Forestry An International Journal of Forest Research

Institute of Chartered Foresters

Forestry 2015; **88**, 237–247, doi:10.1093/forestry/cpu054
Advance Access publication 16 January 2015

Methods for estimating multivariate stand yields and errors using *k*-NN and aerial laser scanning

Jonathan P. Dash[1], Hamish M. Marshall[2] and Brian Rawley[3]

[1]*Scion, 49 Sala Street, Private Bag 3020, Rotorua 3046, New Zealand*
[2]*Interpine Forestry Ltd., 99 Sala Street, Rotorua 3046, New Zealand*
[3]*Silmetra Ltd. 2528, Old Taupo Road, RD1, Tokoroa 3491, New Zealand*

*Corresponding author. Tel.: +64 73435641; E-mail: jonathan.dash@scionresearch.com

Received 30 July 2014

The objective of this study was to investigate techniques for integrating aerial LiDAR data into the forest yield information systems of a forest management company. Nearest neighbour (*k*-NN) estimation was identified as a useful approach and a 4 000 ha case study was undertaken to provide insight into the performance of the technique in a commercial environment. A field dataset was provided by 213 ground plots installed across the study area. Small-footprint discrete aerial LiDAR data were acquired concurrently and processed to derive descriptive metrics across the study area. A simulated annealing algorithm was produced and used to select important variables that were then used during model development. Sampling error for the *k*-NN predictions was estimated using an approach that accounted for spatial correlation in the reference dataset. An independent validation dataset was acquired from the forest manager's conventional high-intensity stand assessments within the study area projected to LiDAR acquisition date. Model validation showed excellent correspondence to the independent dataset with the responses total recoverable volume (relative mean deviation 4.3 per cent, RMSE 44.96 m^3 ha^{-1}), mean top height (relative mean deviation 1.3 per cent, RMSE 1.34 m), basal area (relative mean deviation 4.5 per cent, RMSE 2.06 m^2 ha^{-1}) and stand density (relative mean deviation −1.8 per cent, RMSE 47.39 sph). Log-product volumes were estimated using LiDAR metrics and *k*-NN estimation for all stands with encouraging accuracy and precision. Sampling error was estimated for all stands in the study area, including those with no reference plots. Sampling error estimates were small enough for stand estimates to be useful for the forest manager in all stands. For stands in the validation dataset the sampling errors for stand volume were smaller for the *k*-NN estimates (median confidence interval (CI) 29.13 m^3 ha^{-1}) than for the conventional inventory estimates (median CI 37.9 m^3 ha^{-1}). Imputation model error was examined and found to be insignificant.

Introduction

The use of aerial light detection and ranging (LiDAR) scanning for forest management has been studied since the mid-1980s (Maclean and Krabill, 1986). In recent times, LiDAR scanning has been developed into a tool that is regularly used in the estimation of tree and forest characteristics. There are many examples of the use of LiDAR to estimate forest parameters such as timber volume (Næsset, 1997), stand volume (Watt and Watt, 2013) and carbon stock (Stephens *et al*., 2012). There are also several examples of the use of this technology in an operational forest management context from various regions including European countries (Wallenius *et al*., 2012), North America (Hudak *et al*., 2008a,b) and Australia (Rombouts *et al*., 2010). To date the use of LiDAR for forest assessment purposes in New Zealand has remained largely in the research sphere with the exception of a national carbon inventory that involved a LiDAR component (Stephens *et al*., 2012). Aerial LiDAR provides continuous auxiliary information that is valuable for stand and landscape-level assessments.

Since its initial application to forest resource assessment in the early 1990s (Tomppo and Katila, 1991), *k* nearest neighbour (*k*-NN) imputation has become popular internationally, with peer-reviewed publications on the topic originating from over 20 countries (McRoberts, 2012). Among other factors, the suitability of the technique for mapping and small area estimation problems is likely a major factor in its widespread appeal in the forest industry (Magnussen and Tomppo, 2014). In a forest measurement context, *k*-NN is regularly used to estimate stand parameters for areas that have not been conventionally measured (Falkowski *et al*., 2010). First easily measured, and relatively inexpensive, auxiliary variables are acquired across an area of interest. In the current context this refers to pixels in a raster detailing the distribution of LiDAR metrics. Subsequently detailed measurements of parameters of interest are taken at specific locations within the study area (Moeur and Stage, 1995; Falkowski *et al*., 2010). Using *k*-NN estimation, variables measured in the field can be estimated for unmeasured areas based on detail in the LiDAR dataset obtained from the initial data collection. In a forest management context, the

parameters of interest will typically be stand-level statistics, such as recoverable volume per hectare, stand density or product volumes per hectare at a specific age.

