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

Theme: Radiata Theme

Task No: F10502
Milestone Number: 5.02.6

Report No. : R077

Extension Nearest Neighbour Imputation of Stand Attributes using LiDAR Data Additional Dataset: Tairua Forest

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

Research Provider:
Interpine

Date: 26 June 2013

TABLE OF CONTENTS

EXTENSION: ADDITIONAL DATASET FOR TAIRUA FOREST

A two phase dataset consisting of aerial LiDAR scanning data and field data from ground plots has been made available by Rayonier NZ Ltd, which will allow the novel kNN forest inventory technique to be trialled again. The technique was first implemented in New Zealand in a previous case study in Kaingaroa forest[1], the new dataset allows application of the technique in a forest with very different conditions to the Kaingaroa forest operational trial reported previously.

Interpine had no control or knowledge of sampling design or data collection for this project prior to its implementation, and its subsequent use as an extension dataset for review in the use of kNN Imputation. Therefore limitations were experienced in the application of kNN imputation but this still serves well to validate the usefulness of the approach.

Study Area

Tairua forest is in the vicinity of the Coromandel Peninsula in the North Island of New Zealand. The study area is 89km^2 and was considerably more variable than the Kaingaroa study area in terms of altitude range and relief.

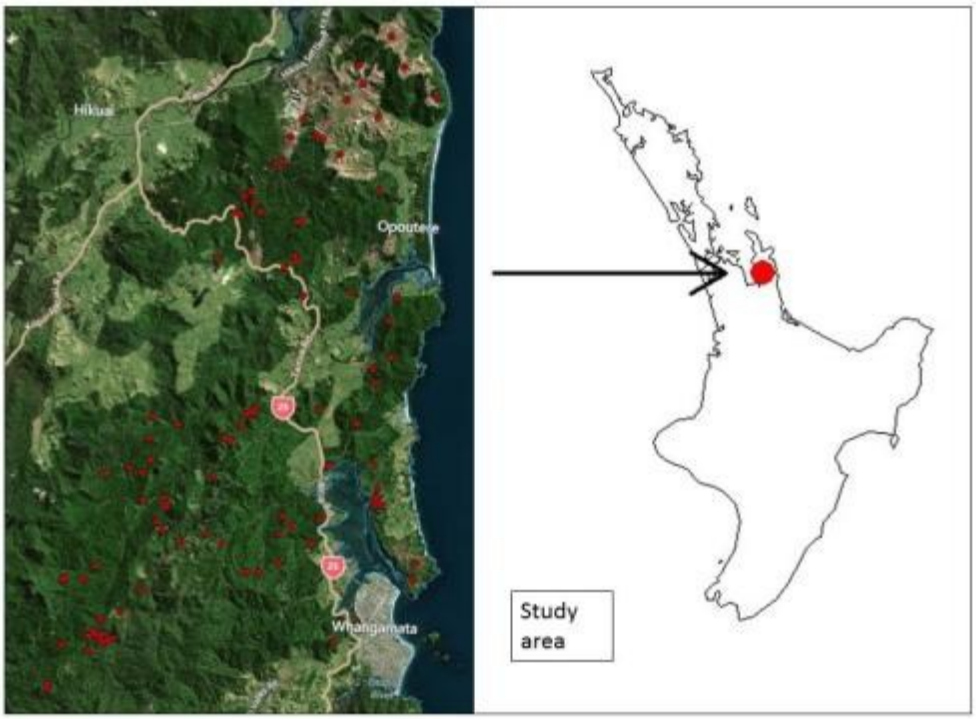


Figure 1. The study area showing allocation of ground plots on the left hand panel.

Field Sampling

The precise methodology of data collection is unknown but sufficient plot information was provided to Interpine to allow the calculation of total recoverable volume (TRV), basal area, top height and stocking in the same manner as for the Kaingaroa trial. Overlapping stem descriptions were also recorded which allow log product volumes to be produced. There were 99 plots with a post differentially corrected high grade GPS fixed plot centre positions which would serve as the

1 Nearest neighbour imputation of stand parameters using aerial LiDAR data – Dash, J.P. Marshall, H.M. and Rawley, B. Prepared for Future Forest Research (2013) Unpublished

reference population for this case study. It is noted that the sample size of 99 ground plots is probably inadequate to serve as a reference dataset for a nearest neighbour approach and that a minimum sample size of 200 plots is probably more appropriate. Furthermore the sampling methodology for plots meant that only 60 could be used to check for bias in the model outputs, these were the only ones where sample inclusion probability could be reasonably calculated.

LiDAR Sampling

As in the Kaingaroa case study 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 7-9 July 2012. An Optech ALTM 3100EA scanner was used at a flying height of 1650m above mean ground level acquiring data at a designed pulse density of 2 per square metre per swath with a 50% swath overlap. The point cloud data was then classified into ground, first and, intermediate returns using automated routines tailored to the project land cover 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 dataset. This editing involved visually checking over the data and changing the classification of points into and out of the ground point dataset. Aerial Surveys reported a resulting mean pulse density of 2.38 points per square metre per swath.

