Department of Primary Industries

Technical Report for Local Land Services

Commercial and Non-Commercial Forest Mapping for Private Native Forest, North and South Coast NSW using LiDAR Data

www.dpi.nsw.gov.au

Published by the NSW Department of Primary Industries

Amrit Kathuria

Title: Commercial and Non-Commercial Forest Mapping for Private Native Forest, North and South Coast NSW using LiDAR Data

First published January 2022

ISBN number

More information

amrit.kathuria@dpird.nsw.gov.au / Forest Science, DPI Forestry, Parramatta, NSW

Acknowledgements

This project was funded by Local Land Services and DPI. Thanks to Nick Cameron for technical advice on forest related issues, Murray Webster for training data, Masoomeh Alaibakhsh for GIS support and Sam Hislop for processing of spatial services Lidar data.

This report is one of 12 deliverables arising from the NSW Local Land Services Private Native Forests (PNF) research and development program with DPI Forest Science, available at https://www.dpi.nsw.gov.au/forestry/science

www.dpi.nsw.gov.au

Your Reference number (generally CM9)

Contents

List of Tables

List of Figures

Executive Summary

Non-commercial forest is defined as native forest that is incapable of producing high-quality log based on its height and form. Commercial forest on the other hand is a part of a forest that has large enough trees to allow it to be harvested by the forest industry.

The aims of this study are to:

a) Analyze the correlative relationship between LiDAR variables and commercial and non-commercial forest

b) Develop a spatially-explicit model (20m resolution) of commercial vs non-commercial forest.

c) Produce a map of commercial and non-commercial forest for North Coast NSW

d) Produce estimates of the commercial forest available for private native forest in North Coast NSW.

Key findings of the project:

- Five variables; average height of the first returns, standard deviation of first returns, cover, 95% percentile height (p95) of first returns and 99% percentile height (p99) of first returns were used as predictor variables for the machine learning model. Two of the most important variables in the classification are p95 and cover. P95 is a surrogate for dominant height.

- A spatially explicit model for commercial vs non-commercial forest was developed at 20m resolution for North Coast NSW.

- The model showed a good agreement with the reference data with an overall accuracy of 86.1%, 95% confidence interval of 85.0% - 87.0%. User's and producer's accuracy were in the range of 84.6% to 87%.

- The model was applied to the LiDAR data from spatial services and a map for the commercial vs non-commercial was produced for North Coast NSW.

- Total PNF area for North Coast where spatial services LiDAR is available is 2,462,607 ha. Area for commercial and non-commercial Forest is 1,133,697 ha (46%) and 1,328,910 ha (54%).

- Total PNF area for South Coast NSW is 3,628,798ha. Of the total PNF area 1,273,565ha(35.1%) is forest. The forest area in South Coast PNF covered by spatial services LiDAR is 1,066,741ha, with 233,466ha (21.9%) Commercial forest and 833,274ha (78.1%) of non-commercial forest.

Introduction

Conventionally forest inventory attributes have been estimated using a representattive sample of plots from the total forested area. Such methods could be costly and time consuming (Wulder et al. 2012). Remotely sensed data have been used extensively for improving estimates and reduce sampling intensity in forestry (McRoberts et al., 2002). Also, accurate maps of the forest provide small area estimates better as sub regions based estimates are generally more accurate with good models. Random forest is a data mining approach used to integrate field measurements with remotely sensed data to produce more reliable results.

LiDAR matrices were used as predictor variables to:

a) Develop a spatially-explicit model (20m resolution) of commercial vs non-commercial forest.

b) Analyze the correlative relationship between LiDAR variables and commercial non-commercial forest

c) Produce a map of commercial and non-commercial forest for North Coast NSW

d) Produce estimates of the commercial forest available for private native forest in North Coast NSW.

Data

Study Area

The study area is private native forest on the NSW North Coast that has Spatial Services LiDAR coverage (as shown at Figure 1).

LiDAR Data

In the first half of 2020, Aerometrex was contracted to supply data over several transects in forestry on NSW's North Coast. The specifications of the LiDAR is presented in Table1.

Table 1: LiDAR setting for the data capture

Altitude	700m
Laser system	VQ780i
Pulse rate frequency (PRF)	2000kHz
Scan rate	300lps
Field of view	+/-30 degrees
Laser footprint	0.126m

The LiDAR data that was collected in 2020 was only for 256 properties. A much more completed LiDAR coverage of the north coast area is available through spatial services. 236

individual LiDAR surveys between 2001 and 2015 covering an area in excess of 245,000 square kilometres is available from spatial services. Figure 1 is the map of the available data, the point density varies from less than 1 point per square meter to 1 or more points per square meter.

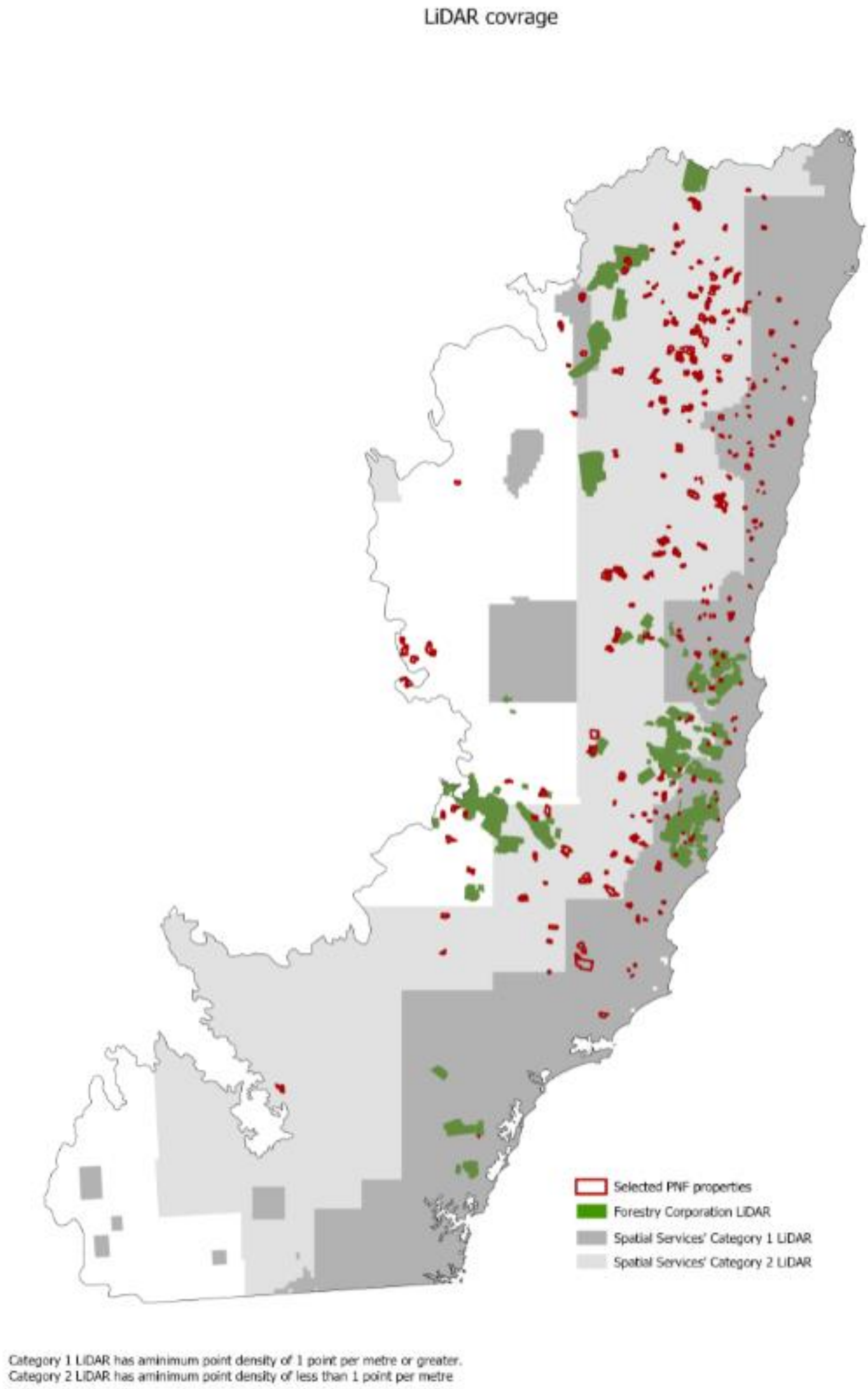

Figure 1: Spatial services LiDAR data.

Training and Validation data

Four properties were selected and API data, orthophotography data, google earth and LiDAR data was used to create the reference data. Field visits were undertaken to check for the accuracy of the reference data.

Methods

Training and Validation data

Training data was collected from four reference properties.

Classification

Non-commercial forest is defined as native forest that is incapable of producing high-quality log based on its height and form. Commercial forest on the other hand is a part of a forest that has large enough trees to allow it to be harvested by the forest industry. Initially it was decided to have three classes; non-commercial, low quality commercial and commercial forest. The low quality commercial forest is difficult to classify using the LiDAR data only. The decision was made to consider only commercial vs non-commercial classes. The commercial forest variable is commercial and low quality commercial forest.

