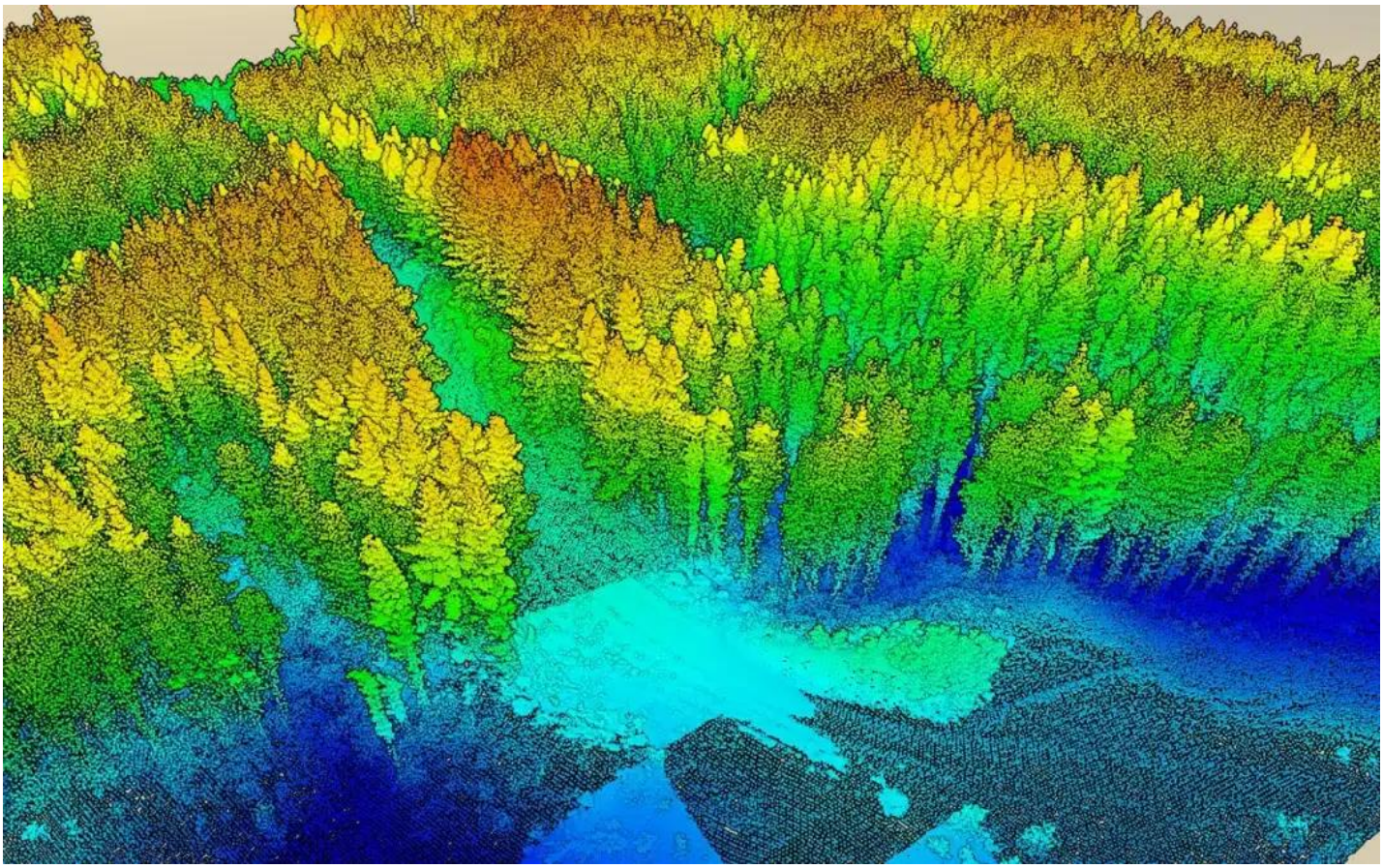
2022



DESCRIBING THE YIELD OF SMALL-SCALE FORESTS IN GISBORNE USING LIDAR

FOREST ENGINEERING DISSERTATION

MILAN CLARKE

ENFO410

Acknowledgements

I would like to thank Dr Vega Xu for her expertise and guidance throughout this project. I would also like to thank Prof. Rien Visser for helping shape the finer details of this project and to the three forestry companies for providing me with stand data used to develop the models of this project.

Abstract

It is estimated that small-scale forests will contribute up to 40% of the nation's wood supply over the next decade therefore it is important that accurate stand characteristics are known. The purpose of this dissertation was to see if plot stand variables for small-scale forests such as mean top height, volume, basal area, and stocking could be related to light detection and ranging (LiDAR) metrics from publicly available LiDAR. The focus area for my research was the Gisborne region as the LiDAR had just recently become available for free and plot data for small-scale forests could easily be obtained from forestry companies. There is little known about the yield of the small-scale forests in Gisborne as owners typically indicate the cost of an inventory crew is not worth the information.

Data from three different forestry companies were used in this research giving a total of 1836 sample plots. A large number of plots meant a wide range of plots from across the region were used to train the models. Multiple linear regression (MLR) models were used to describe the relationship between stand variables and LiDAR metrics as they provided accurate results for a similar study completed in the Wairarapa region of the North Island. LiDAR data was processed through the LAStools programme where 34 different metrics were extracted for model building. The regression models were built using the RStudio programme and the model with the highest R^2 and lowest root mean square error (RMSE) was selected.

The original models including all data were found to be very inaccurate due to the outliers in the data skewing the results. The most accurate models were found when there were no restrictions placed on the number of variables included in the model. This did however highlight a limitation of multiple linear regression (MLR) models which was the risk of collinearity error. All four models had a significant number of collinear variables. Therefore, a limit of seven variables was placed for the final set of models. The models for MTH, basal area, volume and stocking had R^2 values of 0.42, 0.17, 0.21 and 0.07 as well as RMSE values of 7.5%, 17%, 22% and 22% respectively. The MTH, basal area and volume models could be used pragmatically in the Gisborne region and provide ballpark estimations for small-scale forest owners about their yield. However, reliable accuracy in the models was not achieved. The stocking model had no relationship between LiDAR metrics and stems/ha due to the lack of explanatory variables, therefore, the model should not be applied.

Company data was then looked at separately to check accuracy. Company A had the most accurate models with R^2 values of 0.72, 0.63, 0.73 and 0.39 as well as RMSE of 9.7%, 17%, 19% and 22% for MTH, basal area, volume, and stocking respectively. The volume, MTH and basal area models had similar accuracy to the research conducted in the Wairarapa region. Company B had similar RMSE values but much lower R^2 values and Company C had significantly less accurate models than both companies which provided an explanation as to why the combined data models were not as accurate as expected. The accurate models from Company A could be attributed to the smaller sample size which were all measured within two months and most likely by the same crew which provides consistency. A large dataset is not required for accurate models, rather it can be a drawback as there is a greater potential for inconsistency between inventory crews resulting in inaccurate models. By splitting the company data, the point of quality input data will return quality results is evident.

The large dataset provided many challenges when working with the data and created potential sources of error. This research shows the publicly available LiDAR data can be used to create models to estimate stand characteristics given the sample plot data use to train the model is accurate. In future, there should be a large emphasis placed on the quality, not quantity, of sample plot data.

Table of Contents

1. Introduction

The focus of my research will be on small-scale forests in the Gisborne region of New Zealand as small woodlots are now playing a major role in harvest volumes in New Zealand. Small woodlots are responsible for up to 15 million m^3 of harvested logs annually which is slightly over 40% of the nation's total radiata pine supply (MPI, 2016). However, there is limited knowledge and reliability around the inventory of these forests (Xu et al., 2019) as small-scale woodlots require a greater sampling intensity than large production forests, with owners arguing the cost is not worth the information (Goulding & Fritzsch, 2010).

Having accurate forest description data for small-scale forests will be beneficial for many different entities (Bouvier et al., 2015). Forest owners will also have the option of collaborating with nearby forests to satisfy customers who require more logs than can be produced from a single forest. In unison, production forest owners will have an opportunity to supplement log supplies from nearby forests. Harvesting crews will be able to better plan for future work and have greater confidence in yields presented to customers. Mills and wharves will also be able to plan with greater detail on future supply (Manley et al., 2020). Surveying all the individual small-scale forest owners in Gisborne would be impractical and expensive, however, LiDAR provides a unique opportunity for efficient, accurate and low-cost acquisition of forest description data.

Currently, stand characteristics are measured by inventory crews which can be both expensive and inaccurate at times. Inventory crews play a key role in forest management, having accurate forest inventory allows for estate planning, planning for harvesting and sales, keeping the logging crew honest, asset value for accounting and avoiding unexpected surprises (Xu et al., 2018; Bell, 2016). The use of LiDAR to measure forest inventory is now an economically viable option as the availability is widespread and the cost to acquire data has dramatically decreased over recent times (Montaghi et al., 2013). In New Zealand, the Provincial Growth Fund (PGF) LiDAR elevation data capture project has resulted in large portions of the country being airborne laser scanned with the LiDAR point clouds now being made publicly available for free. Using LiDAR in conjunction with ground measurements not only increases the accuracy of the measurements (Holmgren, 2004) but also at least halves the number of ground plots required for the same accuracy as ground-based measurements alone (Melville et al., 2015). The purpose of my dissertation project is to determine whether LiDAR-derived metrics from the PGF LiDAR can be used to accurately determine stand characteristics such as volume, mean top height, stocking, and basal area for small-scale woodlots.

2. Review of Literature

The use of LiDAR to collect data on forest canopy heights has been around since the 1990s (Næsset, 2004) and use for other measurements such as volume and biomass has been around since the early 2000s (Montaghi et al., 2013). LiDAR has been proven in many studies to provide the same, if not better, level of accuracy as ground plots when determining stand characteristics (van Leeuwen & Nieuwenhuis, 2010, Goerndt et al., 2011, Bouvier et al., 2015). LiDAR is now more accessible than ever thanks to programmes such as the PGF LiDAR elevation data capture project which covers the Gisborne region and its small-scale forests. Small-scale forestry has previously been unable to keep pace with LiDAR's technological advancements (van Leeuwen & Nieuwenhuis, 2010). However, with free access to LiDAR data and collaboration with forest management companies in the Gisborne region, LiDAR models to determine stand characteristics are now a possibility.