The Cloudmetrics function in the FUSION LiDAR analysis software product was used to produce various statistical parameters describing the LiDAR dataset in terms of point elevations and intensity. These statistical parameters (Table 1) served as candidates for the predictor variables used in this analysis. These variables would be used in both the target and the reference dataset.

Table 1. The LiDAR metrics which served as candidate predictor variables

Metric	Description	Selected Tairua
Total return count above 0.50	Number of returns above 0.5 height	
Elev minimum	Minimum height	x
Elev maximum	Maximum height	
Elev mean	Mean height	
Elev mode	Modal height	
Elev stddev	Standard deviation of heights	
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	x
Elev L1 – L4	L-moment 1 to 4 of heights	
Elev L CV	L-moment coefficient of variation of heights	x
Elev L skewness	L-moment skewness of heights	
Elev L kurtosis	L-moment kurtosis of heights	
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	

Metric	Description	Selected Tairua
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	x
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	
Percentage first returns above mode	Percentage first returns above modal height	
Percentage all returns above mean	Percentage all returns above mean height	x
Percentage all returns above mode	Percentage all returns above modal height	x
(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	x
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 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	x
Int maximum	Maximum intensity	
Int mean	Mean intensity	x
Int mode	Modal intensity	
Int stddev	Standard deviation of intensity	
Int variance	Variance of intensity	
Int CV	Coefficient of variation of intensities	
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 1st – 99th percentile	x
Surface slope	Surface slope (degrees)	

Metric	Description	Selected Tairua
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	x

Model Development

Model development, variable selection and calculation of sampling error followed the same procedures as detailed for the Kaingaroa case study.

Validation Dataset

As in the Kaingaroa trial a validation dataset has been made available consisting of the stand assessments implemented by the forest manager projected to LiDAR acquisition date using the forest manager's regular yield prediction systems. To provide a measure of predictive accuracy a comparison will be made with pixels aggregated within the stand boundaries also provided by the forest manager. A large number of inventory assessments were made available by the forest managers. The inventoried areas were not consistent with the stand boundaries in all cases and some of the inventories were deemed unsuitable for inclusion in the validation dataset. Stand inventories were excluded from the validation dataset if they failed to not meet the following conditions:

- Stands must have a minimum of 3 plots;
- Inventories must not be older than 8 years old;
- Stand area must be within 2 ha of inventory area as recorded in the inventory population. This requirement was aimed at eliminating mismatches where an inventory included more than one stand or where the inventory from a larger stand in the area had been applied to a subject stand.

Incorporating these filters on the forest managers stand assessments a validation dataset of 23 stands was extracted for validation purposes and grown to LiDAR acquisition date using the forest manager's yield prediction framework.

Results

The results of the application of the kNN technique to the Tairua dataset are summarised in the following sections.

Variable Selection

20 of the 101 candidate predictor variables were selected for inclusion in the calculation of statistical proximity. The selected variables are noted in Table 1.

Sampling Error

Data from the Tairua forest became available after completion of the analysis of the Kaingaroa data. Where possible the same analyses were used and the results of these analyses are reported here.

In the Tairua study, because of the non-contiguous nature of the forest, the area for which LiDAR data was available was much larger than the area within which reference plots were placed. The exact scope of the latter changed as reference plots were placed. The following analysis is restricted to target pixels in stands established from 1977-2001 (inclusive).

The analysis includes 99 reference plots and 46 294 target pixels (4 166 ha).

The selection of point locations for reference plot establishment did not use a simple design. At the time of writing this report, insufficient information on selection probability was available to develop probability-based estimators for comparison with the kNN estimates and these have been omitted. It is likely that this situation could be remedied with access to more information.

Figure 2 provides kNN estimates of total recoverable volume for k=1-99 (orange line) with 95% confidence intervals for k = 2,5,10,20,50 and 98 (blue bars), the mean of the reference plots (black line), and the means and confidence intervals for two probability-based estimators from a subset of the plots. Table 2 provides key values from **Error! Reference source not found.**.

The "Variable probability" entry uses 60 of the reference plots for which a sample inclusion probability could be calculated with reasonable confidence. It is equivalent to the SRS estimator in the Kaingaroa case study but takes into account that reference plot locations were not selected with equal probability.