Procedures of this type result in two separate datasets, a large target dataset that contains only auxiliary variables with complete coverage of the study area, and a small reference dataset where both parameters of interest and auxiliary variables have been observed. A successful imputation will predict the desired response parameters across the target area using the relationship between auxiliary data and variables of interest. To that end, the reference dataset is used to characterize the relationship between the auxiliary variables and the parameters of interest. The variables of interest for each cell in the target dataset are imputed from the nearest neighbours to that cell in the reference dataset, proximity is measured in terms of statistical similarity among the auxiliary variables (Falkowski *et al*., 2010). 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 similarity in covariate space. This approach can be used to estimate the average of any measured or estimated response variable across any arbitrary area of interest (AOI) in the region covered by the target dataset. Examples of AOIs in the context include stands, felling coupes, age classes, riparian strips or even an entire forest. In simple implementations, the target pixels in the AOI are identified. To each target, k donors are assigned. If there are N target pixels, each with k donors, then there are $N \times k$ response variable values and the average of those values is used 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, and then average those across the N target pixels. The weight is based on the similarity between the target and reference plot in terms of the auxiliary metrics. The accuracy of outputs from an imputation are influenced by the selection of the similarity metric, the auxiliary metrics, the scheme for weighting donors and the number of donors (k). This study aimed to provide some insight into the effect of these selections.

In addition to estimates of means and totals for AOIs, confidence intervals around those estimates are essential to their acceptance by forest managers. The fact that the estimates for the multiple target pixels in an AOI are not independent of each other complicates estimation of sampling error. At the very least, two target pixels that share the same donors have predicted values and prediction errors that are highly correlated. Estimation of sampling error for k-NN estimates is an area of ongoing research (Magnussen, 2013; McRoberts *et al*., 2013). The method employed in this study was the model-based approach proposed by McRoberts *et al*. (2007). The technique accounts for the correlation between target observations created by the sharing of donors and by spatial correlation between donors.

Numerous statistical techniques are available for incorporation of remotely sensed data into forest inventory information systems. Following extensive review, k-NN estimation was selected as the preferred technique for use as it has several properties suited to the demands of forest management. A key consideration for the study design was to provide a solution that aligns well within a forest manager's current yield prediction and forest regulation system providing practical outputs that confer immediate benefits to the forest manager. To that end an illustration from a commercial plantation in New Zealand was used.

The planted forest estate in New Zealand comprises 1 719 500 ha constituting ~6 per cent of the total land area. This area of intensively managed industrial forest produces saw logs, pulp logs and residue products for the domestic and export markets and contributes 3.3 per cent to the nation's gross domestic product (NZFOA, 2013). The plantation resource is dominated by *Pinus radiata* D. Don (*P. radiata*), which occupies 90 per cent of the total resource. These plantations are typically grown on a 26–30 year rotation during which stands are thinned to a density of ~350 stems per hectare (sph). Stands may be pruned, usually in three occasions to a height of 6 m, or unpruned. Log production is characterized by a multitude of log-products reflecting the numerous export and domestic markets. The complexity of the market situation means that detailed forest yield information is required.

Professional forest managers utilize forest inventory data for long-term yield projections for planning and resource-valuation purposes. In New Zealand, the accuracy and precision requirements for detailed forest information are high. Most forest managers require a probable limit of error of 10 per cent for stand-level estimates by harvest. This means that novel approaches that can provide precision benefits, improved information, or measurement cost savings are particularly valuable. This paper details the refinement of methods for a forest inventory methodology that incorporates remotely sensed information from airborne laser scanning data.

The principle objectives of this research were to determine the efficacy of the k-NN estimation method at predicting a range of stand attributes using LiDAR. Data obtained over a major New Zealand forest provided a suitable test for this work. Specifically the implementation of methods for estimating sampling error, while accounting for spatial correlation, and efficiently selecting suitable predictor variables were a focus of this study.