2.1 Acquiring Data

Ground-based measurements are the first step in the process of developing a model, sample plots are clipped to the point cloud and used to train the LiDAR model. Accurate ground measurements are essential for a reliable LiDAR model to be developed (Bouvier et al., 2015). Ensuring the centre of the plot is correctly marked out is key, if the centre of the plot when measuring is not the same as the recorded one then this will cause an overlapping error, as shown below in Figure 1, between ground measurements and the LiDAR point cloud (White et al., 2013). Circular plots with a radius of 8 – 14 metres are recommended to minimise edge effects and co-registration error while maximising sampling efficiency and precision and accuracy of target and explanatory variables (Frazer et al., 2011). The selected radius for a plot is determined by the size of the largest tree and desired degree of accuracy in measurements (White et al. 2013). The more plots available for training the LiDAR model the more accurate it will be, however, if the plot measurements themselves are inaccurate then the model will also be inaccurate (Næsset, 2004a).

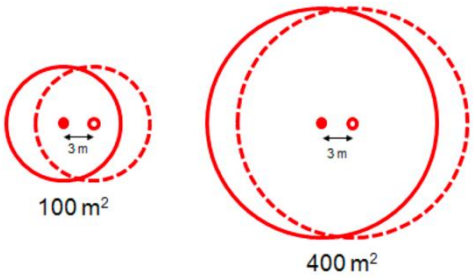


Figure 1. An overlapping error of 3 metres. The overlap for the 100 m2 plot is 63% whereas the overlap for the 400 m2 plot is 82% (White et al., 2013).

LiDAR can be obtained through various methods, the most common form used in forestry is Aerial Laser Scanning (ALS) which are LiDAR scanners attached to fixed-wing aircraft or helicopter platforms (White et al., 2013). ALS has dense sample patterns and a small footprint allowing for detailed recordings of the forest surface at a high area coverage rate (Sabol et al., 2016). Unmanned aerial vehicles (UAV) are a rapidly developing technology with flight range and payload capacity increasing making them a viable option for collecting

LiDAR data. UAVs are a cheaper alternative to sourcing LiDAR when compared to ALS and provide flexible data acquisition as data can be collected more often (White et al., 2013). However, the Civil Aviation Authority (CAA) has regulations in place that state that UAVs can only be flown when in line of sight of the operator and during daylight hours despite the availability of onboard cameras for flying a UAV (Civil Aviation Authority, 2022). Professional UAV operators can apply for exemption from these regulations and with UAVs rapidly developing it is expected the regulations will be relaxed which will allow the adoption of UAVs into forest management (Dash et al., 2016).

The area of small-scale forests in New Zealand varies with different sources. . In a study by Manley et al. 2020 where small-scale forests were mapped for several different regions in the North Island, it was found existing mapping from the National Exotic Forest Description (NEFD), Land Cover Database (LCDB) and Land Use and Carbon Analysis System (LUCAS) overestimated the net forest area by 8%, 17% and 27% respectively. LCDB and LUCAS were found to overestimate forested area due to basing on gross area rather than net area as well as misclassification of some forests. Identifying and mapping a forest stand was based on aerial imagery, NEFD, LCDB and LUCAS. Up to date, aerial imagery is key for accurate boundary mapping.

2.2 LiDAR Models

Collected data needs to be processed to obtain a model that relates the LiDAR metrics to the stand metrics of interest. The first decision to be made when determining what model will be used is the approach to estimation (Melville et al., 2015). The two approach types are area-based and tree-based. The area-based is a predictive model that is derived and links the LiDAR metrics to the desired stand characteristic from the measured sample plots (White et al., 2013). The purpose of the area-based approach is to create a model that can predict stand characteristics where measurements have not been taken. The area-based approach allows for the identification of within-stand variability which may not be well represented by typical ground measurement and with the whole stand being able to be measured this allows for flexibility in inventory reporting (White et al., 2013). The tree-based approach is where individual trees are identified and their corresponding characteristics are predicted or measured from the LiDAR data (Hyyppä et al., 2012). The tree-based approach can provide additional information at a tree level for forest stands however a dense point cloud and simple tree and canopy structure are required for accuracy, therefore, a tree-based approach can quickly become problematic (Kaartinen et al. 2012). The area-based approach has been used in New Zealand before by the Ministry of the Environment for recording the national carbon inventory and it was found to have increased the precision of these recordings (Dash et al., 2016). The area-based approach is considered fully operational for forestry management applications and is therefore recommended over a tree-based approach (Næsset, 2011).

Various models can be used in conjunction with the area-based approach, with different models best being able to express different stand characteristics. For area-based modelling three types of software are required to form a model (Sabol et al., 2016):

- Software to process and manipulate LiDAR data and generate LiDAR metrics. (LAStools)
- Software to develop models to be applied to determine stand characteristics. (RStudio)
- GIS software to project models over areas of interest, manipulate model output raster and incorporate existing stand information. (ArcGIS)

The software programmes in brackets are available on university computers. Processing of the point cloud involves clipping the point cloud to the plots that will be used to train the sample from polygon shapefiles, normalising the heights for each stand in relation to each stands ground point, and extracting metrics for each stand (White et al., 2013). When extracting the metrics for each stand it is common practice to have a threshold for eliminating LiDAR measurements from the ground and low vegetation (Holmgren, 2004). The threshold varies depending on the forest and can range from 1 to 3 metres (Bouvier, 2015; Montaghi, 2013; Goerndt 2011).

Selecting an appropriate model is influenced by how the data is acquired, the quality of the data and tree species. A study in New Zealand by Watt et al. (2015) compared k-NN and multiple linear regression (MLR) models for predicting the site index of plantation radiata pine forests where it was found that MLR produced more accurate results and smaller RMSE in most trials. In contrast, a study by Fehrmann et al. (2008) found that k-NN models were slightly better than MLR models for predicting spruce and pine single tree biomass. A study from Aertsen et al. (2010) trialled 5 model types (parametric multiple linear regression (MLR); four non-parametric models including classification and regression tree (CART), generalised additive model (GAM), artificial neural networks (ANN) and boosted regression tree (BRT)) for estimating site index where it was found all non-parametric models except CART outperformed MLR for mixed pine and cedar Mediterranean mountain forests. No single model is best suited for predicting all stand characteristics (Xu et al. 2019). A benefit of using MLR models is that it is easy for users to understand as there is a clear relationship between stand characteristics and LiDAR metrics whereas for non-parametric modelling the model can be seen as a 'black box' to the user (White et al., 2013). MLR is recommended for estimating small-scale forest stand variables when comprehensive field data are lacking (Xu et al., 2018).

2.3 Accuracy Achieved in LiDAR-Based Inventory

The standard for current inventory crews in New Zealand is $\pm$ 10% of the true harvest volume (Bell, 2016), therefore, this can be used as a benchmark to determine if previous research on this topic was accurate. Accuracy for a LiDAR model can be determined from two measurements, root-mean-square-error (RMSE) and coefficient of determination (R^2) (Sabol et al., 2016).

In a study by Xu et al. 2018 it was found that, when using an MLR model for small-scale radiate pine forests, the R^2 ranged from 0.73 for the basal area (BA) to 0.97 for mean top height (MTH). RMSE for BA was 9.47 m^2/ha, 84.20 m^3/ha for volume, 1.31 m for MTH and 2.11 years for age. In a study by Holmgren et al. 2004, MLR was used to estimate tree heights for Norway spruce and Scots pine which produced an accuracy of 0.92 for R^2 and

0.59 m for RMSE which was an error of 3% of the mean tree height. The basal area was also modelled with an R^2 value of 0.88 and an RMSE value of 2.7 m^2/ha which was an error of 10% of the average basal area. In a study from Næsset, 2004 for modelling Scots pine and Birch it was found to have an R^2 of 0.83 for BA and 0.90 for volume. RMSE for BA was 2.51m^2/ha, which was 8.4% of the average BA, and 16.1 m^3/ha for volume, which was 5.6% of the average volume.

The accuracy of the LiDAR metrics recorded can be influenced by topography, silviculture history, soil type and tree species (Melville et al., 2015). A study by Saremi et al. (2014) found that for radiata pine stands topography and the aspect of stands affect the metrics recorded for height. In a study from Takahashi et al. (2005) on the accuracy of modelling mature Sugi trees, it was found that tree heights were accurately estimated 74% of the time for steep slopes, 86% for gentle slopes and 92% for relatively flat slopes and slopes with no gradient. The model was found to underestimate tree heights on gentle slopes due to a higher percentage of the first point returns therefore increasing the sampling of the treetops. In contrast, the model was found to overestimate tree heights on steep slopes due to trees leaning towards the slope, therefore, distorting readings. In a study by Bouvier et al. (2015), it was found that models for coniferous trees are more accurate than models for deciduous trees. R^2 for predicting the basal area of coniferous trees was 0.67 whereas for deciduous trees it was found to be 0.52. A study from Montaghi et al (2013) found similar results when looking into individual tree detection with errors being greater for mixed-species forests compared to coniferous forests.

When combining ground-based measurements with LiDAR, the number of sample plots required halves for the same precision as field measurements alone (Melville et al., 2015). Melville et al. (2015) also found the relative efficiencies for inventory measurements at least double when using LiDAR models to aid estimations in comparison to field measurements alone. A study from Holmgren (2004) found that low-density point clouds can be used to produce models with similar accuracy to traditional field-based measurements.

3. Objectives

The objective of my dissertation will be to determine if publicly available LiDAR data can be used to develop accurate MLR models to predict stand characteristics. The stand characteristics that models will be developed for are MTH, basal area, volume, and stocking. The desired accuracy of the models is $\pm$10% of the plotted value as this is what is expected of inventory crews and the purpose of these models is to complete the same job. Therefore, the same expected accuracy level should be set. It should be noted that the accuracy of these models is based on the inventory crew data which may not be 100% accurate.

The number of LiDAR metrics (variables) used to determine volume, mean top height and stocking will be decided at a later stage throughout the research process.

4. Method

The Gisborne region covers 838,580 ha of land, of which approximately 67,050 ha is classified as small-scale forests. Small-scale forests are defined as forests greater than 1 hectare in size and less than 1000 hectares and the forest is not owned by a large corporate forestry company.

4.1 Software

The software programmes used for this project are ArcMap, LAStools and RStudio. ArcMap 10.8.1 is used for mapping, linking spatial and stand information, and model building. LAStools is a suite of command line tools used to process LiDAR point clouds and extract relevant data. RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. In this research, RStudio will be used to undertake multiple linear regression model building.