A sample of 26000 sample points were selected from the reference properties and the corresponding LiDAR data was extracted for those points. The scatter plot of the data is presented in Figure 2. As can be seen the light green points are for the low quality commercial forest. It is clear that the low quality commercial falls between the commercial and the non-commercial, the same is reflected by the pair wise plots in figure 4. None of the LiDAR variables can distinguish between the low quality commercial and commercial. The decision was made to consider only commercial vs non-commercial classes. The commercial forest variable is commercial and low quality commercial forest.

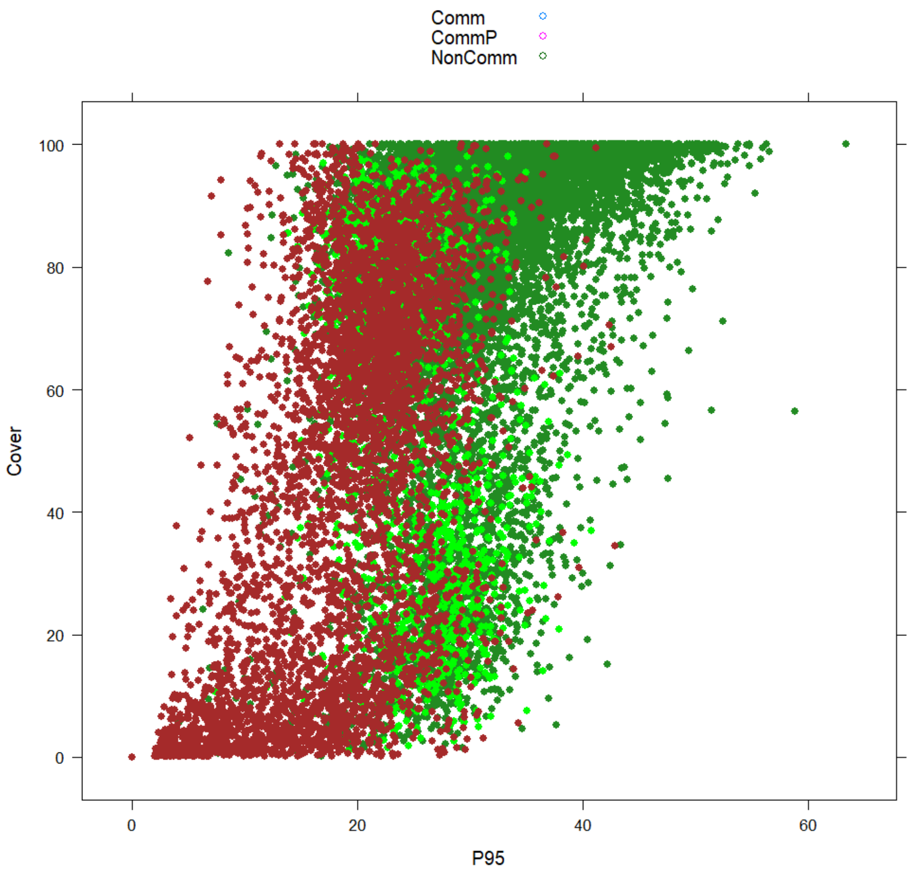

Figure 2: Scatter plot of LiDAR matrices P95 and cover. The brown colour points are non-commercial, light green are the low quality commercial and the dark green is for the commercial forest.

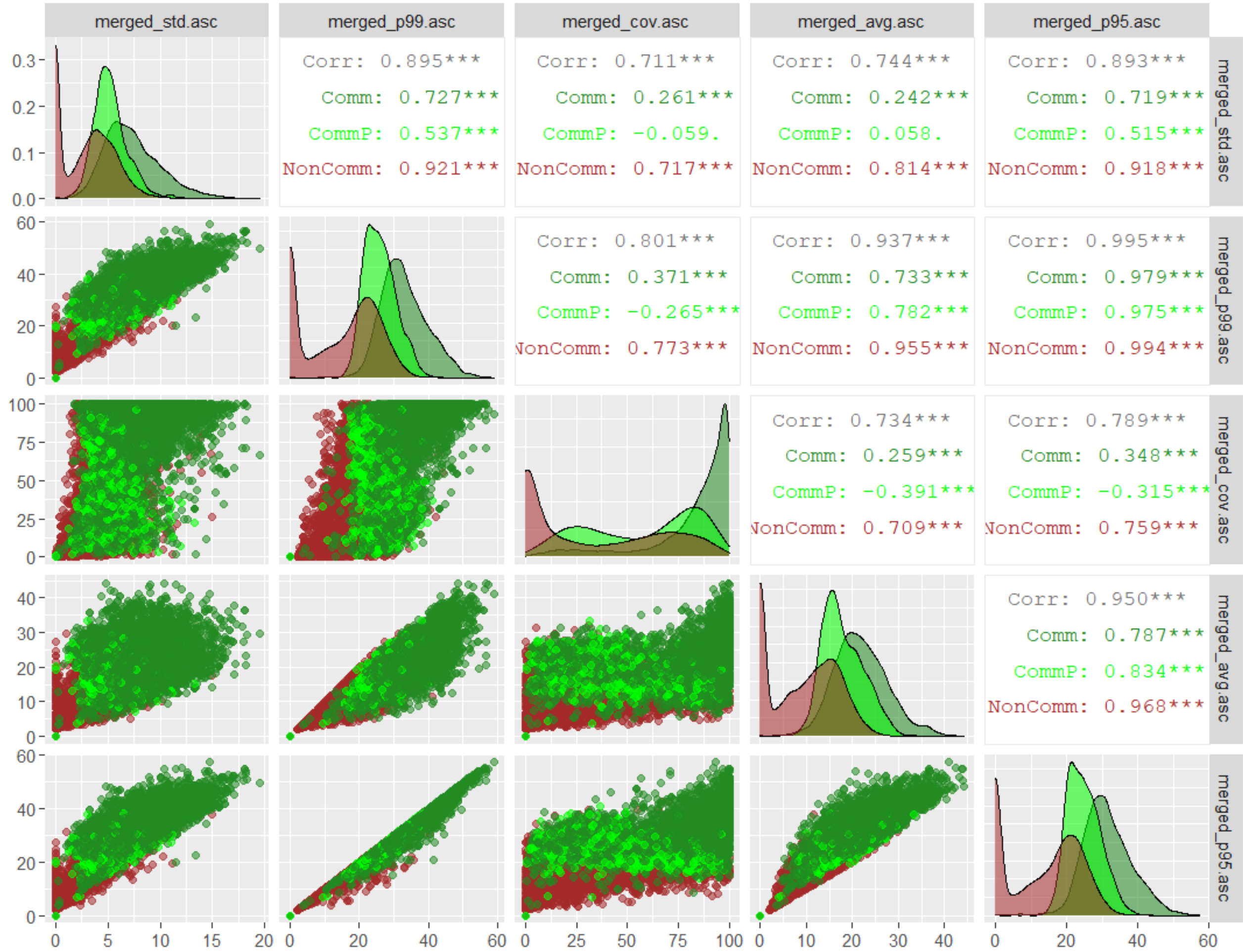

Figure 3: Pairwise scatter plot of all the five LiDAR matrices used. The brown colour points are non-commercial, light green are the low quality commercial and the dark green is for the commercial forest. The numbers in the right are the correlations for the three classes.

Similar plots as figures 3 for two classes; commercial and non-commercial are presented in figure 4. The commercial class is commercial and the low quality commercial. As can be seen from the plots cover and p95 are the best variables for discriminating the two classes.