Table 2 Comparison of methods for estimating mean total recoverable volume for entire Tairua study area

Method	Mean (m^3/ha)	Standard error (m^3/ha)	Standard error / mean (%)	PLE (%)
Variable probability (60 plots)	341.3	37.0	10.8	21.7
Regression on age (60 plots)	367.9	12.9	3.5	7.0
kNN (k=2) (99 Plots)	350.3	9.8	2.8	5.6
kNN (k=5) (99 Plots)	353.1	10.2	2.9	5.8
kNN (k=10) (99 Plots)	356.5	10.4	2.9	5.9
kNN (k=20) (99 Plots)	364.2	12.0	3.3	6.6
kNN (k=50) (99 Plots)	408.5	18.0	4.4	8.8
kNN (k=98) (99 Plots)	447.4	23.7	5.3	10.6

Salient points of **Error! Reference source not found.** and Table 2:

- **Error! Reference source not found.** clearly shows how the kNN estimate becomes increasingly biased as the value of k increases. The mean of the reference plot volumes is not an unbiased estimate for the study area. It is an over-estimate because greater weight was placed on older stands. The kNN estimate with k = 98 must therefore also be biased.
- The PLE is higher than for the Kaingaroa study but this is consistent with the use of half the number of reference plots. All things being equal, halving the number of plots should result in an increase in PLE by a factor of 1.4. For k=2, the increase is 1.47.
- Sampling error increases with k.
- The difference between the PLE of 5.6% for k=2 and the PLE of 7% for the regression estimator is likely to be attributable to the smaller number of plots available for the latter.

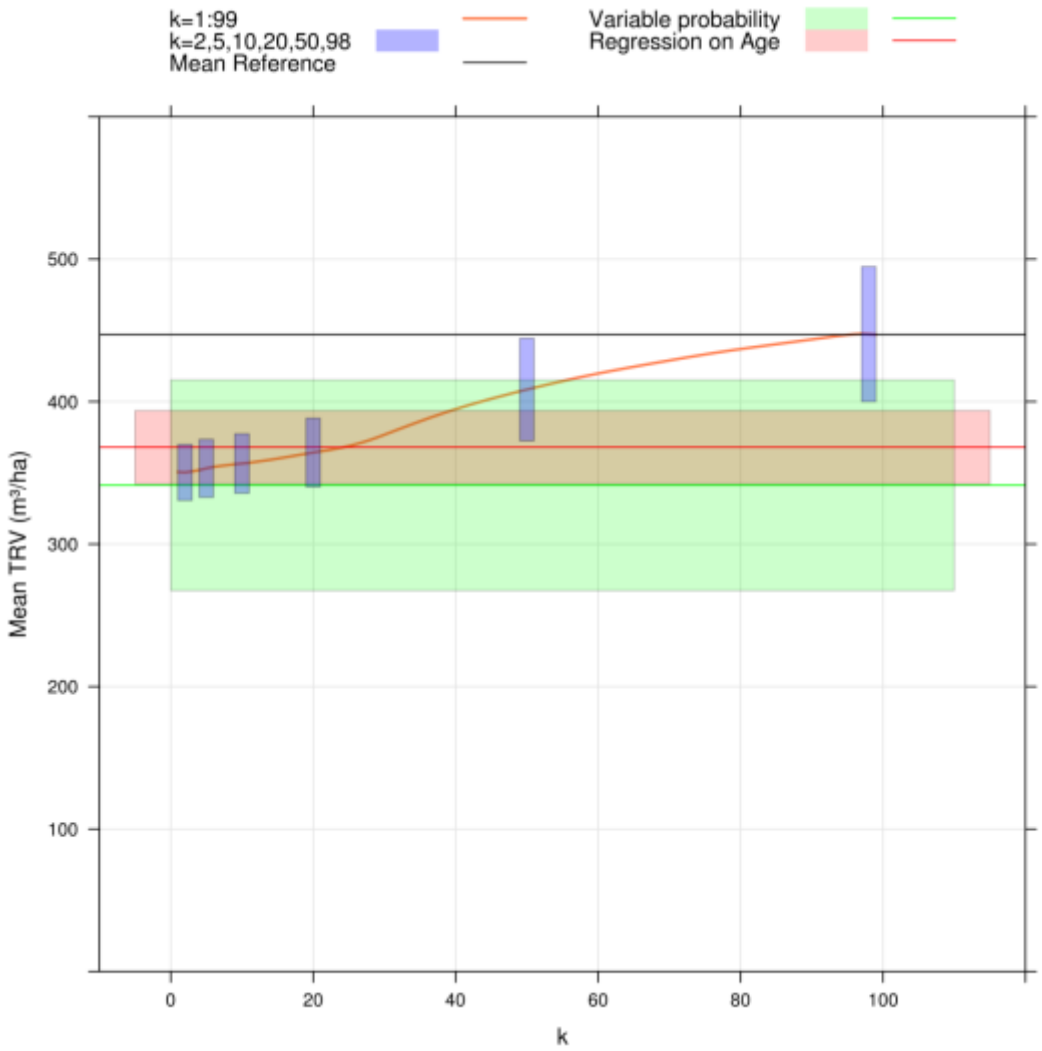


Figure 2 Estimates of total recoverable volume for Tairua study area obtained using a range of methodologies.