Methodology

Study area

The study area for this project encompassed a 2000 m by 20 000 m swath with a total area of 4000 ha . It was situated in Kaingaroa forest covering a wide range of site characteristics and age classes. Kaingaroa forest, in the Central North Island, is New Zealand's largest contiguous plantation occupying ~180 000 ha. The primary forestry species in Kaingaroa is *P. radiata* which occupies >95 per cent of the stocked area. Stands in the study area are grown on a ~28 year rotation. Typical regimes include an initial stand density of ~1000 sph and thinning(s) down to a final density of 300–400 sph prior to harvest. Some stands were grown on a clear-wood regime and pruned to ~6 m (m) high whereas others remained unpruned. The study area was exclusively planted with *P radiata*.

Ground sampling

A total of 213 field plots were installed throughout the study area. The ground sampling design utilized systematic sampling for the majority of plots with the remainder located with adjusted sampling probability. A 400 m grid, with a randomized start point and orientation, was overlaid onto the study area and used to locate 187 plots with one at each grid intersection. The remaining plots were placed in order to target the range of LiDAR metrics that had not been covered by the original 187 plots. Areas covered by raster cells with values that had not been well sampled by the original 187 plots were identified in a GIS and plots were placed at random in these areas. This approach was designed to ensure that the

Table 1 Summary of ground plot and validation dataset, showing the mean, range and standard deviation of age and key stand attributes

Source	Parameter	Age (years)	Stand density (sph)	Basal area (m^2 ha^{-1})	Mean top height (m)	Total recoverable volume (m^3ha^{-1})
Ground plots	Range	2–33	220–1026	7.8–59	5.5–46.5	0–892
	Mean	15	533.5	29.6	23.8	260
	Standard deviation	8.8	310	17.6	12.9	240.5
Validation	Range	17–32	224–564	35–59	27–46	303–774
	Mean	26	347	47.5	35.2	499.4
	Standard deviation	5.4	84.5	7.5	6.7	167.1

full range in the LiDAR predictor variables was sampled by the ground plots. This is important when using *k*-NN estimation as the technique cannot be used for extrapolation. As the sample was not probabilistic, design-based methods could not be used for inference.

The sampling unit was a slope-adjusted 0.06 ha circular bounded plot. Plots were geo-located with a survey-grade global positioning system (GPS). At least 300 points were collected at the centre point of each ground plot and post-differentially corrected. 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, that were free from major growth defects, and selected from across the dbh range present. A minimum of eight tree heights were measured in each plot and used to fit a regression between dbh and height, this was subsequently used to predict the heights of unmeasured trees. Descriptions of stem form, branching and other notable features were collected for mature trees in a manner that allowed log-product segregation to be estimated. Using these measurements, plot level metrics were calculated to provide a set of response variables (Table 1).

LiDAR sampling

The candidate predictor variables used in this analysis were produced from a discrete small-footprint aerial LiDAR dataset covering the study area. LiDAR data were acquired in June 2012 using an Optech ALTM 3100EA scanner. The scanner was located in a fixed-wing aircraft flown at a height of 950 m above mean ground level. Data were acquired with a designed density per swath of a minimum 4 pulses per square metre, and a swath overlap of 50 per cent. The resultant point cloud was classified by the supplier into ground, and non-ground, returns using automated routines. This process was undertaken using the TerraSolid LiDAR processing software module TerraScan. Subsequent manual editing of the LiDAR point cloud data was used to increase the quality of the automated classifications. This editing involved visual inspection and adjustment of the data were required.

The software product FUSION (McGaughey, 2013) was used to produce a set of LiDAR metrics as rasters covering the study area at a 30 m × 30 m resolution. This pixel size was selected because it was consistent with the ground plot size. In a second processing step, the same metrics were produced for locations spatially concurrent with the ground plots. These metrics, paired with the measurements from the ground plots, formed the reference dataset. The LiDAR metrics served as candidate predictor variables during modelling.

Model development and variable selection

All model development and graphical analysis was undertaken in the R statistical software (R Core Team, 2013) and made use of the yaImpute (Crookston and Finley, 2008) package, with semi-variograms fitted using the gstat package (Pebesma, 2004), and random effects with the nlme package (Pinheiro *et al*., 2014).