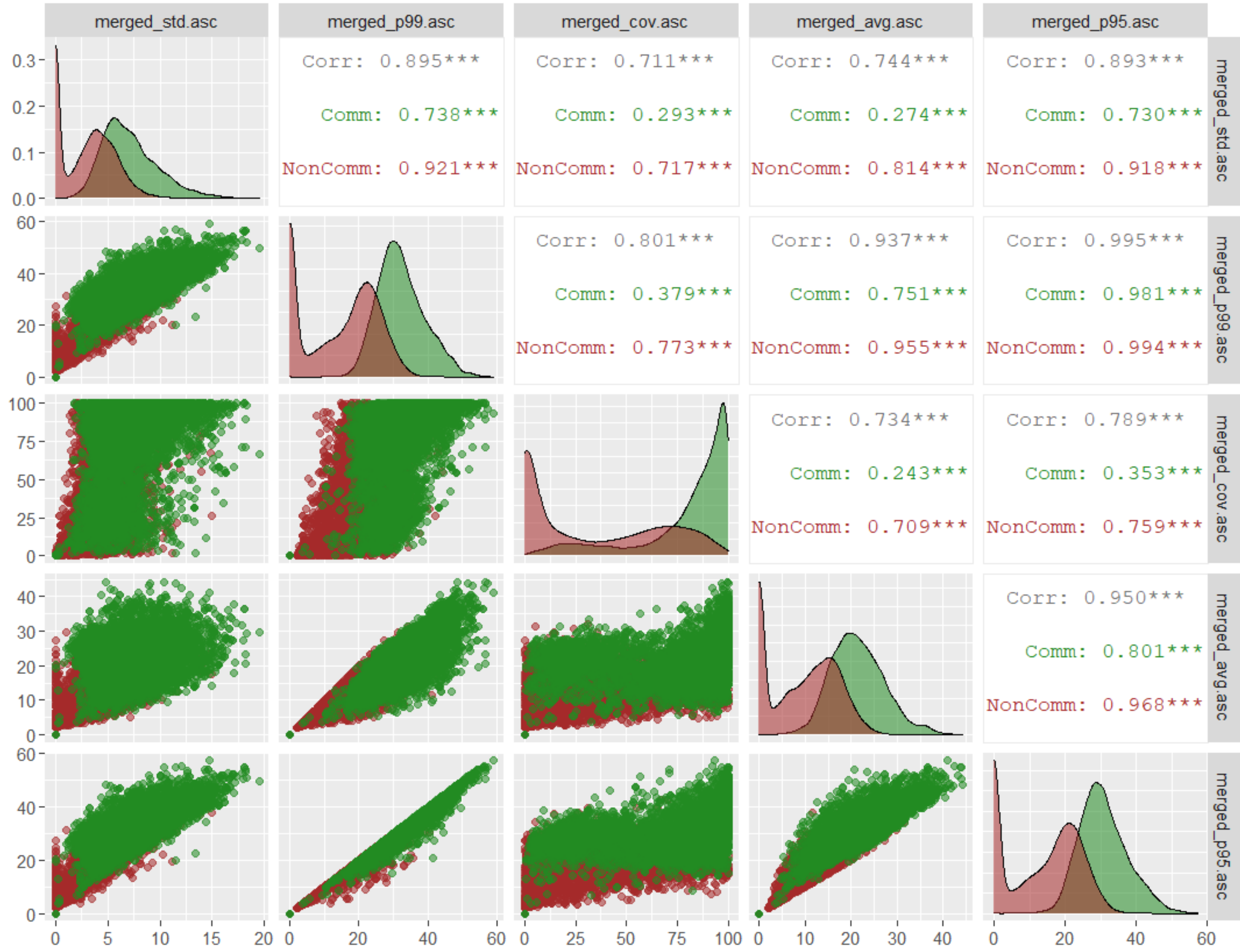

Figure 4: Pairwise scatter plot of all the five LiDAR matrices used. The brown colour points are non-commercial and the dark green is for the commercial forest. The numbers in the right are the correlations for the two classes.

LiDAR Processing

The point density for the LiDAR point cloud data is more than 40 points per square meter. LAStools (rapidlasso GmbH) software were used to process the LiDAR point cloud data. The noise points were identified using lasnoise, the data was then classified ground and non-ground. Digital terrain models were created and normalised point cloud data was created. Once the LiDAR data were normalized, the point data were processed to extract LiDAR metrics at a 20m resolution.

A total of 50 LiDAR variables were initially selected. The final variables for model were selected after looking at a number of publications and identifying the variables strongly correlated with forest identification. As a result, only five variables; average height of the first returns, standard deviation of first returns, cover, 95% percentile height (p95) of first returns and 99% percentile height (p99) of first returns were retained (Table 1).

The point density for the LLS 2020 data is very high compared to the spatial services data; >40 points per square meter against about 1 point per square meter. The correlation of the five variables from LLS and spatial services were high so data from the spatial services LiDAR would be a god surrogate for LLS 2020 data.

Random forest

Random forest is an algorithm for classification and regression developed by Leo Breiman (Breiman, 2001) that uses an ensemble of classification and regression trees (Breiman 1994; Hastie et al., 2001; Ripley, 1996). Random forest uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners and random selection of the independent variables at each node. Bagging involves selecting a bootstrap sample from the original data. Bootstrapping is sampling with replacement from the training data, some data points are left out and some are repeated in this dataset. For each bootstrap sample tree is fully grown, this results in the reduction of tree bias. This process is repeated a number of times (>100) and the predictions from each tree are averaged for the final predicted value. Bagging and random variable selection result in low correlation of the individual trees.

The algorithm yields an ensemble that has a number of desirable characteristics such as good accuracy; robustness to outliers and noise; speed; internal estimation of error and estimation of variable importance. Prediction performance of the random forest algorithm is performed using a type of cross-validation in parallel with the training step by using the so-called out-of-bag (OOB) samples. As in a bootstrap sample, sampling is done with replacement, approximately one third of all the observations are left out of the bootstrap sample; these observations are called "out-of-bag" (OOB) data. The OOB data are then used to estimate prediction accuracy (Liaw and Wiener ,2002).

Variable Selection Method

After training the model, variable importance plot was used to identify which variables have the most predictive power. Variables with high importance are drivers of the outcome and their values have a significant impact on the outcome values. By contrast, variables with low importance might be omitted from a model, making it simpler and faster to fit and predict.

Model Assessment and Validation

Before any model can be used with confidence, adequate validation or assessment of the magnitude of errors that may result from their use should be performed. Model validation, in its simplest form, is a comparison between predicted and observed values. All models were validated using apparent validation (same data used for development and testing). Apparent performance is attractive because it is easy to perform. It however results in optimistic estimates of performance. Internal validation methods were used to evaluate model performance on an underlying population (also labelled 'reproducibility'). Internal validation methods aim to provide a more accurate estimate of model performance for new data that is not used in model building. This was done by dividing the reference data into training and test data and the validation was done using the test data.

Accuracy Assessment

Performance of a classification model such as random forest is evaluated using a Confusion matrix. The matrix compares the actual observed values with those predicted by the machine learning model. This gives us a quantitative measure of how well our classification model is performing and what kinds of errors are made.

	Observed Values	
Predicted Values	**Commercial**	**Non-Commercial**
Commercial	**True Positive (TP)**	**False Positive (FP)**
Non-Commercial	**False Negative (FN)**	**True Negative (TN)**

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Cohen's kappa coefficient (κ) is a statistic (Cohen, 1960) that measures the agreement between the observed and chance agreement. It reflects the difference between actual agreement and the agreement expected by chance. Kappa of 0.80 means there is 80% better agreement than by chance alone.

Cohen's kappa is

$$\kappa = \frac{(\text{Observed} \;\; - \text{ chance agreement})}{(1 - \;\text{ chance agreement})}$$

User's accuracy: User's accuracy corresponds to error of commission (inclusion). From the perspective of the user of the classified map, it quantifies how accurate the map is?. It is calculated as: Number of pixels correctly identified in a given map class / Number of pixels claimed to be in that map class. In other words it defines how many of the pixels on the map are actually what they say they are?

Producer's accuracy: From the perspective of the maker of the classified map, how accurate is the map?. It is calculated as the number correctly identified in reference pixels of a given class/ number of pixels actually in that reference class. This statistics quantifies how many pixels on the map are labelled correctly?

Map Creation

Five LiDAR matrices at 20m resolution were used as predictor variables to produce a map of commercial vs non-commercial forest. As the model is applied at pixel level a single or a small group of cells may be misclassified as an entity different from the sea of cells surrounding it, when in reality, the entity belongs to the group of cells that surrounds it. To remove the single, misclassified cells in the classified image, the Majority Filter tool is applied. All the analysis was done using R statistical package (R Development Core Team 2020).

Results

Reference Data

A sample of 26000 sample points were selected from the reference properties and the corresponding LiDAR data was extracted for those points.

Fitting Random Forest

The data was split into training and test; 70% as the training data and other 30% data is used for testing and validating the model. The data was made balanced for the two classes before a random forest model was fitted. Figure 5 is the plot of the predictor variables and the mean decrease in accuracy, higher the decrease in accuracy the better the predictor. The three most important variables are p95, cover and p99.