Figure 3 provides estimates of total recoverable volume by stand for a random sample of stands in the study area along with 95% confidence intervals for the kNN estimates. There were too many stands to show all of them. The same points may be made about Figure 3 as were made about Figure 9 in the Kaingaroa case study[2]. In addition, it is worthwhile comparing the two graphs. The stand-level confidence intervals in Tairua (Figure 3) are somewhat wider than those in Kaingaroa. The median stand-level PLE in the Tairua study is 12%, compared with 10% in Kaingaroa. The difference at stand-level is smaller than the difference at the study-level. In part this is because the Tairua study excluded stands that were younger than age 10 and the PLE tends to be higher for younger stands. The Kaingaroa study included stands younger than age 10. If stands with mean TRV less than 100 m^3 are excluded from both studies then the comparable median PLE values are 12% for Tairua and 8% for Kaingaroa, which is more in keeping with the difference in the total number of reference plots. It is also broadly consistent with the median number of reference plots used per stand; 45 in Kaingaroa and 22 in Tairua.

Preliminary analysis showed that the expected magnitude of stand-level confidence intervals is a complex function of stand-size, the total number of reference plots available, what range of values for age and the response variable they cover, how many of the reference plots are actually useable in any one stand and, of the reference plots that are used in a stand, how uniformly weighted they

2 Nearest neighbour imputation of stand parameters using aerial LiDAR data – Dash, J.P. Marshall, H.M. and Rawley, B. Prepared for Future Forest Research (2013) Unpublished

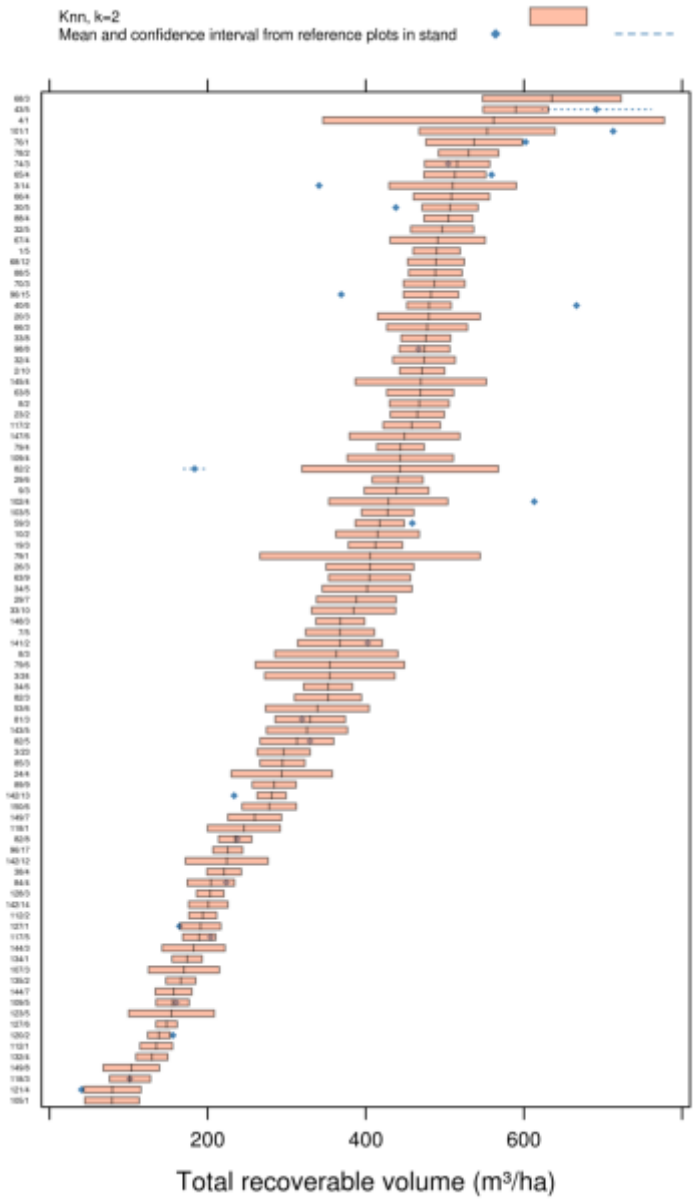
are in terms of the number of target pixels to which each is the nearest neighbour. More work would be required to provide useful guidelines on this.

The key points from this are:

- Having more reference plots is part of keeping the stand-level confidence intervals low, but
- the relationship is not going to be as simple as it is in a conventional inventory, and
- despite the Tairua study having half the number of plots as Kaingaroa, the stand-level confidence intervals in Tairua were still encouragingly small.

It was not possible to examine possible stand-level bias in the Tairua data set because there were too few stands with more than one plot.

Figure 3. Estimates of total recoverable volume by stand for Tairua study



Model Validation

Statistical proximity between all cells in the reference and target datasets was successfully calculated and kNN approach was implemented so that the response variables could be calculated for any area within the study forest and was mapped across the extent of the area of LiDAR acquisition.