It is beneficial to select appropriate predictors for a given response to improve prediction quality when there are many candidate predictor variables. Numerous variable selection procedures that are designed to cull predictors in order to select only those that are most valuable have been studied and documented (Dalponte *et al*., 2008; Packalén *et al*., 2012), and several of these procedures were trialled by the authors.

A comparison of several variable selection procedures (Packalén *et al*., 2012) indicated that using a simulated annealing approach aimed at minimizing model error performed favourably as a method of variable selection for imputation of forest parameters with remotely sensed data. Following the technique described by Packalén *et al*. (2012), a randomized local search method known as simulated annealing (Kirkpatrick *et al*., 1983) was implemented in R. As implemented here variable selection via simulated annealing (VSSA) seeks to minimize model root mean square difference (RMSD) by repeatedly imputing the total recoverable volume (TRV), mean top height (MTH), basal area and stand density responses for the reference dataset, with various sets of predictors. This technique is known to provide a good approximation of the global optimum in a large search space, while avoiding local optima by restricting moves to poorer solutions in a controlled manner. Selected variables included canopy cover, several height percentiles and height distributions, and intensity percentile distributions. The predictors selected using VSSA resulted in smaller values of model error than alternative sets of predictors selected using alternative procedures.

Nearest neighbour imputation

Following variable selection, *k*-NN estimation was implemented using the random forest (Breiman, 2001) classification approach to quantify statistical proximity. Numerous other distance metrics were trialled (including Euclidean, Mahalanobis and *k*-most similar neighbour) but were found to result in higher model error. This finding was consistent with previous studies that showed random forest to provide a robust proximity measure (Hudak *et al*., 2008a,b) producing models with superior performance (Latifi and Koch, 2012).

Using the random forest distance approach, observations are considered similar if they tend to converge in the same terminal node in a suitably constructed collection of classification and regression trees (Breiman, 2001; Liaw and Wiener, 2012). The metric used to define statistical distance 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, 2008). The distance metric selected has an important role in model performance, several comprehensive reviews of these metrics are available (Crookston and Finley, 2008; Hudak *et al*., 2008a,b).

Once donors had been identified, response parameters were imputed for each target pixel. This resulted in several surfaces covering the entire study area and detailing the distribution of each response. A single

imputation was used to predict all responses for a target pixel; once a reference plot was selected as a donor all response variables measured therein were applied to the subject target pixel.

The effect of the additional complexity associated with using VSSA and random forest distance were assessed through a comparison with more simplistic approaches. Unweighted Euclidean distance was used as an alternative approach to random forest distance. To provide an alternative to VSSA the 10 predictor variables that with that were most highly correlated with TRV were tested. The performance of each imputation was compared by calculating RMSD for the TRV response. An optimal value of k for each imputation model was estimated by finding the k value that minimized the mean stand-level deviance from the validation dataset estimate of TRV.

Volumes of log-products

Collection of stem form and branch size information in forest inventory plots is standard practice in New Zealand from mid-rotation assessments onwards. In the study area the forest manager uses optimal log bucking software (Rawley, 2011) to calculate product volumes from these descriptions and stem dimensions. The same method, with the same log-product specifications, was used in the reference plots to calculate product volumes. These were then summarized using the k-NN method for AOIs that coincided with the stand boundaries of the validation dataset. Total recoverable value per hectare, calculated using nominal log prices ($\$ m^{-3}$) and product volumes, was also used for comparison.

Validation dataset and agreement analysis

Detailed stand assessment is undertaken in the study forest three times during a rotation cycle; at mid-rotation (age 15–18), 5 years prior to harvest, and immediately pre-harvest. These assessments represent a substantial investment in measurement reflecting the high value placed on detailed stand-level information. A database of the forest managers' stand assessments was made available to provide a validation dataset. The database was interrogated to extract all stand assessments that fell within the LiDAR swath and that were yet to be harvested. Where an assessed stand was partially within the LiDAR swath, the plots outside the study area were excluded and the stand area was re-calculated. The forest manager used systematic sampling to locate measurement plots within stands; measurements were based on circular bounded plots. The forest manager's yield prediction systems were used to project standing tree assessments to the date of LiDAR acquisition to provide a dataset for comparison. Where stands had multiple measurements, only the most recent was included in the validation and assessments >5 years old were omitted to minimize errors associated with growth predictions. The 29 stand assessments available as a validation dataset for comparison are summarized in Table 1.

