Department of Primary Industries

Technical Report for Local Land Services

Sampling Design for Private Native Forestry R&D Project

www.dpi.nsw.gov.au

Published by the NSW Department of Primary Industries

Amrit Kathuria

Title: Sampling Design for Private Native Forestry R&D Project

First published June 2023

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 and John Samuel for technical advice on private native forest related issues and Masoomeh Alaibakhsh for supplying the PNF properties shapefile.

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

Executive Summary

LLS has engaged DPI to design and implement an R&D project that will help address policy, planning needs and information gaps relating to private native forestry. The project seeks to extrapolate forest attribute and canopy change models to the broader private native forest estate using field plots, LiDAR, Sentinel and Landsat satellite imagery. The assessment of inventory measures such as biomass, basal area, stumps and stags, coarse woody debris, standing timber etc are measured with field plots, which are then used to calibrate the models using the remote sensing data. The document outlines the sampling strategy used for this study. Sampling strategy for any program must consider sampling design, sample size estimations and estimation methods both for the estimates and the variances.

The sampling strategy implemented has the following main elements:

- Stratification of the population by size of the property and date PNF plan were authorised, with varying sampling intensity among strata, so that each stratum is adequately sampled for statistical reliability
- For each property selected in the sample, LiDAR data is captured using transects. Within a subset of properties, ground-based plot are selected, where a range of forest structure, composition and inventory measurements are collected. The plot-based system is complemented by remotely sensed data such as Landsat/Sentinel.
- A network of sampling plots across northern NSW.
- Estimation of attributes using design based, model based or hybrid approaches.

Introduction

LLS has engaged DPI to design and implement an R&D project that will help address policy, planning needs and information gaps relating to private native forestry. The project seeks to extrapolate forest attribute and canopy change models to the broader private native forest estate using field plots, LiDAR, Sentinel and Landsat satellite imagery. The assessment of inventory measures such as biomass, basal area, stumps and stags, coarse woody debris, standing timber etc are measured with field plots, which are then used to calibrate the models using the remote sensing data.

Assessing all properties belonging to the target population is generally too costly, which is why sampling techniques are used select representative samples instead. The general principle of sampling is to select a subset from a population and draw inferences from the sample for the entire population. The selection of the most appropriate sampling design is subject to several considerations, respondent burden cost and safety are some of them. To ensure the validity and reliability of the estimates, the sample selected should be representative of the population. A non-representative sample will lead to inaccurate and misleading estimates – a situation which will be unknown as the true estimate of the whole population is unknown. The best way to avoid a biased result is by using scientifically rigorous rules to pick up a representative sample and minimising errors in measuring each sample. “Design-unbiasedness” and “sampling precision” are often used to summarize the characteristics of a good statistical sampling design (Dumais and Gough 2012). Sampling strategies developed for this study ensure that: (1) probability sampling design; this ensures that both the design based and model based approaches could be used for estimation; (2) given the cost consideration, population estimates are close to what a complete census would have given and precise enough for the intended purposes; (3) the data could be used for the model based inference using LiDAR and satellite data; (4) variance estimation can be done; and (5) the sample is spatially well distributed.

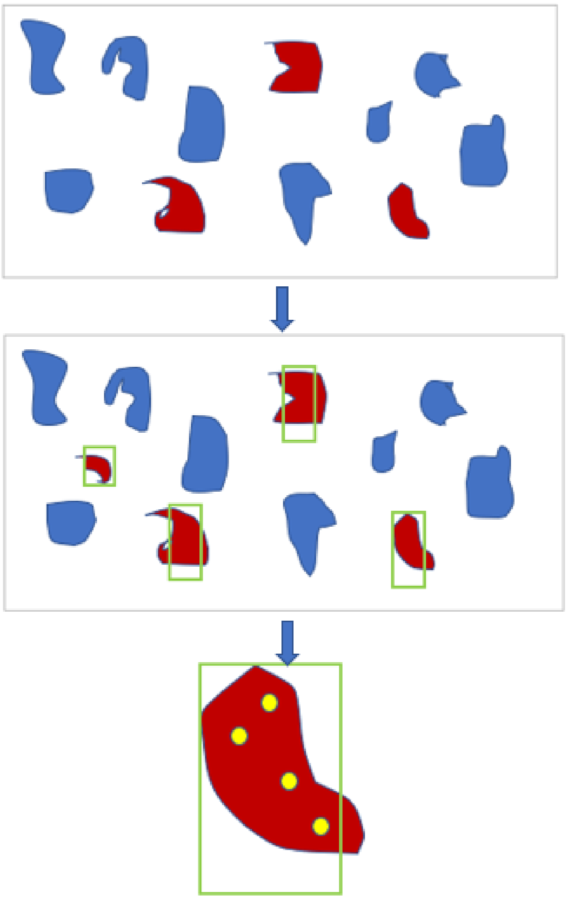
The sampling strategy implemented has the following main elements:

- Stratification of the population by size of the property and date PNF plan was authorised, with varying sampling intensity among strata, so that each stratum is adequately sampled for statistical reliability
- For each property selected in the sample, LiDAR data is captured using transects. Within a subset of properties, ground-based plot are selected, where a range of forest structure, composition and inventory measurements are collected. The plot-based system is complemented by remotely sensed data such as Landsat/Sentinel.
- A network of sampling plots across northern NSW.
- Estimation of attributes using design based and model based.

Sampling Design

The project is focused on the region covered by the PNF Code of Practice for Northern NSW (loosely described as the north coast). This region encompasses ~11M hectares of land and ~3.4M hectares of private native forest. The project seeks to specifically examine private native forests that are being subject to timber harvesting (i.e. those that have a 'live' PNF Plan) in the Northern Code region where ~ 80% of all PNF activity occurs.

A three-stage probability sampling design is used to select the properties from the NSW's North Coast area. In a multi-stage sampling design, the sample units at each stage are sub-sampled from the sample units selected at the previous stage (Dodge, 2006). Figure 1 presents the sampling design. The three stages used in PNF study design are described below:



First Stage: Primary Stage units (PSU): 2705 properties with approved PNF plan in November 2020
Blue are properties in the population and the red are the selected properties.

Second Stage: Secondary Stage units (SSU): 256 selected properties with Lidar transects (light green)

Third Stage: Tertiary Stage units (TSU): Field plots on subset of selected properties for inventory and forest monitoring variables measurement.

Figure1: Schematic diagram for the three-stage sampling design used for the PNF R&D project.

First Stage

First-stage units are commonly referred to as primary sample units. As at March 2020, there were 3,185 individual properties authorised for timber harvesting in the Northern NSW region (Figure2). 480 (15%) of these properties were less than 15 hectares in size. Area of <15 ha properties 4,012.6 ha, and total area of 3185 properties is 455,554.2 ha. These small properties were removed from the sample population as they accounted for only 0.9% of the authorised forest. This left a sample population of 2,705 properties. A stratified sampling design is used at selection of properties at the first stage. In stratified sampling, the population is partitioned into non-overlapping groups, called strata and a sample is selected by some design within each stratum. The 2,705 properties were stratified according to their size and the date that they were authorised. There were four size strata: 15-50 ha, 50-200 ha, 200-500 ha and >500 ha. The date the plan was approved was the second strata with three classes: 2019-2020, 2016-2019 and pre 2016. So, there are 12 strata, table1 summarises the number of properties in each strata.

Stratification generally increases the precision of the population estimates. The second reason for the stratification is the avoidance of bias as there is better representation of the population in smaller subpopulations (Cochran, 1977). Another reason for stratification is to have estimates of population parameters for groups within the population, e.g. comparison by size or year the plan was approved. A simple random sample is then drawn from each stratum.

The number of samples in each of the strata are selected by using number of properties in the strata, area of the properties in the strata and the year the plan was approved. More samples were selected from larger properties with the PNF plan that was recently approved as it was hypothesised it is more probable that the properties harvest soon after the plan is approved.