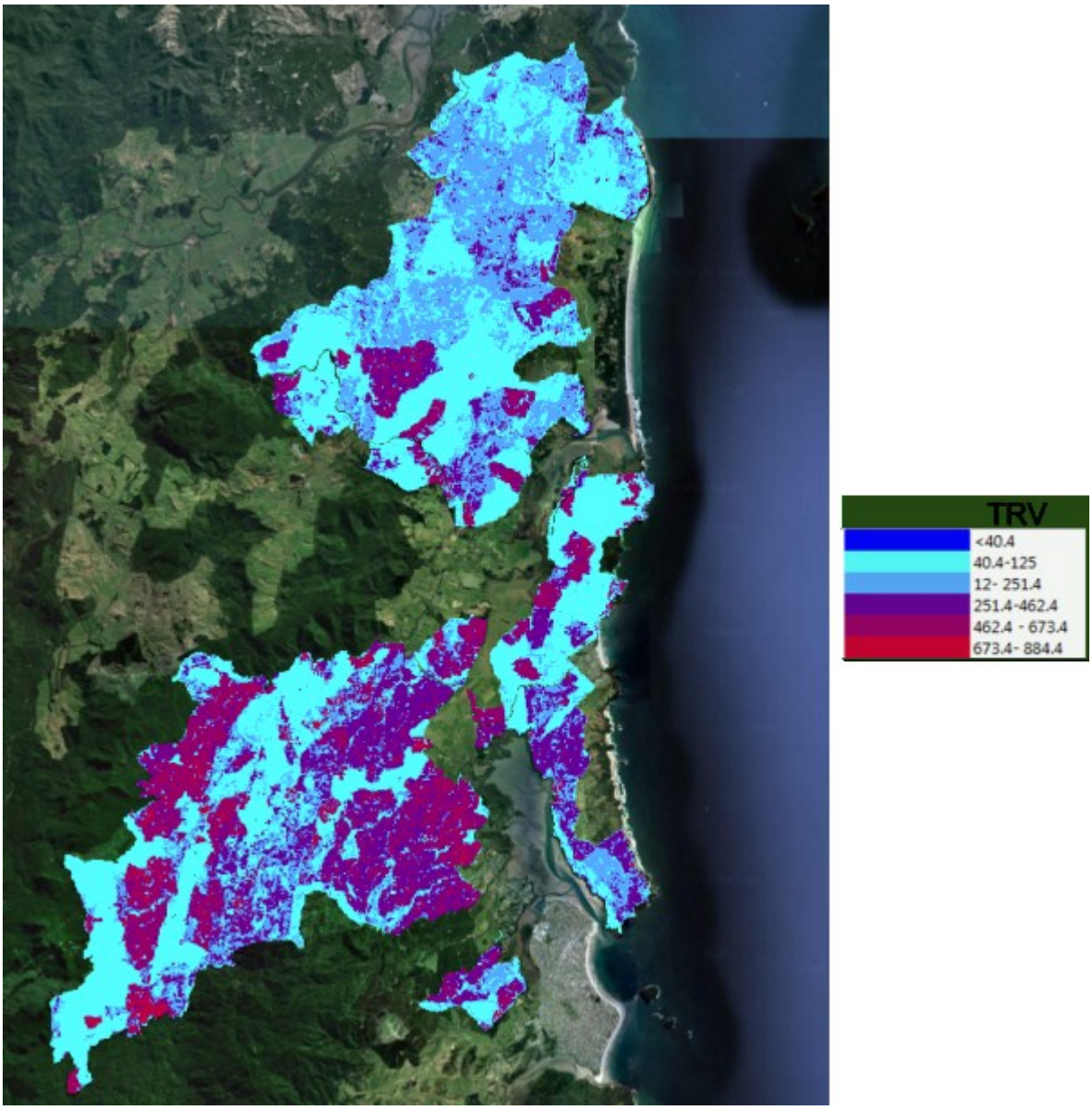


Figure 4. The imputed TRV surface across Tairua forest. A surface of this type was created for all response variables.

An indication of the predictive quality of the model is provided by the observed and imputed vales for the reference dataset with the red line showing a (1:1) correspondence between observed and imputed (Figure 5). These figures indicate that the model is doing a reasonable job of accounting for the variation in the reference plots for TRV and top height. The relationship for basal area is worse, but still adequate whereas the figure for the stocking reflects poorly on the model predictions.