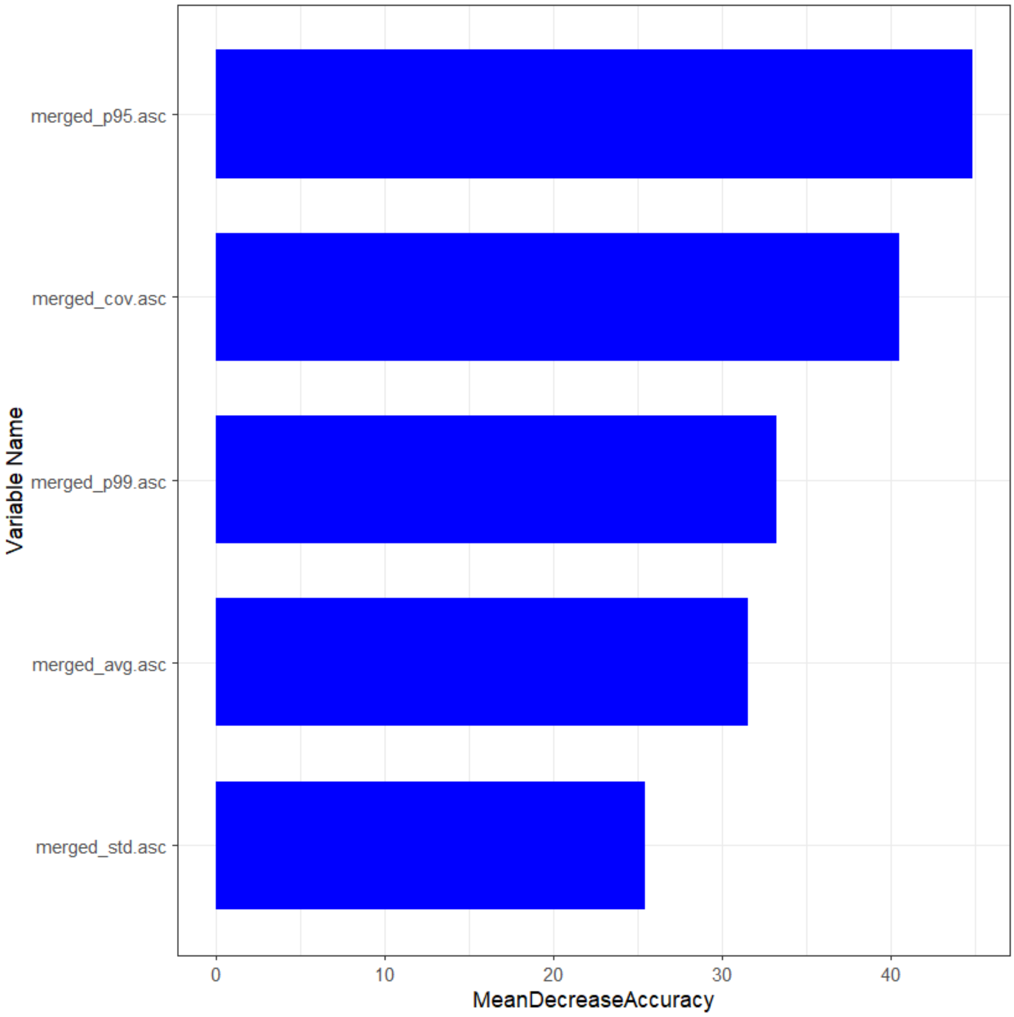

Figure 5: Variable importance plot. The greater the Mean Decrease in Accuracy score, the more important the predictor is in classifying commercial forest.

The estimate of the out of bag error for the training data is 14.9%, with error for commercial class 13.4% and for the non-commercial class was 16.5%. The effect of the each of the predictor variables can be assessed using partial dependence plot. The partial dependence plot shows the marginal effect one or two features have on the predicted outcome of a machine learning model (Friedman 2001). A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic or more complex. Figure 6 is partial plot of the predictor variables for commercial forest. From the plot it is clear that if the if cover is less than 10% the forest is non-commercial but the chances of a commercial forest increase as the cover increases and the chances of the forest to be commercial is very high for cover greater than 80%. The chance of forest to be commercial is really high if p95 is greater than 30m.

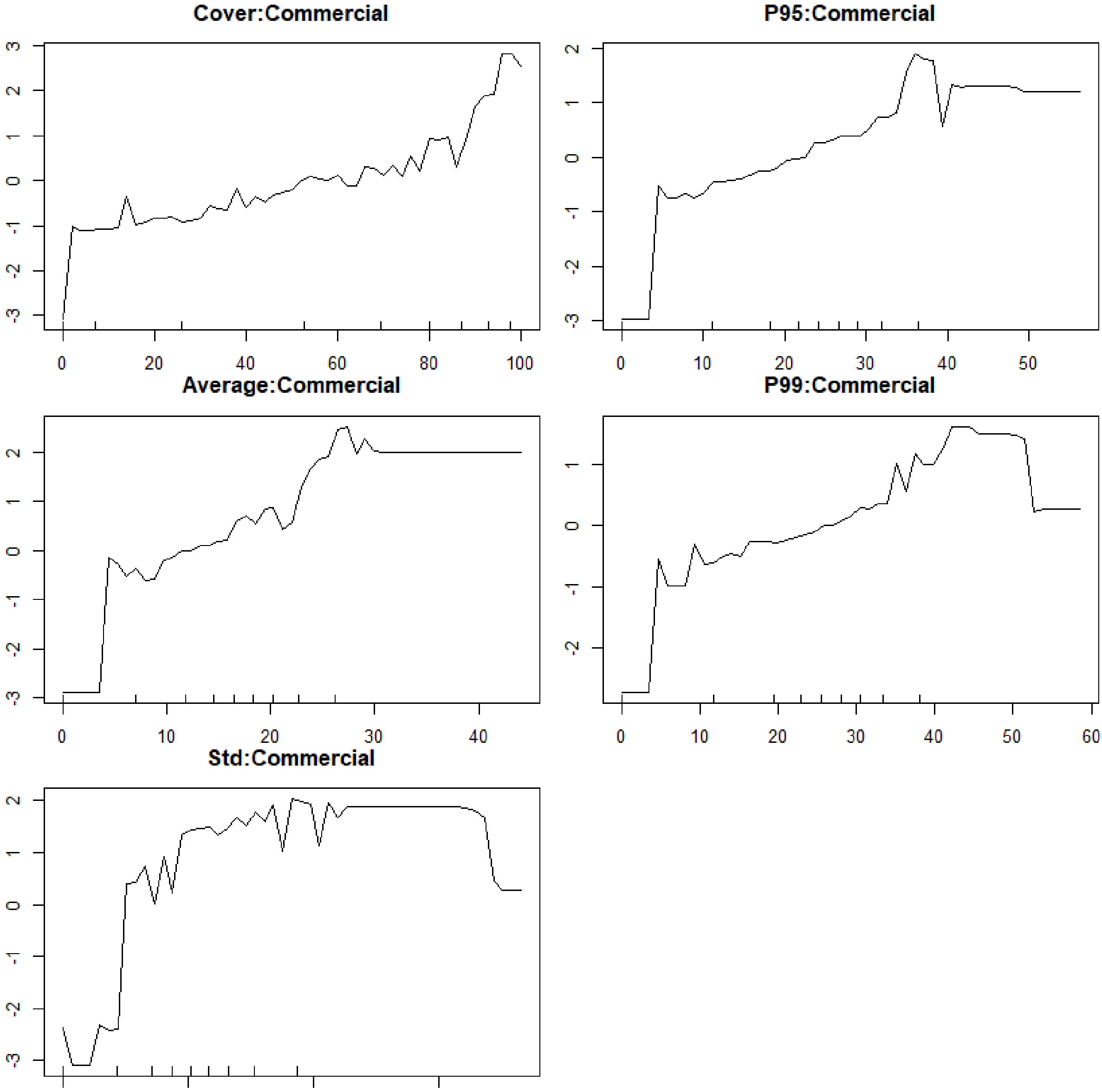

Figure 6: Partial dependence plots showing the effect of the predictors on commercial forest classification.

Two of the most important variables in the classification are p95 and cover. P95 is a surrogate for dominant height. The partial dependence plot of the commercial forest with the two variables is presented in figure 7. The probability of the commercial forest is really low if the cover is low and p95 is less than 10m. The probability of pixel classified as commercial forest is very high for cover greater than 80% and p95 more than 30m. The probability is still very high for cover greater than 20% and p95 greater than 40m.

