PO Box 1127
Rotorua 3040
Ph: + 64 7 921 1883
Fax: + 64 7 921 1020
Emil: info@ffr.co.nz
Web: www.ffr.co.nz

Theme: Radiata

Task No: F10502 **Report No. : R076**
Milestone Number: 5.02.6

Nearest Neighbour Imputation of Stand Attributes using LiDAR Data

Authors:
J P Dash, B Rawley and H M Marshall

Research Provider:
Interpine

Date: 26 June 2013

TABLE OF CONTENTS

Disclaimer

This report has been prepared by New Zealand Forest Research Institute Limited (Scion) for Future Forests Research Limited (FFR) subject to the terms and conditions of a Services Agreement dated 1 October 2008.

The opinions and information provided in this report have been provided in good faith and on the basis that every endeavour has been made to be accurate and not misleading and to exercise reasonable care, skill and judgement in providing such opinions and information.

Under the terms of the Services Agreement, Scion's liability to FFR in relation to the services provided to produce this report is limited to the value of those services. Neither Scion nor any of its employees, contractors, agents or other persons acting on its behalf or under its control accept any responsibility to any person or organisation in respect of any information or opinion provided in this report in excess of that amount.

Disclaimer

The information in this document has been prepared and approved by Interpine Forestry Limited (Interpine). Access to the information in this document is being given by Interpine specifically to the person(s) to which it was intended. The information contained in this document remains the intellectual property of Interpine and may not be reproduced, distributed or published by any recipient for any purpose without the prior written consent of Interpine.

Although all reasonable care has been taken to ensure that the information contained in this document is accurate, neither Interpine nor its respective officers, advisers or agents makes any representation or warranty, express or implied as to the accuracy, completeness, currency or reliability of such information or any other information provided whether in writing or orally to any recipient or its officers, advisers or agents.

Interpine and its respective officers, advisers, or agents do not accept: any responsibility arising in any way for any errors in or omissions from any information contained in this document or for any lack of accuracy, completeness, currency or reliability of any information made available to any recipient, its officers, advisers, or agents; or any liability for any director or consequential loss, damage or injury suffered or incurred by the recipient, or any other person as a result of or arising out of that person placing any reliance on the information or its accuracy, completeness, currency or reliability.

EXECUTIVE SUMMARY

The objective of this study was to investigate potential approaches for integrating aerial LiDAR scanning data into the current forest yield information systems of a forest management company. Once a favourable approach was identified a case study was to be implemented across a 4000ha study area. This implementation should provide insight into the probable performance of the system in a production environment. Estimating the sampling error of the estimates produced and using the selected approach to predict forest growth were also key requirements of the project.

The literature review phase of this project identified that a k-nearest neighbour (kNN) approach was deemed to be the most appropriate for this study. kNN has many favourable properties including being free from distributional assumptions, robust, and suitable for integration into the forest managers' current forest yield information systems. The approach uses a measure of the statistical proximity between a target area of forest, for which LiDAR information is available, to assign the measurements from ground plots in a reference population that contains both LiDAR data and on-ground measurements. 213 circular bounded on-ground plots were measured to provide response variables for the modelling process. A LiDAR dataset was collected across the study area to provide canopy metrics for both the reference and the target datasets.

The kNN approach was implemented across a 4000ha swath of Kaingaroa forest and the resulting modelling approach meant that a surface mapping the distribution of several forest variables was produced across the study area. Model evaluation suggested that the models produced performed well. A completely independent validation dataset was produced from the forest managers' pre-existing stand assessments within the study area grown on to the date of LiDAR acquisition. The validation dataset contained 43 individual stands of various ages. The kNN imputation values for each stand were derived and compared to the validation dataset. This analysis showed that the imputation stand estimates and the traditional stand assessment estimates for the stands in the validation dataset were very similar for total recoverable volume (m3/ha) and top height (m). The correspondence for basal area (m2/ha) and stocking (sph) were somewhat worse but accurate for model outputs derived from a small number of plots and fit for the intended use of these statistics in a forest management context. This work indicated that key stand parameters can be accurately calculated for a large number of stands using a small number of ground plots (213) when LiDAR data is available and the kNN approach is followed.

Calculating sampling error for the estimates derived was a key technical challenge for this project. To achieve accurate estimates of sampling error the spatial correlation in the reference plots must be accounted for and procedures were implemented to achieve this. Sampling error was calculated for all stands in the study area and the measures of the confidence interval produced were encouragingly small. For stands in the validation dataset the estimates of sampling error were smaller than those achieved using traditional inventory in most cases. Very little evidence of bias or problems with the implementation of the kNN approach were found during this process. Processing time for the sampling error calculation at the whole study area level was extended and an alternative approach may be required for a larger scale application of the approach.

The kNN technique was used in conjunction with the forest manager's yield prediction systems to produce estimates of future stand volumes for stands in the study area. A comparison with the yield tables provided by the forest manager for stands in the validation dataset indicated that these yield predictions were broadly accurate. Product mix was also produced for all stands in the study area using the kNN technique. A comparison with the stands in the validation dataset revealed that this was also broadly accurate. The derivation of future yield and log product mix are meant as a proof of concept for the use of the kNN technique in this study. Some flaws in these estimates were encountered but nothing that is deemed insurmountable during a practical implementation of the approach.

In conclusion the case study reported on herein was a highly successful practical implementation of the integration of LiDAR data for use in forest resource assessment. Key challenges that may have hindered the uptake of LiDAR technology, including integration with current yield prediction frameworks and calculation of sampling error, have been addressed. Considerable steps have also been made towards the development of a fully functional production system that should now be the next goal for this approach.

INTRODUCTION

Project Scope and Objectives

The purpose of this project was to implement an inventory methodology that could take advantage of the remotely sensed information available from airborne Light Detection and Ranging (LiDAR) scanning data collected over a trial area in Kaingaroa forest, New Zealand. A key objective of the project was to provide a solution that fits within the forest manager's current yield prediction system and to provide practical outputs that can be utilised by the forest manager.

Background Information

The following section details some of the key information collated as part of a literature review for this project and is aimed at guiding the research work undertaken and to provide context for the methodology implemented in the case study.

LiDAR in Forest Inventory

The application of light detection and ranging (LiDAR) to forest management has been studied since the late 1970s. Airborne LiDAR has been developed into a tool that can potentially produce direct and indirect measurement of trees and forests which has considerable implications for forest resource assessment. There are many examples of the use of LiDAR to derive timber volume estimates (Naesset, 1997, Means et al., 2000 and Parker and Evans, 2004 etc.) and some examples of operationalization of the technology in Scandinavian countries and Australia (e.g. Musk 2011, Rombouts et al. 2008). To date in New Zealand the use of aerial LiDAR scanning for forest resource assessment purposes remains firmly in the research sphere. A number of the characteristics of LiDAR data allow generation of additional information suitable for stand and landscape-level management. Numerous studies have shown that LiDAR metrics can be related to both total standing and recoverable volume and other forest parameters that are of interest to forest managers (e.g. Watt et al., 2011, Naesset, 1997).

There are a number of statistical techniques available for incorporation of remote sensing data into forest inventory information systems. Specifically regression techniques and nearest neighbour imputation are reviewed in the following sections.

Regression Techniques

Regression and ratio sampling, like stratification, were developed to increase the precision and efficiency of sampling when applied in a forest inventory context. This is achieved through the use of an auxiliary variable that is measured on each sample unit in addition to the variable of interest (Freese, 1962). Regression or ratio sampling can only be used when the population mean or total for the auxiliary variable are known without error. This is different to double sampling where the true population mean of the auxiliary variable is not known (Husch et al., 2003) and needs to be estimated from a sample. To successfully utilise regression or ratio sampling a strong relationship between the auxiliary variable and the variable of interest is required. Ratio sampling can be used instead of regression estimation where this relationship has a y intercept of zero (Avery and Burkhart, 1994). The precision benefit gained through using regression estimation is proportional to the strength of the correlation between the variable of interest and the auxiliary variable. Impressive improvements in precision are possible with regression estimation when the correlation is close to one (Cochran 1977, Avery and Burkhart, 1994).

Although regression techniques clearly offer significant benefits to foresters working with remote sensing technologies in many cases there are a number of limitations to regression techniques that indicate that a different approach might be favourable. These factors include:

- The need for a minimum sample size within each area of interest when using regression estimation is a key limitation of this technique. In the majority of commercial forest management scenarios in New Zealand resource information is required at the stand level. To use regression estimation effectively a minimum of around 30 sample plots would be required per stand if the stand was the area of interest. Regression estimation is an excellent technique when increasing the precision of information at the forest level is of interest. However, in these cases using auxiliary variables other than those derived from LiDAR may be more cost effective.
- Regression techniques require an assumption of a normal distribution of co-variates. Violation of this assumption can invalidate the predictions made using regression approaches. This limits the type of data that can be used in a regression approach. For example categorical and count data do not follow a normal distribution and so may be inappropriate for use.
- There is no simple way of integrating LiDAR data using a regression modelling approach into the current yield prediction systems that are currently used in New Zealand.

While popular in the research environment it is likely that these factors are limiting the uptake of LiDAR technologies into a production environment such as that found in the New Zealand forest industry.

Nearest Neighbour (kNN) imputation

Since its initial application to forest resource assessment in the early 1990s (Tomppo 1991, 1996) k nearest neighbour (kNN) imputation has become extremely popular internationally with peer reviewed publications on the topic originating from over twenty countries (McRoberts 2012). In forest inventory kNN imputation is generally used to assign forest attributes for areas that have not been inventoried often based on a two phase sampling design. The first phase consists of the acquisition of easily measured auxiliary variables across an area of interest (e.g. LiDAR data or other remotely sensed information). The second phase consists of detailed measurements of variables of interest at specific locations within the study area (Moeur and Stage 1995, Fallkowski et al. 2009). In a forestry specific setting the variables of interest will typically be forest level parameters such as stand volume, stocking or product volumes per hectare at a reference age. This produces two separate datasets, the first a reference dataset containing both variables of interest and auxiliary variables, the second, a target dataset containing only auxiliary variables. The objective of an imputation is to predict, or impute, the variables of interest across the area of interest using the relationship between auxiliary data and variables of interest in the reference dataset. To achieve this the reference dataset is used to characterise the relationship between the auxiliary variables and the variable of interest. The variables of interest within the target dataset are then estimated by imputing them from the (k) nearest neighbours where proximity is measured in terms of statistical similarity (Fallkowski et al. 2009). One of the key functional advantages of the kNN technique is the ability to extrapolate beyond regions of the forest that were part of the original sampling design and therefore take advantage of the extensive auxiliary information available via remote sensing techniques such as aerial LiDAR scanning. The reference observations used to provide measurements of the variable of interest are known as donors and statistical proximity is described using some metric of distance in covariate space. The number of donors used to impute the target cell values is referred to as k, where k=1 the variable of interest is simply taken from the nearest neighbour. Where k>1 then either a simple average or an average weighted according to the statistical distance between the target and its donors may be used.

The distance metric selected plays an important role in a kNN imputation model and also contributes to calculation performance. There are numerous distance metrics implemented in the yaImpute R library (Crookston and Finley 2008) that have been investigated as part of this project

and are worthy of a brief summary here. For all measures of nearness, with the exception of random forests proximity matrix, nearness is defined using Euclidean distance weighted using some form of weighting matrix. The “Raw” distance metric is calculated using distance based on the untransformed, or raw, values of the auxiliary variables. By contrast “Euclidean” distance uses normalised auxiliary variable values to define distance in multivariate predictor variable space. Predictor variables are normalised by subtracting the mean and dividing by the standard deviation of each predictor variable. The method “mahalanobis” (Mahalanobis 1938) uses the dimensional components of Euclidean distance transformed by the inverse of the covariance matrix of the predictor variables. The distance method “ICA” is similar to mahalanobis but based on independent component analysis where distance is computed in a projected space defined by components that are statistically independent and are assumed to have a non-Gaussian distribution (Hyvarinen & Oja, 2000 as cited in Hudak et al., 2008). The most similar neighbour methods (msn and msn2) provide a measure of distance computed in a projected canonical space where as for the method gradient nearest neighbour (GNN) (Ohman and Gregory 2002) distance is computed using a projected ordination of predictors based on canonical correspondence analysis.

Random forests proximity is calculated quite differently to the other proximity metrics mentioned above. Observations are considered similar if they occupy the same terminal node in a suitable constructed classification and regression tree. (See Breiman 2001 and Liaw and Wiener 2002 for detail) Distance is then calculated as one minus the proportion of trees where a target observation is in the same terminal node as a reference observation. (Crookston and Finlay 2008)

Each distance metric listed above have been used and investigated and implemented as part of this study and the random forests distance was selected for application in the production of model outputs. Random Forests distance has two notable advantages over the alternatives; it is non-parametric and free from distributional assumptions and predictor variables can be a mixture of categorical and continuous variables whereas the other distance methods require continuous variables to define the search space axes (Crookston and Finlay 2008). Furthermore provisional analysis on the model outputs indicated that models imputed using random forest distance were the best performing and this is consistent with the findings of Hudak et al (2008) who noted that kNN models using random forests distance produced the lowest values of model error and were the most stable. A German research team (Latifi et al 2010) also compared the various distance techniques implemented in yaImpute in a kNN imputation setting and found that the random forests distance measure was superior to the others. Further detail on the random forest methodology is therefore required and is reviewed in section 3.2.4.

Random Forests

The random forest method for describing statistical distance has been selected for use in this study as it has a number of properties that are favourable for the intended use here. As a relatively unusual approach in an imputation context some additional detail on the technique is warranted. Developed by Breiman (2001) a random forest is the result of a machine learning algorithm which randomly iterates through samples of the data it is exposed to generate a large group, or forest, of classification and regression trees. The iterative nature of random forest provides a distinct advantage over other imputation methods as bootstrapping of the data leads to more robust predictions (Hudak et al. 2008). The inclusion of random subsets of predictor variables means that problems associated with inter correlated predictor variables and over fitting are alleviated (Breiman 2001) and that, if required, measures of variable importance can be generated (Hudak et al. 2008).

Although strictly a classification tool the random forest method can be used to derive statistics that are analogous to statistical distance which is calculated as one minus the proportion of shared terminal nodes in the forest of classification trees. The random forest method can be summarised as follows: Each continuous variable is discretised and reference observations are then classified with respect to the other response variables and the predictor variables. The resulting classification

trees are then concatenated to calculate the proportion of shared terminal nodes across all the response variables for the purpose of identifying nearest neighbours (Hudak et al. 2008).

Sampling Error for kNN Estimates

In section 0 it was discussed how the k nearest neighbour (kNN) technique can be used to calculate the average of any measured or calculated response variable across any arbitrary area of interest (AOI) in the study area. Examples of response variables are basal area, total recoverable volume and the volume of small sawlogs 3 years in the future. Examples of areas of interest are stands, coupes, age classes, riparian strips or even the entire study area. In the simplest implementation, we identify the target pixels in the area of interest. To each target pixel we assign k reference plots. If we have N target pixels, each with k reference plots, then we have N*k response variable values and can use the average of those N*k values as the best estimate of the average for the AOI. More complicated implementations first calculate a weighted average of the k reference plot values for each of the N target pixels, then average across the N target pixels. The weight is based on the similarity between the target and reference plot in terms of LiDAR metrics.

Having calculated the average for a response variable, a potential user of this approach might be interested in knowing:

1. How does one calculate the sampling error of the kNN estimate for the AOI?
2. How big/small is that sampling error relative to the alternatives?
3. What are the practical issues in calculating sampling error?

This section addresses these questions.

Calculating Sampling Error for Areas of Interest

Supposing we have already calculated the response variable value for each of the N target pixels that represent an AOI by using the average of the values for k nearest neighbours to the target pixel. We could then calculate the sampling error using those N values in the same way that we would if we had established a plot on top of each of the N target pixels. In other words we could assume that the N values represent a simple random sample (SRS). But that would grossly under-estimate the true sampling error. The problem is that we don't really have N independent values to work with. We have instead a much smaller number of reference plots, used to greater or lesser degree, and sometimes not at all, in any one AOI.

McRoberts et al, 2007 provide a method for calculating the sampling error across the N target pixels from a much smaller number of reference plots. The variance of the sum of the N target pixel values includes both the sum of the target pixel variances and the sum of the co-variances between each target pixel. If we analysed the data using an SRS approach then we would be doing exactly the same thing except that in the SRS approach we assume that the co-variances are zero.