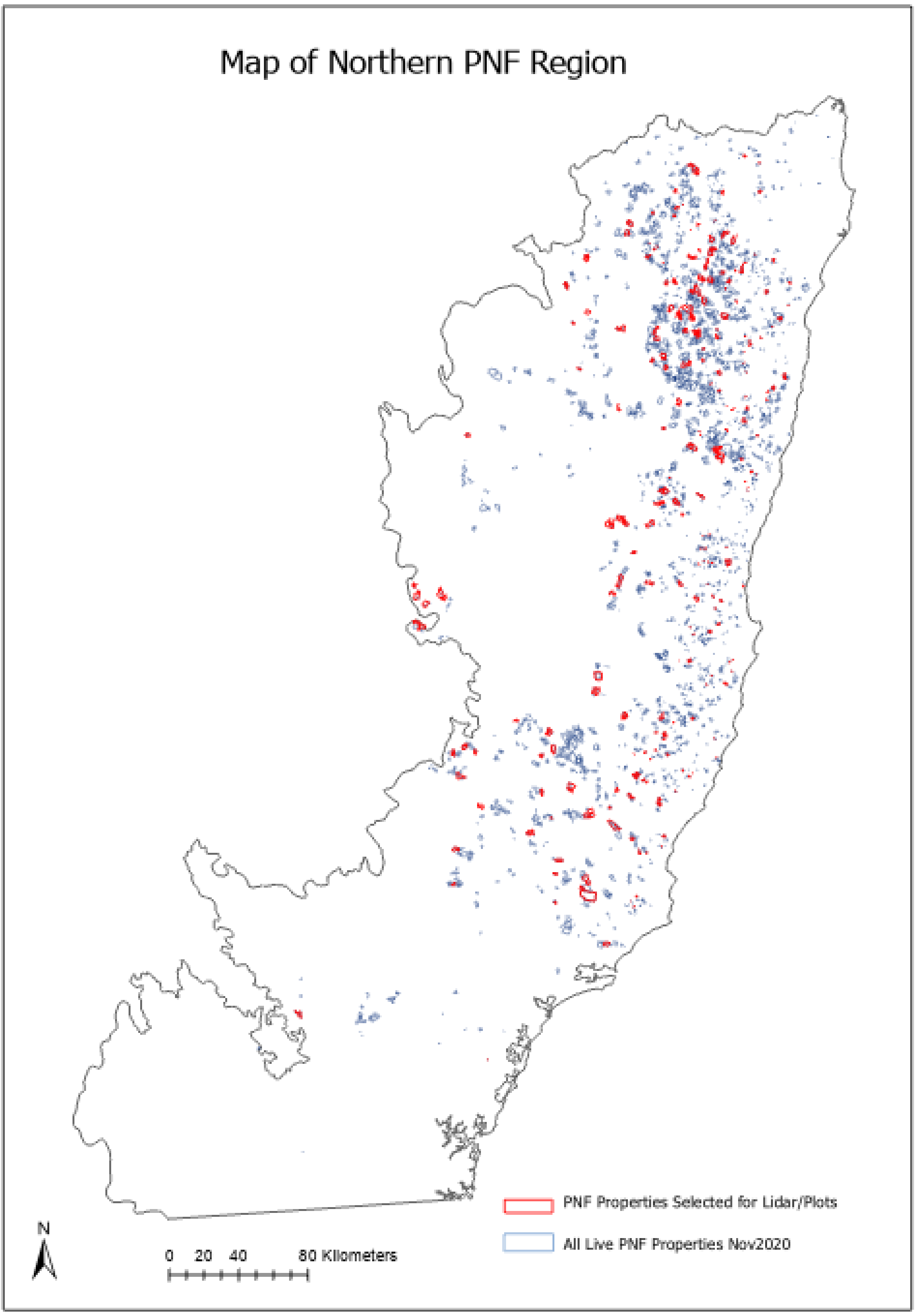


Figure 2: Map of North Coast NSW Private Native Forest area, all the properties with PNF plan on Nov2020 (blue) and the selected properties in the sample (red).

Table 1: Summary of the number and area under the properties in the population and in the sample

Area Strata	Year Strata	No of properties	No of sampled properties	Area of properties (ha)	Area of sampled properties (ha)
15 to 50 ha	Pre 2016	804	21	25264	732
15 to 50 ha	2016-2019	278	20	8874	636
15 to 50 ha	2019-20	79	40	2559	1287
15 to 50 ha	Total	**1161**	**81**	**36696**	**2655**
50 to 200 ha	Pre 2016	699	22	71291	2157
50 to 200 ha	2016-2019	220	22	22349	1974
50 to 200 ha	2019-20	46	22	4522	2108
50 to 200 ha	Total	**965**	**66**	**98162**	**6239**
200 to 500 ha	Pre 2016	280	21	86848	5877
200 to 500 ha	2016-2019	88	24	26597	7116
200 to 500 ha	2019-20	19	19	6048	6048
200 to 500 ha	Total	**387**	**64**	**119493**	**19041**
>500 ha	Pre 2016	156	21	169414	17363
>500 ha	2016-2019	30	18	22932	12442
>500 ha	2019-20	6	6	4845	4845
>500 ha	Total	**192**	**45**	**197191**	**34650**
Total		**2705**	**256**	**451542**	**62585**

Second Stage

For each of the 256 properties selected in the first stage data from LiDAR transects was collected. The properties in red in figure 2; are the selected properties. There is no apparent clustering of selected properties; they are spatially well distributed. The transect were drawn such that we have the maximum area covered from each of the property, making sure that most of the property is covered by LiDAR data.

The airborne capture of LiDAR and orthophotography using a Riegel sensor covers ~750 metres(m) of land per pass (swath). Properties between 15 and 200 ha in size were flown over once generating a 750 metre wide transect (Figure3). Properties 200 -500 ha in size had two flight transects that overlapped by ~30% (total width of ~1,275 m). Large properties >500 ha had a minimum of three overlapping transects (total width of ~1,950 m).