4.2 Mapping

Before LiDAR processing can begin all the mapping of small-scale forests in Gisborne has to be updated. Mapping has previously been done in 2017, however, more recent aerial images have been made available. Forests have that not been replanted since 2017 will be assumed to no longer be land in forestry and will be removed from the shapefile database. When mapping small-scale forests, the following guidelines from Manley et al, 2020 were applied:

- The area had to be over 1 ha and greater than 30 m wide, but the 1 ha rule was relaxed when there were contiguous small blocks that added to over 1 ha
- Gaps over 0.1 ha were excluded from the forest area polygons
- All mapping on ArcGIS was done at a maximum scale of 1:4,000.

This allows for consistent and accurate results and exclusions of smaller areas that could potentially affect accuracy. Young forest outline accuracy can be checked by importing the shapefile into google earth and comparing it to the previous year's aerial photographs from when the stand was a mature forest.

4.3 Plot data

Plot data was made available from three forestry companies in the Gisborne region, meaning 1836 plots were available to use in the research. The plot data was collected using consumer-grade GPS with dates of recordings ranging from 2017 through to 2021. The relevant plot data includes tree height, DBH, slope, coordinates of the centre, and area of the plot. The locations of sample plots are shown below in Figure 2.

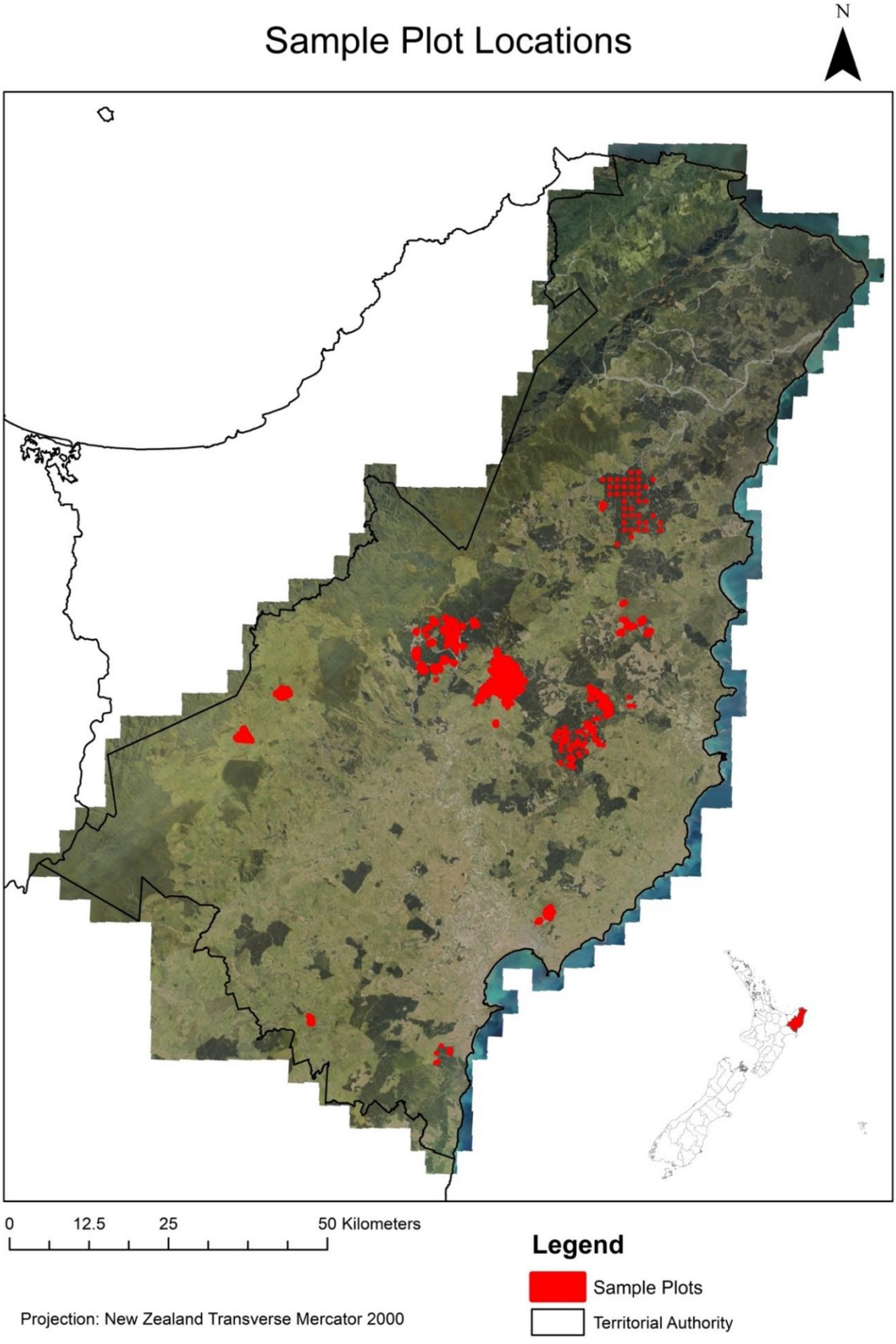


Figure 2. Location of sample plots.

If plot data contained missing heights of trees the Petterson model will be used to determine the unknown heights (Petterson, 1955):

$$Height \ = \ 1.4 \ + \ (a + b \ / \ D)^{2.5}$$

Where a and b are coefficients estimated from the known heights and diameters of the trees in the plot and D is the diameter of the tree. From the plot data the MTH, basal area, volume and stocking for the stand can then be calculated using the following formula:

$$MTH \ = Average\ height\ of\ tallest\ [100 \ \times Plot\ size\ (ha)]\ trees$$

$$Basal\ Area \ = \frac{\Sigma Basal\ Area}{Plot\ size\ (ha)}$$

$$Volume = Basal\ Area \ \times (0.9 + 0.3 \ \times MTH)$$

$$Stocking = \frac{Number\ of\ Trees}{Plot\ size\ (ha)}$$

The 2D plot area for shapefiles in ArcGIS can then be calculated using the equation:

$$Plot\ area = \sqrt{\frac{10000 \ \times \ Plot\ size\ (ha)}{\pi \cos\left(\frac{Slope\ (degrees)}{180}\right)}}$$

Getting the 2D area for the plot minimises the risk of an overlapping error occurring. A 2d plot area is required as a sample plot on a slope will have a different area when looking directly down on it in the 2D ArcGIS map. The received data must then be formatted so all companies' information is presented the same way and it can be appended to the shapefiles of the corresponding plot in ArcGIS by indexing each stand and joining the files.

4.4 LiDAR data

The LiDAR point cloud was formed from Aerial Surveys between 2018 and 2020. The point cloud has a density of 10.07 points/m^2 and uses the New Zealand Transverse Mercator coordinate system. There are 5 classifications used for the points: unclassified, ground, low point, water, and high noise.

The LiDAR point cloud will then be processed through LAStools using the code shown in the appendix. The LiDAR point cloud comes in tiles from the LINZ database with a size of 7000 metres by 3000 metres. Firstly, the quality of the LiDAR data is to be checked through the lasinfo tool, this allows for information on the number of points in each classification, coordinate system used, and any issues with the data.

Processing of the data can then begin through the tool lasnoise. Lasnoise reclassifies cells of size 4 m by 4 m by 2 m that contain 3 or fewer points into a low noise classification. This

allows for more effective processing and precision when further using the LiDAR data as the noise contamination has been removed.

Next, the lastile tool is used to create new tiles of size 2000 by 2000 metres with a buffer of 20 metres around the edge of the square. Reshaping the tiles and adding buffers eliminates the risk of edge effects along tile boundaries. This was a risk with the original tiles as some may have small gaps in between them which can cause issues if a forest polygon falls in that region. The buffers are flagged as withheld so they can be removed easily if not required in other processing stages.

The lasheight tool is then used to normalise the height of the LiDAR. The points classified as ground points in the point cloud are given an elevation of zero to represent the ground and then all other point heights are recalculated in relation to the zeroed ground points. This tool helps when extracting the height intensity metrics.

The tiles are then indexed using the lasindex tool to create an LAX file containing spatial indexing information on the tile. The LAX file is used to speed up access to relevant areas of the LAZ file whenever there is a spatial query.

Plot metrics can then be extracted using the lascanopy tool. Firstly, all tiles are merged followed by the low and high noise classification points being dropped. The merged tile file can then be clipped to the boundaries of the test plots to get the point clouds of each plot. The metrics are then extracted for each plot, the canopy cover cut-off height is set at 3 metres and the height cut-off is set at 2 metres. The metrics extracted are shown in Table 2 below.

Table 1. Metrics extracted from the lascanopy tool.

Variable Code	**Description**
cov	Canopy cover
dns	Canopy density
max	Max height
avg	Average height
qav	Average square height
std	Standard deviation of heights
ske	Skewness of heights
kur	Kurtosis of heights
p(50, 75, 90, 95, 99)	Height percentiles
int_(max, avg, qav, std, ske, kur, p)	Intensities of metrics
b(30, 50, 80, 90)	Bicentiles
c(00, 01, 02)	Height count raster
d(00, 01, 02)	Height density raster

4.5 Model Building

The most suitable approach for model building in the Gisborne region is the area-based approach as this allows for predicting characteristics for areas where ground measurements

have not been recorded and have been previously used successfully for a similar project. An area-based approach is also suited as the LiDAR is around 10 pts per m2, which isn't dense enough for delineating individual trees, and the co-registration error needs to be minimal for a tree-based approach. An MLR model will also be best suited for the small-scale forests in Gisborne as it is an easily understood model that has a small accuracy advantage over other models for radiata pine stands in New Zealand.

The metrics for each stand are presented in an excel spreadsheet where they are then lined up with their corresponding stand metrics. The spreadsheet is then run through RStudio using the code shown in the appendix.

Firstly, the 'regsubsets' tool is used to determine the best fitting variables given a maximum number of allowable variables. The regsubsets tool selects the combination of variables that give the smallest residual sum of squares (RSS) value for the given number of variables. To begin with, all variables will be considered. The regsubsets tool then displays the best fitting variables for models with 1 through to 34 variables.

The models are then run through a model selection criterion, using the summary tool, where the best number of variables is given for a required statistical measurement. The statistical measurements are adjusted R^2, Capability Potential (CP), and Bayesian Information Criterion (BIC). R^2 and Residual Sum of Squares (RSS) are not included as they will always favour the model with the highest number of variables.

The average cross-validation error is then computed as the model prediction error. CV errors can be predicted for all 1 through to 34 variable models with the which.min(cv.errors) tool then applied to show what number variable model has the smallest CV error.