When k > 1, then the variance of the value for a single target pixel can be calculated by treating the k nearest neighbours as a sample from a super-population of possible nearest neighbour values.
The co-variance between any two target pixels is calculated from the variance for each of the two pixels and the correlation between them. This correlation has two components. The first arises when the two target pixels share the same reference plot; or plots if k > 1. If two target pixels share the same reference pixels then their predicted values and prediction errors are highly correlated. This component is a given and easy to calculate. The second component arises from the possibility that the prediction errors from two reference plots, where the reference plots are close together, might be correlated. This spatial correlation component is not a given and depends on the characteristics of the study area and the placement of reference plots. One way to think about the spatial correlation component is that if all of the reference plots having a specific set of LiDAR

metrics come from the same part of the same stand then they are not providing independent estimates for a target pixel with the same set of LiDAR metrics in a different stand. If one of those reference plots over-predicts for another part of the study area then they probably all do to some extent.

More detail on the method is provided in Appendix 1.

METHODOLOGY

The methodology detailed in the following sections was implemented as a case study for the use of k nearest neighbour (kNN) imputation approach for predicting stand attributes from an aerial LiDAR dataset and corresponding ground data.

Study Area

The study area for this project encompassed a 4000 hectare (ha) swath of Kaingaroa forest in the Central North Island of New Zealand (Figure 1). Kaingaroa is New Zealand's largest contiguous plantation area occupying over 180000 ha of the volcanic pumice plateau south of Rotorua. The primary forestry species in Kaingaroa is *P. radiata* (*Pinus radiata D. Don*) which occupies more than 95% of the stocked area. *P. radiata* is grown on a ca. 28 year rotation. Typical regimes include an initial stocking of around 1000 stems per hectare and thinning(s) down to a final stocking of around 300 – 400 sph prior to harvest. Some stands are grown on a clearwood regime and pruned to approximately 6m whereas others remain unpruned. The study area is exclusively planted in P. radiata. The range of stand conditions present in the study area are summarised in Table 1.

Table 1. The range of stand conditions in the study area.

Age (y)	Stocking (sph)	Basal Area ($m^{2/}$ha)	Top Height (m)	TRV (m^3/ha)
1 – 33	220 - 1026	7.8 - 59	5.5- 46.5	1 - 892

Ground Sampling

Response variables for the imputation approach developed in this study came from 213 field plots located throughout the study area. The ground sampling design followed a "hybrid" approach that utilised simple random sampling for the majority of plots with the remainder located with adjusted sampling probability. A 400m grid with a randomised start point and orientation was overlaid onto the study area and used to locate 183 field plots with one allocated at each grid intersection. Subsequently rasters detailing the distribution of predictor variables used in previous regression sampling work, and found to have a strong relationship with TRV, were produced and the remaining 30 plots were placed into the study area targeting the range of these metrics that had not been sampled by the original 183 plots. The purpose of this sampling exercise was to ensure that the full range of the predictor variables was sampled by the ground sampling units. The grid layout was used initially so that a simple random sampling (SRS) estimate could easily be derived from the ground plots for comparison with other estimates.

The sampling unit for this project was a slope adjusted 0.06 ha circular bounded plot. Plots were geolocated with a Trimble GeoXH 6000 – global positioning system (GPS). At least 300 points were collected at the centre point of each plot which was post-differentially corrected to give on average a sub-0.5m accuracy. Within each plot tree diameter at breast height (dbh) was measured on all trees, and total tree height was measured on a sub-sample of plot trees. Sufficient tree heights were measured to fit a diameter height regression for each plot which was used to predict the heights of unmeasured trees. Overlapping feature tree description was also recorded on mature trees according to the RAD05 tree cruising dictionary (YTGEN User Group 2007). Using these measurements plot level statistics were calculated to provide response variables in the reference data set used for nearest neighbour imputation.

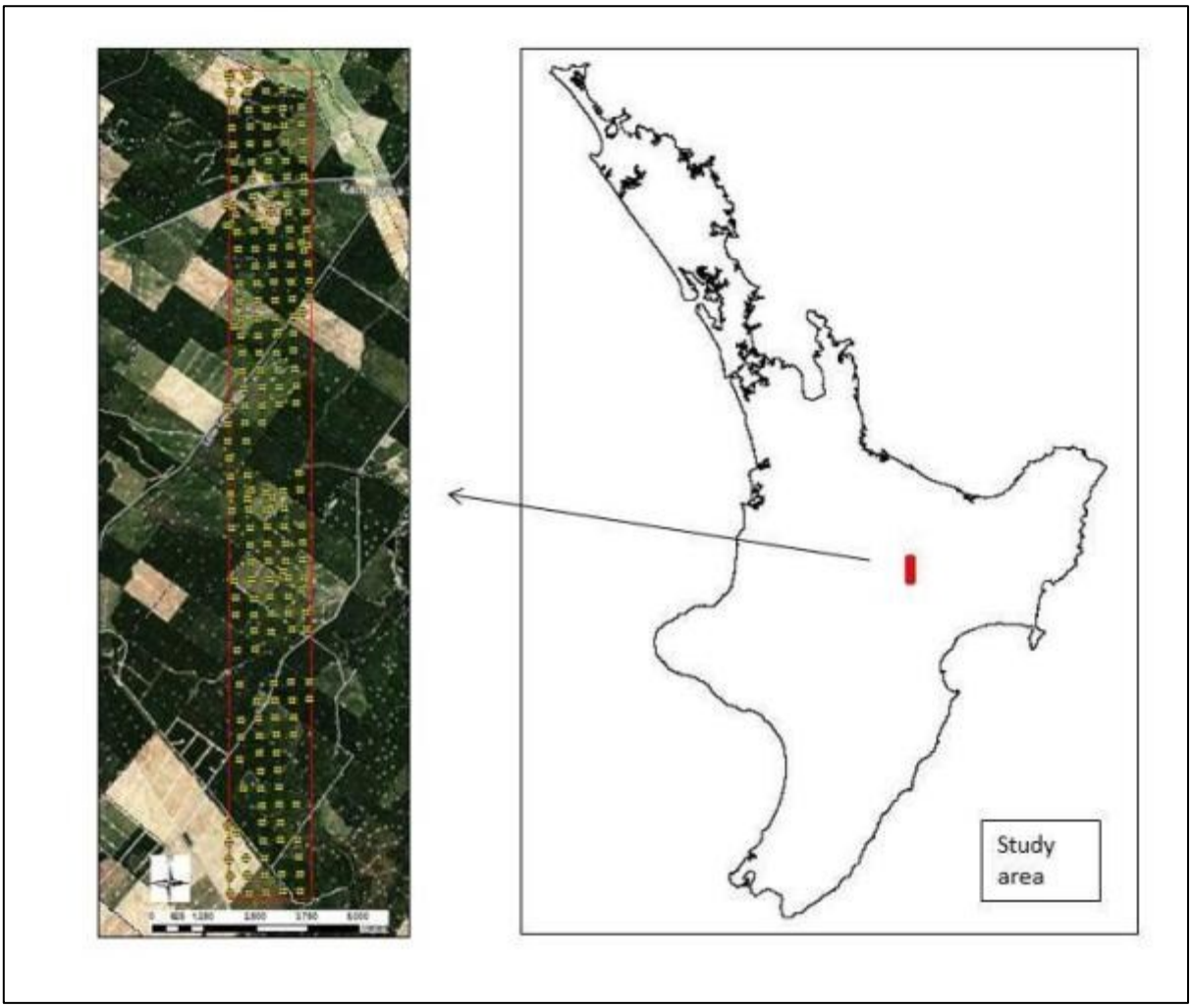

Figure 1. Study area and field plot locations. The left panel shows the 4000 ha study area (outlined in red) with installed plots shown as light yellow circles.

LiDAR Sampling

Candidate predictor variables used in this analysis were derived from airborne LiDAR scanning of the study area. LiDAR acquisition was carried out by Aerial Surveys Ltd. using a fixed wing aircraft on the 28 June 2012. An Optech ALTM 3100EA scanner was used at a flying height of 950m above mean ground level acquiring data with a designed pulse density per swath of a minimum 4 pulses per square metre, and a swath overlap of 50%. The point cloud data was then classified into ground, first and, intermediate returns using automated routines tailored to the project landcover and terrain. The subsequent steps were undertaken using TerraSolid LiDAR processing software module TerraScan. Manual editing of the LiDAR point cloud data was undertaken to increase the quality of the automatically classified ground and above ground point datasets. This editing involved visually checking over the data and changing the classification of points into and out of the ground point dataset.

The Cloudmetrics function in the FUSION LiDAR analysis software product (Macgaughy 2010) was used to produce various statistical parameters describing the LiDAR dataset in terms of point elevations and intensity. These metrics (

Table 2) which were spatially concurrent with the ground plots served as candidates for the predictor variables used in this analysis. These variables would be used in both the target and the reference dataset and would need to be calculated independently for subsequent LiDAR datasets.

Table 2. The metrics used as candidate predictor variables

Metric	Description	Selected Kaingaroa
Total return count above 0.50	Number of returns above 0.5 height	
Elev minimum	Minimum height	
Elev maximum	Maximum height	x
Elev mean	Mean height	
Elev mode	Modal height	
Elev stddev	Standard deviation of heights	x
Elev variance	Variance of heights	
Elev CV	Coefficient of variation of heights	
Elev IQ	75th percentile minus 25th percentile of heights	
Elev skewness	Skewness of heights	
Elev kurtosis	Kurtosis of heights	
Elev AAD	Average absolute deviation from mean of heights	
Elev L1 – L4	L-moment 1 to 4 of heights	
Elev L CV	L-moment coefficient of variation of heights	
Elev L skewness	L-moment skewness of heights	
Elev L kurtosis	L-moment kurtosis of heights	x
Elev P01 – P99	Heights 1st to 99th percentile	x
Return 1 - 9 count above 0.50	Count of return 1 – return 9 points above 0.5m height	
Other return count above 0.50	Count of other returns above 0.5 height	
Percentage first returns above 2.00	Percentage first returns above 2m height	
Percentage all returns above 2.00	Percentage all returns above 2m height	
(All returns above 2.00) / (Total first returns) * 100	(All returns above 2m height) / (Total first returns) * 100	
First returns above 2.00	First returns above 2m height	
All returns above 2.00	All returns above 2m height	
Percentage first returns above mean	Percentage first returns above mean height	x
Percentage first returns above mode	Percentage first returns above modal height	x
Percentage all returns above mean	Percentage all returns above mean height	
Percentage all returns above mode	Percentage all returns above modal height	
(All returns above mean) / (Total first returns) * 100	(All returns above mean height) / (Total first returns) * 100	
(All returns above mode) / (Total first returns) * 100	(All returns above modal height) / (Total first returns) * 100	
First returns above mean	Number of first returns above mean height	
First returns above mode	Number of first returns above modal height	
All returns above mean	Number of returns above mean height	
All returns above mode	Number of returns above modal height	
Total first returns	Total number of 1st returns	
Total all returns	Total number of returns	
Elev MAD median	Median of the absolute deviations from the overall median	
Elev MAD mode	Median of the absolute deviations from the	

Metric	Description	Selected Kaingaroa
	overall mode	
Canopy relief ratio	((mean height - minimum height) / (maximum height – minimum height))	
Elev quadratic mean	Generalized means for the 2nd power	
Elev cubic mean	Generalized means for the 3rd power	
Int minimum	Minimum intensity	
Int maximum	Maximum intensity	
Int mean	Mean intensity	x
Int mode	Modal intensity	x
Int stddev	Standard deviation of intensity	
Int variance	Variance of intensity	
Int CV	Coefficient of variation of intensities	x
Int IQ	75th percentile minus 25th percentile of intensities	
Int skewness	Skewness of intensities	
Int kurtosis	Kurtosis of intensities	x
Int AAD	Average absolute deviation from mean of intensities	
Int L1 – L4	L-moment 1-4 of intensities	x
Int L CV	L-moment coefficient of variation of intensities	
Int L skewness	L-moment skewness of intensities	
Int L kurtosis	L-moment kurtosis of intensities	
Int P01 – P99	Intensities 1^{st} – 99^{th} percentile	
Surface slope	Surface slope (degrees)	
Surface aspect	Surface aspect (degrees azimuth, 0 degrees at 12 o'clock, increasing clockwise)	
Profile curvature * 100	Profile curvature * 100 (in direction of slope)	
Plan curvature * 100	Plan curvature * 100 (perpendicular to slope)	
Solar radiation index	Solar radiation index	
Age	Crop age at time of LiDAR acquisition	

Model Development

All imputation and model development was implemented in the R statistical software package (R development core team 2012) and made use of the yaImpute package (Crookston and Finlay 2008) as well as the randomForest and RODBC (Ripley 2002) packages.

Variable Selection

With many candidate predictor variables it may be beneficial to select the optimal predictors appropriately for a given response. Numerous variable selection procedures designed at thinning LiDAR metrics to only those which are most valuable are documented in the research literature and several approaches were implemented during this study. Implemented approaches including the varSelRF R package (Diaz-Uriarte 2012) were used for variable selection. varSelRF selects important predictor variables through an iterative, backwards aggressive variable elimination process which is designed to minimise the RF out-of-bag (oob) error rate without creating bias in the final model (Fallkowski et al. 2010). The technique implemented by varSelRF uses a genetic algorithm based on the principle of evolution by natural selection. Variable selection by varSelRF was found to be extremely volatile when used with the random forests distance metric. This result

is consistent with the findings documented elsewhere (Lafiti et al. 2010) and resulted in alternative variable selection approaches being favoured. Packalen et al. 2012 described a comparison of several variable selection techniques and found that a simulated annealing approach aimed at minimising model error was the most accurate method of variable selection for kNN imputation of forest variables with remotely sensed data. Variable selection via simulated annealing (VSSA) as implemented in this study sought to minimise model root mean square difference (RMSD) by repeatedly solving the kNN imputation model. Following the technique of Packalen et al. (2012) a randomised local search method known as simulated annealing (Kirkpatrick et al 1983, Aarts and Lenstra 1997) was used. This technique is known to provide a good approximation of the global optimum in a large search space whilst avoiding local optima by preventing moves to poorer solutions. The technique is often analogised as similar to the annealing of metal with control achieved by a parameter called temperature which can be gradually decreased according to a cooling schedule (Packalen et al 2012). A variable selection by simulated annealing algorithm was developed by the study authors for the purposes of this project.

kNN Imputation

Following variable selection imputation procedures were implemented using the various functions of the yaImpute R Package (Crookston and Finley 2012). There are several imputation methods available through yaImpute. In this case the random forest classification approach was used to calculate statistical proximity as it has been shown to be a very robust approach proximity measure in previous studies (Hudak et al 2008a, Hudak et al 2008b) and is free from distributional assumptions (Crookston and Finley 2012). Under the random forest approach observations are considered similar if they tend to end up in the same terminal node in a suitably constructed collection of classification and regression trees (Breiman 2001, Liaw and Wiener 2002). The distance metric used to define the k nearest neighbours is calculated as one minus the proportion of trees where a target observation is in the same terminal node as a reference observation (Crookston and Finley 2012).

For the imputation process the number (k) of reference observations to be used can be specified. The kNN prediction for the continuous response variable is then calculated as the average of the k nearest neighbours. Increasing k effectively moves the prediction towards the population mean which is unrealistic for non-normal or skewed datasets. Increasing k also reduces the pure error which is useful in describing the variability in the response values (Hudak 2008a).Numerous values of k were implemented as part of this project (see Mcroberts. 2012 for a review of this). A random forest tree ensemble was produced for distance calculation. The ensemble consisted of 1000 bootstrap replicates with only the candidate predictors selected by VSSA included.

A target dataset was derived from rasters of the predictor variables generated at a 30m x 30m resolution. The yaImpute package was used to assign values of the response variable to all cells across the landscape where predictor variables were available. This resulted in surfaces across the entire study area for detailing the distribution of each of the response variables (Figure 2).

