Survey designs which maximize efficiency gains in ALS-based forestry plot imputation

Gavin Melville[1], Christine Stone[2], Jan Rombouts[3]

[1]Trangie Agricultural Research Centre, Mitchell Hwy, Trangie 2823, Australia
[2]NSW Department of Primary Industries Forest Research, Level 12, 10 Valentine Ave, Parramatta 2124, Australia
[3]One FortyOne Plantations, 152 Jubilee Hwy E., Mt Gambier 5290 Australia

*Corresponding author: Gavin.Melville@dpi.nsw.gov.au

ABSTRACT

The use of airborne laser scanner (ALS) data to estimate forest resource inventory variables is now becoming widespread in Australia (Rombouts et al., 2010; Stone et al., 2011). ALS data is combined with survey plot data to construct model-based estimates of timber volume. In particular, volumes of timber products, sourced from plantations of *Pinus radiata* have successfully been estimated using nearest neighbor methods. One of the challenges in this approach is to construct samples which capture the full efficiency gains which are achievable using the multiplicity of variables which can be derived from the ALS data. Some of the sampling design approaches that have been investigated include random sampling, grid sampling, stratification, systematic selection and balanced sampling. Estimates have been examined from both a design-based and model-based perspective (Melville et al., 2015). This talk will present results based on several approaches including a novel method specifically designed for imputation which optimizes the survey design by using the distance properties of the sample in the space defined by the auxiliary variables.

I INTRODUCTION

The use of remote sensing to measure key inventory metrics over large areas of forest has become widespread in recent years (e.g. Breidenbach et al., 2010; Latifi et al., 2010; McRoberts, 2012; Hudak et al., 2014; Rombouts et al., 2014; Dash et al., 2015). One of the remote sensing techniques currently being employed in Australian softwood plantations is LiDAR, also referred to as airborne laser scanning (ALS). It has been established that the use of LiDAR data in conjunction with ground measurement leads to efficient prediction of commercially important attributes including timber volume, basal area and stems per hectare (Rombouts et al., 2010).

LiDAR and/or other auxiliary data have been used as covariates in a range of prediction methods including imputation, linear regression and machine learning methods such as random forest. With imputation, for example, forest plots which have been measured on the ground are linked to non-measured plots according to their similarity in the covariate space defined by the auxiliary variables. Specific information relating to key attributes, such as timber volume, is assigned to non-measured plots from the most closely related measured plots as defined by the similarity structure.

Measurement of field plots is a time consuming and costly process requiring specialized inventory crews who often need to access remote and challenging terrain. In order to be effective the field plots must cover the full range of variability in the key attributes. Therefore, appropriate

methods of selecting the field plots are essential in constructing imputation estimates which are efficient, robust and economical.

There are a variety of sampling designs which can lead to a good spread in the attributes of interest. These include methods which are focused on structural attributes such as stratification, and spatially systematic designs such as grid sampling. Methods have also been developed which utilise the auxiliary data explicitly including stratification based on LiDAR variables (Hawbaker et al., 2009), multivariate methods such as sensor-directed response surfaces (SRDS - Lesch, 2005), and methods employing geographical coordinates such as generalized random tessellation stratified sampling (GRTS - Stevens and Olsen, 2004). More recently methods have become available which construct a sample which is simultaneously balanced with respect to multiple design variables and these have also been applied to forest inventory (Grafström et al., 2014). The present study investigates a new type of sample, termed nearest centroid (NC - Melville and Stone, 2016), which uses a multivariate clustering algorithm to select a sample of field plots.

In this paper the NC sampling method is presented, and data from a *Pinus radiata* plantation in South Australia are used to illustrate the technique. Comparisons are provided with other types of plot selection strategies.