Knowing the best number of variables for the model is given criteria for adjusted R2, CP, BIC, and CV the most suited number of variables and what variables they are can be determined. The coefficients for each variable in the model can then be found. In a separate excel spreadsheet, the stand characteristic is lined up with the best-suited variables for the next stage of processing. If there is uncertainty around the best number of variables the next stage can be repeated for models with a different number of variables and the model with the lowest R^2 and RMSE is the best-suited model.

The validation set approach is used to determine the coefficients for the model. The data is firstly split into 70% training and 30% validation to help evaluate model performance. The model is then built from the training data and the R^2 and RMSE errors are given. A summary of the model is then printed out to give the coefficients of the model.

Once these steps have all been completed four equations expressing stand volume, MTH, stocking and basal area will be completed. If the accuracy of the respective models is not within $\pm$10% of the true value restrictions on the model variables will be placed followed by splitting of data into age class bins and by companies to see how accuracy is affected.

5. Results

5.1 Summary of Plot Data

Table 2 below shows a summary of the characteristics of the sample plots. The 5th and 95th percentile values are shown instead of the range as there were some extreme values in the original data. A total of 1836 sample plots were included when training the model. Majority of the sample plots were recorded for pre-harvest inventory.

Table 2. Summary of plot characteristics.

Stand Variable	**Mean**	**5th Percentile**	**95th Percentile**
Plot size (ha)	0.05	0.03	0.06
Slope (°)	22	7	37
Age (years)	27	25	29
MTH (m)	36.6	30.9	42.1
Basal Area (m^2/ha)	69.6	39.8	91.0
Volume (m^3/ha)	843	447	1132
Stocking (stems/ha)	392	240	600

5.2 Summary of LiDAR Data

Table 3 below shows a summary of the data for the LiDAR metric outputs from the lascanopy tool. Not all metrics have been included in the summary table below as the general trend can be observed from one metric that belongs to a group e.g., height percentiles (p50, p75, p90, p95, p99).

Table 3. Summary of LiDAR metric outputs.

Variable Code	**Mean**	**5th Percentile**	**95th Percentile**
cov	91	70	99
dns	80	64	91
max	40	31	46
avg	26	18	32
qav	730	392	1053
std	6	3	10
ske	-1	-2	0
kur	5	2	9
p50	27	19	33
int_max	3741	2160	6422
b30	8	1	25
c00	182	0	654
d00	2	0	9

5.3 Original Models

To begin with, all 1836 sample plots were included when building the models and no restrictions were placed on the variables included in the model. The accuracy of these models was then reviewed for the next step of improving the accuracy of the model. Many sets of models were run for each stand variable and the model with the highest R^2 and lowest RMSE was selected. Variables are considered important if they have a significance level of 0.1

Table 4 below shows the accuracy of the models and the number of significant variables out of the total. MTH had the most accurate model with the RMSE value of 7.3% being within the desired 10% accuracy. Basal Area, volume and stocking models had poor R^2 values as well as high RMSE values making the models unusable.

Table 4. Original model results.

Characteristic	R^2	RMSE	RMSE (%)	Significant Variables
MTH	0.34	2.68 m	7.3	12/16
Basal Area	0.18	49.7 m^2/ha	70	12/14
Volume	0.15	743 m^3/ha	91	11/25
Stocking	0.07	106 stems/ha	27	7/7

The low R^2 for the models shows a poor ability to predict results over the whole range of input data. In the MTH model, the input data ranged from 20.98m to 113.38m whereas in the predicted results the data ranged from 28.45m to 44.06m. The MTH, Basal area and volume models all contained insignificant variables with over half of the variables in the volume model being insignificant. All the variables included in the stocking model were considered significant. This is due to the lower number of variables included in the model.

As all the original data was included in these models there were outliers for all stand characteristics. For example, in the volume data, one stand had a volume of 24,238 m^3/ha. This is impossible in real life and can therefore be removed from the data used to train the model. Values like this may come because of an error when recording the data, some plots had an area of 0.01 hectares and 20 + trees recorded. Outliers can easily skew results and create inaccuracies as shown by the original results. For the next step of developing the models, the outliers were removed from the data used to train the models to see if the accuracy would improve.

5.2 Removal of Outliers

There were still no restrictions placed on the variables used in the equations when these models were built. Table 5 below shows the upper limits to the data included in the models. No lower limits were set as all data across the stand characteristics were considered to have no values too low to be included.

Table 5. Upper limits for removing outliers.

Characteristic	Maximum
MTH	45.0 m
Basal Area	110 m^2/ha
Volume	1500 m^3/ha
Stocking	650 stems/ha

A maximum of 45m was set for MTH as the oldest sample plot was 37 years old and it is unlikely Radiata Pine would be taller. A maximum of 110 m^3/ha was set for basal area and 1500 m^3/ha was set for volume as it is very unlikely for a production forest to be much greater. A maximum of 650 stems/ha was set for stocking as the youngest sample plot was 20 years old so it was likely that it had been through at least one stage of thinning.

Table 6 below shows the accuracy of the models and the number of significant variables out of the total. MTH again had the most accurate model with an RMSE of 7.54% which is within the desired 10% accuracy.

Table 6. Removed outlier model results.

Characteristic	R^2	RMSE	RMSE (%)	Significant Variables
MTH	0.45	2.76 m	7.5	16/19
Basal Area	0.20	12.7 m^2/ha	18	16/22
Volume	0.24	171 m^3/ha	20	12/18
Stocking	0.09	92 stems/ha	23	15/19

For MTH, R^2 increased by 0.1039 and the RMSE increased by 0.4m. The number of variables included in the model increased by 3 however the number of insignificant variables decreased by 1. For Basal area R^2 increased by 0.0181 and the RMSE significantly decreased by 36.9m^2/ha. The number of variables included in the model increased by 8 and the number of insignificant variables remained the same. For volume, R^2 increased by 0.0884 and the RMSE significantly decreased by 572.2m^2/ha. The number of variables in the model decreased by 7 and the number of insignificant variables decreased by 8. For stocking, R^2 increased by 0.0216 and the RMSE decreased by 14 stems/ha. The number of variables in the model increased by 12 and the number of insignificant variables increased by 4.

The accuracy for the basal area, volume and stocking models were still not within the desired accuracy of 10%. A poor R^2 value for all models again reflects their inability to predict values over the full range of input data. The number of insignificant variables was still relatively high for all models. Therefore, for the next step of developing the model the insignificant variables were removed from the models to see how that would increase accuracy.

5.3 Removal of Insignificant Variables

Table 7 below shows the accuracy of the models and the number of significant variables out of the total. MTH again was the most accurate model and still the only model within the desired accuracy of 10%

Table 7. Removed insignificant variable model results.

Characteristic	R^2	RMSE	RMSE (%)	Significant Variables
MTH	0.44	2.75 m	6.7	14/16
Basal Area	0.17	12.9 m^2/ha	20	12/16
Volume	0.22	175 m^3/ha	23	10/12
Stocking	0.07	94 stems/ha	24	12/15

For MTH the R^2 decreased by 0.0028, the RMSE decreased by 0.01m and the model contained 2 insignificant variables. For basal area the R^2 decreased by 0.0279 the RMSE increased by 0.2 m^2/ha and the model contained 4 insignificant variables. For volume the R^2 decreased by 0.0140 the RMSE increased by 4 m^3/ha and the model contained 3 insignificant variables. For stocking the R^2 decreased by 0.0138, the RMSE decreased by 0.01m and the model contained 3 insignificant variables.

All the models had a decrease in accuracy because of their insignificant variables being removed. Interestingly, all the models still contained insignificant variables. Further models were run where the insignificant variables from these models were removed, and the same trend of decreasing accuracy and more insignificant variables occurred.

The high number of variables included in the models developed leaves them at risk to collinearity error. The number of collinear variables for each characteristic from the models above is shown below in Table 8.

Table 8. Collinear variables in the models.

Characteristic	Number of Collinear Variables
MTH	8
Basal Area	11
Volume	10
Stocking	10

Table 8 shows the model's vulnerability to collinearity error with the MTH model having the lowest number of collinear variables at 8. To minimise the risk of collinearity error a maximum variable limit was placed on the models for the next step of development. By running test models it was found when the models contain seven explanatory variables the increase in accuracy from fewer variable models was satisfactory and the risk of collinearity error was low. Therefore, the models with seven variables were used as the final models for estimating stand variables.

5.4 Final models

The final models produced with all the combined companies' data were these models. To further investigate the accuracy of the models the data needs to be split up. This section covers each model separately.

5.4.1 Mean Top Height

The R^2 for the model was 0.42 and the RMSE was 2.76m which was 7.5% of the average. All variables in the model had a significance level of 0. Figure 3 below shows the plotted MTH from the sample plots against the predicted MTH from the model. A relatively accurate model was expected due to MTH's direct relationship with some LiDAR metrics.

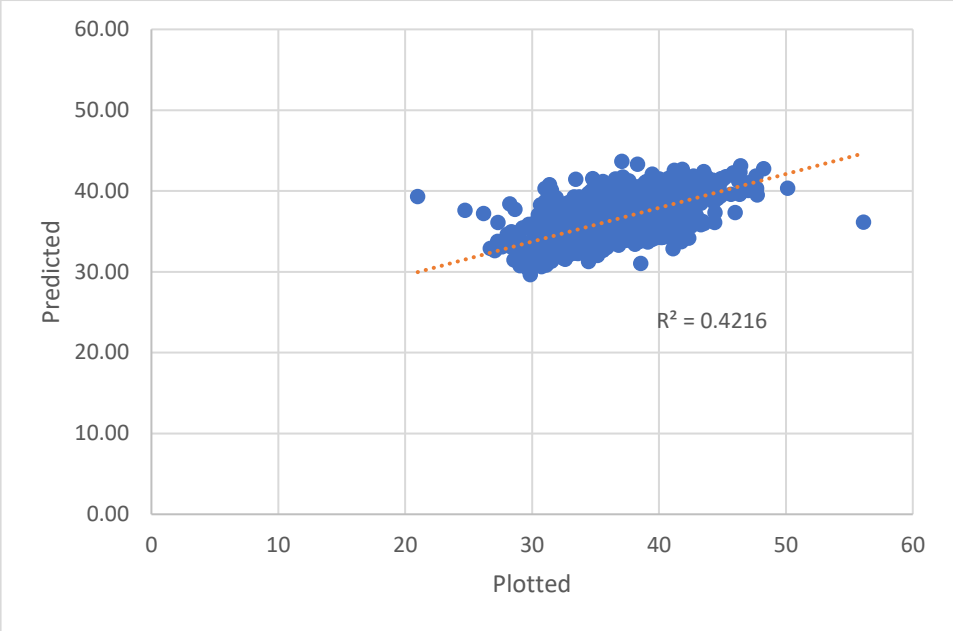


Figure 3.Final MLR model for MTH.