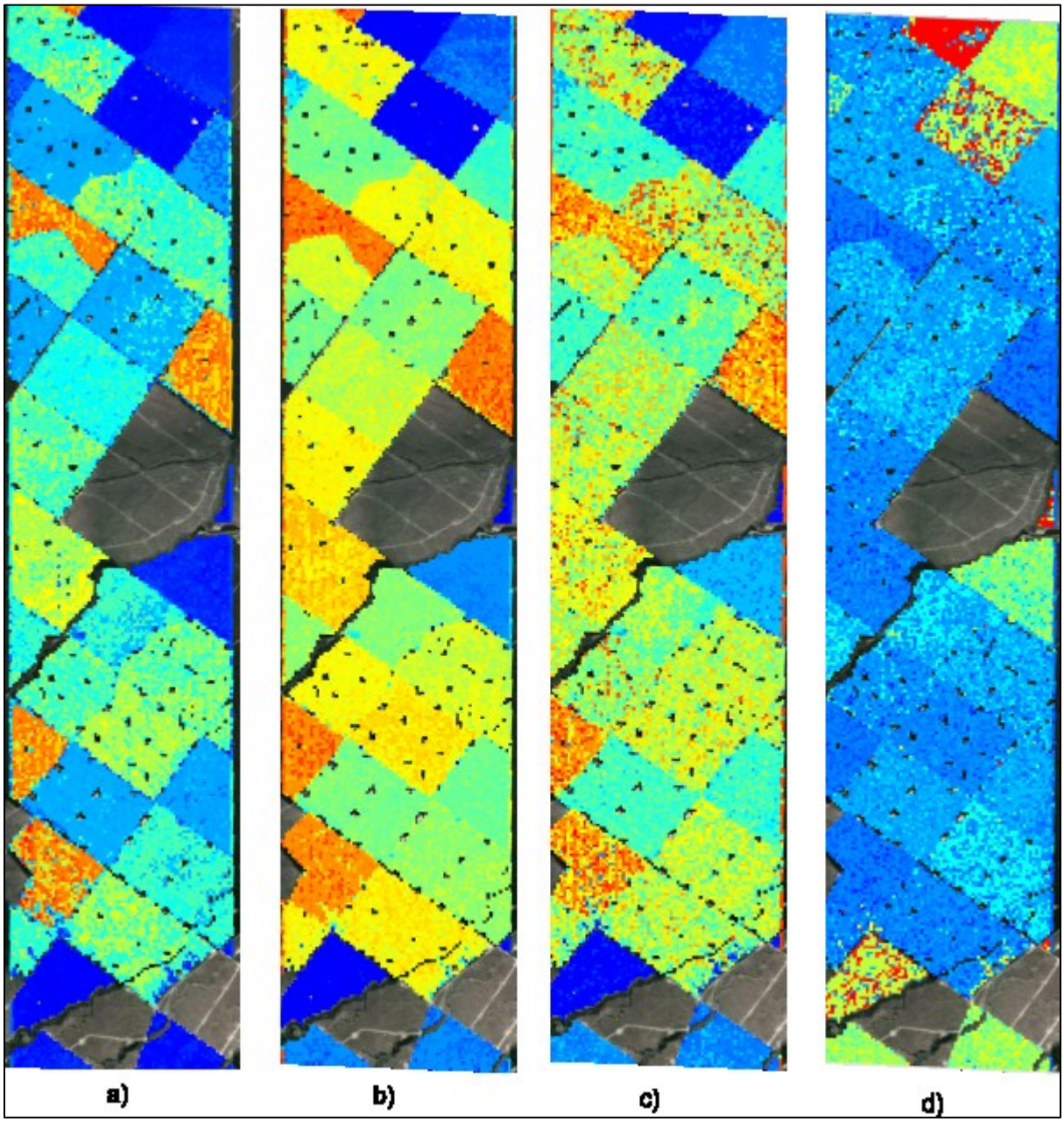

Figure 2. Rasters showing the distribution of response variables of a) TRV, b) top height, c) basal area and d)stocking across the study area. These response pixels can be aggregated to provide estimates for any area of interest.

Yield Table Development

A key strength of the kNN approach and one of the main drivers for the selection of this methodology for use in this study is the ability to incorporate the statistical approach within the forest manager's current yield prediction systems. Under a kNN methodology once a neighbour(s) is selected for a target cell any values calculated for the donor(s) can be applied to the target cell. This means that a yield table derived for a reference cell can be used to predict values at a specific point in time for a target cell. Yield tables for each reference cell were produced and used to provide yield estimates, including log product volumes, for target cells covering an age range of 15-30 years. The target pixel level yield

tables can be aggregated at any area of interest to provide estimates of yield development. Aggregates at the stand level can be compared with the validation dataset to provide a measure of the accuracy of imputed yield table performance compared with traditional yield tables. The attribution of future yields is demonstrated in Figure 3. TRV is used in this example but any output of the yield prediction system, such as stocking or log product volume at a specific age, could be used. In the scenario in the image below the target stand contains three red pixels, one yellow pixel, and two purple pixels giving a TRV of 383.33m^3/ha at measurement date based on a weighted average of the reference values. Similarly the yield tables for the donor reference plots can be amalgamated according to the number of occurrences of each donor within an area of interest to produce a stand yield table.

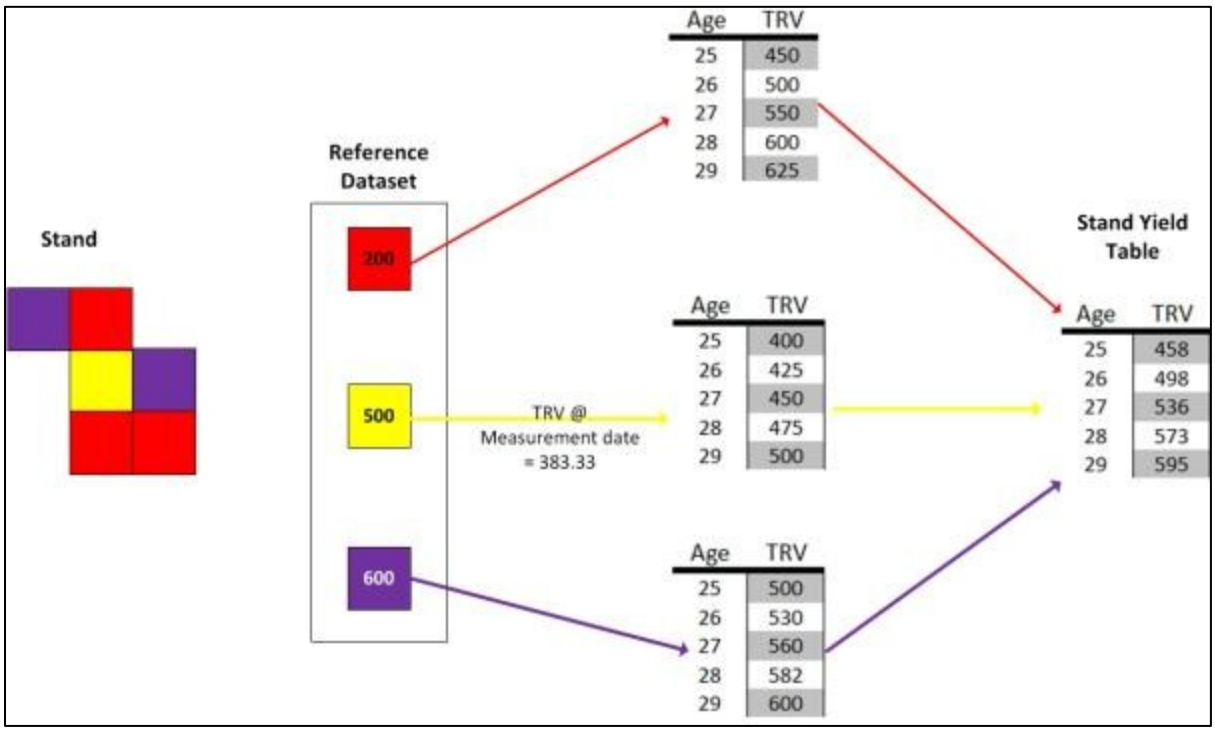

Figure 3. Yield table development under a kNN approach

Validation Dataset

A database of the forest managers' stand assessments was made available to provide a validation dataset for this project. The database was interrogated to extract all inventories that fell within the LiDAR swath and were yet to be harvested. Where an inventory was partially within the LiDAR swath the plots outside the study area were excluded. The forest managers yield prediction systems were used to project standing tree assessments to the date of LiDAR acquisition to provide a comparison. To increase the age range of stands available for validation, information was also acquired from silvicultural quality control check type measurements which record stand characteristics but for which there is no way of calculating the sampling error.

The validation dataset available for comparison is summarised in Table 3.

Table 3. A summary of the validation dataset available for the Kaingaroa case study. All figures shown are at LiDAR acquisition date.

N	Age range (y)	TRV range (m^3/ha)	Top Height Range (m)	Basal Area range (m^2/ha)	Stocking range (sph)
43	5 -32	12-900	7.1-46	9.23-58.5	224-961

RESULTS

Variable Selection

Using the VSSA algorithm predictor variables were selected for inclusion in the model the variables selected for inclusion are noted in

Table 2. In total, 19 of the 101 candidate predictor variables were selected for use in the Kaingaroa case study.

Sampling Error Results

The method of McRoberts et al, 2007 was used to calculate sampling error for the entire study area and for stands within the study area. Use of this method necessitated a prior analysis for spatial correlation.

Spatial Correlation

In this context, spatial correlation is about the tendency for reference plots that are close together to generate similar imputation errors. The reference plots in a closely spaced cluster of reference plots will tend to all predict a little higher or lower than the average for all reference plots with the same LiDAR characteristics. A target pixel having multiple nearest neighbours from the same cluster will tend to have lower variation in the range of imputed response values and this lower variation, if not allowed for, can result in an under-estimate of sampling error. The extreme case is where two reference plots are exactly coincident, in which case their correlation is 1.

Spatial correlation can be visually assessed by plotting the variance of the differences between imputation errors of pairs of reference plots against the distance between the plots, for all possible pairs of reference plots. Figure 4 provides an example for several values of k (number of nearest neighbours) where the response variable is total recoverable volume. A trend in which variation increases with distance to a plateau would be taken as evidence of spatial correlation. Such a trend is not evident in Figure 4 at moderate distances but, with a good imagination, can be seen in Figure 4 Semivariogram for Total Recoverable Volume over medium distances

where the same data is shown over a limited range of distances. Correlation is calculated from the fitted curve (dashed line) in Figure 4 Semivariogram for Total Recoverable Volume over medium distances

, as 1 – y/asymptote, where y is a value from the curve and asymptote is the asymptotic maximum of the curve. The effective range of spatial correlation, i.e. the point at which the curve gets very close to the asymptotic value and correlation gets very close to zero is a few hundred metres. To put this in perspective, most of the plots were on a 400m grid and the spatial correlation is really only material when additional plots were placed off the grid. The short range of spatial correlation is consistent with the findings of McRoberts et al, 2007 and Magnussen et al, 2009. It is tempting to conclude that spatial correlation could be ignored in production inventory systems if plot spacing is kept above 400m. This might, in fact, be the case. However, it would be premature to arrive at this conclusion on the basis of a single study.

Figure 6 shows that different response variables can have different patterns of spatial correlation; including undetectable for top height with k=5.

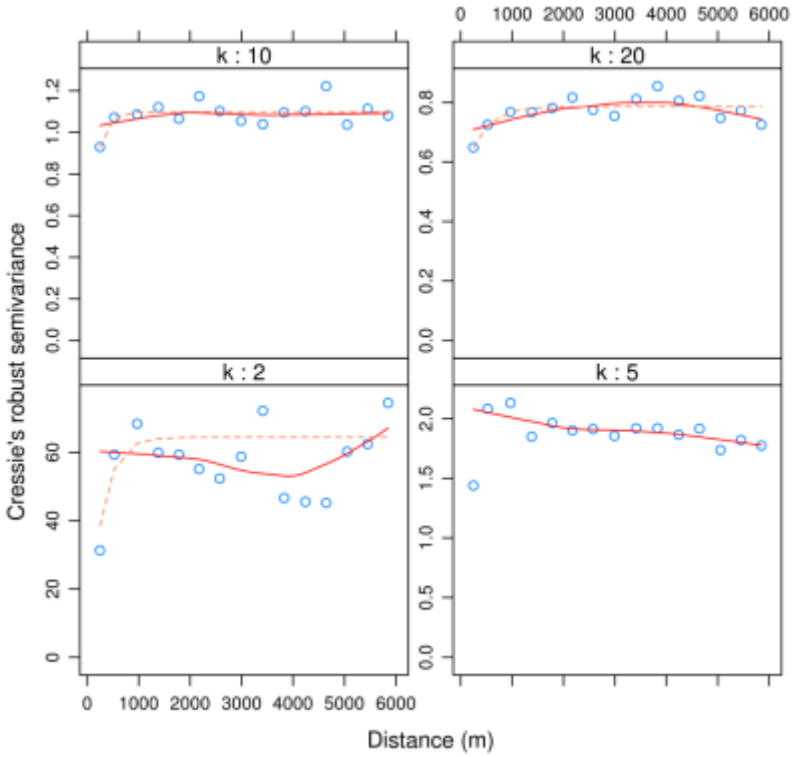

Figure 4 Semivariogram for Total Recoverable Volume over medium distances

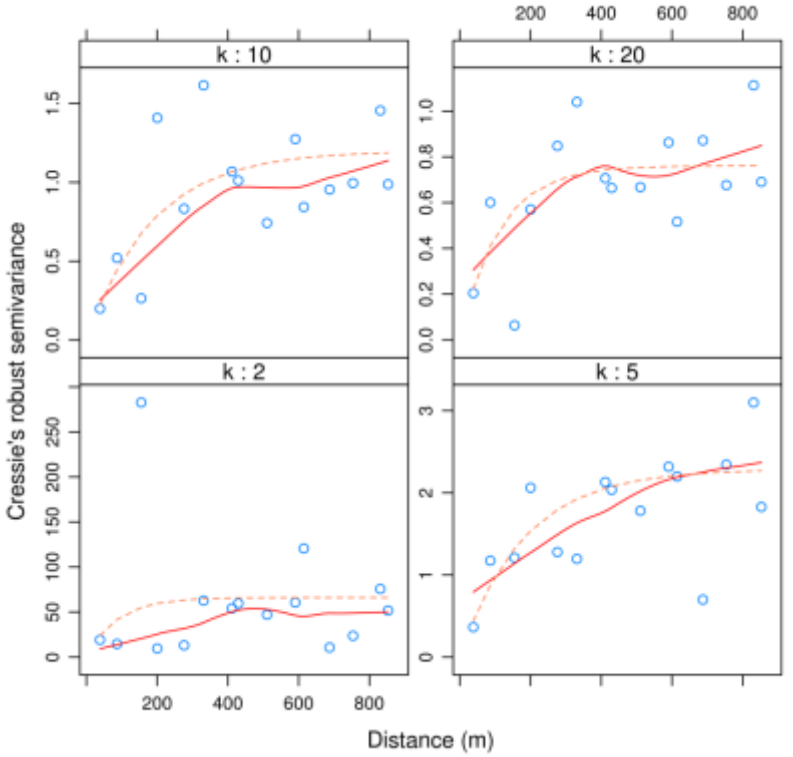

Figure 5 Semivariogram for total recoverable volume over short distances

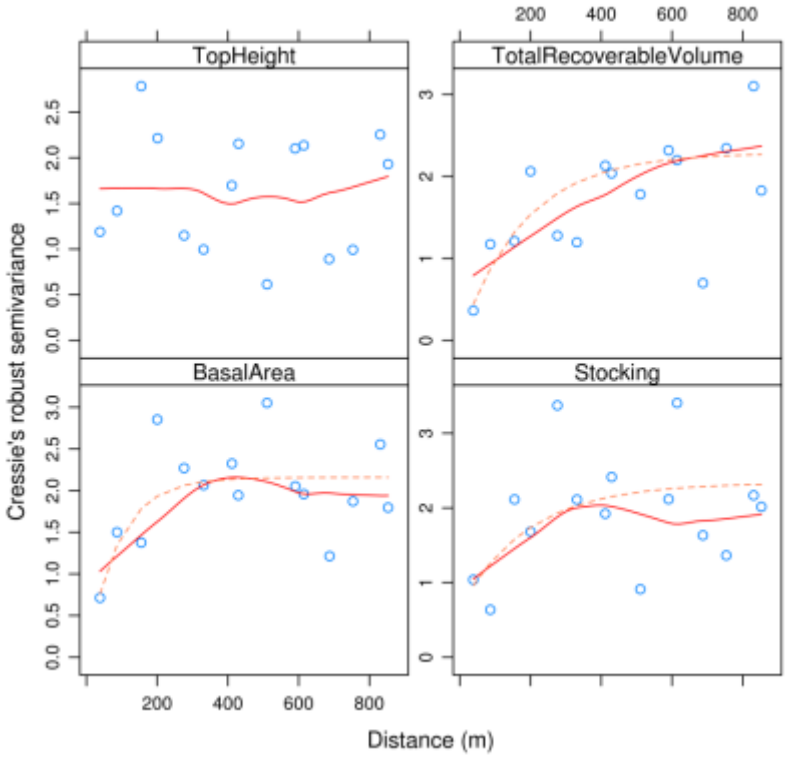

Figure 6. Semivariogram for several response variables for k=5 over short distances

The noise in Figure 4, Figure 4 Semivariogram for Total Recoverable Volume over medium distances

and Figure 6 creates uncertainty about the exact magnitude and range of spatial correlation. Rather than dwelling on whether the patterns are real, it is more instructive to see what effect the assumed spatial correlation has on estimates of sampling error. Figure 7 compares relative sampling error calculated both with and without recognition of spatial correlation for total recoverable volume. Each point represents a single stand in the study area. Relative error is the standard error of the estimate of the stand mean divided by the stand mean total recoverable volume (m^3/ha). The dotted lines in Figure 7 have a slope of 1. When spatial correlation of the magnitude shown in Figure 4 Semivariogram for Total Recoverable Volume over medium distances

is incorporated in estimates of the sampling error then the sampling error is higher than when spatial correlation is ignored; the solid lines in Figure 7. Over all stands, the solid lines in Figure 7 represent an increase[1] in sampling error for total recoverable volume of 15%; i.e. a probable limit of error (PLE) of 10% would increase to 11.5%.
The effect of incorporating spatial correlation in the estimates of sampling error depend upon the response variable as well as the value of k. The equivalent increases for other response variables are 19% for basal area, 4% for stocking and 12% for top height.