The sample area captured by LiDAR and orthophotography represents 10% of the properties forests that have a 'live' PNF plan within the Northern NSW PNF Code of Practice region (and 8% of the properties by number).

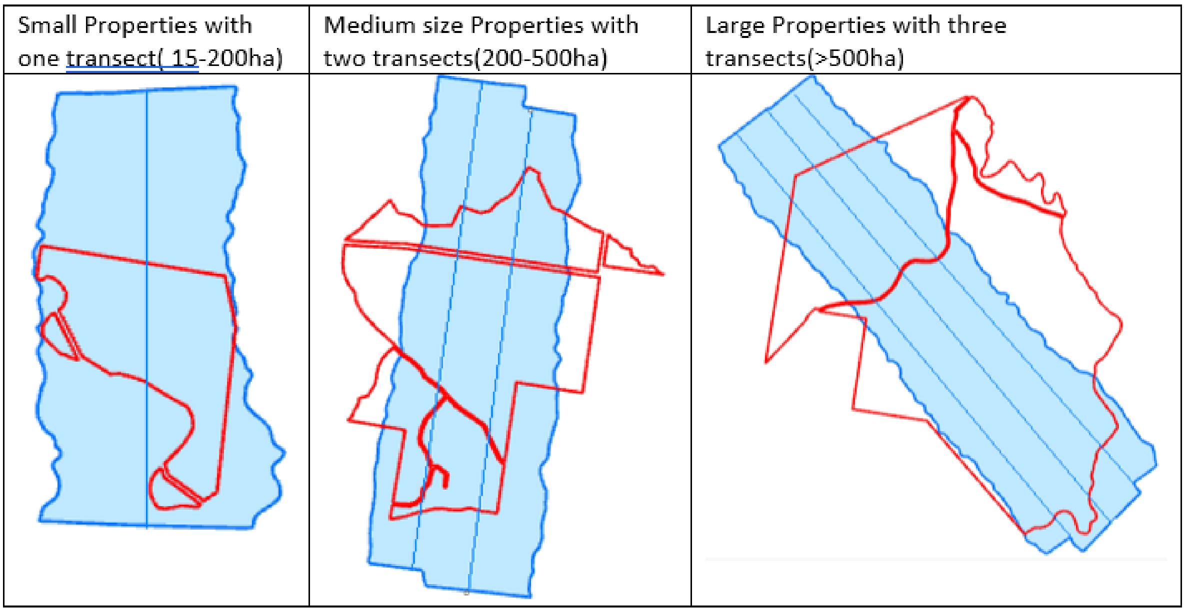


Figure 3: The orientation of LiDAR transects (examples).

Third Stage

The third stage sampling units are the tertiary units. The plot level field data for the inventory variables was collected at this stage. A random selection of properties is the ideal but as property owners' consent is needed to collect plot information the plots could only be collected from the owners who responded positively for being in the field survey. Table 2 presents the distribution of 34 properties which were available for plot sampling.

Table 2: The distribution of properties by Year of PNF approval and Size class.

	LT50	50to200	200to500	GT500	Total
2019-20	12	2	2	2	18
2016-2019		2	3	3	8
LT2016		1	3	4	8
Total	12	5	8	9	34

The spatial distribution of the properties is presented in map below. The spread of the properties spatially is quite good (figure 4) except for the middle region from Kempsey to Bellingen where no property is available for survey.