II MATERIALS AND METHODS

2.1 Inventory approach

We define the "area of interest" (AOI) as the finite region over which predictions will be made together with (if separate) the finite region over which ground-based measurements will be made. The AOI is tessellated into a set of non-overlapping contiguous pixels called "virtual" plots which are used to define the population. Virtual plots are constructed to have similar dimensions to the ground-based plots and each virtual plot is associated with a set of metrics which is calculated from the LiDAR data and used as auxiliary information. Virtual plots are further characterised as "reference" plots - the set of plots which are selected for ground-based measurement, "candidate" plots - the set of plots from which the reference plots are chosen, and "target" plots - the set of plots for which predications are made. Therefore the approach which is used in an actual forest inventory can be separated into the following steps:-

1. Define the area of interest
2. Define the area containing the candidate plots
3. Construct a population of virtual plots in areas (1) and (2)
4. Select a sample of reference plots from area (2)

2.2 Imputation

The k-nearest neighbour approach involves calculating the distance or similarity in the auxiliary space between measured reference plots and target plots in order to determine which reference plots are most similar each target plot. The auxiliary variables are chosen because of their ability to predict the variable of interest which, in commercial forests, is typically timber volume. For every target plot, the k-nearest neighbour imputation estimate is the weighted mean of the variable of interest from the k most-similar reference plots, calculated as

$$\tilde{y}_i = \frac{\sum_{j=1}^{k} w_{ij}\, y_j^i}{\sum_{j=1}^{k} w_{ij}}$$

where $\tilde{y}_i$ is the prediction for target plot i , y_j^i is the jth nearest reference plot to plot i, and w_{ij} is the weight assigned to plot j (McRoberts et al., 2007). In this study $w_{ij} = 1$ was employed throughout.

2.3 NC sample

The NC sample operates by partitioning the target population into clusters of plots having minimum sums of squares of distances, from plots to the cluster centroids. The number of clusters is chosen to be equal to the required sample size, n, and the clusters themselves are not required to be spatially contiguous. The clustering procedure which was employed in this study is termed k-means clustering (Hartigan and Wong, 1979) and was performed using the standardized auxiliary variables and a Euclidean distance metric. K-means clustering is normally used for multivariate analysis of complex datasets. After calculating the cluster centroids one then finds the plots in the candidate set which are closest to the centroids in the auxiliary space and these become the reference plots.

An example is shown in Figure 1 which illustrates the selection of 8 reference plots by forming the target set into 8 clusters in the space defined by the auxiliary LiDAR variables p1m (proportion of heights greater than 1m) and mqh (mean quadratic height). The plots closest to the cluster centroids (centroids shown as black stars) are then selected as the reference plots. The term "nearest centroid" derives from the fact that the reference plots are the nearest plots to the target plot centroids. During imputation these plots become the nearest neighbours to target plots in the same cluster (provided they are sufficiently close to the plot centroids). Generally there would more than two auxiliary variables.