The effect of incorporating spatial correlation into estimates of sampling error are relatively consistent across many stands with the notable exception of a single stand; see top left of k=2 panel in Figure 7. When k=2, the kNN estimate for that particular stand is dominated by two reference plots that are only 9.7 m apart and assumed to be highly correlated. When k is higher than two the influence of these two reference plots is less.

The results in the subsequent sections incorporate the effects of spatial correlation.

1

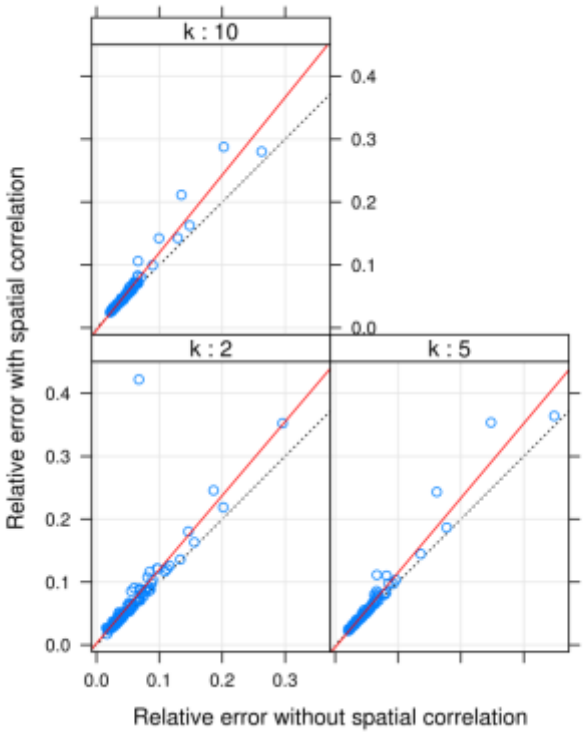

Figure 7 Effect of spatial correlation on calculated sampling error

Sampling Error Estimates

Entire Study Area

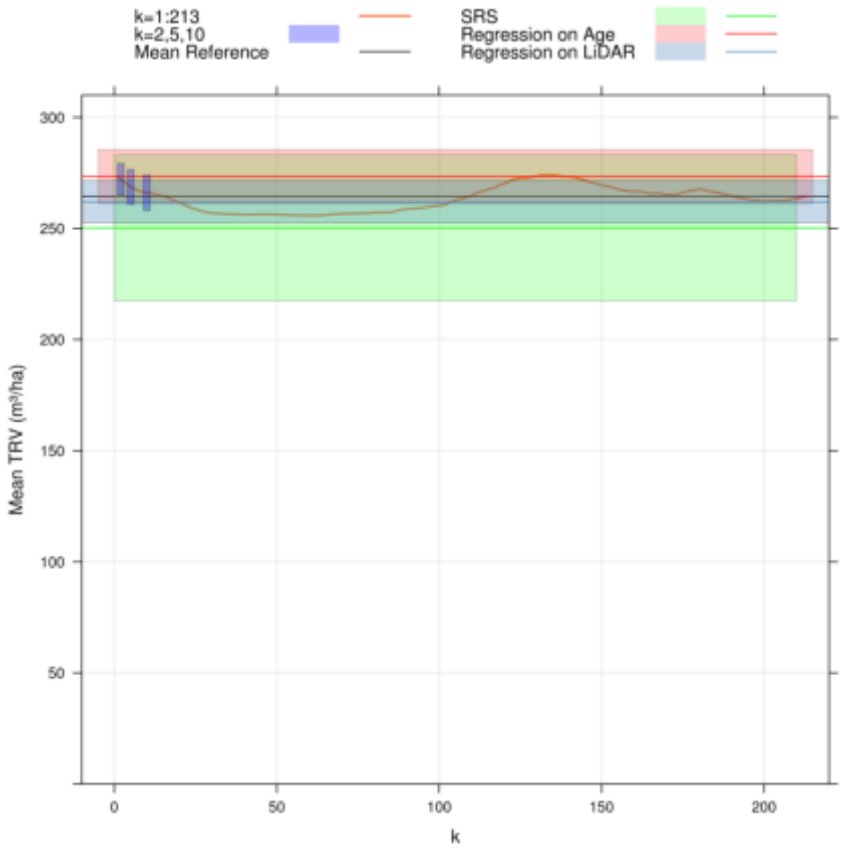

Figure 8. Estimates of total recoverable volume for entire study area

Error! Reference source not found.Figure 8 provides several estimates of the average total recoverable volume per hectare for the entire study area along with 95% confidence intervals on the estimates of the mean for each. “Mean Reference” is the average of the 213 reference plots. These were not all established with known probability so the average of all 213 does not represent an unbiased estimate of the average for the study area. The value is provided for reference only.

Of the 213 reference plots, 187 were established systematically on a 400m grid. They can be used in a conventional systematic or simple random sample estimator of the mean and the variance of the mean; the “SRS” line and band in Figure 8. The SRS estimator is the least efficient estimator of those examined because it makes no use of auxiliary data.

Using stand age as an auxiliary variable in a regression estimator decreases the size of the confidence interval significantly; the “Regression on Age” line in Figure 8. This doesn’t make any use of the LiDAR data. It could be improved by adding other known auxiliary variables but that wasn’t considered necessary for this study.
The simplest way to use the LiDAR data for estimating average total recoverable volume over the entire study area is to use a regression estimator with LiDAR metrics as the auxiliary variables. This is also a fairly robust and efficient way to make use of the LiDAR data. The “Regression on LiDAR” set in Figure 8 used this approach. The first 12 principal components of the available LiDAR metrics were used as auxiliary variables. Counts were first converted to proportions of total counts and metrics with no variation were eliminated. Age and other auxiliary variables that were not sourced from the LiDAR point cloud were not included. The first 12 components represent 96% of the variance in the LiDAR metrics. Backward elimination was used to remove components that did not contribute to prediction of TRV, leaving 6 auxiliary variables as predictors. This approach was not chosen for elegance. It was a simple, brute force way of seeing how much predictive

power could be squeezed out of the LiDAR metrics. The LiDAR data performed a little better than age as a predictor of volume.

The meandering line in Figure 8, labelled "k=1:213" in the legend, provides a kNN estimate for the entire study area for each value of k from 1 through to 213. When every target pixel has 213 nearest neighbours, every reference pixel is used equally and the kNN estimate is the same as the average of the 213 reference pixels. The 95% confidence intervals for k=2, 5 and 10 are provided as the shaded rectangles described as "k=2,5,10" in the legend.

The salient points of Figure 8 are:

- None of the estimators for which confidence intervals were calculated were significantly different to any of the others. This is a useful check of the kNN approach. A difference would suggest that the method had been implemented incorrectly or was fundamentally biased.
- The confidence intervals for the kNN estimates were about the same as for the design-based regression estimator that used the same LiDAR data as auxiliary data. Again, this provides some reassurance that the kNN method was implemented well because there was no expectation that it would provide markedly different confidence intervals.
- If the objective of the inventory was to estimate the total or the average volume over the entire 4000ha of the study area then the use of lower cost auxiliary variables, like age, might provide for more cost-effective estimate than the use of LiDAR. That doesn't mean to say that a smaller number of ground plots combined with LiDAR would not have provided a more cost effective approach than 213 plots without LiDAR. This option was not examined[2].

Table 4 provides the key values from Figure 8.

Table 4 Comparison of methods for estimating mean total recoverable volume for entire study area

Method	Mean (m^3/ha)	Standard error (m^3/ha)	Standard error / mean (%)	PLE (%)
Simple random sample (187 plots)	250.3	16.5	7.6	15.2
Regression on age (213 plots)	273.3	6.1	2.2	4.4
Regression on LiDAR metrics (213 plots)	262.0	4.7	1.8	3.6
kNN (k=2)	272.2	5.2	1.9	3.8
kNN (k=5)	268.7	4.5	1.7	3.4
kNN (k=10)	266.0	4.9	1.8	3.6

[2] The option of using fewer plots with LiDAR could be examined with the data that are available but was

By Stand

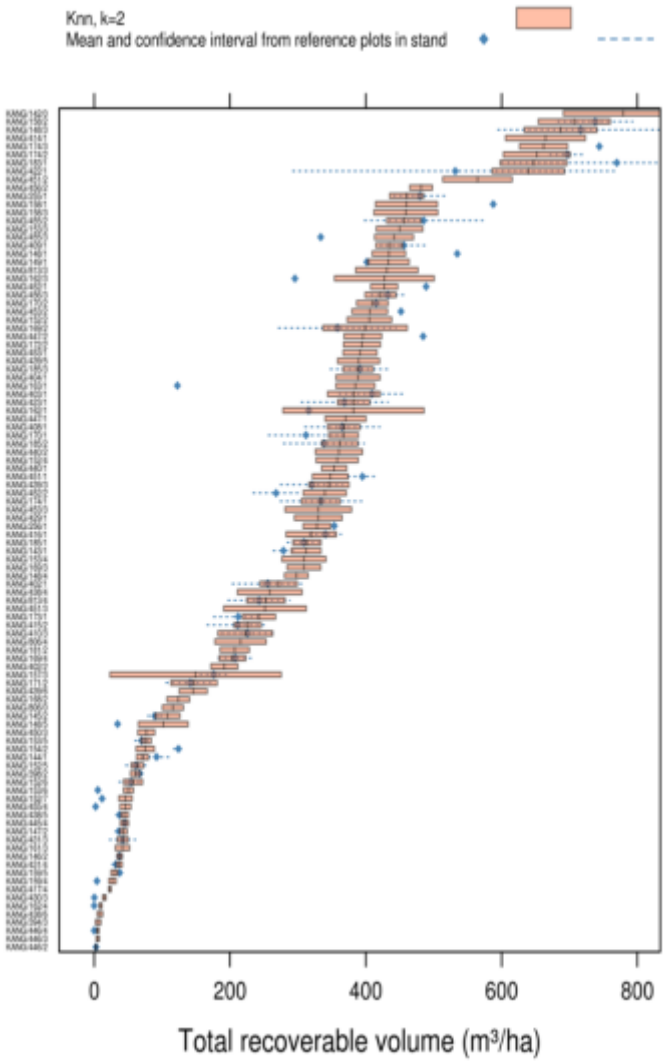

Figure 9. Estimates of total recoverable volume by stand

Figure 9 provides estimates of total recoverable volume by stand for all stands in the study area along with 95% confidence intervals where these were available. The shaded rectangles represent kNN estimates where k=2. The width of the bar represents the 95% confidence interval. The diamond shaped points represent the average of the reference plot values in the stand and the dashed lines show the 95% confidence interval calculated by treating the reference plots in a stand as a sub-population of the study area. The underlying assumption in doing so, that the reference plots represent a simple random sample, is not strictly true but is useful for illustrative purposes.

The salient points of Figure 9 are:

- With only 213 plots it is not possible to provide an estimate of the mean for all stands because many stands do not contain any plots. When the stand contains only a single plot, it is not possible to estimate a confidence interval.
- Where it is possible to estimate a confidence interval from the plots contained in a stand, that confidence interval is almost always wider than the confidence interval estimated from the k nearest neighbours.
- The kNN method provides a mean and confidence interval for every stand in the study area, whether or not that stand contains reference plots. In most cases the confidence intervals are narrow enough for the stand-level estimates to be useful.

Bias

Estimates of sampling error from the kNN method start with the assumption that the underlying model is complete and correct and provides an unbiased estimate at each target pixel and for each AOI. The sampling errors don't take into account model bias for an AOI. For example, a single eucalypt stand in a study area that is predominantly *P. radiata* might not be well served by *P. radiata* reference plots if the relationship between LiDAR metrics and a response variable is species-dependent. In this case, with enough data, a nearest-neighbour model might be expanded to include species but, in general, one should assume that for some areas of interest the estimates will be biased.

In the current study the imputation error for the reference plots themselves was used to provide an indication of the potential magnitude of bias at the stand-level. The imputation error is the difference between the observed value of a response variable at the reference plot and the value imputed from the k nearest neighbours. In this analysis, a reference plot was not allowed to have itself as a nearest neighbour and additionally, because the interest was in stand-level imputation errors, it was not allowed to have any nearest neighbours in the same stand. The second condition provides a better indication of potential bias for stands with no reference plots but was not otherwise used during kNN calculations for an AOI.

Figure 10 shows raw imputation errors by stand for the reference plots that fell within that stand. The response variable is total recoverable volume and k = 1. Each point represents a single reference plot and the solid line is the average for all the reference plots in the stand. The solid line has two components and only one of them is bias. The other is random variation. With enough stands, even without bias some stands will have all their plots on one side of the line.