The equation for the MTH model was:

$$MTH = 34.17 + 0.02qav - 0.24kur - 0.54p50 + 0.06b80 + 0.18d00 - 0.01int\_\max + 0.09int\_kur$$

5.4.2 Basal Area

The R^2 for the model was 0.17 and the RMSE was 12.6 m^2/ha which was 17% of the average. Six variables in the model had a significance level of 0 and one variable having a significance level of 0.001. Figure 4 below shows the plotted basal area from the sample plots against the predicted basal area from the model. The poor R^2 reflects the model's inability to predict the full range of results from the plotted data. A plot which has a basal area of 40 m^2/ha is predicted to have 60 m^2/ha and a plot with 90 m^2/ha is predicted to have 70 m^2/ha. This is not a desirable outcome for the model and is therefore unusable.

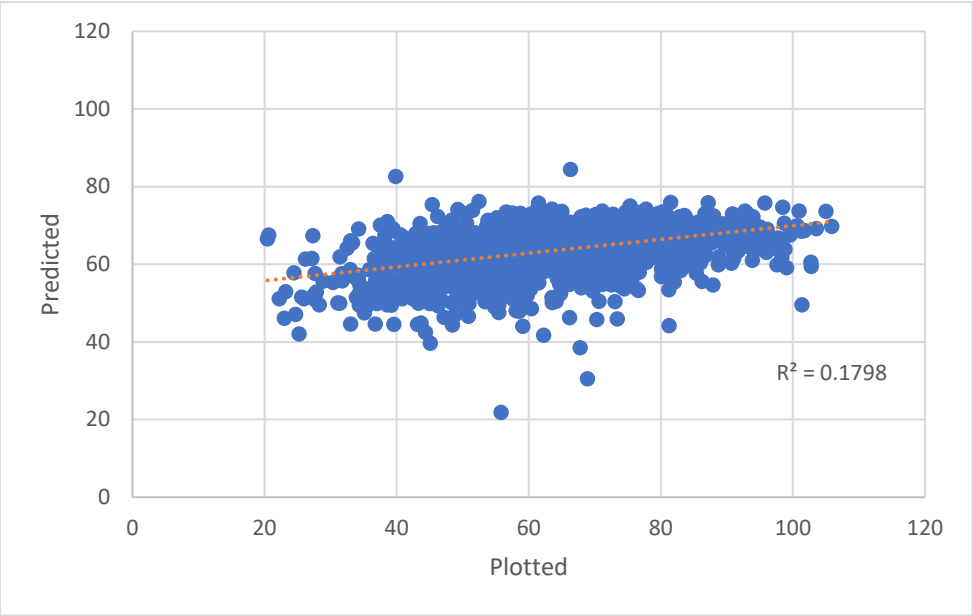


Figure 4. Final MLR model for basal area.

The equation for basal area was:

$$BA = -3.14 + 0.02qav - 1.04std + 0.65b30 + 0.54d02 - 0.01int\_\max - 0.04\, int\_p50 - 0.01int\_p90$$

5.4.3 Volume

The R^2 for the model was 0.21 and the RMSE was 172 m^3/ha which was 22% of the average. All variables in the model had a significance level of 0. Figure 5 below shows the plotted volume from the sample plots against the predicted volume from the model. Again, the poor R^2 reflects the model's inability to predict the full range of results from the plotted data. A plot which has a basal area of 400 m^3/ha is predicted to have 700 m^3/ha and a plot with 1200 m^3/ha is predicted to have 850 m^3/ha. This is not a desirable outcome for the model and is therefore unusable.

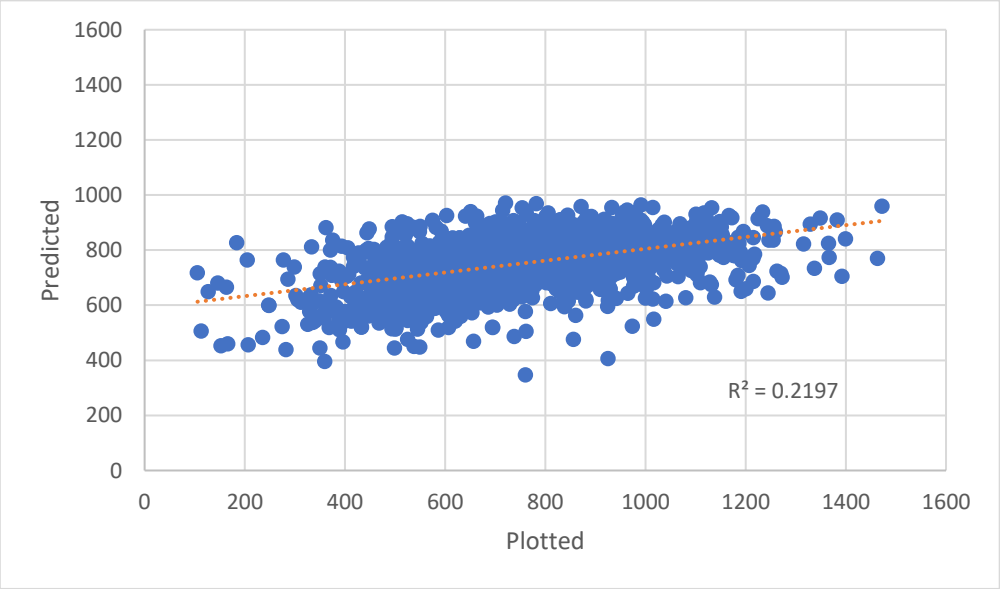


Figure 5. Final MLR model for volume.

The equation for volume was:

$$V = 513.54 + 0.89qav - 34.16p50 + 5.91d00 - 0.03int\_max + 0.53int\_p50 - 0.25int\_p75 + 6.21dns$$

5.4.4 Stocking

The R^2 for the model was 0.0701 and the RMSE was 91 stems/ha which was 22% of the average. Five of the variables had a significance level of 0 with the other two variables having a significance level of 0.001. Figure 6 below shows the plotted stocking from the sample plots against the predicted stocking from the model. Again, the poor R^2 reflects the model's inability to predict the full range of results from the plotted data. A plot with 200 stems/ha is predicted to have 350 stems/ha and a plot with 600 stems/ha is predicted to have 400 stems/ha. This shows there is no relationship between stocking and LiDAR metrics.

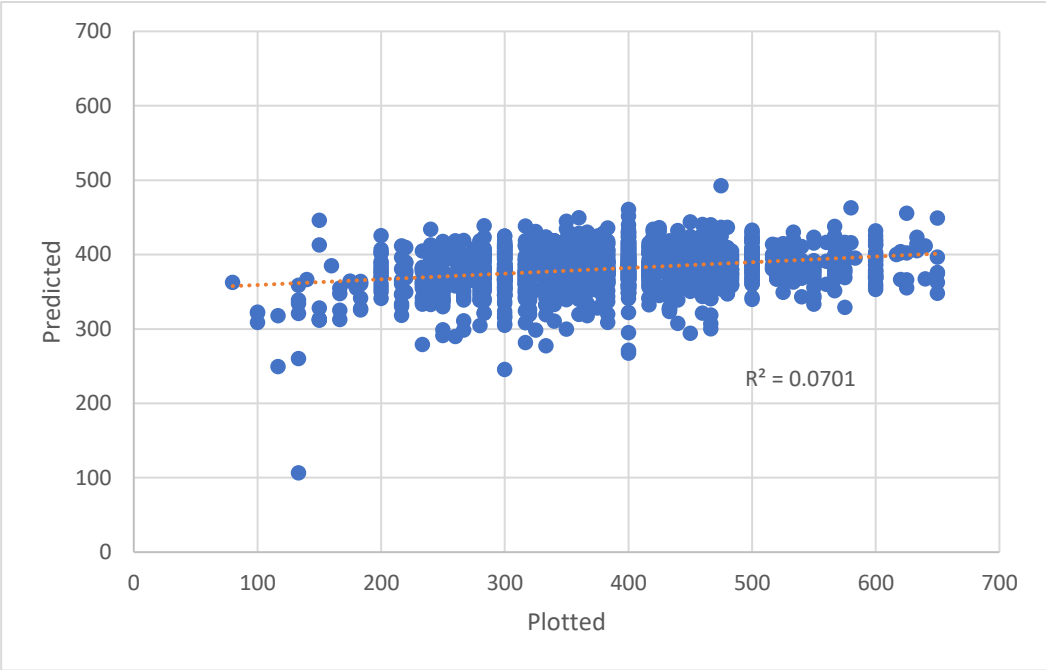


Figure 6. Final MLR model for stocking.

The equation for stocking was:

$$S = 262.12 + 6.52kur - 14.42p75 + 28.65p90 - 15.52p99 - 2.77d02 - 0.08int\_std + 4.94dns$$

5.3 Splitting of Data

The desired level of accuracy had been achieved through the final MTH model where the RMSE error reached 7.5% of the average which was within the desired 10% accuracy threshold. However, as the R^2 value for the model was 0.42, this resulted in a poor spread of predicted MTH's. The range of plotted MTH's was 20.98m to 36.14m and the final model had a range from 29.85m to 44.66m which means the model is unsuitable as it fails to predict MTH's for the whole range of input data. The volume, basal area and stocking

models did not meet the desired accuracy threshold of 10% therefore all the models are considered unsuitable.

Originally, the plan was to further break down the data by age if the models were unsuitable to see if that would increase accuracy. However, looking at the age class distribution below in Table 9, 89% of the sample plots were planted within the space of three years (1996, 1997, 1998). This meant breaking the plots down by age class was no longer suitable therefore it was decided that the plots would be broken by the three companies that provided the plot data.

Table 9. Year of sample plot planting.