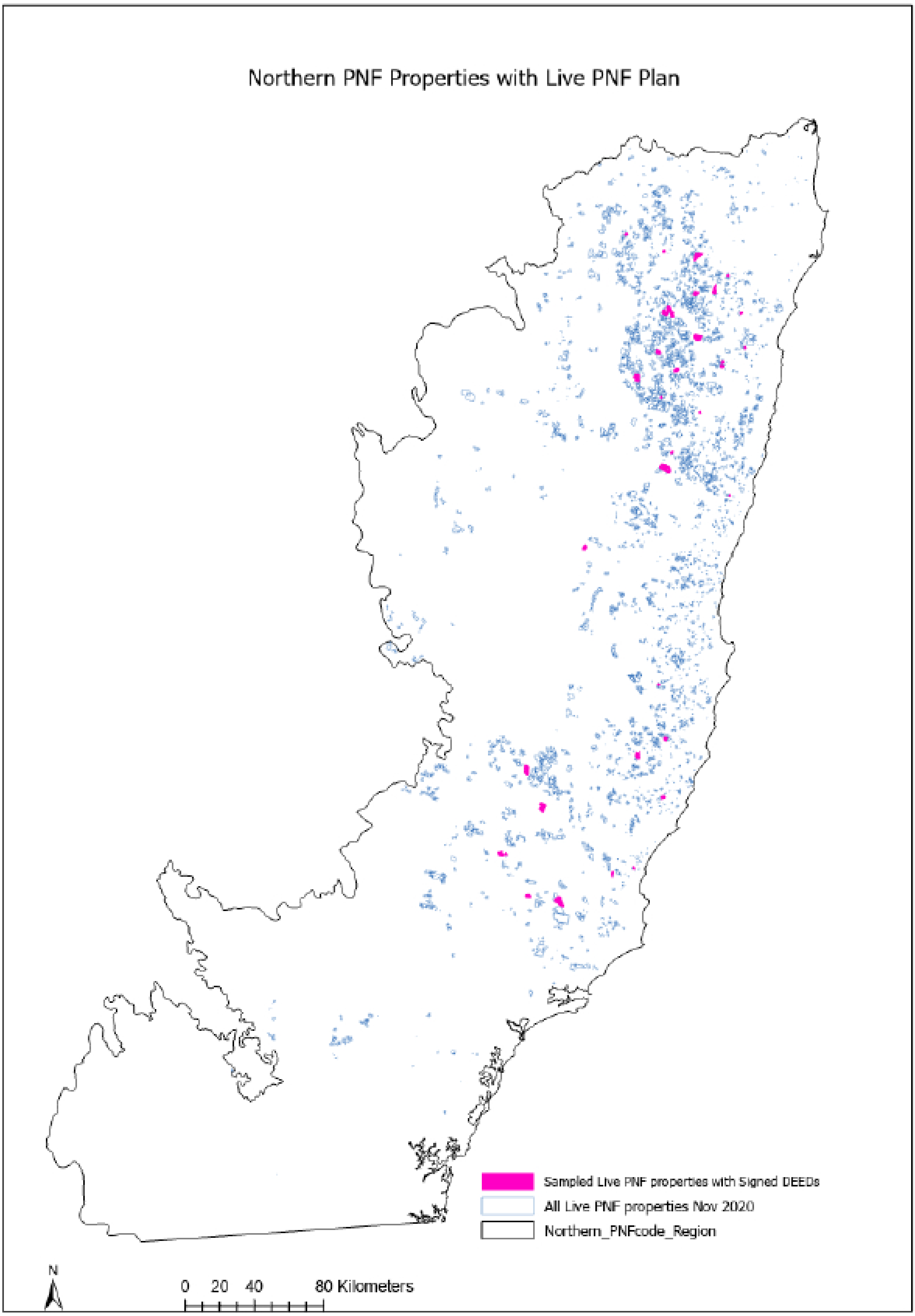


Figure 4: Map of North Coast NSW, all the properties with PNF plan on Nov2020 (blue) and the selected properties with signed deed (fuscia).

All together a sample of 60 plots is to be collected for inventory sampling. The population is commercial forest in PNF properties. YAG map created was used as strata. YAGs are yield association groups defining forest types. The plots were placed in the non-exclusion areas of the property. Also, the plots were placed in the area of the property which is covered by the LiDAR transect. The plots that are on the edge of two YAG classes where moved slightly.

The sample sizes for the plots were estimated based on the % of area in each YAG. YAG information was missing from some of the Northern Tableland properties, adjustment of the area in these YAGs was done based on the Northern Tableland YAG for the whole of North Coast. Also, Blackbutt, Spotted Gum and Moist Coastal YAGs are very important (Nick Cameron, personal communication), so they were given more weight. A minimum sample size of 5 was chosen and 2 for the very small YAG classes such as Swamp Sclerophyll and Viney invasive scrub. So, the column 'Modified to keep a minimum sample in strata' is the final sample size from each of the YAG strata, giving a total of 60 samples. The samples are to be collected over a period of three years.

Table 3: Proportion of area in commercial, non-commercial forest for all of north coast and for properties with PNF plan.