The agreement between the k-NN estimates and the validation dataset was examined by calculating RMSE

$$\mathrm{RMSE} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

where y_i is the stand value of response y in the validation dataset, $\hat{y}$ is the k-NN estimate for the response y and n is the number of stands assessed. The mean deviation (MD)

$$\mathrm{MD} = \frac{\sum (y_i - \hat{y}_i)}{n} \quad (2)$$

was also calculated to investigate any systematic differences in the comparison.

The MD and RMSE statistics provide useful information on the precision and accuracy of the k-NN estimates. However, from an operational perspective, these statistics may not be particularly informative. Bland and Altman (1986) developed a set of procedures for assessing agreement between clinical measurement techniques, including the Bland–Altman graph. An instructive form of Bland–Altman graph (Wallenius *et al.*, 2012) that included a measure of tolerable discrepancy between the two inventory estimation techniques was used to provide additional insight. These graphs can be considered objective assessments as neither method is assumed to reflect the absolute truth. Wallenius *et al.* (2012) used a discrepancy value of 20 per cent as a tolerable rate following the findings of a study that included interviews with large-scale forests owners in Finland about their information needs for operational planning (Laamanen and Kangas, 2011). No such study has been undertaken in New Zealand but the accuracy requirements among managers of large forests are likely to be more stringent. Suitable estimates of agreement based on the estimates obtained were included in the graphical analysis.

Estimating sampling error

The sampling error estimation method employed (McRoberts *et al.*, 2007) required a prior analysis for spatial correlation. In this context, spatial correlation relates to the tendency for reference plots that are close together to generate similar imputation errors. A target pixel with multiple donors from the same spatial cluster will tend to have low variation in the range of imputed response values. If spatial correlation is not accounted for, this low variation can result in an underestimate of sampling error.

The method used for calculating sampling error for an area of interest follows that described in detail by McRoberts *et al.* (2007). Briefly the method involves:

Given an AOI with N target pixels

(1) Prepare a correlation matrix between reference plots using several iterations. In the first iteration variance ($\hat{\sigma}_i^2$) was constructed under the assumption that no spatial correlation existed and iteration stopped when no cell in the correlation matrix changed by >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 rationality of this assumption was checked with visual inspections of semi-variogram plots.
(2) Calculate local variance ($\hat{\sigma}_i^2$) for each target pixel from using the correlation matrix from step 1.
(3) Calculate variance of k-NN estimators using equations.

Empirical semi-variograms were fitted using the gstat package of R (Pebesma, 2004).

Assessing model error

Estimates of sampling error for k-NN estimates assume that the underlying model is complete and correct, and provides an unbiased estimate at each target pixel and for each AOI. The sampling errors do not take into account any lack of fit in model predictions for an AOI. The imputation error for the reference plots themselves were used to provide an indication of the magnitude of lack of fit 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 donors. The imputation error for an AOI has two components. One is the sampling error, discussed above, that can be estimated for an AOI. The other is model lack of fit, which arises because the imputation model is imperfect. This lack of fit cannot be estimated for a single AOI but its distribution is of interest.

In this analysis, a reference cell could not have itself as a donor. Additionally, because the interest was in stand-level imputation errors, a reference plot could not have any donors in the same stand. These conditions

provided a useful indication of potential model lack of fit for stands containing no reference plots. However, such techniques were not otherwise used during *k*-NN estimation for an AOI. A random effects model was used to partition imputation error for each reference plot between a stand-level component (model lack of fit) and remaining noise. Analysis was limited to stands with more than one reference plot and the imputed value for a reference plot used the single nearest neighbour that was not in the same stand.

Results

Model development

Following variable selection, a value for each response was imputed for all pixels in the target dataset and the reference dataset. A comparison of RMSD between imputed and observed values for all pixels that have both (provides insight into model performance (Table 2). In this analysis RMSD was calculated using a leave-one-out approach where a reference observation's nearest neighbour in the reference dataset was used for comparison with the measured value. This statistic can be scaled by dividing by the standard deviation of the reference values to provide a means of comparing response with different units (Table 2). The RMSD values for all stand attributes within the reference plots were relatively low and these values show predictions of MTH and TRV to be more precise than that of basal area and stand density.