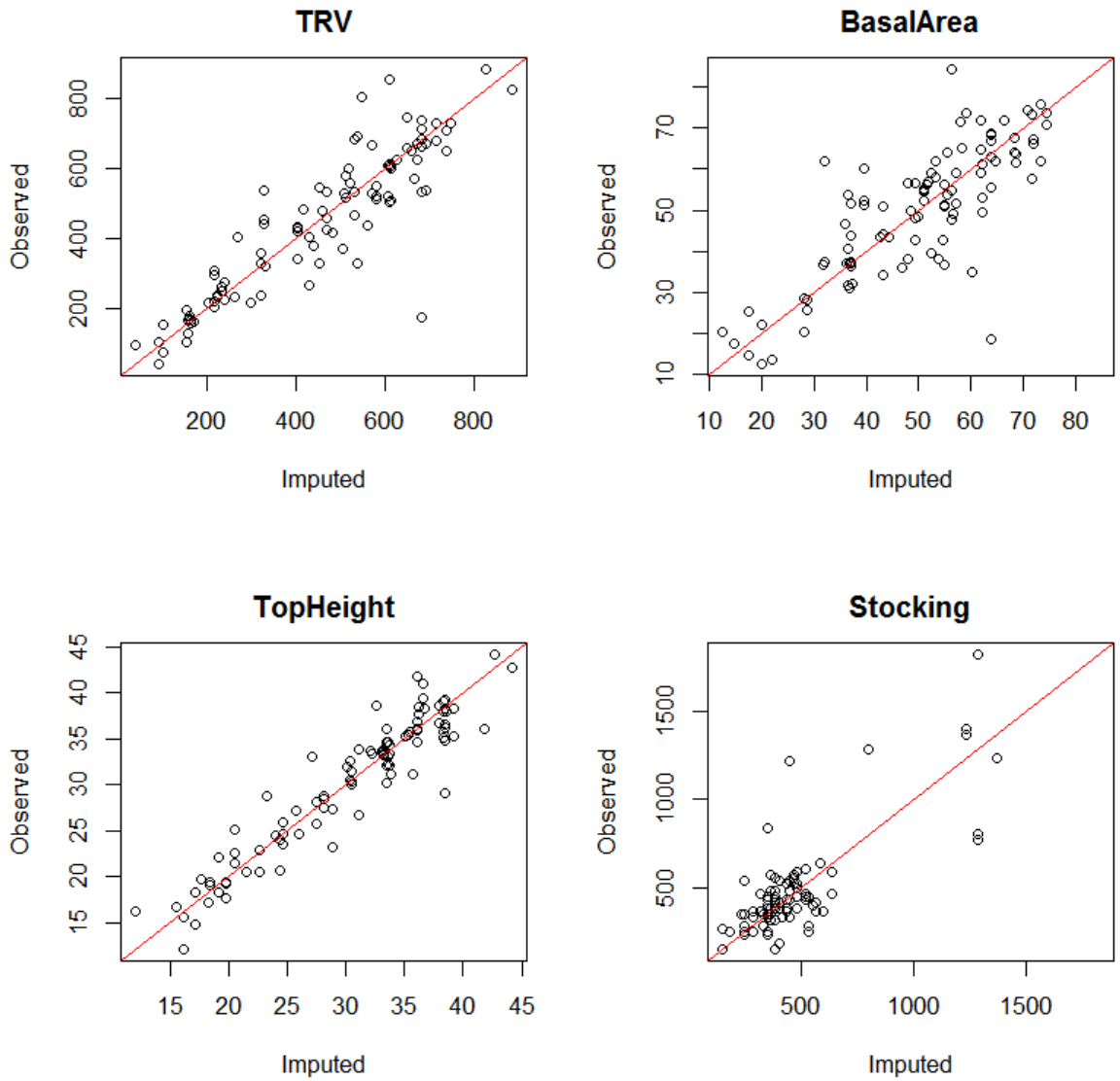


Figure 5. The observed and imputed values for the reference plots

Table 3 details the root mean square difference (RMSD) for 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. The 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. In this instance we can see that the model performance for TRV and top height are considerably better than those for top height and stocking.

Table 3. RMSD for the response variables

Response	RMSD	Scaled RMSD
TRV	94.21	0.45
Top Height	2.52	0.33
Basal Area	9.90	0.61
Stocking	168.76	0.62

This section provides details of a comparison of the stand level aggregation of the imputed values compared with the validation dataset. The imputed (k=5) and inventory values are shown in Figure

of 1. From Figure 6 it seems that the imputed values for TRV and top height correlate fairly well with the inventory values but the stocking and basal area variables do not. For the TRV response it seems that for larger values of Inventory TRV the imputed values tend to be lower than the traditional inventory. The stocking result is particularly poor with the imputed stocking values clustered just below the mean of the reference plots (466 sph). This result indicates that for the stands in the validation dataset stocking cannot be predicted well in this instance.