	Commercial Forest				
YAG	PNF properties	North Coast	Modified Area % for Table land YAGS	Sample size based on proportional allocation	Modified to Keep a minimum sample in each strata
Blackbutt	1.1%	2.0%	1.1%	1	5
Spotted Gum	18.6%	8.6%	18.6%	11	13
Moist Coastal Eucalypts	11.6%	13.0%	11.6%	7	9
Semi-moist and Taller Dry Eucalypts	31.6%	28.8%	31.6%	20	13
Dry Sclerophyll	19.3%	14.8%	19.3%	12	8
Tableland Eucalypts - Moist	6.6%	12.8%	9.7%	6	5
Swamp Sclerophyll	1.3%	1.5%	1.3%	1	2
Tableland Eucalypts - Dry	2.2%	3.1%	2.6%	2	3
Viney invasive scrub / degraded	1.3%	1.9%	1.3%	1	2

A good surrogate for dominant height is the 90th percentile of height (p90) using LiDAR data. The distribution of the dominant height is presented below for the commercial forest in PNF properties. The dominant height of the selected plots for 2,021 sample is presented in figure 5. The histogram is the dominant height for commercial forest of all PNF properties and the red dotted lines are the dominant heights of the 2,021 selected plots. In the plot data a good spread of dominant heights is represented.

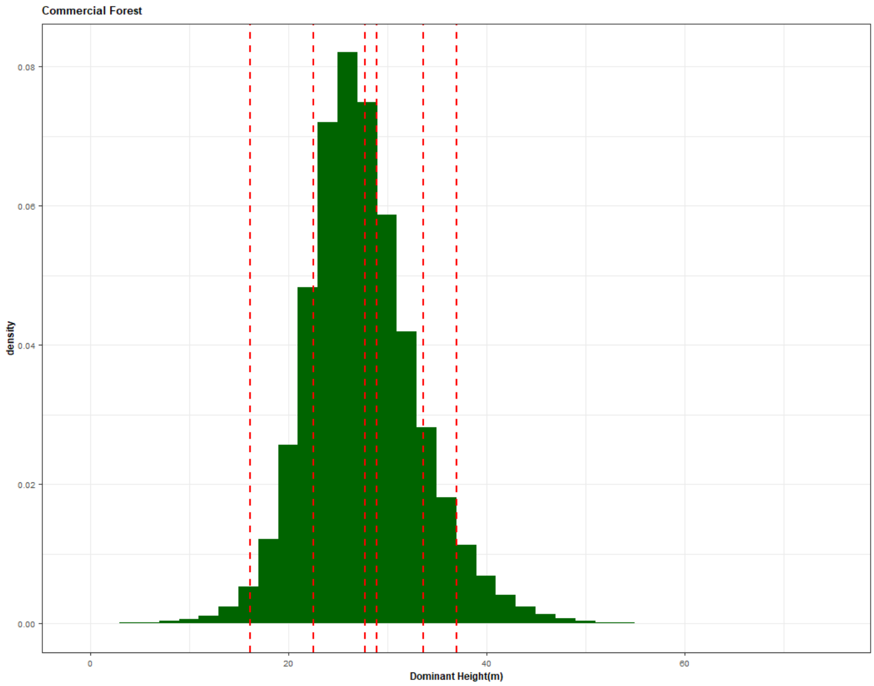


Figure 5: The histogram of dominant height for commercial forest of all PNF properties and the red dotted lines are the dominant heights of the 2,021 selected plots.

Estimation methods

Design Based Inference

Design-based inference typically assumes a finite population of elements to which one or more fixed target quantities are linked. The objective normally is to estimate some fixed population parameter, such as the total or the mean of these quantities (Gregoire and Valentine 2008). In order to estimate the fixed but unknown parameters a probability sample is selected from the population according to some appropriate sampling design, which assigns positive inclusion probabilities to each element. Mathematical formulas (estimators) are used for estimating the parameters based on the sample data.

Estimation at LiDAR transects level

N is the total number of properties in the population (2,705)

N_h Number of properties in each stratum $\Sigma N_h = N$, h is for strata varying from 1 to L strata

n is the number of properties selected in the sample

n_h is the sampled properties from each strata $\Sigma n_h = n$

y_{hi} is the variable of interest for the ith unit in the hth stratum

Estimation of the mean

$$\bar{y} = \frac{\sum_{i=1}^{L} N_h \bar{y}_h}{N} = \sum_{h=1}^{L} W_h \bar{y}_h$$

Where $W_h = \frac{N_h}{N}$

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{i=1}^{n_h} w_{hi}}$$

where w_{hi} is

$$w_{hi} = \frac{A_{hip}}{A_{hiL}}$$

A_{hip} is the area of the ith property in the hth stratum, A_{hiL} is the area under LiDAR transect of the ith property in the hth stratum.

w_{hi} is the inverse of fraction of the property covered by LiDAR transect.