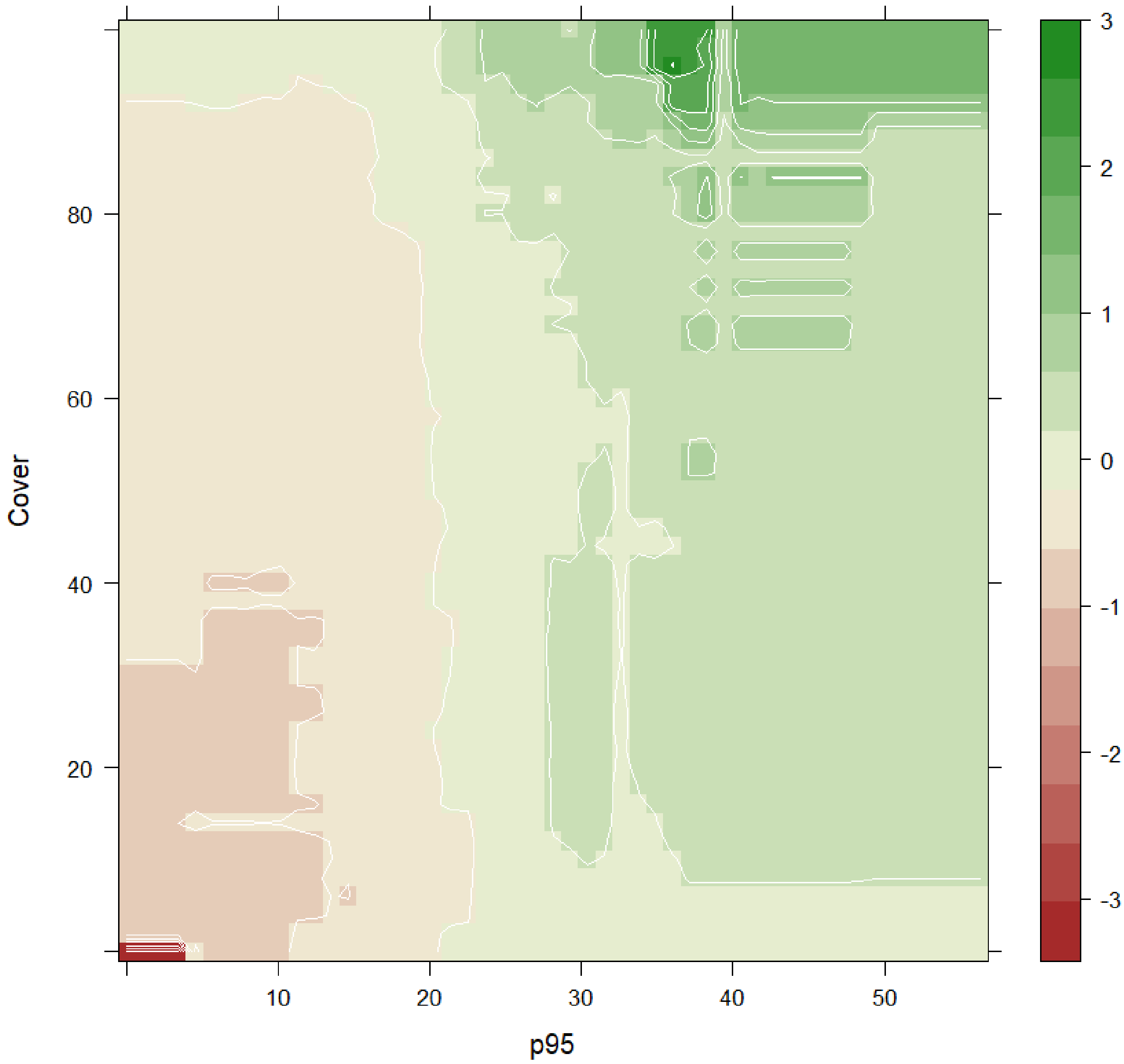

Figure 7: Partial dependence plots of p95 and cover showing the effect of the predictors on commercial forest classification.

Spatial services LiDAR data

After fitting the model using the LLS LiDAR data from 2020, the model was fitted to the spatial services data, as this data is more complete for the north coast. The five variables used for modelling were compared between the two LiDAR data collection methods. The point density for the LLS 2020 data is very high compared to the spatial services data; >40 points per square meter against about 1 point per square meter. The pair wise plots of the five variables from LLS and spatial services for a property is presented in figure 8. It is clear from these plots that the correlation between the two datasets is really high. Correlations between the two measures for p95 was 0.96, for p99 it was 0.96, average height 0.94, standard deviations 0.9 and for cover 0.94. Figure 9 is another such plot for another property. The correlations for all other variables are greater than 0.9 except for cover. The correlation for cover is 0.67, the correlations of cover in the two datasets is not as high as the height matrices, This could be because of harvesting, drought effect or the time the year data was collected along with fire effects.

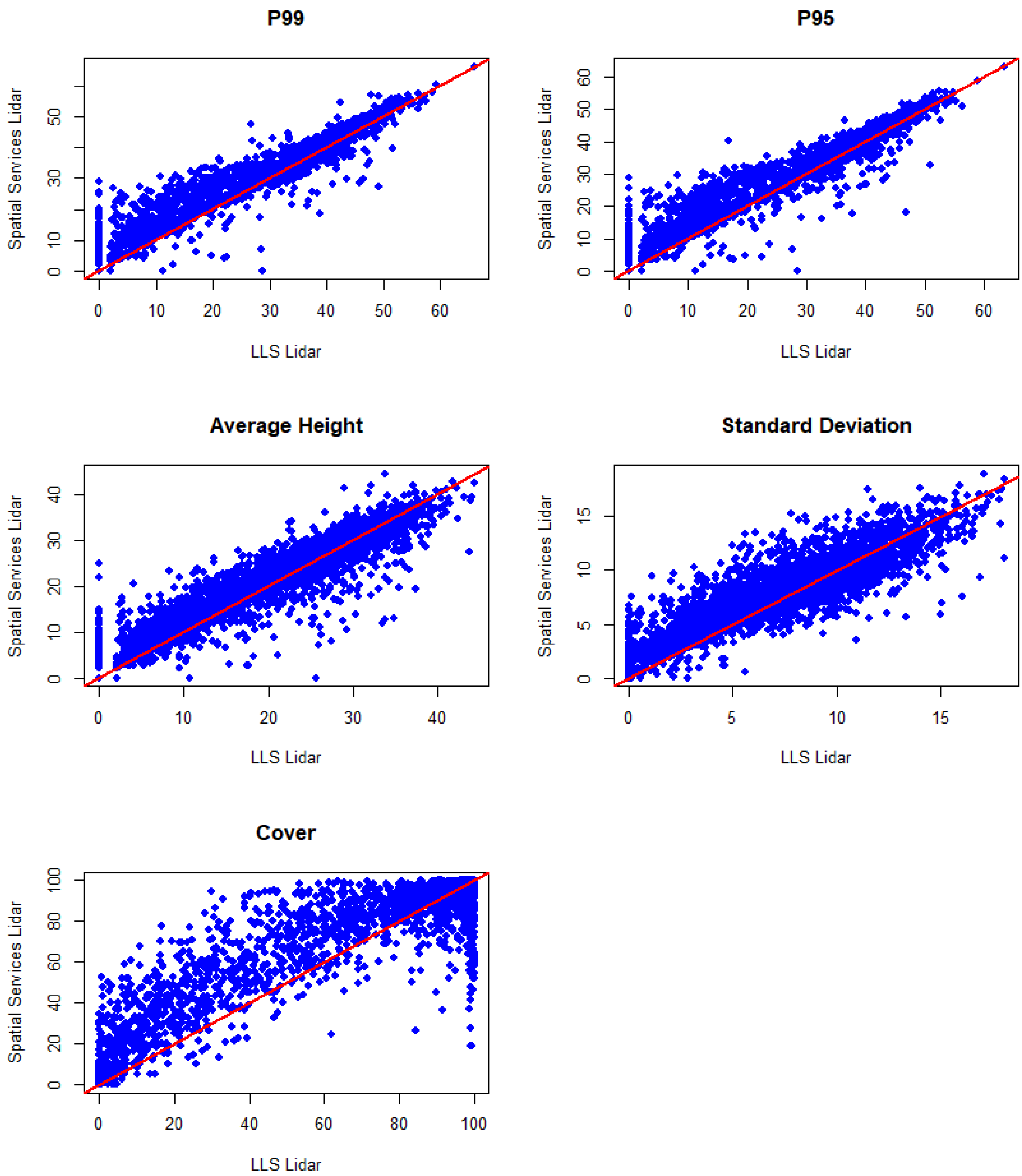

Figure 8: Scatter plot of the five LiDAR matrices using the data for a property from LLS 2020 and spatial services data. The red line is the linear fit with 0 intercept and a slope value of 1. If the relationship between the two variables is perfect then the points would lie on this line.