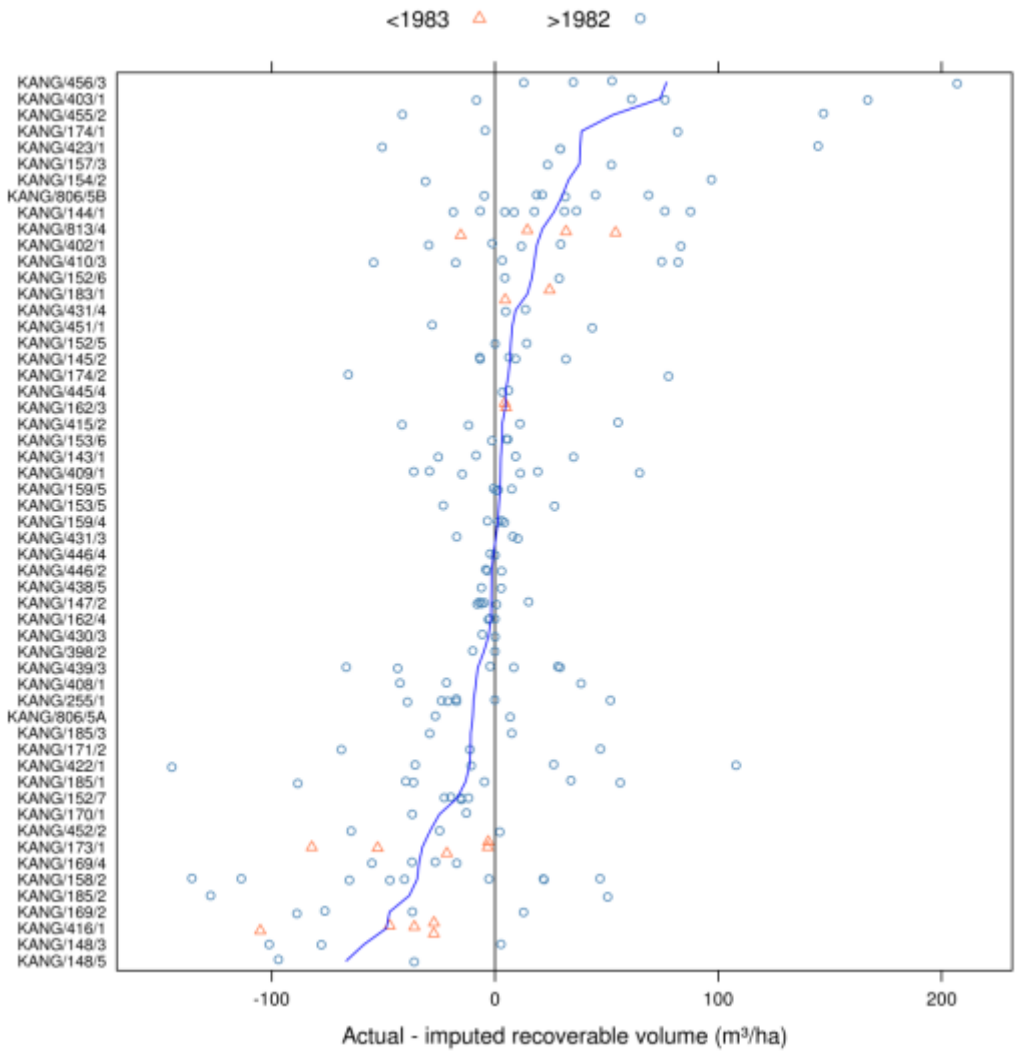

Figure 10. Imputation errors by stand for total recoverable volume where k = 1

A random effects model[3] was used to partition the variation in Figure 10 between a stand-level component (model bias) and remaining noise. The between-stand variation, calculated in this way, has a standard deviation of about 6% of the stand mean. This is not significantly different to zero in a statistical sense (p = 0.09) and could be ignored. On the other hand, it does provide a best estimate of the potential magnitude of stand-level bias; one in which for 95% of stands the absolute bias would be less than 12% of the mean. This is on top of the estimated sampling error but, unlike sampling error, can't be calculated at a stand-level because most stands do not have sufficient reference plots.

Model Evaluation

Following variable selection random forest distance was used to impute a value for all cells (k=5) in the target and reference dataset. Table 5 details the root mean square difference (RMSD) for the

[3] A random-slopes model was the most informative in that the slopes are directly interpretable as a proportional error. However, multiple models were tried with the same conclusion about the magnitude and significance of the stand-level effect. The random-slopes model is $imputed_{ij} = (\beta + b_i)\, observed_{ij} + \varepsilon_{ij}$ where i represents stands, j represents plots within stands, $\beta$is a fixed effect and b_i is a normally distributed stand-level random effect. A power relationship between weights and observed values was used to remove heteroscedasticity. Significance tests relied on a likelihood ratio test of nested models with the

four response variables which can be thought of as analogous to root mean square error in an imputation setting and can be used as a method of assessing model quality. In this instance it is clear that for the reference plots the predictive quality for top height and TRV is superior to basal area and stocking. Scaled RMSD is the RMSD divided by the standard deviation of the reference observations and provides a means of comparing RMSD between responses with different units.

Table 5. RMSD for the response variables

Response	RMSD	Scaled RMSD
TRV	51.10	0.21
Top Height	1.81	0.13
Basal Area	4.89	0.27
Stocking	116.67	0.37

A comparison of the relationship between observed and imputed values in the reference dataset (

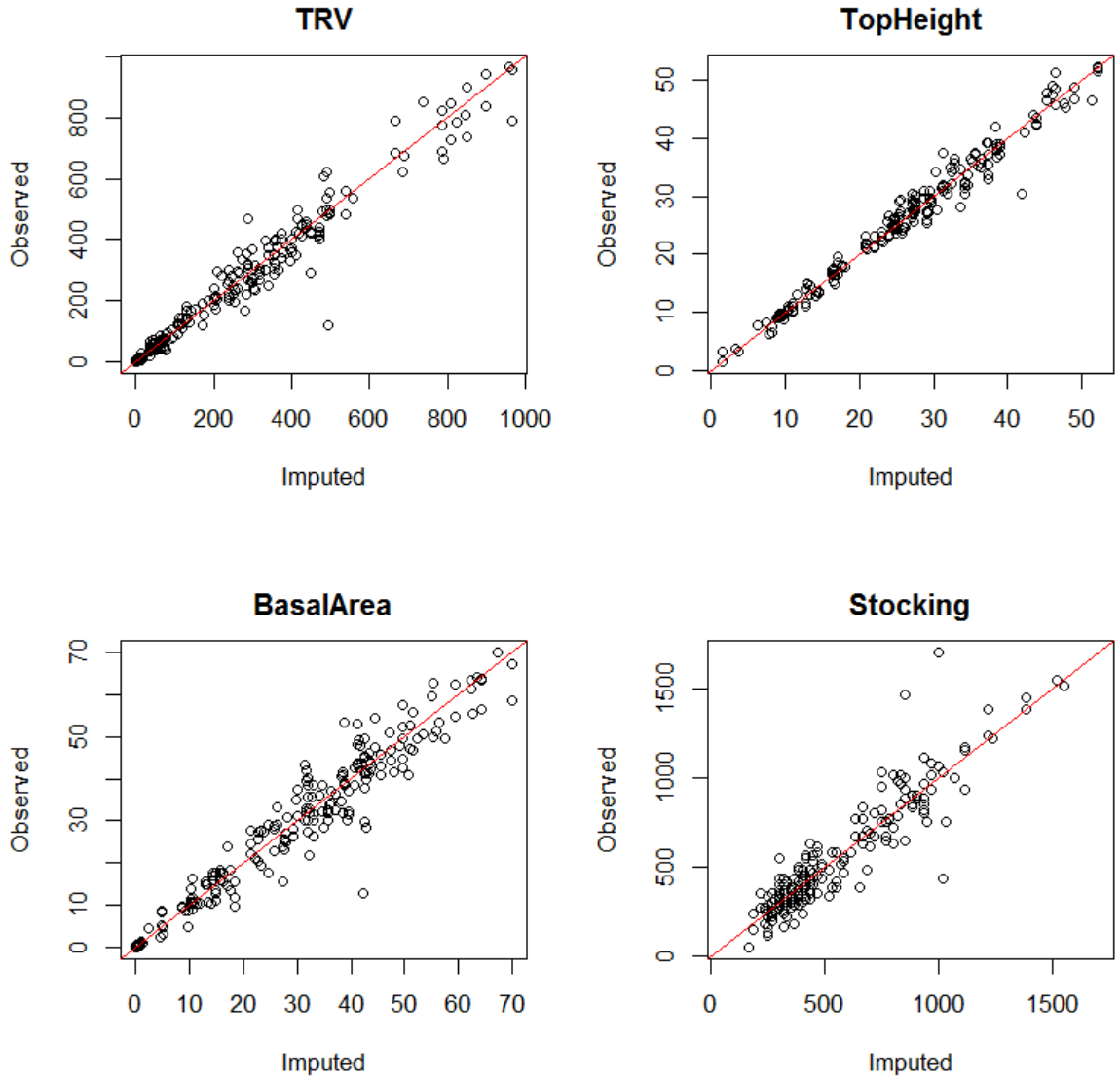

Figure 11) provides some insight into the imputation's predictive quality for each response variable. The prediction accuracy for the response variables is encouraging and there is no evidence of bias for any of the response variables. The predictive quality for basal area and for stocking is worse for the larger values.

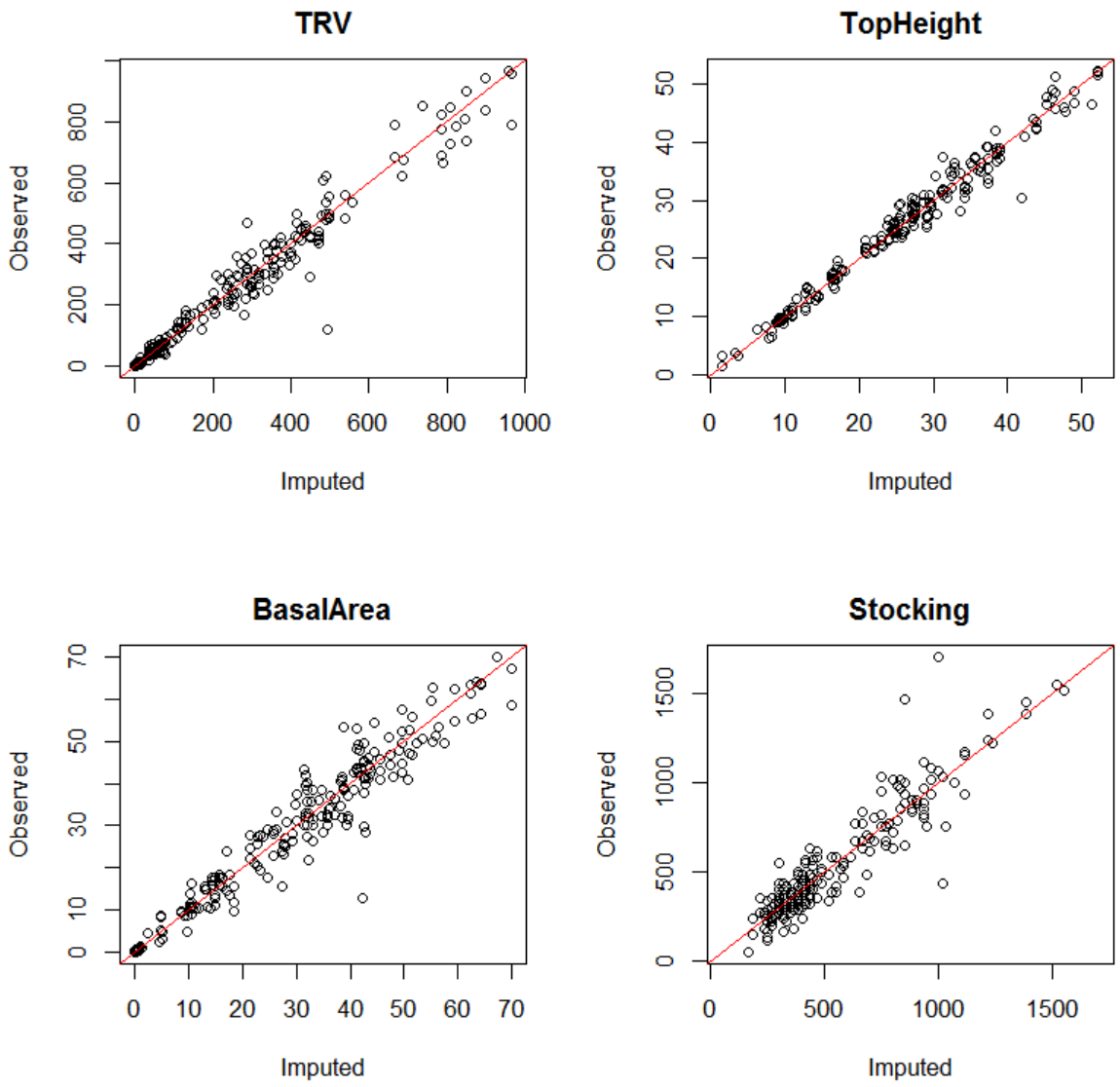

Figure 11. The observed and imputed values for the reference dataset

Graphical analysis of model residuals (Figure 12) suggest that there is no systematic error in model predictions when compared to observations in the reference dataset.

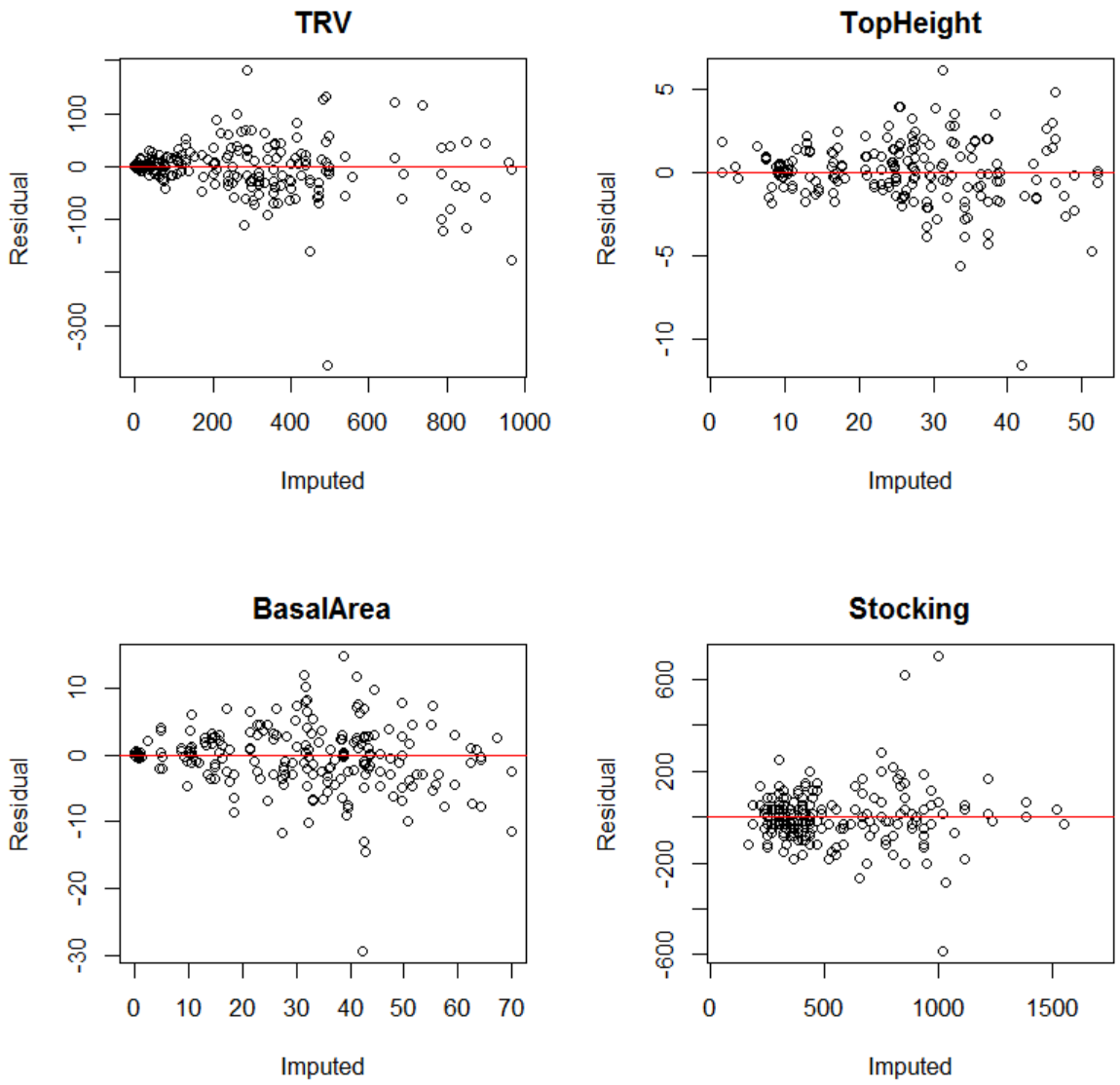

Figure 12. Model residuals (imputed-observed)

Model Validation

The validation dataset was used to compare aggregations of imputed pixels from within the forest manager's stand boundaries with the forest manager's predictions from forest inventory data. A comparison is made for each response variable in the following sections. Where presented in the graphical output error bars represent the 95% confidence interval calculated as documented in previous sections.