Variance of the mean

$$V(\bar{y}) = V(\sum_{h=1}^{L} W_h \bar{y}_h)$$

$$V(\bar{y}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) + 2 \sum_{h=1}^{L} \sum_{j>h}^{L} W_h W_j \, Cov(\bar{y}_h \bar{y}_j)$$

As all the samples are drawn independently in the different strata all covariance terms are 0. So, the variance of $\bar{y}$ is

$$V(\bar{y}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$$

Cochran (1977) outlines the formula for the variance for the strata mean when sample in each stratum is selected at random

$$V(\bar{y}_h) = \frac{S_h^2}{n_h}\frac{(N_h - n_h)}{N_h}$$

Here

$$S_h^2 = \frac{\sum_{i=1}^{n_h} w_{hi}(y_{hi} - \bar{y}_h)^2}{\sum_{i=1}^{n_h} w_{hi}} \times \frac{n_h}{n_h - 1}$$

$$V_h = \frac{S_h^2}{n_h} = \frac{\sum_{i=1}^{n_h} w_{hi}(y_{hi} - \bar{y}_h)^2}{(n_h - 1)\sum_{i=1}^{n_h} w_{hi}}$$

So the variance for the mean is

$$V(\bar{y}) = \sum_{h=1}^{L} W_h^2\, V_h \frac{(N_h - n_h)}{N_h}$$

The formula for 95% confidence interval is

$$\bar{y} + 1.96\sqrt{V(\bar{y})}$$

Estimates of population totals $N\bar{y}$ and the variance of the total is $N^2V(\bar{y})$.

Resampling Approaches for Variance Estimation

For complex sample data, sampling variance could be estimated using replication techniques such as Balanced Repeated Replication (BRR), Jackknife Repeated Replication (JRR) and the Bootstrap methods. In each of these methods, multiple replicates (also called subsamples) are drawn repeatedly from a full sample according to a specific resampling scheme. The original weights for each replicate are then modified for the primary stage units to create replicate weights. The statistics of interest are estimated by using the replicate weights for each replicate. Variability among the estimates derived from these replicates is an estimate of the variance. Let y be the value of our statistic as calculated from the full sample; let y_i(i= 1,...,n) be the corresponding statistics calculated for replicates samples (n is the number of replicate samples). The estimate for the sampling variance of the statistic is the average of (y_i

− y)^2. Wolter (1985) and Lohr (2010) outline the necessary computational formulas for BRR. Variance formulas for the delete-1, jackknife method is given by Shao and Tu(1995), Berger, (2007) and Cambell(1980).

With improvements in computational flexibility and speed, jackknife and bootstrap methods for sampling error estimation and inference have become more common (Rao and Wu, 1988). These methods provide unbiased estimates of the sampling error arising from complex sample selection procedures. The estimates capture the effects of stratification, clustering, and unequal probabilities of selection.

Model-Based Inference

For Model-based inference, the first step is to formulate a statistical model. Field data using plots is still needed for calibration and validation of the model. Auxiliary data that is correlated with the variable of interest (biomass, volume, carbon, stems per hectare etc.) is essential for the construction of a model. Remote sensing data could be used as auxiliary data. The field data is used to develop a model, and that model can then be used to make predictions for the whole of the area rather than just estimates of totals or mean values as is estimated by a design-based approach. Figure 6 is a diagrammatic representation of the model-based approach that can be used for producing inventory variable maps for this study.