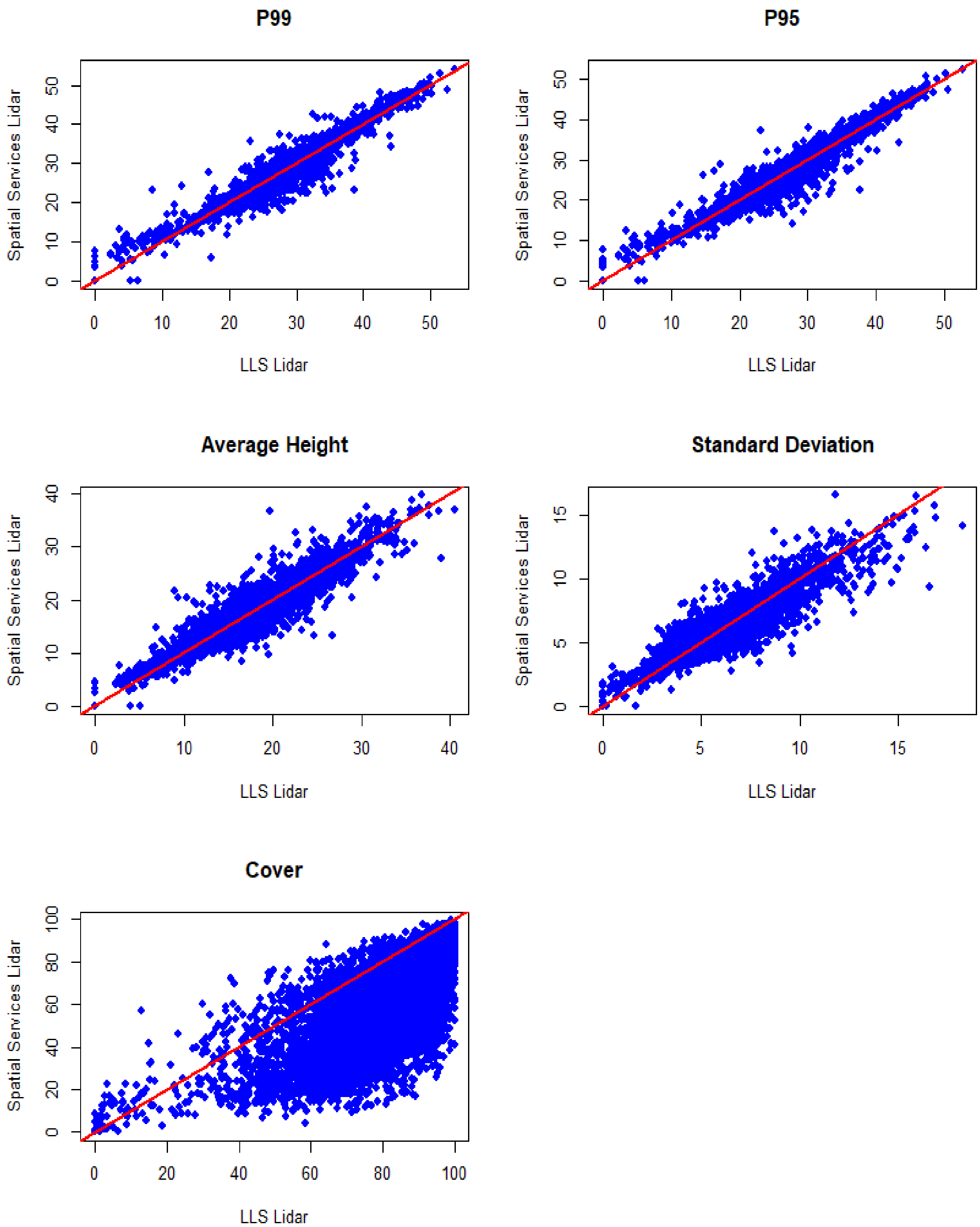

Figure 9: Scatter plot of the five LiDAR matrices using the data for another property from LLS 2020 and spatial services data. The red line is the linear fit with 0 intercept and a slope value of 1. If the relationship between the two variables is perfect then the points would lie on this line.

Accuracy Assessment

Test data was used to check the accuracy of the model. The study area was classified into two classes; areas that are commercial forest and areas that are non-commercial forest. The map showed a good agreement with the reference data, an overall accuracy of 86.1%, 95% confidence interval of 85.0% - 87.0%(Table 2). The kappa value calculated is 0.72. Landis and Koch (1977) have proposed the following as standards for strength of agreement for the kappa coefficient: ≤0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect. So, the agreement between the predicted values from the random forest and reference data is substantial.

Table 2: Accuracy assessment of the map using spatial services data.

	Classes	
	Commercial	Non-Commercial
User's Accuracy	87.8	84.3
Producer's Accuracy	85.3	86.9

Map and area estimates of Commercial and Non-commercial Forest

Map Commercial and Non-commercial Forest

Five LiDAR matrices at 20m resolution were used as predictor variables to produce a map of commercial vs non-commercial forest. A focal filtering is applied using a 5 by 5 cell window. A focal operation applies an aggregation function to all cells within the specified neighbourhood, uses the corresponding output as the new value for the central cell, and moves on to the next central cell. A modal value was used an aggregation method. The results of the smoothing are shown in figure11.

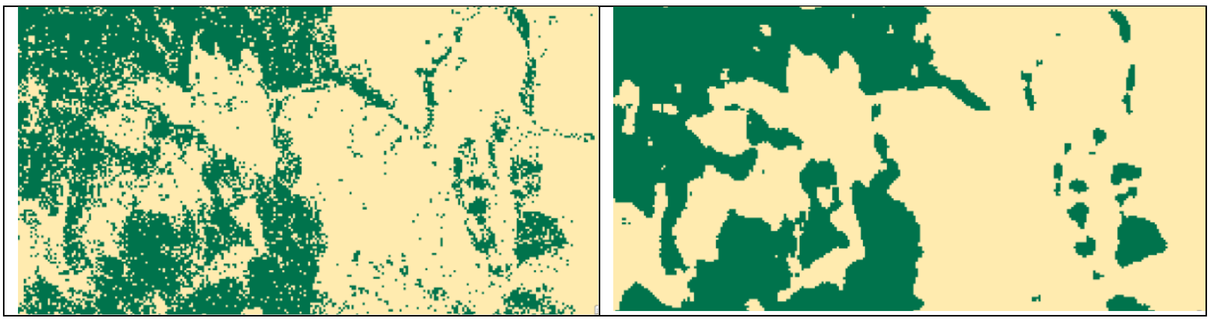

Figure 10: The map on the left is without the smoothing and on the right is the smoothed data.

The model was applied to the data from spatial services for North and South Coast and a map for the commercial vs non-commercial was produced for North Coast NSW, the map is presented in figure 11.

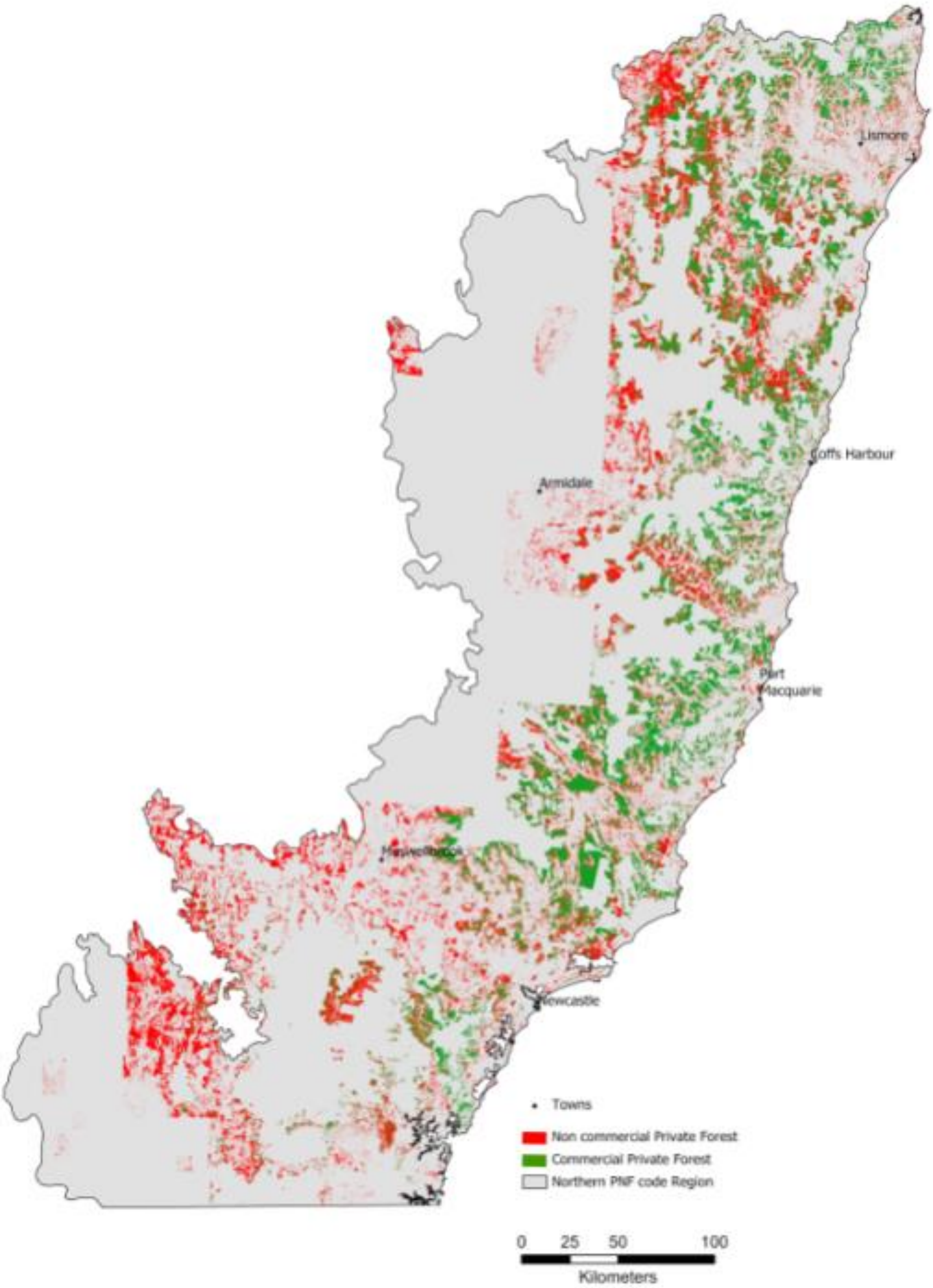

Figure 11: Map of commercial vs non-commercial forest for North Coast NSW PNF .

Area estimates of Commercial and Non-commercial Forest

The area estimates for the commercial and non-commercial PNF forest for North coast were produced using the above map. Total PNF area for North Coast where spatial services LiDAR is available is 2,462,607 ha. Area for commercial and non-commercial Forest is 1,133,697 ha (46%) and 1,328,910 ha (54%).

The model was then applied to the data from spatial services for South Coast and a map for the commercial vs non-commercial was produced for South Coast NSW, the map is presented in figure 12.