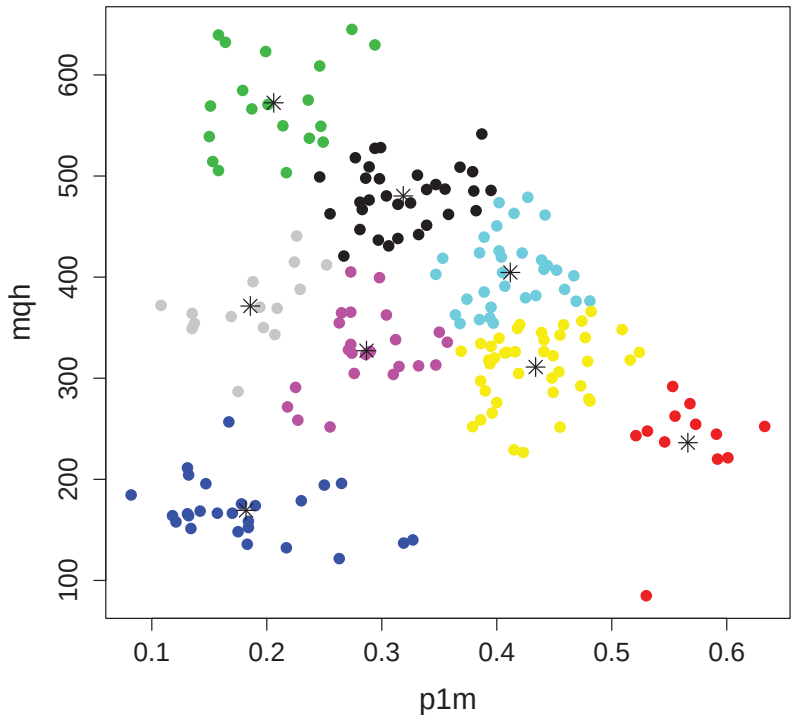

Figure 1: Illustration of NC sample using two auxiliary variables

2.4 Data

The *P. radiata* plantation used in this study is located in South Australia and occupies around 3300 hectares. The plantation is surveyed periodically to provide key product and management information. This study was based on data from 304 ground plots measured in 2012 and a list of plot attributes is given in Table 5.1 of the Final Project Report presented by Rombouts et al. (2014). The essential characteristics of the ground plots are that they have an area of 1000 m^2 and contain trees with an age range of 14-32 years and a mean age of 26.1.

LiDAR data were acquired in January 2012 by flying an aircraft equipped with an ALTM Orion device at an altitude of 800 m across the estate. The LiDAR first return data had a mean point density of 5.9 pulses m^{-2}. A detailed description of the LiDAR data specifications is provided in Table 5.2 of Rombouts et al. (2014).

Around 120 separate variables were extracted from the LiDAR point cloud and are summarized in Table 3. These variables were available for each virtual plot. Ten of these variables were eventually selected for imputation according to how well they were able to predicted the key attributes of interest using an approach described in Rombouts et al. (2014).

LiDAR	first return	last return	description
pground	Y	Y	proportion of ground returns
p>x	Y	Y	proportion of heights > x (x=1,2,5,10m)
sd	Y	Y	standard deviation of heights
skew	Y	Y	skewness of height distribution
kurtosis	Y	Y	kurtosis of height distribution
h	Y	Y	mean height
hmax	Y	Y	maximum height
hmax4	Y	Y	four highest in each plot quadrant
hx	Y		x% percentile height (x=10%∼100%)
d0.x	Y	Y	proportion of heights between 0% and x% hmax (x=10∼100%)
h>0	Y		mean height (heights > 0m i.e. vegetation)
mqh>0	Y		mean quadratic height (heights > 0m)
mqh>1	Y		mean quadratic height (heights > 1m)
scanangle	Y		mean scan angle in the plot
non LiDAR			
lop			thinning status (last operation)
nsq			site quality index
age			plantation age

Table 1: Auxiliary variables available for imputation models

2.5 Simulation approach

Three sampling methods were examined as part of this study. An approach often used in forest inventories is to place a grid over the AOI and select plots at the grid intersection points. LiDAR data are not required to construct either a grid sample or a completely random sample. The use of LIDAR data as *a priori* information in the sample design is aimed at obtaining efficiency gains from the plot selection process. In this study the sampling methods which were evaluated, in addition to random sampling, were locally balanced sampling (Grafström et al., 2014) and the proposed NC method.

In the sampling simulations below, the 304 study plots were divided at random into a target set (200 plots) and a candidate set (104 plots). Reference plots were selected from the candidate plots using each of three sampling strategies i.e. random, locally balanced, and NC. The prediction method used throughout was Euclidean imputation with $k = 1$. The number of reference plots was fixed at either 10, 25, 50 or 75 and the variable of interest was the timber volume in each plot.