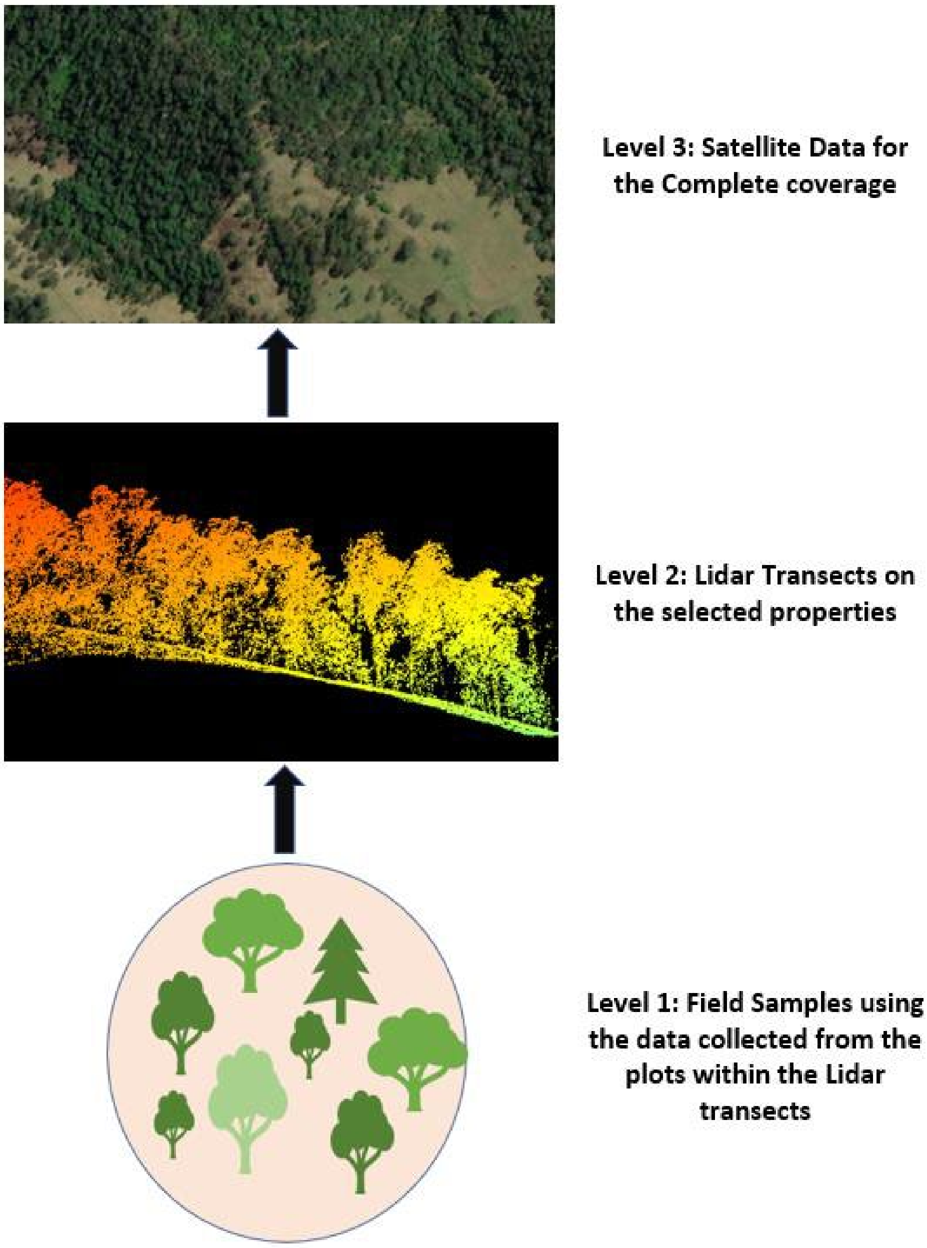


Figure 6: Hierarchical model representing the steps in model-based estimation for this study.

Variance estimation for purposes of quantifying precision is often complex and computationally intensive, particularly when estimates are obtained from remote sensing-based maps. In fact, parametric estimators in the form of simple equations may not be available or may be very difficult to incorporate into estimation software. An increasingly popular alternative is to use resampling estimators such as the jackknife or the bootstrap (Efron & Tibshirani, 1994). However, caution is advised, because these resampling estimators quantify only the portion of uncertainty resulting from sampling, not the uncertainty resulting from model misspecification or lack of fit. In addition, the resampling procedures must mimic the original sampling features such as clustering.

References

Berger, Y. G. (2007): A jackknife variance estimator for unistage stratifed samples with unequal probabilities. Biometrika, 94 (4), pp. 953 - 964.

Cochran, W. G. (1977). Sampling techniques (3rd ed.). New York: John Wiley & Sons.

Efron, B. and Tibshirani, R.J. (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC, UK.

Corona P., Fattorini L., Franceschi S., Scrinzi G., Torresan C. (2014) Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: model-based, design-based, and hybrid perspectives. Can J of For Res 44:1303– 1311.

Dodge, Y. (Ed.). (2006). The Oxford dictionary of statistical terms. New York: Oxford University Press Inc.

Dumais, J., & Gough, H. (2012). School sampling methodology. In V. Greaney & T. Kellaghan (Eds.), Implementing a national assessment of educational achievement (Vol. 3). Washington, DC: The World Bank.

Gregoire, T.G., and Valentine, H.T. 2008. Sampling strategies for natural resources and the environment. Chapman & Hall/CRC, Boca Raton, Fla.

Lohr, S. L. (2010). Sampling: Design and Analysis. 2nd ed. Boston: Brooks/Cole.
Rao, J.N.K & Wu, C.F.J. (1988.), "Resampling inference with complex sample data," Journal of the American Statistical Association, 83, pp. 231-239.

Saarela, S., Sören Holm, Anton Grafström, Sebastian Schnell, Erik Næsset, et al., 2016. Hierarchical model-based inference for forest inventory utilizing three sources of information. Annals of Forest Science, Springer Verlag/EDP Sciences, 73 (4), pp.895-910.

Shao, J. and Tu, D. (1995): The jackknife and bootstrap. New York: Springer.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York.