The stand level imputed and inventory forest parameter values for each stand in the validation dataset are displayed in (Figures 13– 17). In Figure 13 each datum represents the imputed and inventory value for a single stand. The dashed line shows a unbiased correlation between inventory and imputation and the solid line shows the linear relationship between the values. There is no conclusive evidence of bias in the model residuals for the reference dataset.

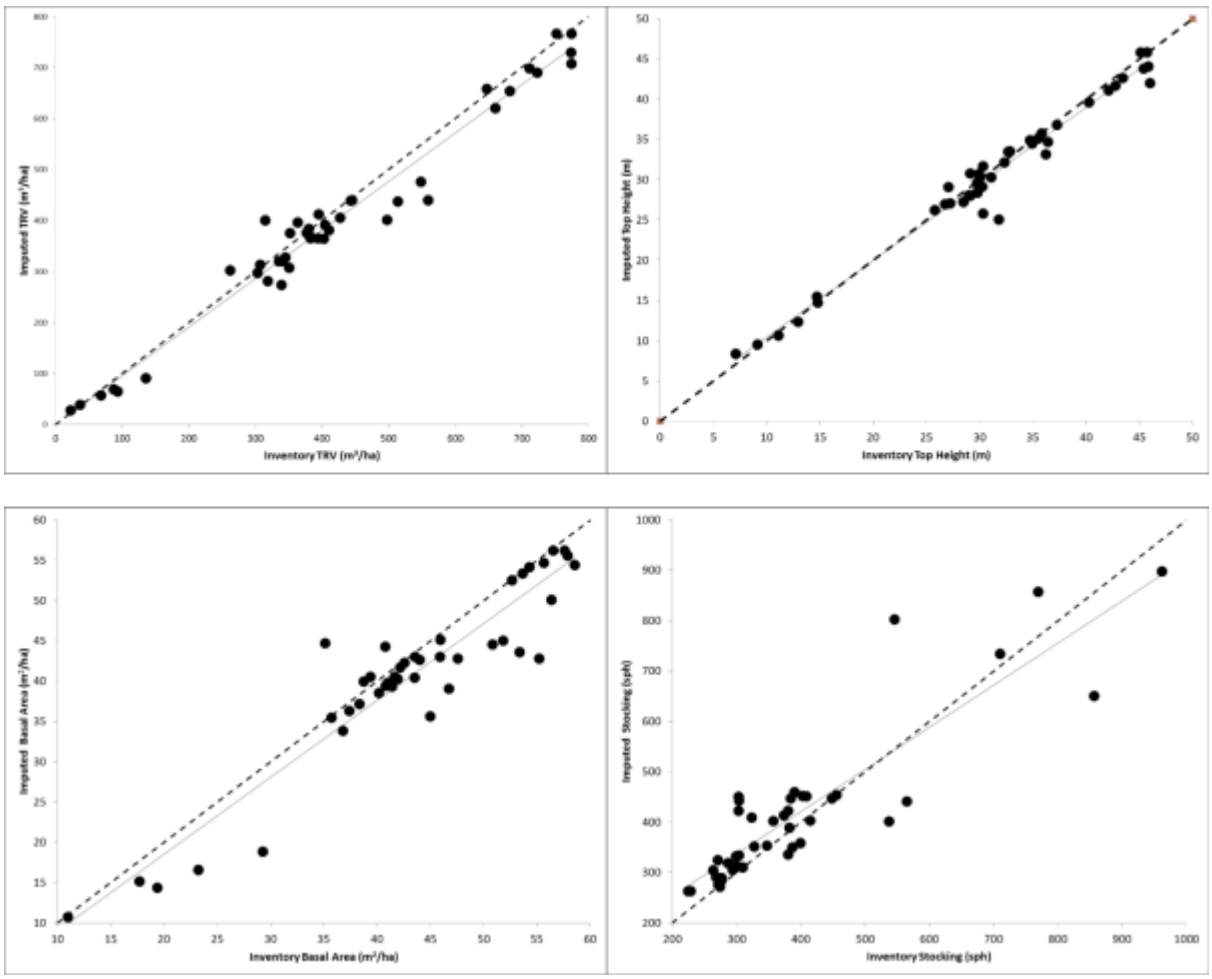

Figure 13. The relationship between imputed and inventory stands for the validation dataset

Figures 14-17 show the imputed and inventory values for each stand in the validation dataset. The imputed and inventory values are very similar for TRV and top height but slightly less well correlated for the basal area and stocking responses. The error bars indicate that for the TRV response the precision of the kNN approach is equivalent to or tighter than the validation dataset in most cases. All stands where there is a considerable discrepancy between imputed and validation TRV were visited and investigated in detail. Through this process it was found that several stands had experienced considerable wind damage in between the forest manager's stand assessment and LiDAR acquisition date. As a result the imputed values in these stands were considerably lower than the validation values which is a very encouraging result as it is indicative of the accuracy gains available using this approach. When compared with the validation dataset the imputed stocking value appears to perform worse for stands with very high stockings. This result in unsurprising as the highly stocked stands are young and prior to full canopy closure there may be insufficient consistency and detail in the LiDAR data to adequately assign neighbours.

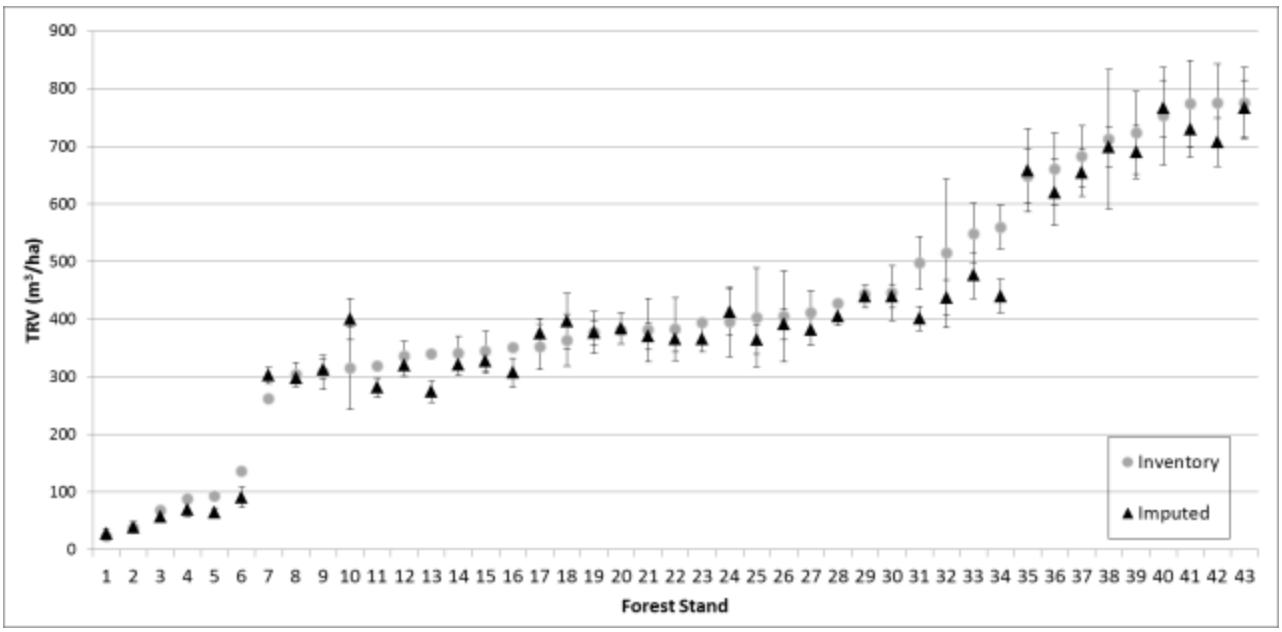

Figure 14. The imputed and inventory TRV for stands in the validation dataset

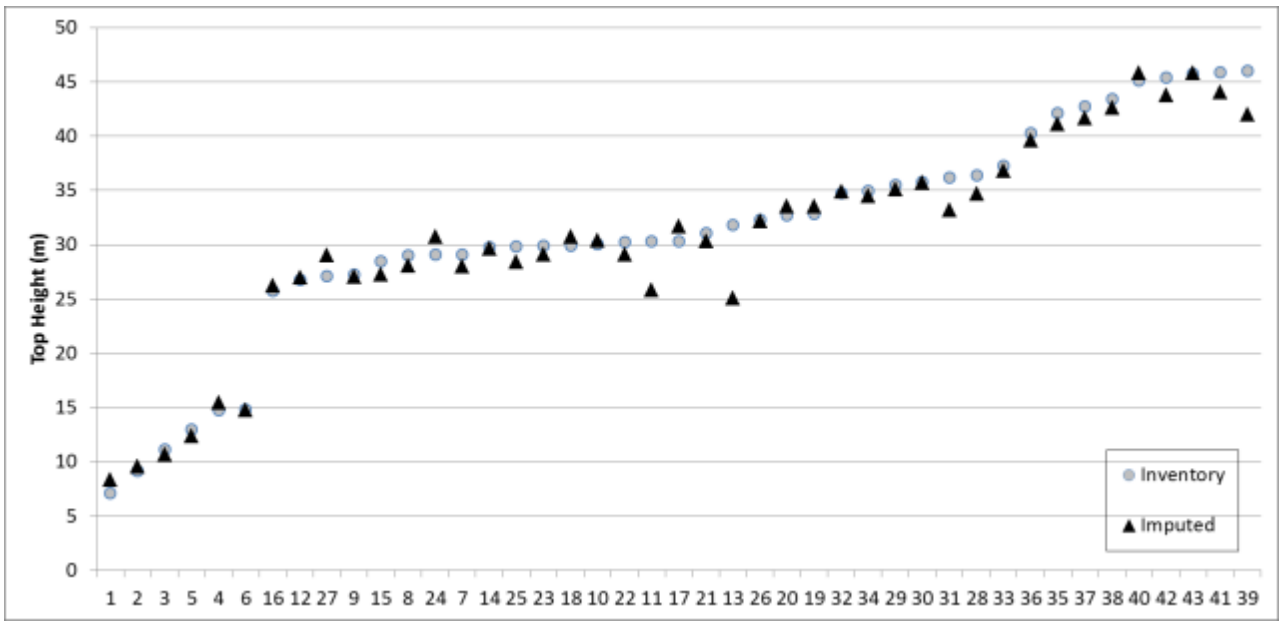

Figure 15. The imputed and inventory top height for stands in the validation dataset

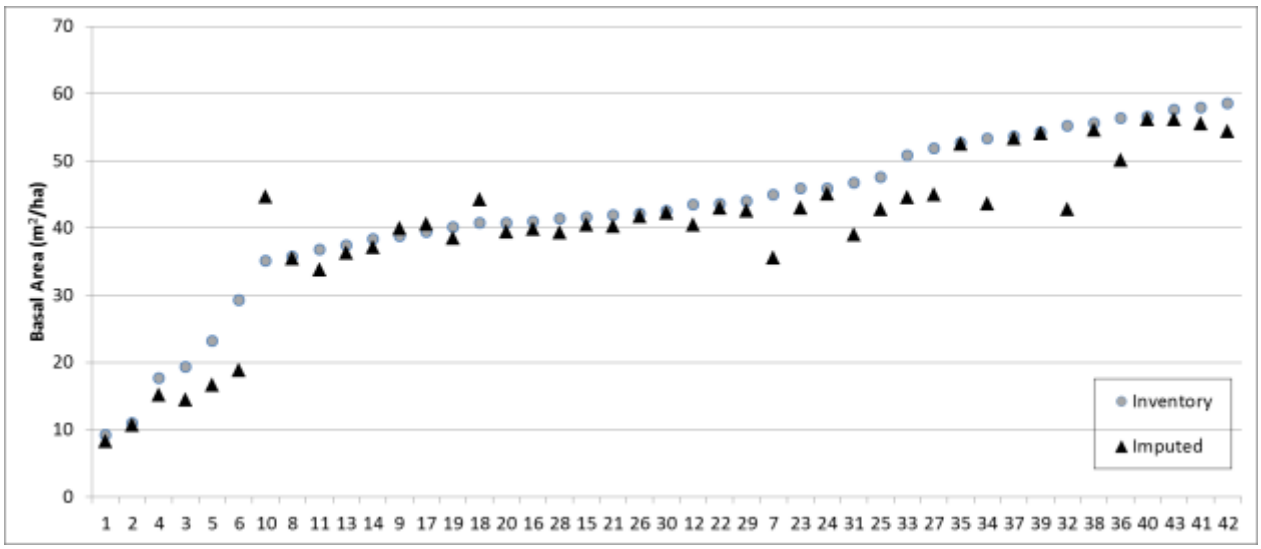

Figure 16. The imputed and inventory basal area for stands in the validation dataset

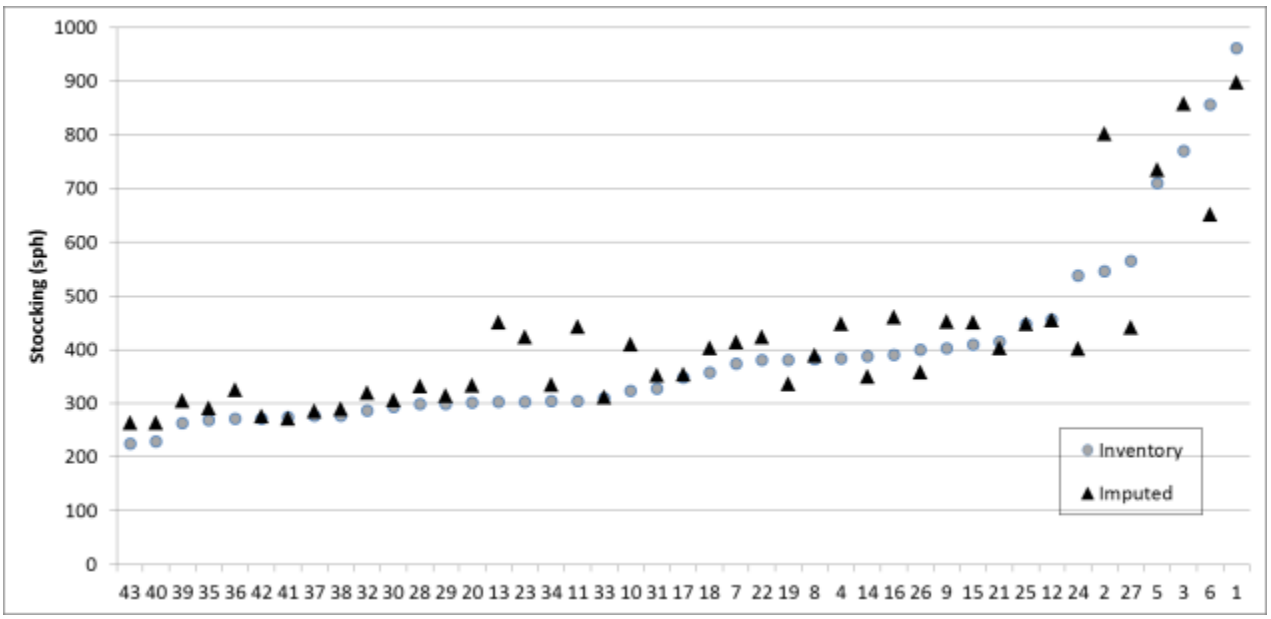

Figure 17. The imputed and inventory stocking for stands in the validation dataset

Grade Mix

The collection of tree descriptions in forest inventory plots is standard practice in the study forest from mid rotation inventory onwards. These stem descriptions are used to estimate product volumes for stands using an optimal log bucking algorithm in a yield analysis software product (Silmetra 2011). The same tree descriptions were recorded in the reference plots where appropriate and so product volumes could be imputed for all target pixels based on the descriptions in the donor plots. A generic cutting strategy used by the forest management company for planning purposes was used to provide a product mix. The cutting strategy used was for comparative purposes only and was not intended to provide accurate predictions of log product recovery. The same cutting strategy was used in both the processing of the reference plots and the validation dataset. To provide a comparison with the validation dataset target pixels within stand boundaries were aggregated in the same manner as in earlier sections.

A comparison of the imputed product mix and the product mix from the validation dataset is presented in Figure 18. From this comparison it seems that the imputed product mix is consistent for the most part with the product mix produced from the validation dataset.