The various sampling strategies were evaluated in terms of how well the known variable of interest was predicted. The simulations were repeated 10,000 times with each realisation comprising a new set of target plots, candidate plots and reference plots. Comparisons were done in terms of the relative bias (RB %) and the relative root mean squared error (RMSE %). The relative bias was calculated at the AOI level and is defined as

$$RB = \frac{\sum_j (\hat{Y}_j - Y_j)}{\sum_j Y_j},$$

where $\hat{Y}_j$ is the estimate of total timber volume over the AOI for the j'th realisation and Y_j is the actual total timber volume over the AOI for the j'th realisation. At the AOI level the relative root mean squared error is defined as

$$RRMSE = \frac{\sqrt{\frac{1}{B}\sum_j (\hat{Y}_j - Y_j)^2}}{\frac{1}{B}\sum_j Y_j},$$

where B is the number of realizations. Relative efficiency measures were also calculated for each sampling strategy, using the simple random sample as a benchmark. For any pair of sampling strategies the relative efficiency is calculated as the squared ratio of the AOI-level RMSE values.

III RESULTS

Table 2 presents the simulation results for the three sampling strategies and the four sample sizes which were investigated.

Sampling method	Sample size	Relative Bias (%)	Relative RMSE (%)	Relative efficiency
Random	10	-0.1	4.7	1.0
Balanced		0.2	3.8	1.5
NC		-0.1	3.1	2.2
Random	25	0.2	2.8	1.0
Balanced		0.1	2.3	1.4
NC		0.5	2.1	1.7
Random	50	0.3	1.8	1.0
Balanced		0.2	1.7	1.1
NC		0.7	1.6	1.2
Random	75	0.2	1.6	1.0
Balanced		0.3	1.6	1.1
NC		0.5	1.5	1.2

Table 2: Summary comparisons of bias, accuracy and relative efficiency of three sampling strategies for predicting mean timber volume using field plot measures and LiDAR data

 [ISBN: 978-2-9105-4510-5]

The relative bias is very small with these data, irrespective of the sampling method. With small sized samples the relative RMSE varied from 4.7% using a random sample, to 3.1% using the NC sample. Hence the relative efficiency of the NC sample is 2.2 times that of the random sample. The efficiency gains are most pronounced when the sample size is small. For larger samples, e.g. n= 75, the relative efficiency of the NC sample, compared to the random sample, is 1.2. The relative RMSE results are also presented graphically in Figure 2.

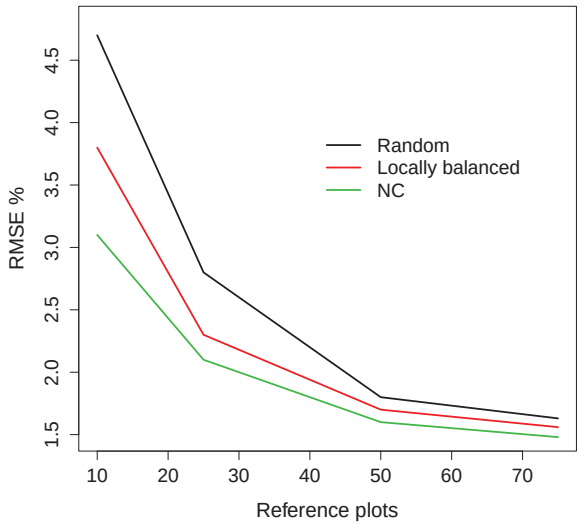

Figure 2: Relative RMSE % vs sample size for three sampling schemes

IV DISCUSSION

This paper describes an approach to inventory design which is specifically aimed at imputation and essentially different to other sampling methods. The results in Table 2 illustrate efficiency gains which are more than double that of random samples. The implications for inventory design are that surveys can be constructed with around half the number of plots that would be required using a conventional sampling approach. The method can be employed either within or across strata although the results in this paper were achieved without stratification.

Note that using remotely sensed data for survey design necessitates having the data available prior to plot selection. Therefore either the data need to be newly acquired or need to be available from a previous campaign. Where existing data are used, they need to be sufficiently recent. The key criteria in this respect is the extent of the correlation between the remotely sensed data and the variables of interest.