Year Planted	Plots
1985	4
1989	12
1990	3
1991	3
1992	51
1993	40
1994	54
1995	558
1996	838
1997	271
1998	12
1999	11
2000	7
2001	7
2002	1

Company A had 47 sample plots planted from 1989 to 2002, Company B had 903 sample plots planted from 1985 to 1997 and Company C had 882 plots planted from 1989 to 2000. The results from the individual company models are shown below in Table 10. A significant increase in the accuracy of models was achieved for Companies A and B whereas Company C had poor performing models. This highlights an issue with the combined company data models as the data from Company C brings down the accuracy of those models. By splitting the companies data, there is more consistency around the plot data used as it is likely collected in the same manner and potentially by the same inventory crew. The accuracy achieved from company A was likely achieved because of a smaller, quality data set used to train the model which again highlights an issue for the combined data models. A large dataset is not required for accurate models, rather it can be a drawback as there is a greater potential for inconsistency between inventory crews resulting in inaccurate models. The largest sample plot dataset came from Company B so this rules out simply having too much data as a reason for inaccurate models.

Table 10. Results of models from different companies.

Model	Company	R^2	RMSE (%)	RMSE (unit)
MTH	A	0.72	9.7	3.83 m
	B	0.70	5.1	1.86 m
	C	0.23	8.1	3.01 m
Basal Area	A	0.63	17	11.3 m^2/ha
	B	0.35	16	10.2 m^2/ha
	C	0.16	21	14.4 m^2/ha
Volume	A	0.73	19	154 m^3/ha
	B	0.37	20	145 m^3/ha
	C	0.19	23	190 m^3/ha
Stocking	A	0.39	22	65 stems/ha
	B	0.08	20	77 stems/ha
	C	0.12	23	100 stems/ha

The only models that could be considered as within the desired accuracy are the MTH models from Company A and B. The RMSE is less than 10% and the R^2 values mean the model can predict over a good range of the input data. Company C's MTH model did have an RMSE of less than 10% however the low R^2 shows a poor ability to predict over a full range of input data.

Company B had the lower RMSE values for all models except stocking where Company A had the lowest. Company A had significantly higher R^2 values for all the models and is therefore considered to have the most accurate models of all companies. The combination of R^2 and RMSE on accuracy is highlighted in Figures 7,8 and 9 below which compare the MTH, basal area and volume models for Company A and B.

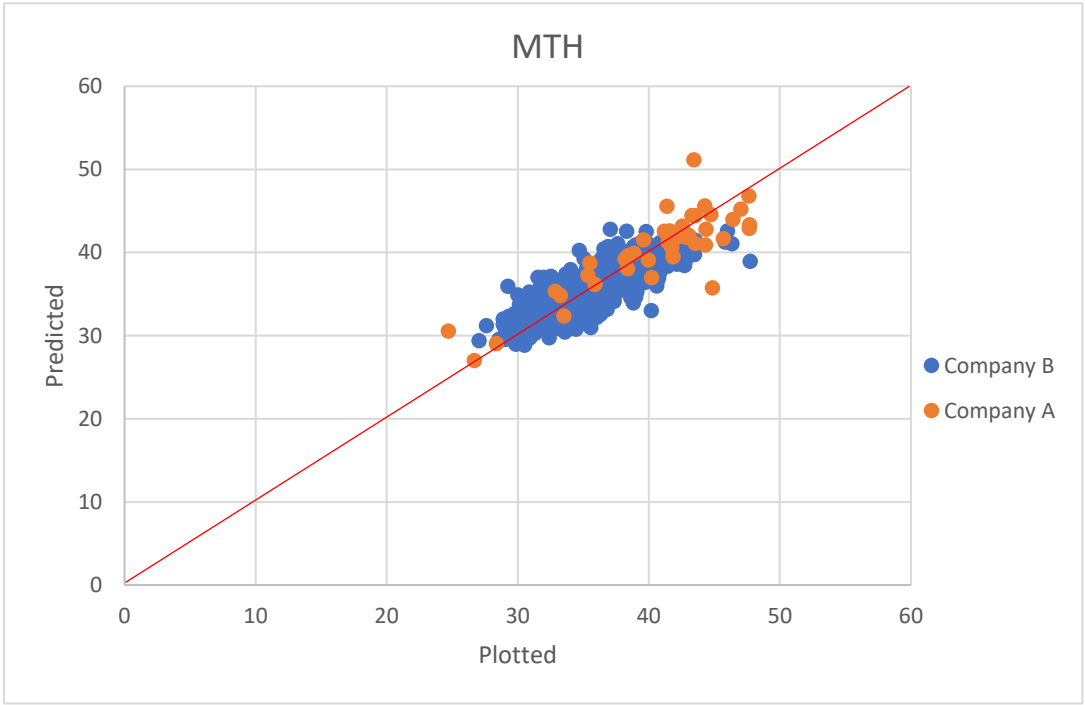


Figure 7. Comparison of company MTH model.

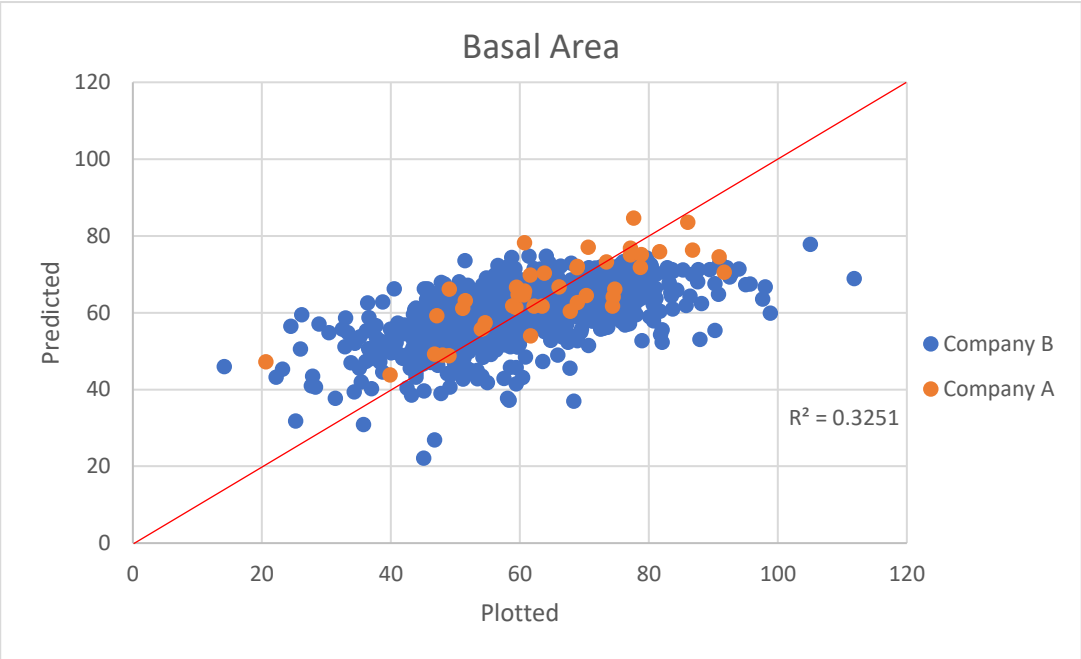


Figure 8. Comparison of company basal area model.

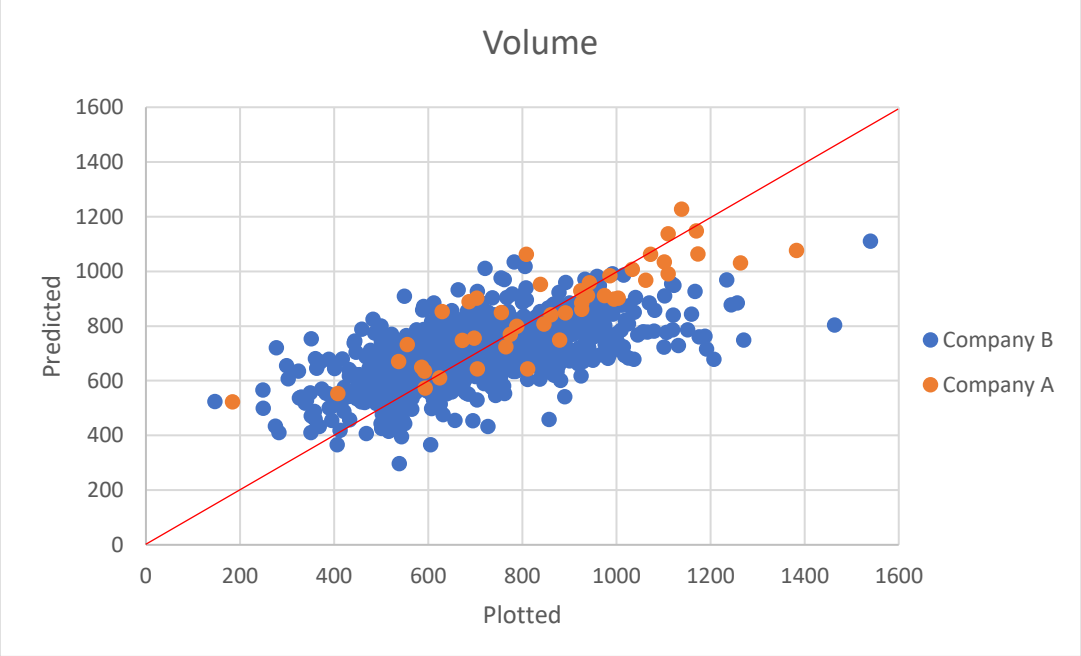


Figure 9. Comparison of company volume model.

6. Discussion

The large data set used in this research provided both benefits and challenges when building the models. The desired accuracy of 10% was nowhere close to being achieved for volume, basal area, and stocking models. Research completed on this topic usually involves smaller datasets where one crew measures all the plots and higher quality LiDAR data is used. The desired accuracy of 10% was optimistic for this research. However, it was set as this is the expected standard of inventory crews and the objective of these models was to be as accurate as them.

The original models for basal area, volume and stocking with all data included had large RMSE errors of 70%, 91% and 27% respectively. The large RMSE values for basal area and volume can be attributed to the extreme outliers in the data. The low RMSE of 7.3% for MTH was expected due to previous research having errors as low as 3% (Holgrem et al., 2004). This is typically due to the direct relationship with height percentile variables (Næsset, 2004) such as p90. The original model developed included three height percentiles. All models had poor R^2 values of 0.34, 0.18, 0.15 and 0.07 respectively meaning their ability to predict values over the whole range of input data was poor therefore making the models even less useable. The poor R^2 values could be attributed to the poor relationships for single variables, no single variable had an R^2 value greater than 0.3. The large disconnect between the accuracy of the original models and the desired accuracy can mostly be attributed to the inclusion of outliers and therefore the results from these models can be ignored.