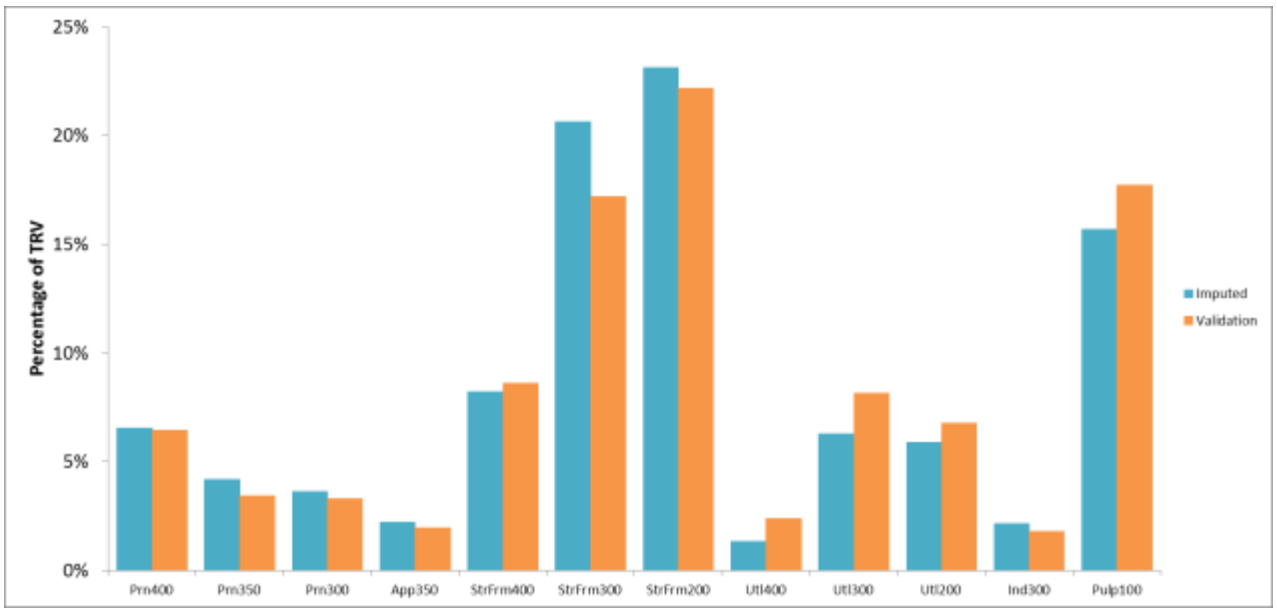

Figure 18. Product mix from the imputed stands and the validation dataset

To compare product mix on a stand by stand basis the product volumes per hectare were multiplied by a nominal value ($/m^3). The values applied were not meant to reflect real world market prices but simply to provide a statistic with which to compare the imputed and validation product volumes. This comparison is shown in Figure 19 and although there is a discrepancy in some stands for the most part the values produced by both methods are comparable.

In some stands the imputed value is considerably higher than the validation dataset and this is due to the prediction of pruned volumes in un-pruned stands. This has occurred because the reference cells providing the product mix are derived from pruned stands whereas in fact the target stands are un-pruned. In a production setting this would be a cause for concern. This could be remedied either by including pruning status as a covariate in the calculation of neighbour proximity or through separate sampling design for the pruned and un-pruned component of the area of interest.

As the purpose of this study is to provide a proof of concept for the technique rather than to produce a final solution the implementation of these solutions is beyond the scope of this project and will be addressed in detail during further work.

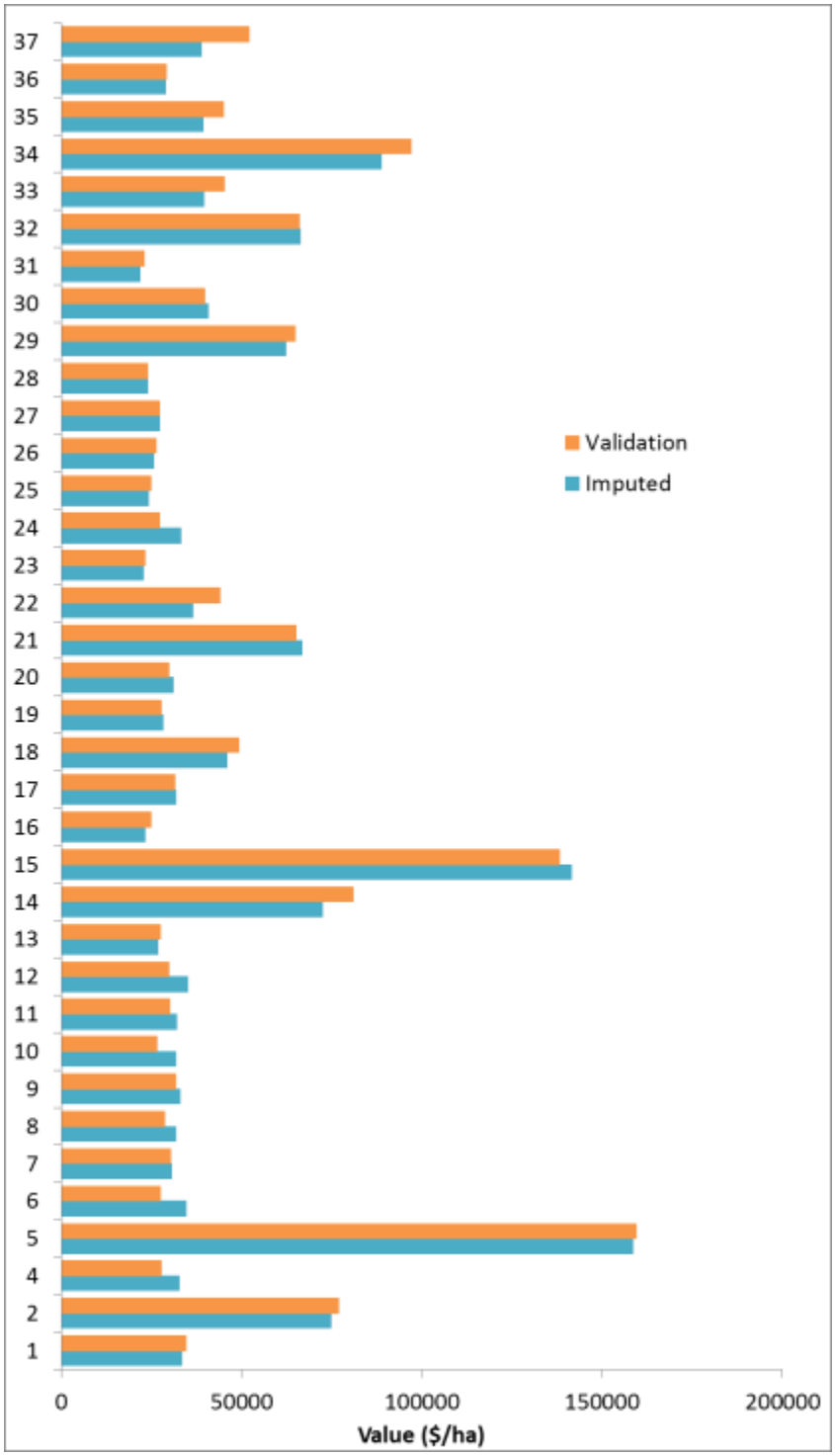

Figure 19. The nominal values of stands in the validation dataset

Yield Table Development

Following the methodology outlined in section 0 yield tables based on the kNN imputation model were developed for stands in the study area of interest. Yield tables based on the imputed values were produced for ages 20-35 and these were compared with yield tables from inventories within the validation dataset (Figure 20).

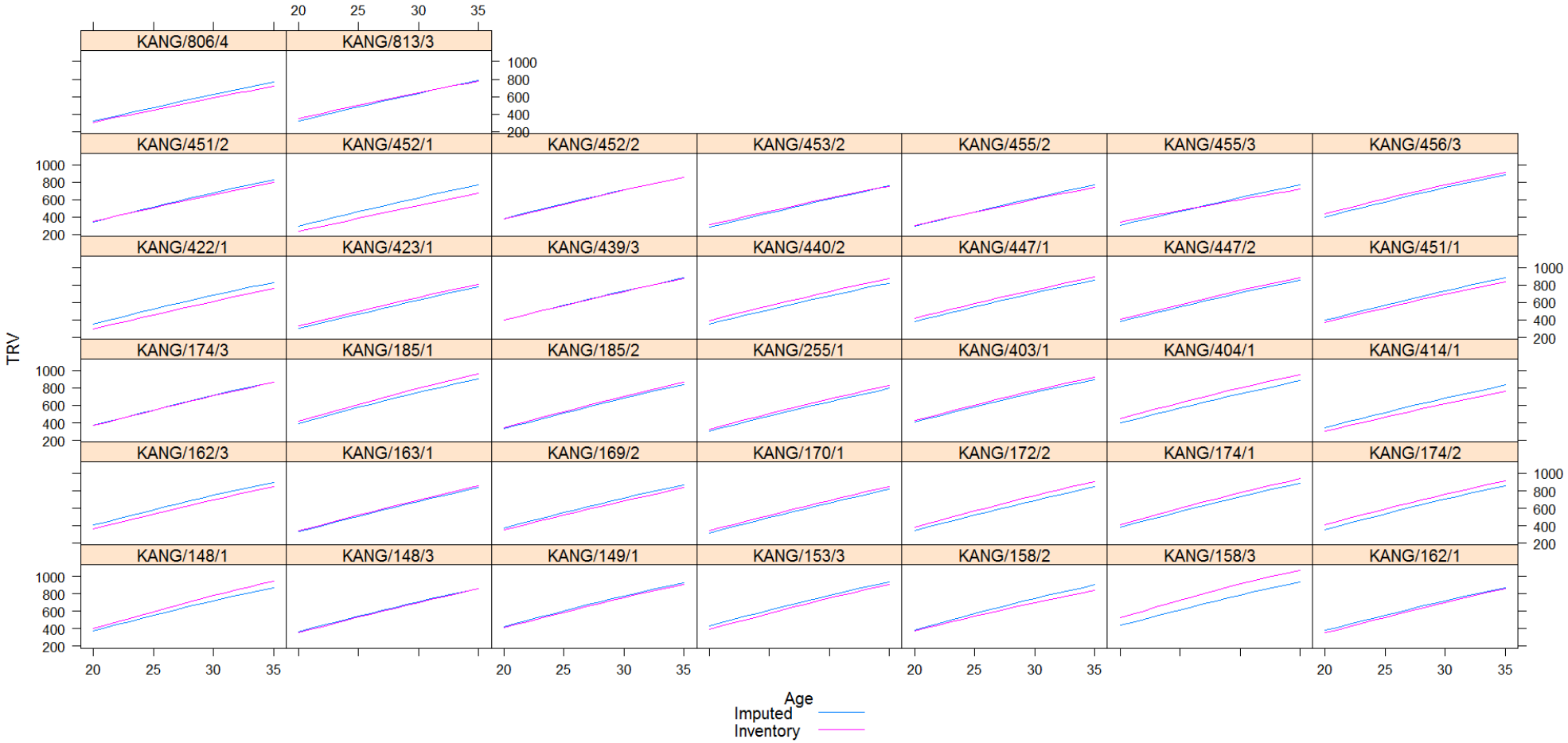

Figure 20. TRV development based on the imputed values and validation (inventory) dataset

Figure 20 suggests that the imputed yield tables do not differ significantly from the yield tables produced from the validation dataset. Figure 20 shows that for the vast majority of stands available for comparison there is only a small discrepancy between the lines representing the TRV development projected by the imputation and that from the validation dataset. There are two stands where the yield projections differ by a considerable margin and the reason for this is unclear. Figure 21 displays a residuals view of the yield tables plotted against the Inventory TRV. This figure suggests there is no conclusive pattern in the residuals indicating that the imputed values of future TRV at specific ages are unbiased. The stands with very different imputed and validation values are visible at the top and bottom of the figure.

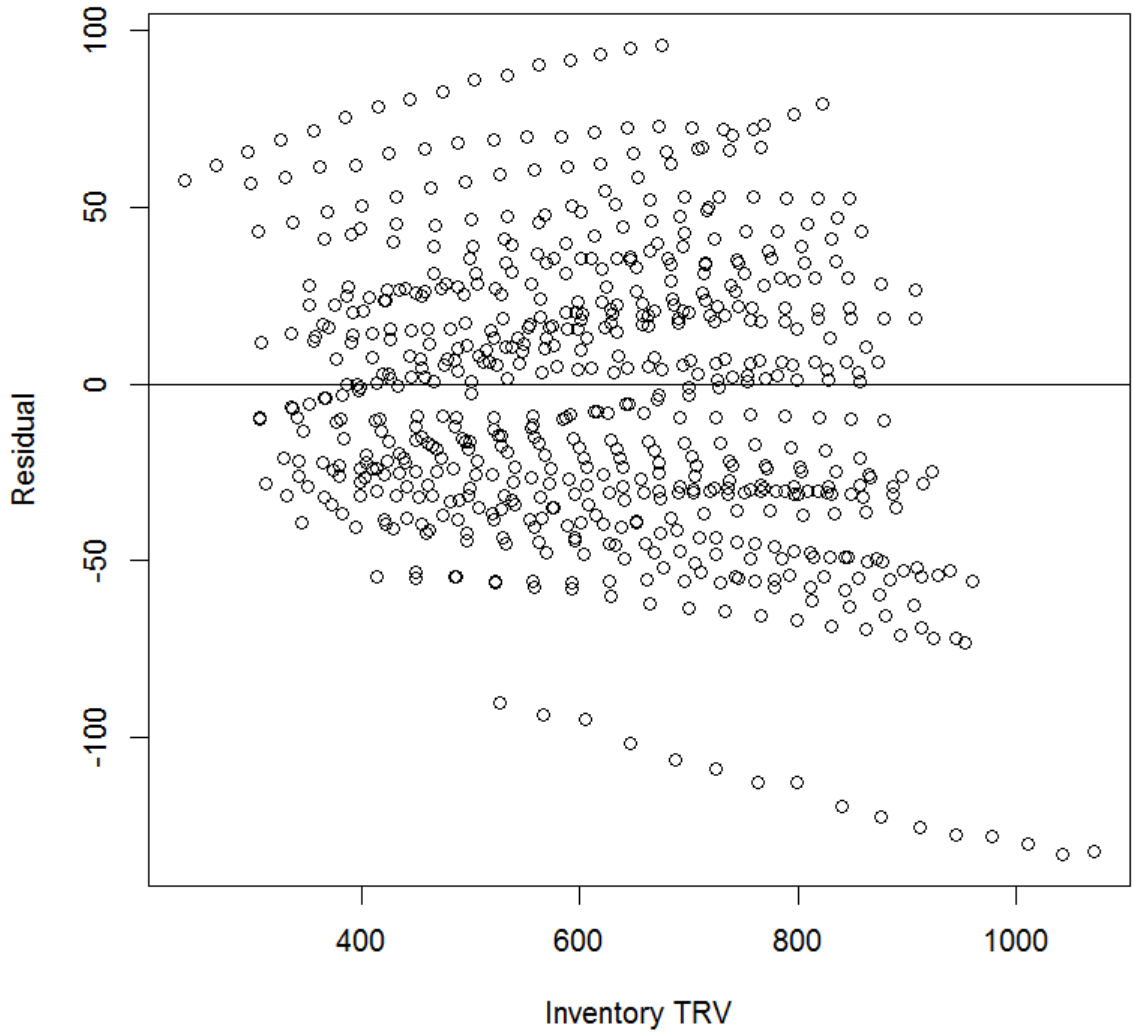

Figure 21. The residuals (imputed-inventory) for the TRV projection

DISCUSSION

The key points arising from this study are discussed in the following sections.

The kNN Approach

In the Kaingaroa case study reported on in this document the kNN approach has been successfully implemented and shows considerable promise. The technique has been harnessed as a method for integrating aerial LiDAR scanning data into a forest inventory system for resource assessment purposes. The results reported on here suggest that the technique can be used to provide accurate results when compared to a validation data set derived from pre-existing stand assessments in the study area. This result indicates that the distance metric used to assess the nearness of different parts of the study area can provide valid and meaningful measures of their similarity.