One of the primary advantages of the NC sampling method is its flexibility. Small area estimation (SAE) provides a good illustration of this. SAE uses reference plots which are mostly (or completely) outside the AOI to make predictions within the AOI. Most of the existing sampling strategies, including balanced sampling, are not suited to SAE because it is not possible to target the sample specifically to the AOI. There may or may not be reference plots in the sample which are similar enough to the target plots to provide acceptable imputation estimates. By partitioning the small area into defined clusters, the NC sample permits the selection of reference plots which are closely matched to the target plots.

It is proposed to make the clustering approach available as an R function (R Core Team, 2015) to enable sample selection in any application where auxiliary data are available for a population of discrete units such as forestry plots.

V CONCLUSIONS

The sampling method presented in this paper is an intuitive approach to forest inventory where imputation methods are to be used. The method has been trialled on *P. Radiata* datasets from four separate locations in eastern Australia. In every case it has proved to be superior to other sampling methods. With small samples in particular, relative efficiencies are substantially higher than random sampling methods and moderately higher than balanced sampling strategies.

References

Breidenbach J., Nothdurft A., Kändler G. (2010). Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest using airborne laser scanner data. *European Journal of Forest Research 129*, 833–846.

Dash J. P., Marshall H. M., Rawley B. (2015). Methods for estimating multivariate stand yields and errors using k-nn and aerial laser scanning. *Forestry 88*, 237–247.

Grafström A., Saarela S., Ene L. T. (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research 44*, 1156–1164.

Hartigan J. A., Wong M. A. (1979). A k-means clustering algorithm. *Applied Statistics 28*, 104–108.

Hawbaker T. J., Keuler N. S., Lesak A. A., Gobakken T., Contrucci K., Radeloff V. C. (2009). Improved estimates of forest vegetation structure and biomass with a lidar-optimized sampling design. *Journal of Geophysical Research 114*, GE00E04.

Hudak A. T., Haren A. T., Crookston N. L., Liebermann R. J., Ohmann J. L. (2014). Imputing forest structure attributes from stand inventory and remotely sensed data in western Oregon, *Forest Science 60*, 253–269.

Latifi H., Nothdurft A., Koch B. (2010). Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: apredictors. *Forestry 83*, 395–407.

Lesch S. M. (2005). Sensor-directed response surface sampling designs for characterizing spatial variation in soil p *Computers and Electronics in Agriculture 46*, 153–179.

McRoberts R. E. (2012). Estimating forest attributes for small areas using nearest neighbors techniques. *Forest Ecology and Management 272*, 3–12.

McRoberts R. E., Tomppo E. O., Finley A. O., Heikkinen J. (2007). Estimating aerial means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. *Remote Sensing of Environment 111*, 466–480.

Melville G., Stone C. (2016). Optimizing nearest neighbour information - a simple, efficient sampling strategy for forestry plot imputation using remotely sensed data. *Australian Forestry*, to appear.

Melville G., Stone C., Turner R. (2015). Application of lidar data to maximize the efficiency of inventory plots in softwood plantations. *New Zealand Journal of Forestry Science 45:9*, 1–16.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rombouts J., Ferguson I., Leech J., Culvenor D. (2010, September 14–17). An evaluation of the field sampling design of the first operational lidar based site quality survey of radiata pine plantations in South Australia. In *Proceedings of the 2010 Silvilaser Conference*, Freiburg, Germany.

Rombouts J., Melville G., Kathuria A., Stone C. (2014). Operational deployment of lidar derived information into softwood resource systems. *Final Report for Project PNC305-1213, http://www.fwpa.com.au/rd-and-e/resources/611-operational-deployment-of-lidar-derived-information-into-softwood-resource-systems.html*.

Stevens D. L. J., Olsen A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association 99*, 262–278.

Stone C., Penman T., Turner R. (2011). Determining an optimal model for processing lidar data at the plot level: results for a *Pinus radiata* plantation in New South Wales, Australia. *New Zealand Journal of Forestry Science 41*, 191–205.