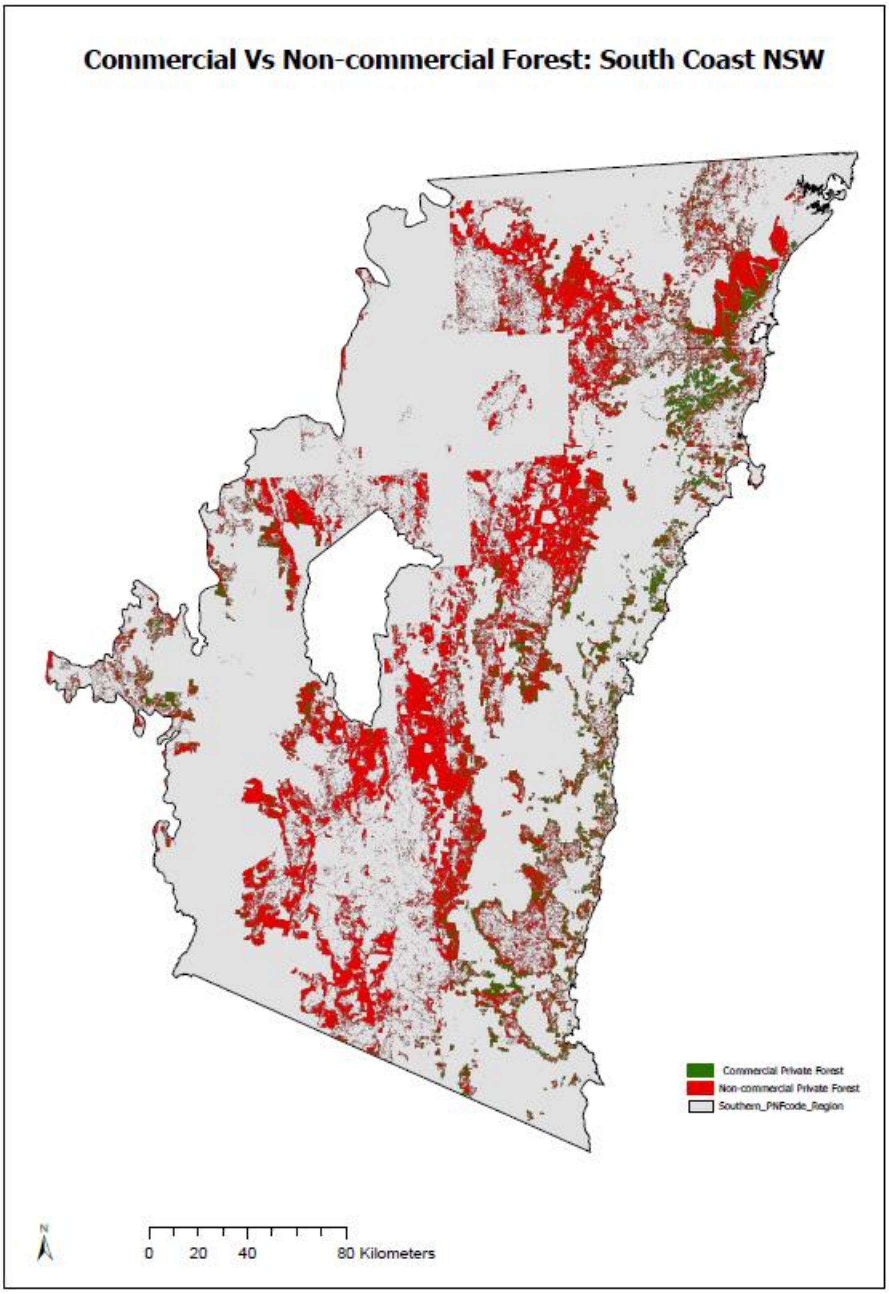

Figure 12: Map of commercial vs non-commercial forest for South Coast NSW PNF .

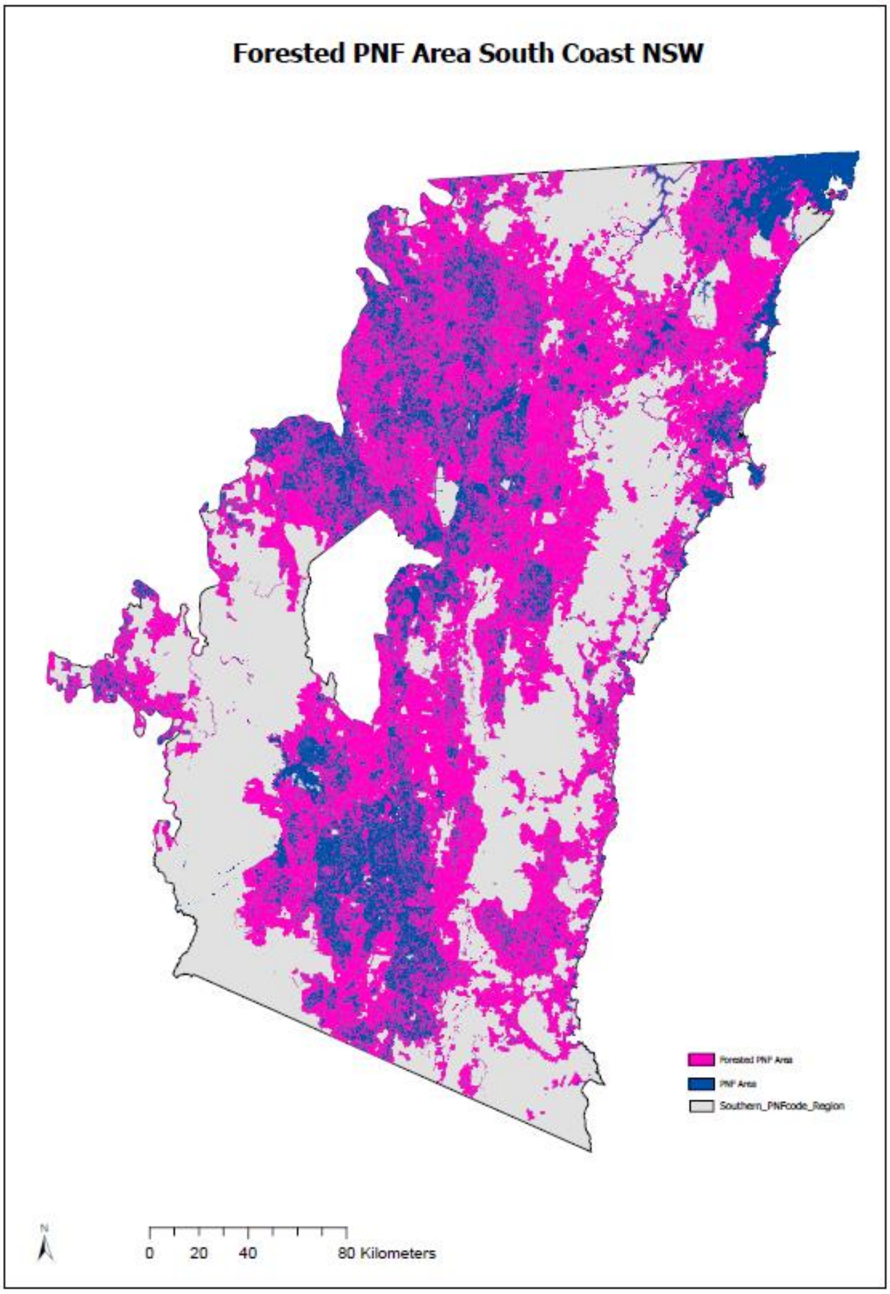

Figure 13: Map of all PNF area and Forested area for South Coast NSW.

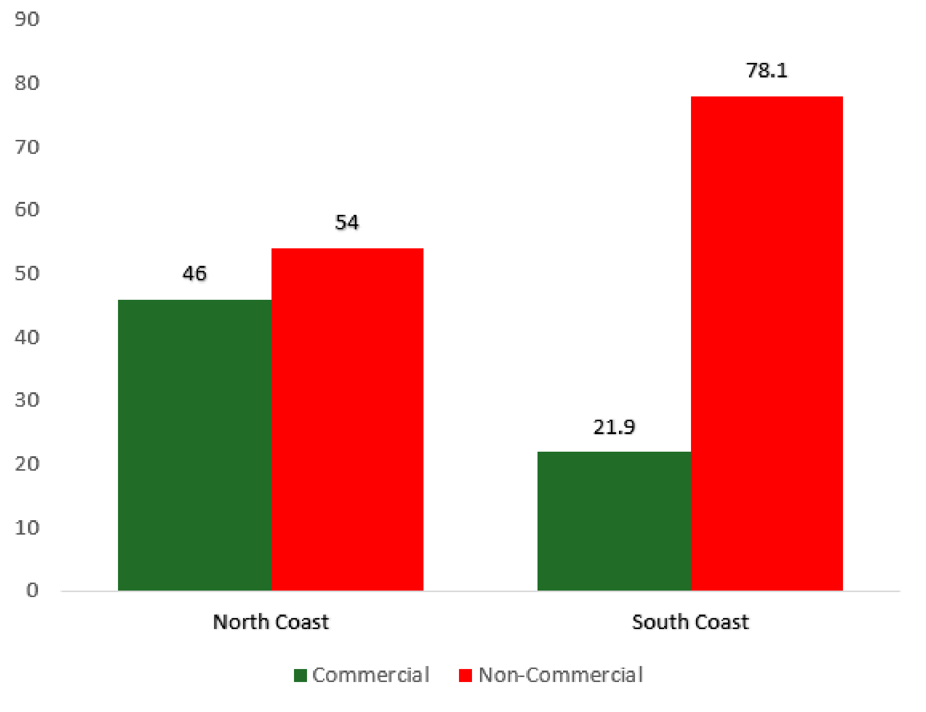

Figure 14: Percentage of commercial vs non-commercial areas covered by spatial services LiDAR for private native forest in North and South Coast of NSW.