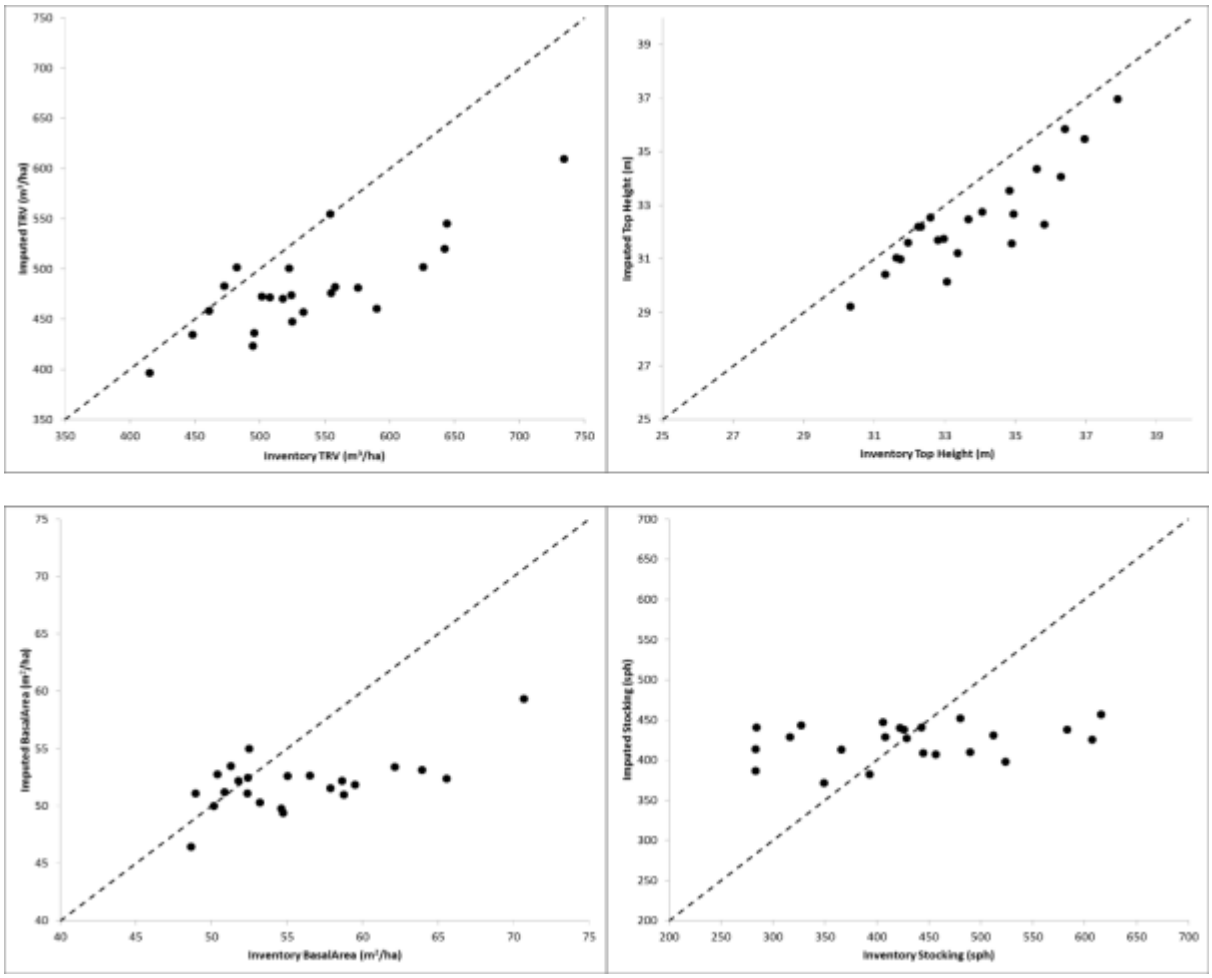


Figure 6. Imputed and Inventory values for stands in the validation dataset

Figure 7 is a box and whisker plot of the scaled residuals ($\frac{Imputed-Inventory}{Inventory}$) for the four response variables, residuals have been scaled to allow a comparison between responses. The residuals for top height are clustered just below zero indicating consistency with the inventory values and a slight under prediction. The residuals for stocking are considerably worse indicating that there is little correspondence between the imputed and the inventory values for stocking.