The accuracy of the models with different restrictions is shown in Table 11 below. As expected, the removal of outliers led to an increase in accuracy for all models with increasing R^2 values and decreasing RMSE. The unlimited variable models were found to contain insignificant variables; therefore, they were removed in the next step of the model building. This step reduced the model's risk of overfitting as the unlimited variable models contained many variables. Accuracy was slightly decreased for all models, and it was found the risk of collinearity error was still high. This highlights a limitation to using MLR models as they are susceptible to this error (Aertsen et al. 2010). A seven-variable limit was set to reduce the likelihood of this error and provide the final models for all the companies data. The accuracy of these models was relatively close to the best accuracies achieved in previous iterations. The final models achieved the lowest RMSE in the basal area and stocking models.

Table 11. Results from models including all companies' data.

Model	Step	R^2	RMSE (%)	RMSE (unit)
MTH	Original	0.34	7.3	2.68 m
	Removed Outliers	0.45	7.5	2.76 m
	Significant	0.44	6.7	2.75 m
	Final	0.42	7.5	2.77 m
Basal Area	Original	0.18	70	49.7 m^3/ha
	Removed Outliers	0.20	18	12.7 m^3/ha
	Significant	0.17	20	12.9 m^3/ha
	Final	0.17	17	12.5 m^3/ha
Volume	Original	0.15	91	743 m^2/ha
	Removed Outliers	0.24	20	171 m^2/ha
	Significant	0.22	23	175 m^2/ha
	Final	0.21	22	172 m^2/ha
Stocking	Original	0.07	27	106 stems/ha
	Removed Outliers	0.09	23	92 stems/ha
	Significant	0.07	24	94 stems/ha
	Final	0.07	22	91 stems/ha

The desired accuracy was achieved for the MTH model and could be applied pragmatically throughout the Gisborne region. The basal area and volume models could be used to provide small-scale forest owners with a rough estimation for their stands and gain some knowledge on what to expect from a plotting crew. The stocking model provided no accurate results and should not be applied. As there is little known about the yields of small-scale forests in the Gisborne area the MTH, basal area and volume models can provide ballpark figures for owners and forest managers to work off.

From these results it can be taken that developing accurate MLR models for large regions is difficult and not up to the same quality that is expected from a typical plotting crew. Consistent, quality input data is the only way that accurate models can be developed. This was reflected in the increase in accuracy of the models once the outliers were removed. However, for these models to be more accurate a standard should be set on how and what to measure for each sample plot. Although the LiDAR data used in this study was not as high of quality as used in other research, it would not have as much of a detrimental effect on accuracy as seen in these models if the plot data used was entirely accurate.

Once the data was split by companies there was a clear trend in the accuracy of the models produced. The R^2 values for company A and the RMSE values for Companies A and B were like those found in a study from Xu et al. (2018), as shown in Table 12 below. The study from Xu et al. (2018) used a survey grade GPS to locate plot centres, measured more tree heights per plot and did not investigate MLR models for stocking. The substantial difference between the companies provided some insight as to why the original models were not as accurate as previous research. There was a clear relationship between LiDAR metrics and stand characteristics in Company A's models unlike the models produced from the combined data. This was likely due to a smaller sample size and data for the plots being collected all within 2 months.

Table 12. Comparison of Xu et al. (2018) study to company models.

Model	Type	R^2	RMSE (%)	RMSE (unit)
MTH	Company A	0.72	9.7	3.83 m
	Company B	0.70	5.1	1.86 m
	Xu et al. (2018)	0.97	5.2	1.31 m
Basal Area	Company A	0.63	17	11.3 m^2/ha
	Company B	0.35	16	10.2 m^2/ha
	Xu et al. (2018)	0.73	19	9.4 m^2/ha
Volume	Company A	0.73	19	154 m^3/ha
	Company B	0.37	20	145 m^3/ha
	Xu et al. (2018)	0.88	19	84 m^3/ha

A small number of quality sample plots can produce just results as accurate as a large dataset (Melville et al., 2015). This point is shown in Company A and B's results, Company A had 47 sample plots whereas Company B had 903 sample plots. In the study by Xu et al. (2018) 112 sample plots were used to train the model. The company results show that given a quality set of input data, accurate models can be produced using publicly available LiDAR data. This is something small-scale woodlot owners should take note of. If a small-scale

woodlot owner gathered quality sample plots where the height of all trees was measured, then accurate and useable models can easily be produced for their forest. As fewer sample plots are required to create an accurate model to predict stand characteristics, it would likely cost the owner less to create accurate models for their small-scale woodlot. The split of the data by companies shows that one model for all small-scale forests in Gisborne is not practical.

The stocking model was the poorest performing model out of all the stand characteristics, this was likely due to a lack of explanatory LiDAR metrics and the area-based approach used to develop the models. A tree-based approach would likely result in more accurate models for predicting stocking (Hyyppä et al., 2012) however a dense point cloud and a consistent canopy cover are required which would be an issue in this case.

6.1 Sources of Error

Given the large data set and mix of companies, there were many areas where errors could arise causing inaccuracy in results. Errors in the sample plot data could have come from incorrect plot measurements and differing dates between measuring plots and LiDAR collection. Errors in the processing of the data could have come from the equations used to estimate stand characteristics and misalignment of plot data

6.1.1 Plot measurements

The desired accuracy of 10% for the models was selected as this is the expected accuracy of plotting crews in New Zealand (Bell, 2016) however if there are errors of +/- 10% of the true value in the sample plot data, this can cause issues when creating the MLR models. For the models to be accurate the input data must be very accurate (Bouvier et al., 2015), and errors from the plotting data will be compounded and create errors in the MLR models. Having data from three different companies also meant three different data collection crews/ methods further creating variation in the data.

6.1.2 GPS Receiver

Consumer-grade GPS's were used by the crews when collecting plot data which means the position they recorded as the plot centre may be different from the position they were really in. This is due to the forest canopy distorting signal (Tomaštík, Jr. *et al* 2017) causing positioning errors of up to 10.2 metres (Anders, 2018). Incorrect positioning can cause overlapping errors where the measured plots and the LiDAR data are not measuring all the same trees.

6.1.3 Time Difference

The sample plots were measured from 2017 through to 2021 and the LiDAR data was collected from 2019 to 2020. This means that there was a potential for a 3-year gap between a plot being measured and the LiDAR being scanned over the same plot area. The time difference will mean the plot characteristics have changed and they will no longer be

the same as what the LiDAR is measuring. Therefore, the data being used to train the model from that plot is inaccurate leading to inaccurate results.

6.1.4 Equations Used

Plot data from the companies consisted of the DBH of all trees and the height of some. This meant a height-diameter relationship equation had to be used to determine the remaining height of the other trees. While there is a strong relationship between the height and diameter of trees it is not completely accurate. The predicted heights were then used to predict the MTH of the stand as well as the volume. By using equations to predict the stand metrics the resultant value may be slightly different to the true value further leading to potentially inaccurate results.

6.2 Future Research

If this research was to be continued in future a greater focus on the quality of the input data should be focused to eliminate the sources of error outlined above. It is almost impossible for a plotting crew to be 100% accurate when measuring plots however if the same crew measures all the plots used to train the model then there is consistency in the measurements and if there is a slight error made when measuring it would be across all plots and not some. A greater emphasis should also be placed on collecting heights for more trees as the more tree heights recorded the more accurate the height estimates will be for other trees. Having a large data set made it hard to comb through all the plot data and look for errors in each plot, with other steps such as variable limits and 70/30 training and validation split, a smaller sample set could be used to train the model where the quality can be better monitored. Ideally, the plots should be measured at the same time the LiDAR is scanned, this can be an issue for large-scale areas of study like in this report. However, there could be more of an attempt to line up measurement dates to avoid large periods between collection dates.

7. Conclusion

This study showed that LiDAR-derived metrics can produce accurate models for MTH, reasonable models for basal area and volume, and low accuracy models for stocking using publicly available LiDAR data. The desired accuracy level of 10% was best suited for the MTH models given the direct relationship of the height percentile metrics. However, it was optimistic for the basal area, volume and stocking models given the lack of good explanatory metrics and accuracy of previous research. A limitation to MLR models was shown with the most accurate models being achieved with unlimited variables included but this came with a risk of overfitting the model and collinearity error.

The combined dataset provided challenges for creating accurate models given the variations between plotting crews and dates for the sample plots. Splitting the data by companies showed that an increase in accuracy can be achieved when the sample plots are consistent. Companies A and B had similar RMSE values for all their models however Company A

achieved much higher R^2 values for their models. This can be attributed to the difference in samples used with company A having 47 sample plots whereas company B had 903 sample plots. The RMSE values from the combined data models were like those found in the individual company models however the R^2 values were much lower. This created a limitation to the application of the models as they had a poor ability to predict results over the full range of input data.

The results show that there is potential for using MLR models to estimate stand characteristics for small-scale forests. Creating one model for the whole Gisborne region was impractical. However, on a company and forest-by-forest basis accurate models can be developed to estimate stand characteristics.

There are still many potential sources of error that can compound quickly when collecting data and creating these models, therefore, a greater emphasis should be placed on using accurate sample plots to train the model in future applications. The use of the free LiDAR data to create models to predict stand variables for small-scale forests should continually be investigated in future as there are benefits to those such as forest owners, contractors, trucking companies, sawmills, and ports.

References

Aertsen, W., Kint, V., van Orshoven, J., Özkan, K., & Muys, B. (2010). Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological Modelling*, *221*(8), 1119–1130. https://doi.org/10.1016/j.ecolmodel.2010.01.007

Bell, A. (2016, August). *NZ Farm Forestry - The science or art of forest inventory*. NZ Farm Forestry. https://www.nzffa.org.nz/farm-forestry-model/resource-centre/tree-grower-articles/august-2016/the-science-or-art-of-forest-inventory/#

Bouvier, M., Durrieu, S., Fournier, R. A., & Renaud, J. P. (2015). Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. *Remote Sensing of Environment*, *156*, 322–334. https://doi.org/10.1016/j.rse.2014.10.004

Civil Aviation Authority. (2022). *Certification for Unmanned Aircraft*. www.aviation.govt.nz. https://www.aviation.govt.nz/drones/part-102-certification-for-drones/

Dash, J., Pont, D., Brownlie, R., Dunningham, A., Watt, D., & Pearse, G. (2016). Remote sensing for precision forestry. *NZ Journal of Forestry*, *60*(4), 15–24.