Any parameter of interest which has been recorded in the reference plots can be assigned to any target pixel in the study area. These predictions could be compared through aggregation of pixels within the forest managers stand boundaries and comparing with the validation dataset. This analysis indicated that the total recoverable volume (m^3/ha) and top height (m) estimated using the kNN imputation approach was extremely similar to the parameter estimates at LiDAR date produced from the validation dataset. The imputation estimates for basal area (m^2/ha) and stocking (sph) were less similar to the validation dataset but still probably adequate for the majority of uses of these parameters in the forest managers information system. The results of the comparison with the validation dataset are particularly encouraging as the number of plots used to derive the kNN estimate is far smaller than that in the validation dataset and the results are comparable. Furthermore the kNN technique can be used to provide estimates for all stands in the LiDAR area, based on the 213 reference plots and LiDAR data, and there is no evidence to suggest that the estimates would be less accurate and precise than those reported on during model validation.

A variable selection algorithm has been developed which has been used to select important predictor variables and discard those which are unimportant based on model prediction error. The variable selection algorithm can be used to selected variables in an unsupervised manner and has been used for the current study dataset and others available to this study's authors. Although a full validation of the performance of the algorithm is beyond the scope of this project it has been observed that using the variables selected by the algorithm produced better predictions than alternative techniques. The quality of prediction was assessed both in terms of the model statistic root mean square difference (RMSD) but also by comparing the stand level model outputs with stand estimates from the validation dataset for various forest parameters.

Log Product Mix

Log product volumes at LiDAR acquisition date, or at a user defined specific stand age, have been calculated using the kNN approach for all pixels in the study area. These can be summarised to provide log volume estimates for any area of interest. To provide a measure of the quality of these predictions target pixels were aggregated at the stand level and compared with estimates from the validation dataset. This process indicated that for the whole study area the grade mix estimates produced by kNN imputation were similar to those produced by the forest manager's traditional stand assessments. In a minority of cases the imputed product mix for a stand differed from the validation values. There could be several causes for this and if the kNN technique is to be installed into a production environment then these issues have to be worked through. For example some stands which the forest manager's stand assessment showed to be unpruned contained pruned volume in the imputed product mix because the donor cells selected for use were pruned. This issue can be overcome in a number of ways by accounting for treatments which have a strong

effect on the log products produced in the modelling process. This can be incorporated into the methodology of this study in a number of ways:

1. Inclusion of silviculltural treatment as a categorical predictor in the proximity calculation. This technique would rely on the accuracy of the forest manager's stand records and previous knowledge of treatments which have a significant effect on certain product volumes. In most cases this would be achievable.
2. Including a volume of a certain product of interest (e.g. pruned log volume) or a statistic which serves as an index of product mix (e.g. plot value) as a response variable during variable selection. Following this methodology predictor variables would be selected which adequately accounted for patterns in the predictor variables to ensure appropriate neighbour selection.
3. Where a silvicultural treatment is known to result in the production, or otherwise, of a product of interest (e.g. pruned/unpruned) this can be accounted for in the project sampling design by splitting the area based on this treatment distinction. This would have the effect of ensuring that target pixels from pruned stands can only acquire pruned plots as nearest neighbours and the same would be true of unpruned plots.

Some challenges remain in the imputation of product mix but the case study implemented here serves as a proof of concept and suggests that the technique can be extended to produce log product volumes in a production setting. Given that the study objectives were focussed on stand information around the early and mid-rotation periods it is likely that the accuracy of the log product volumes produced will be acceptable for their intended use. While some issues remain in the use of this technique for producing log product mix the authors feel they are not insurmountable and that they technique shows considerable potential to deliver log product mix in a production setting.

Yield Table Development

A process for using kNN imputation to derive yield table has been successfully implemented and validated, to some extent, in this case study. The imputed yield tables appeared to be accurate and unbiased when compared to the validation yield tables. These results indicate that the technique implemented here can be used to derive future yields in a manner which can be easily integrated with the forest manager's current yield prediction systems.

Sampling Error Calculation Summary

As part of the study the sampling error for response variable averages was calculated using the method of McRoberts et al, 2007. This method accounts for correlation between target pixels that share the same reference pixel(s) and for correlation between reference pixels that are close together.

At the level of the entire study area, the use of LiDAR metrics as auxiliary variables resulted in a modest improvement in sampling error when compared with the use of stand age as an alternate and lower cost auxiliary variable.

The real advantage of using kNN estimates[4] based on LiDAR metrics is that it is possible to impute values with useful and consistent precision for every stand in the study area, even for stands containing no reference plots.

The sampling error does not incorporate uncertainty due to the kNN model being potentially biased at the stand level for some stands. Analysis of stands with multiple reference plots showed that this bias is small ($\pm 6\%$) and not significantly different to zero in this study area.

[4] This advantage potentially applies to other model-based estimators that do not use a nearest-neighbour

There are some practical issues with calculating sampling error for kNN estimates but none that would be insurmountable in the development of a production system.

Practical Considerations for Sampling Error Calculation

There are three main practical issues associated with calculating and using the sampling errors for areas of interest using the kNN approach:

- The size of the computation
- Estimating spatial correlation
- The errors are not additive

None of these issues is insurmountable.

Size of the Computation

Calculation of sampling error requires computation of the interactions between each pair of target pixels and their k reference pixels in the area of interest[5]. When the area of interest is the entire study area and has 43,548 target pixels then, with k=5, that means computation and summarisation of $(5*43{,}548)^2$ which equates to over 50 billion values. The size of the computational problem has two implications. Firstly, it can't all be fitted into computer memory at the same time so that the calculations must be staged which adds to complexity. This is not something that would be done in a spreadsheet. Secondly, it takes time; about an hour to compute the sampling error for the whole study area, for one value of k and one response variable, with 4 CPU cores running at 100%.

McRoberts et al., (2007), show that the size of the problem can be reduced by sampling. That approach wasn't used here because it would have added extra complexity to the calculations. Sampling would be required in a production implementation.

The size of the problem reduces dramatically when the areas of interest are smaller. For example, dividing the study area into 102 stands and calculating the sampling error for each stand takes about a minute for all 102 stands with one response variable and one value of k. The reason that 102 stands takes a lot less time than one study area is that, in the former case, the covariance between target pixels that are not in the same stand need not be calculated.

Spatial Correlation

Preparing a spatial correlation matrix between the reference plots is complex but not particularly time-consuming. It is complex in that it requires multiple iterations of the process of fitting a semivariogram. All of this can be automated, and was automated for the study. The problem is that the spatial data tends to be noisy so that fitting a semivariogram can fail, or worse can silently produce an implausible outcome. Failure in the context of the study required manual inspection and/or intervention. Automating this for a production system will require more experience with the modes of failure and how best to deal with them.

McRoberts et al, (2007), recommend that reference plots are placed far enough apart that spatial correlation can be ignored. This is sound advice but, depending on the size of the study area, will not always be possible. At a minimum, some check on whether spatial correlation can be ignored in any specific inventory may be required.

[5] In other words the covariance matrix is NxN where N is the number of target pixels but each cell draws

Additivity of Errors and Combining Areas of Interest

In a conventional stand-based inventory programme, each stand inventory is independent and the variances are additive. For example, an annual cut plan might contain 100 stands each with a PLE of 10%. The PLE for the cut-plan can be calculated by combining the 100 independent stand-level variances. Assuming approximately equal stand sizes it would be about 1%; $10/\sqrt{100}$.

If the kNN approach is used then the stand-level errors are not independent and not additive. This means that the sampling error for the wood in an annual cut-plan, or any other higher-level AOI, must be calculated from scratch. In general, the PLE of the total will be higher than the PLE calculated from the sums of stand-level variances.

The non-additivity of errors has implications for downstream information systems that currently assume additivity and for the design of future information systems to cover areas of interest at multiple levels of aggregation.

CONCLUSION

The objectives of this case study have been met through the implementation of a kNN imputation approach which has allowed calculation of accurate stand level forest parameters across the study area. The use of LiDAR data in this manner shows considerable promise and it is likely that with further research and development even more information can be extracted and used. From the results presented here it seems that the technique detailed in this document could replace some components of traditional forest inventory in New Zealand and should serve as a pilot study to a larger scale implementation. The technique has also been used to successfully produce log product mix for stands in the study area although further refinement of the technique will be required for a production implementation. Stand yield tables have also been developed using the kNN technique without significant re-development of the forest manager's information systems and produced in the forest manager's required format. Calculation of sampling error for the kNN values was the key technical challenge addressed during this project and this has been successfully achieved.

Whilst technical challenges remain the authors believe that none are insurmountable and the technique shows great potential as a mechanism for incorporating aerial LiDAR scanning data for forest inventory purposes in a production forest environment.

APPENDIX 1 – SAMPLING ERROR ESTIMATES

The method for calculating imputation error for an area of interest is described by McRoberts et al. (2007). This section does not duplicate the equations or proofs in the original journal article to which the reader is referred. It sets out only to clarify which steps were used and to some extent how they were implemented. References to equations are to those in the journal article.

Given an area of interest with N target pixels

1. Prepare a correlation matrix between reference plots using several iterations through equations 6a, steps 1-4 on p 471 and 7b. In the first iteration variance ($\hat{\sigma}_i^2$) was constructed under the assumption that no spatial correlation existed (equation 6b) and iteration stopped when no cell in the correlation matrix changed by more than 0.01 between the penultimate and ultimate iteration. If convergence failed on the fitting of the empirical semi-variogram then spatial correlation was assumed to be zero. The reasonableness of this assumption was checked with visual inspections of semivariogram plots.
2. Calculate local variance ($\hat{\sigma}_i^2$) for each target pixel from equation 6a using the correlation matrix from step 1.
3. Calculate variance of kNN estimators using equations 8a, 9a, 14a and 14b.

These steps were implemented using the R programming language (R Development Core Team 2012). Empirical semi-variograms were fitted using the gstat package of R (Pebesma 2004).

REFERENCES

Aarts, E. and Lenstra, J.K. 1997 Introduction. In Local search in combinatorial optimisation. Edited by Aarts, E., and Lenstra, J.K. John Wiley and Sons, New York, pp1-16

Avery, T.E., and Burkhart, H.E. (1994) Forest Measurements. 4th Edition. McGraw-Hill, Boston. pp 408

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

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

Crookston, N.L. and Finley, A. (2008). yaImpute: An R package for kNN imputation. Journal of Statistical Software 23, 10.

Diaz-Uriarte, D. (2012) R Package varSelRF. Variable Selection using Random Forests.

Fallkowski, M.J., Hudak, A.T. Crookston, N.L., Gessler, P.E., Uebler, E.H., and Smith, A.M.S. (2009). Landscape-scale parameterization of a tree-level forest growth model: a k-nearest neighbour imputation approach incorporating LiDAR data. Can. J. For. Res. **40** 184-199

Freese, F. (1962). Elementary Forest Sampling. US Dept. Agriculture Handbook No. 232.

Hudak, A. T., Crookston, N.L., Evans, J.S., Hall, D.E., Fallkowski, M.J. (2008) Nearest Neighbour Imputation of Species-Level, plot scale structure attributes from LiDAR Data. Remote Sensing of Environment 112. Pp2232-2245.

Hudak, A.T. Evans, J.S., Crookston, N.L., Falkowski, M.J., Steigers, B.K., Taylor, R., Hemingway, H. 2008. Aggregating Pixel level basal area predictions derived from LiDAR data to industrial forest stands in North-Central Idaho. USDA Forest Service Proceedings RMRS-P-00

Husch, B., Beers, T.W., Kershaw, J.A. (2003) Forest Mensuration, 4th Edition, John Wiley & Sons, Inc, New Jersey. pp 443.

Hyvarinen, A. and Oja, E. (2008). Independent component analysis: Algorithms and applications. Neural Networks, 13 411-430.

Kirkpatrick, S. Gelatt, C.D., Vecchi, M.P. 1983. Optimisation by Simulated Annealing. Science, Vol. 220. No. 4598, pp. 671-680 doi: 10.1126/science.220.4598.671.

Latifi, H., Niothdurft, A., and Koch, B. (2010) Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR derived predictors. Forestry, Vol. 83 No. 4

Liaw, A. and Wiener, M. (2013) Breiman and Cutler's random forests for classification and regression. R package. http://stat-www.berkeley.edu/users/breiman/RandomForests

MacGaughy, R.J. (2010) FUSION/LDV: Software for LiDAR Data Analysis and Visualization. Pacific Northwest Research Station. Forest Serice. USA. pp. 150.

Magnussen, Steen, McRoberts, Ronald E. and Tomppo, Erkki O. (2009) *Model-based mean square error estimators for k-nearest neighbour predictions and application using remotely sensed data for forest inventories*, Remote Sensing of Environment, Vol. 113.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. Proceedings of the National Institute of Science of India, 12 (pp.49-55)

McRoberts, Ronald E., et al., et al (2007). Estimating areal means and variances of forest attributes using the k-Nearest Neighbours technique and satellite imagery, Remote Sensing of Environment 111, pp. 466-480

Mcroberts, R.E. (2012) Estimating forest attribute parameters for small areas using nearest neighbour techniques. Forest Ecology and Management 272 (2012) 3-12

Means, J.E., Acker, S.A., Fitt, J.B., Renslow, M., Emerson, L., Hendrix, C.J. (2000) Predicting forest stand characteristics with airborne scanning lidar. Photogrammetric Engineering & Remote Sensing. 66, 1367-1371

Moeur, M. and Stage, A.R. 1995. Most similar neighbour: An improved sampling inference for natural resource planning. Forest Science 41, 337-359

Musk, R. 2011. Stand level inventory of eucalypt plantations using small footprint LiDAR in Tasmania, Australia. Proceedings Silvilaser 2011, Oct 16-20 2011. Hobart, Australia.

Naesset, E. (1997). Estimating timber volume of forest stands using airborne laser scanner data. Remote Sensing Of Environment, 61, 246-253

Ohman, J.L., and Gregory, M.J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbour imputation in coastal Oregon, U.S.A. Canadian Journal of Forest Research 32 pp.725-741.

Packalen, P., Temesgen, H., and Maltamo, M. (2012) Variable selection for nearest neighbour imputation methods used in remote seeing based forest inventory. Can. J. Remote Sensing, Vol 38 No 5 pp1-13.

Parker, R.C., Evans, D.L. (2004). An application of LIDAR in a double-sample forest inventory. Western Journal of Applied Forestry. 19, 95-101.

Pebesma, E.J. (2004). Multivariable geostatistics in S: the gstat package.

Pont, D. Tairua LiDAR Case Study - Stand Selection. Scion. 2013. p. 10.
R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 390051-07-0, URLhttp://www.R-project.org/.

Ripley, B. 2002. Package RODBC An ODBC database interface for R.

Rombouts, J., Ferguson, I.S., Leech, J.W. (2008) Variability of LiDAR volume prediction models for productivity assessment of radiata pine plantations in South Australia. Proceedings Silvilaser Sept 17 – 18 2008, Edinburgh U.K.

Silmetra Limited, (2011), YTGen User Manual, version 2.9.0

Silmetra Limited, (2011), Plotsafe User Manual, version 1.5.0

Tomppo, E. 1991. Satelite image based national forest inventory of Finland. In: Proceedings of the symposium on Global and Environmental Monitoring Techniques and Impcats. 17-21 September 1990, Victoria, British Columbia, Canada. International Archives of Photogrammetry and Remote Sensing 28, 419-424

Tomppo, E. 1996. Multi-source national forest inventory of Finland In: Palvinen, R. Vanclay, J., Miina, S. (Eds) New thrusts in Forest Inventory, Proceedings of the Subject Group S4.02-00. Forest Resource Inventory and Monitoring and Subject Group S4.12-00. Remote Sensing Technologies. Vol 1, IUFRO XX World Congress, Tampere Finland, 6-12 August 1995. EFI Proceedings 7, 27-41

YTGEN User Group, (2007). PlotSafe Overlapping Feature Cruising Forest Inventory Procedures. Interpine Forestry Ltd. Available from https://www.interpine.co.nz/PlotSafe%20and%20YTGEN%20Download%20Files/PlotSafe%20Cruising%20Manuals%20and%20Information/PLOTSAFE%20Overlapping%20Feature%20Cruising%20Forest%20Inventory%20Procedures%20-%20Feb%202007.pdf

For further information please contact:

Jonathan Dash
Interpine Forestry Limited
Email : jonathan.dash@interpine.co.nz
Telephone : 07 345 7573
Facsimile : 07 345 7571