Limitations of the Study

Coverage

The spatial services LiDAR layer does not cover the whole of the of PNF properties. The map in Figure 15 shows the spatial location of the properties. The white area is where there is no spatial services LiDAR.

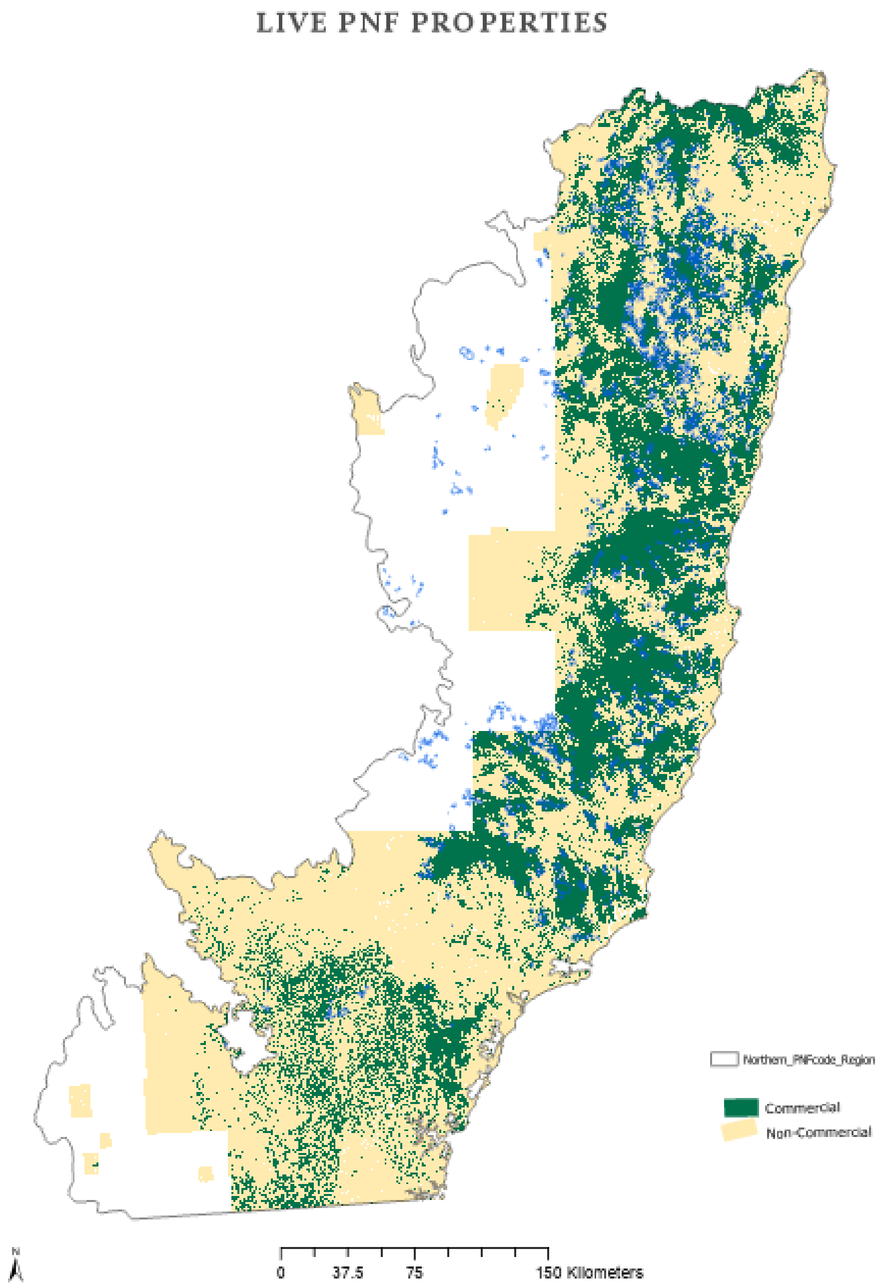

Figure 15: Map of the spatial location of the PNF properties. White area is where there's no LiDAR data.

LiDAR

The LiDAR data used to produce the map is spatial services data, the point density for this data is much less than data collected in 2020 for LLS project. The data was collected over a long period of time, mostly from year 2001 to 2015. So, if harvesting or any other disturbance happened from 2015 to now that is not captured by this map. Spatial services data does not cover the whole of North Coast. A plot of cover and p95 by the different reasons for non-commercial forest is presented in figure 16. As can be seen some of the classes have high cover and high p95 values, such as dense mesic canopy so this type of forest would most likely be classified as commercial forest.

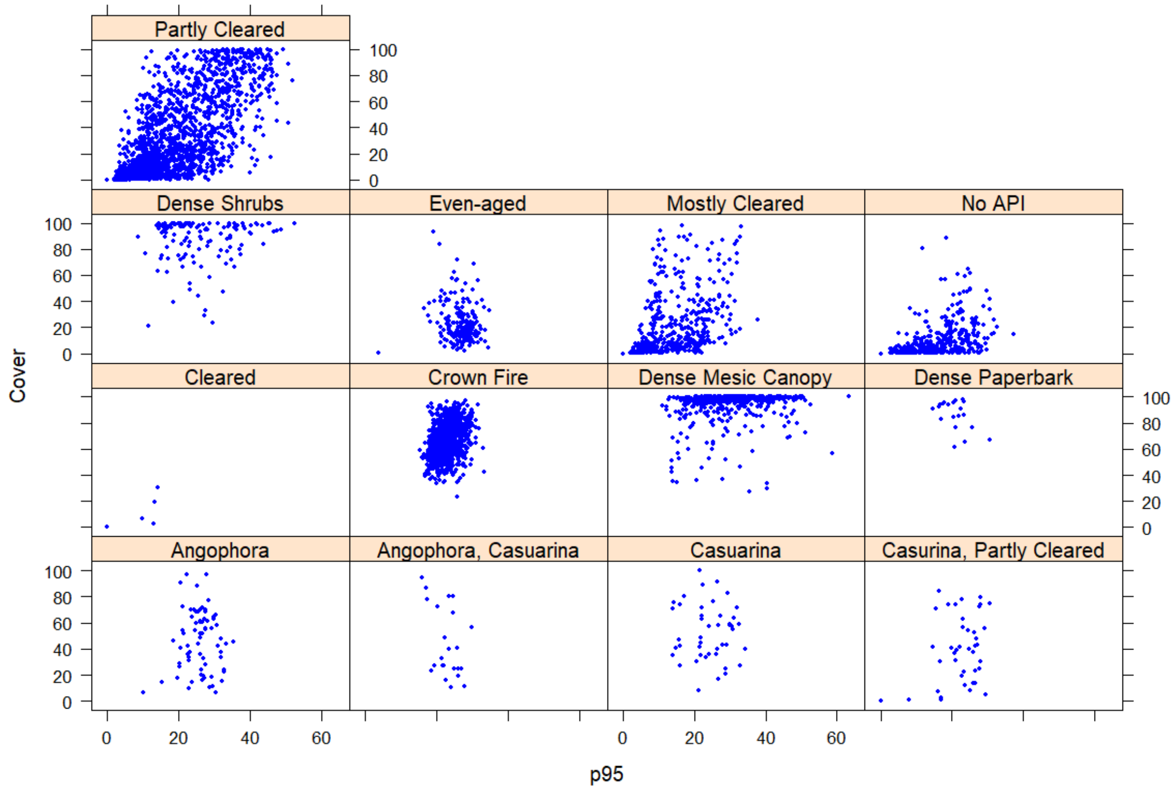

Figure 16: A scatter plot of cover and p95 is by the different reasons for non-commercial forest

Future Work

The commercial vs non -commercial classification might improve by the use of topographic and satellite data. Forest Yield Association mapping is another sub project which should further improve the accuracy of the map as some genera such as Angophora, Casuarina and mesic dense canopy etc. that are classified as non-commercial could be identified in that project and that map could improve the accuracy of classification. Site index and disturbance mapping could also improve this classification. Another refinement of this classification is creating three classes instead of two. The three class classification would be non-commercial, commercial and low quality commercial.

References

Breiman, L. (1994) Bagging predictors. Technical Report 421, Department of Statistics, University of California at Berkeley.

Breiman, L. (2001) Random Forests, Machine Learning 45(1), 5-32.

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas ; 20 37-46.

Friedman, Jerome H.(2001) "Greedy function approximation: A gradient boosting machine." Annals of statistics: 1189-1232

Liaw A and Wiener M (2002) Classification and Regression by randomForest. R News 2: 18-22.

McRoberts, R.E., Wendt, D.G., Nelson, M.D., and Hansen, M.H. (2002). Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. Remote Sens. Environ. 81: 36–44.

R Development Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge.

Wulder, M.A., White, J.C., Nelson, R.F., Næsset, E., Ørka, H.O., Coops, N.C., Hilker, T., Bater, C.W., and Gobakken, T. (2012). “LiDAR sampling for large-area forest characterization: a review.” Remote Sensing of Environment, Vol. 121: pp. 196–209.