Fehrmann, L., Lehtonen, A., Kleinn, C., & Tomppo, E. (2008). Comparison of linear and mixed-effect regression models and a *k*-nearest neighbour approach for estimation of single-tree biomass. *Canadian Journal of Forest Research*, *38*(1), 1–9. https://doi.org/10.1139/x07-119

Frazer, G., Magnussen, S., Wulder, M., & Niemann, K. (2011). Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. *Remote Sensing of Environment*, *115*(2), 636–649. https://doi.org/10.1016/j.rse.2010.10.008

Goerndt, M. E., Monleon, V. J., & Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Canadian Journal of Forest Research*, *41*(6), 1189–1201. https://doi.org/10.1139/x11-033

Goulding, C., & Fritzsch, M. (2010, May). *NZ Farm Forestry - Improving forest inventory for the woodlot owner*. NZ Farm Forestry. https://www.nzffa.org.nz/farm-forestry-model/resource-centre/tree-grower-articles/may-2010/improving-forest-inventory-for-the-woodlot-owner/

Holmgren, J. (2004). Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. *Scandinavian Journal of Forest Research*, *19*(6), 543–553. https://doi.org/10.1080/02827580410019472

Hyyppä, J., Yu, X., Hyyppä, H., Vastaranta, M., Holopainen, M., Kukko, A., Kaartinen, H., Jaakkola, A., Vaaja, M., Koskinen, J., & Alho, P. (2012). Advances in Forest Inventory Using Airborne Laser Scanning. *Remote Sensing*, *4*(5), 1190–1207. https://doi.org/10.3390/rs4051190

Anders, J. (2018). On the accuracy of GNSS in forests : a test of consumer-grade GNSSequipment, smartphones and open-source postprocessing software under forest canopies, for mapping of forest species. *Master's Thesis Norwegian University of Life Sciences*.

Tomaštík, J., Salon, S., & Rastislav, P. (2017). Horizontal accuracy and applicability of smartphone GNSS positioning in forests. *Forestry: An International Journal of Forest Research*, *90*(2), 187–198.

Kaartinen, H., Hyyppä, J., Yu, X., Vastaranta, M., Hyyppä, H., Kukko, A., Holopainen, M., Heipke, C., Hirschmugl, M., Morsdorf, F., Næsset, E., Pitkänen, J., Popescu, S., Solberg, S., Wolf, B. M., & Wu, J. C. (2012). An International Comparison of Individual Tree Detection and Extraction Using Airborne Laser Scanning. *Remote Sensing*, *4*(4), 950–974. https://doi.org/10.3390/rs4040950

Manley, B., Morgenroth, J., & Xu, C. (2020). Quantifying the area of the small-scale owners' forest estate in the East Coast, Hawke's Bay and Southern North Island. *New Zealand Journal of Forestry*, *6*(4).

Melville, G., Stone, C., & Turner, R. (2015). Application of LiDAR data to maximise the efficiency of inventory plots in softwood plantations. *New Zealand Journal of Forestry Science*, *45*(1). https://doi.org/10.1186/s40490-015-0038-7

Montaghi, A., Corona, P., Dalponte, M., Gianelle, D., Chirici, G., & Olsson, H. (2013). Airborne laser scanning of forest resources: An overview of research in Italy as a commentary case study. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 288–300. https://doi.org/10.1016/j.jag.2012.10.002

MPI. (2016). *Wood Availability Forecasts – New Zealand 2014–2050.* https://www.mpi.govt.nz/dmsdocument/

Næsset, E. (2004). Accuracy of forest inventory using airborne laser scanning: evaluating the first nordic full-scale operational project. *Scandinavian Journal of Forest Research*, *19*(6), 554–557. https://doi.org/10.1080/02827580410019544

Næsset, E. (2011). Estimating above-ground biomass in young forests with airborne laser scanning. *International Journal of Remote Sensing*, *32*(2), 473–501. https://doi.org/10.1080/01431160903474970

Petterson, H. (1953). Yield of Coniferous Forests. *Medd. Stat Skogsforsoksanct*, *45*(IB).

Sabol, J., Procházka, D., & Patočka, Z. (2016). Development of models for forest variable estimation from airborne laser scanning data using an area-based approach at a plot level. *Journal of Forest Science*, *62*(No. 3), 137–142. https://doi.org/10.17221/73/2015-jfs

Saremi, H., Kumar, L., Turner, R., & Stone, C. (2014). Airborne LiDAR derived canopy height model reveals a significant difference in radiata pine (Pinus radiata D. Don) heights based on slope and aspect of sites. *Trees*, *28*(3), 733–744. https://doi.org/10.1007/s00468-014-0985-2

Takahashi, T., Yamamoto, K., Senda, Y., & Tsuzuku, M. (2005). Estimating individual tree heights of sugi (*Cryptomeria japonica* D. Don) plantations in mountainous areas using small-footprint airborne LiDAR. *Journal of Forest Research*, *10*(2), 135–142. https://doi.org/10.1007/s10310-004-0125-8

van Leeuwen, M., & Nieuwenhuis, M. (2010). Retrieval of forest structural parameters using LiDAR remote sensing. *European Journal of Forest Research*, *129*(4), 749–770. https://doi.org/10.1007/s10342-010-0381-4

Watt, M. S., Dash, J. P., Bhandari, S., & Watt, P. (2015). Comparing parametric and non-parametric methods of predicting Site Index for radiata pine using combinations of data derived from environmental surfaces, satellite imagery and airborne laser scanning. *Forest Ecology and Management*, *357*, 1–9. https://doi.org/10.1016/j.foreco.2015.08.001

White, J. C., Wulder, M. A., Varhola, A., Vastaranta, M., Coops, N. C., Cook, B. D., Pitt, D., & Woods, M. (2013). A best practices guide for generating forest inventory attributes

from airborne laser scanning data using an area-based approach. *The Forestry Chronicle*, *89*(06), 722–723. https://doi.org/10.5558/tfc2013-132

Xu, C., Manley, B., & Morgenroth, J. (2018). Evaluation of modelling approaches in predicting forest volume and stand age for small-scale plantation forests in New Zealand with RapidEye and LiDAR. *International Journal of Applied Earth Observation and Geoinformation*, *73*, 386–396. https://doi.org/10.1016/j.jag.2018.06.021

Xu, C., Manley, B., & Morgenroth, J. (2019). Describing area and yield for small-scale plantation forests in Wairarapa region of New Zealand using RapidEye and LiDAR. *New Zealand Journal of Forestry Science*, *49*. https://doi.org/10.33494/nzjfs492019x16x

Appendix

LAStools Code

```
:: Check the details of LiDAR data
:: cd /d D:\LAStools\bin
::lasinfo -i”E: \Gisborne\LiDAR\Point_cloud\*.laz ^
::      -cd ^
::      -merged ^
::      -odir quality -o lidar_info.txt

:: Classify noise points as class 7- for each cell 4 by 4 by 2, 3 or
fewer points are classified as noise. Note this will create many
denoised .laz files :: and take up a lot of space

lasnoise -i E:\Gisborne\LiDAR\Point_cloud\*.laz ^
         -step_xy 4 -step_z 2 -isolated 3^
        -ignore_class 2 ^
         -odir D:\denoise -o denoise.laz

:: Create new tiles with buffers
lastile -i D:\denoise\*.laz ^
        -o "tile.laz" -tile_size 1000 -buffer 20 -flag_as_withheld -faf
^
        -odir D:\buffer -o buffer.laz

::Normalise height of each point
lasheight -i D:\buffer\*.laz^
         -replace_z ^
          -odir D:\height -o normal.laz ^

:: Add lndex to each .laz file
lasindex -i D:\height\*.laz ^


:: Plot metrics in a csv
lascanopy -i D:\height\*.laz -merged ^
          -drop_class 7 18^
          -lop
E:\Gisborne\Data_received\Working\LiDAR\las\All_stands_fixed_overlap.sh
p ^
          -cover_cutoff 3.0 ^
          -cov -dns ^
          -height_cutoff 2.0 ^
          -c 2.0 5.0 10.0 50.0 ^
          -max -avg -qav -std -kur ^
          -p 50 75 95 99 ^
         -int_avg -int_qav -int_std -int_ske -int_kur ^
         -int_p 50 95 99 ^
          -b 30 50 80 90 ^
          -d 2.0 5.0 10.0 50.0 ^
         -odir D:\data -o plots.csv
```

RStudio Code

```
"basal area example"
#basal area = G

library(leaps)
library(tidyverse)
library(caret)
install.packages("caret")

models <- regsubsets(G~., data = regression_test, nvmax = 5) # nvmmax
is max numbver of variables you allow
summary(models) # shows which variables to used based on allowed
variables

res.sum <- summary(models)
data.frame(
Adj.R2 = which.max(res.sum$adjr2),
CP = which.min(res.sum$cp),
BIC = which.min(res.sum$bic)
)

# spits out suitable number of variables from different (R2,CP, BIC)
criteria

get_model_formula <- function(id, object, outcome){
  models <- summary(object)$which[id,-1]
  predictors <- names(which(models == TRUE))
  predictors <- paste(predictors, collapse = "+")
  as.formula(paste0(outcome, "~", predictors))
}

get_model_formula(N, models, "G") # shows best "N" variable model
formula (i.e N = 3,4,5)

get_cv_error <- function(model.formula, data){
  set.seed(1)
  train.control <- trainControl(method = "cv", number = 5)
  cv <- train(model.formula, data = data, method = "lm",
              trControl = train.control)
  cv$results$RMSE
}
model.ids <- 1:5 #number of variables
cv.errors <-  map(model.ids, get_model_formula, models, "G") %>%
map(get_cv_error, data = regression_test) %>%
unlist()
cv.errors

# gives cv errors for models with differing amounts of variables (shows
up to 5 variables in this case)

which.min(cv.errors) # result shows the suitable amount of variables to
use

### create seperate excel sheet with only selected variables to then
put into training/ validation
```

```
set.seed(123)
training.samples <- regression_test$G %>%
  createDataPartition(p = 0.7, list = FALSE) # 70% training, 30%
validation
train.data  <- regression_test[training.samples, ]
test.data <- regression_test[-training.samples, ]
model <- lm(G ~., data = train.data)
predictions <- model %>% predict(test.data)
data.frame( R2 = R2(predictions, test.data$G),
            RMSE = RMSE(predictions, test.data$G),
            MAE = MAE(predictions, test.data$G))

RMSE(predictions, test.data$G)/mean(test.data$G)

# whatever model gives lowest RMSE is the most suitable model

summary(model)

#gives the coefficients for the model
```