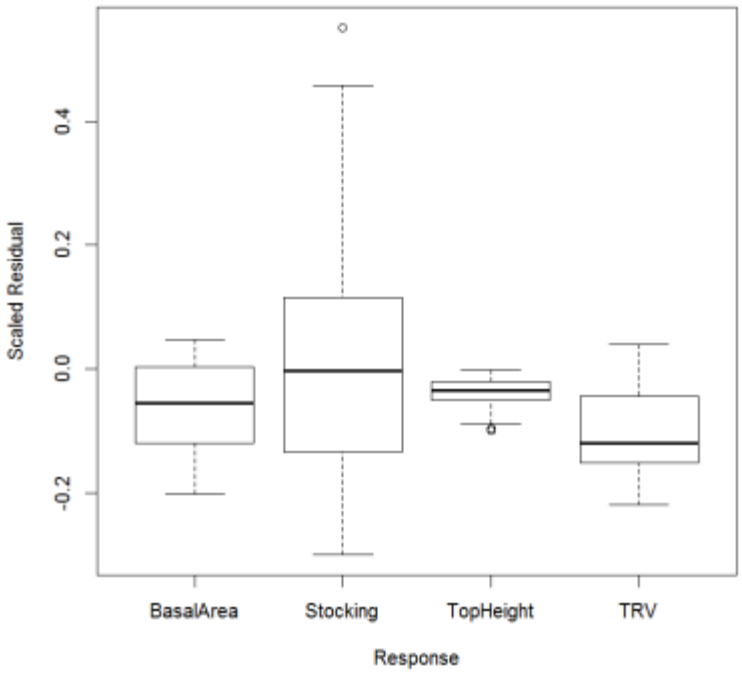


Figure 7. Box plot of the scaled residuals for stands in the validation dataset

As detailed in Section 0 the sampling error for any response variable can now be calculated for any area of interest. Figure 8 shows the imputed and inventory values of TRV for stands in the validation dataset. This figure shows that the imputed TRV values for the validation stands are encouragingly similar to the inventory values. The error bars show the 95% confidence interval for the stand estimates, the imputed confidence intervals are generally smaller than those from the inventory values. This is impressive when it is considered that these values were derived from only 99 reference plots whereas 209 traditional inventory plots were included in the validation dataset. It should also be noted that from the 99 plots in the reference dataset stand level estimates and confidence intervals can now be calculated for every stand in Tairua forest.

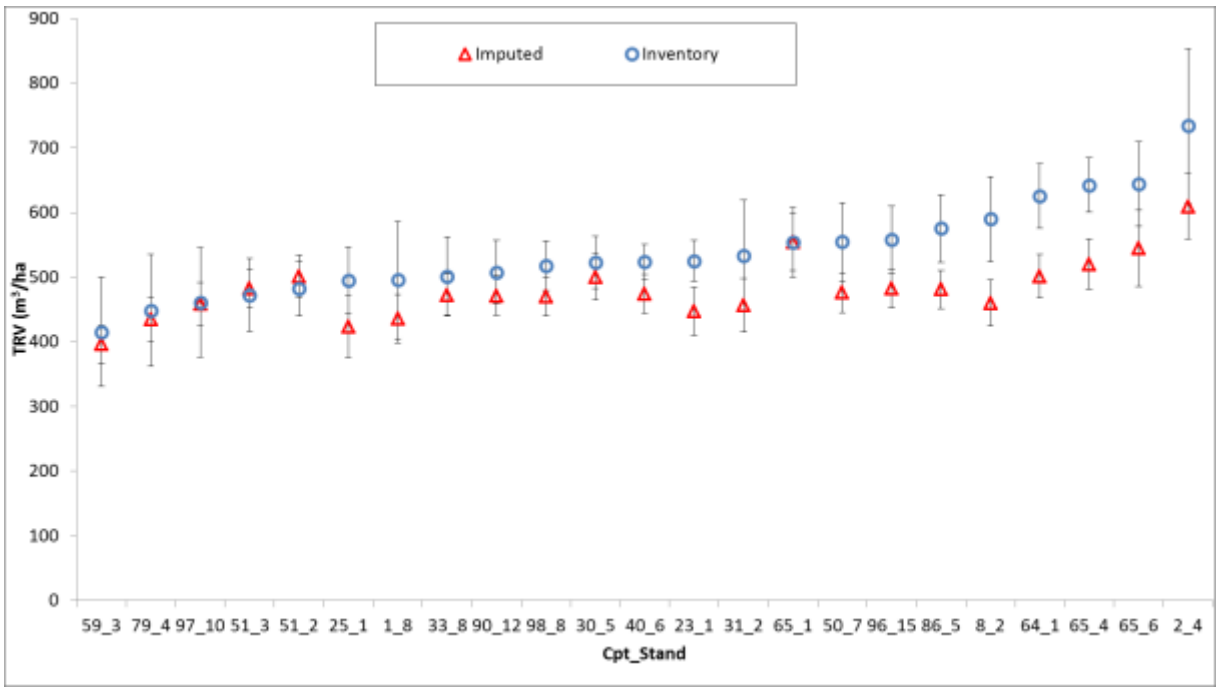


Figure 8. The Imputed and inventory values for the validation dataset. Error bars show the 95% confidence interval

Discussion

The case study reported on here provides a demonstration of the ability to integrate LiDAR data using a kNN imputation methodology to arrive at stand parameters for an area of interest. Important variables from the candidate LiDAR predictor variables have been selected using the variable selection algorithm produced by the study authors. This allows the pruning of unimportant metrics. The statistical proximity between all reference and target pixels has been successfully calculated using the random forest distance measure and appears to provide a reasonably robust measure of the similarity between pixels. Using the same techniques produced in the Kaingaroa case study the sampling error for the kNN estimates over any area of interest can now be calculated and a kNN estimate of the four response variables and an estimate of sampling error has been produced for all stands in the study area. A sampling technique using the covariate matrix produced during sampling error calculation has also been implemented which will prove useful when applying this methodology to a production environment. For further practical considerations and implications on the calculation of sampling error for kNN estimates and to avoid replication readers are referred to the Kaingaroa case study report prepared by the same authors for Future Forest Research.

The results reported in this study are encouraging and serve as a further proof of concept for the application of the kNN technique for integrating LiDAR data for forest research assessment purposes. Given the constraints of the trial dataset (fewer reference plots and lower intensity LiDAR) the results of this study are excellent.