The relationship between observed and imputed values in the reference dataset provides insight into the predictive quality of the imputation for each response (Figure 1). Prediction accuracy was encouraging and there was little evidence of bias for any response based on these figures. The predictive quality for basal area and for stand density deteriorated at higher values. There were a small number of outliers where there was poor correspondence between the observed and predicted values. These arose when a reference cell's nearest neighbour did not provide similar response parameters to the reference cell itself. The relative scarcity of these outliers suggests that this only occurred infrequently.

A comparison of the modelling techniques employed *vs* more simplistic methods of predictor variable and neighbour selection revealed that the additional complexity improved model performance (Table 3). The combination of VSSA and random forest distance was found to be the best performing combination trialled

Table 2 The RMSD and scaled RMSD values for the response parameters associated with the imputation model calculated through a leave one out cross-validation. MD and RMSE summarize the agreement between the conventional and *k*-NN estimates for the validation dataset

Variable	MD	Relative MD	RMSE	RMSD	Scaled RMSD
Total recoverable volume	20.4	4.3%	44.9	51.1	0.2
Mean top height	0.4	1.3%	1.3	1.8	0.1
Basal area	2.1	4.5%	4.6	4.9	0.3
Stand density	−6.7	−1.8%	47.4	116.7	0.4

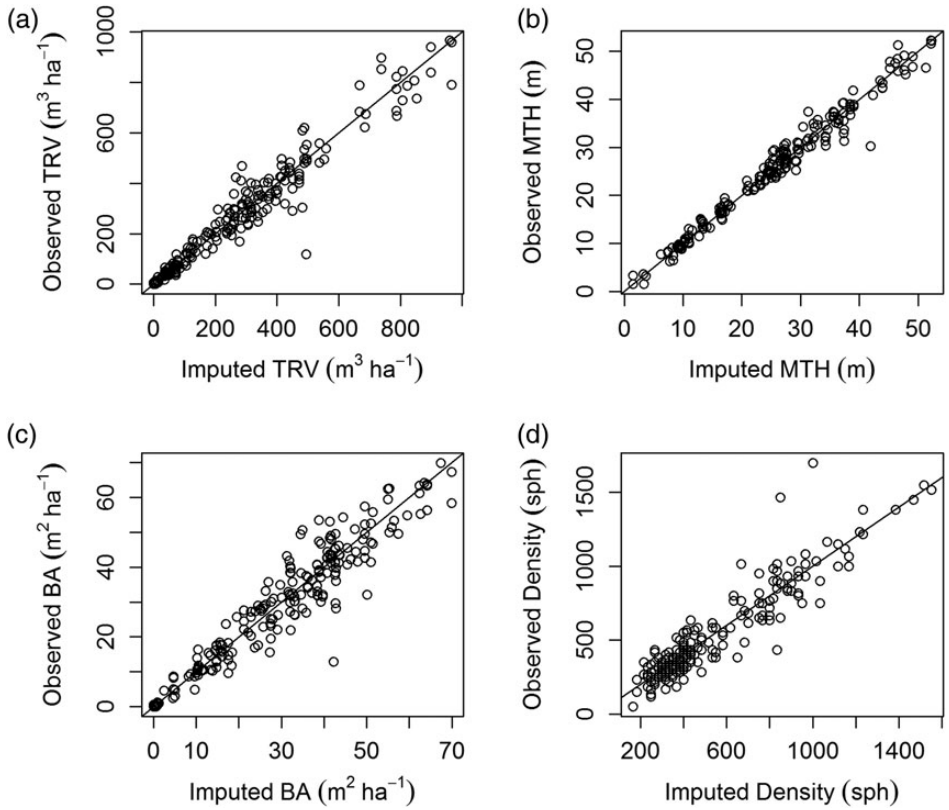

Figure 1 Relationship between the observed and imputed values for the reference dataset (*k*=1) for (a) Total recoverable volume (TRV), (b) Mean top height (MTH), (c) Basal Area (BA), and (d) Stand density. The diagonal line shows the 1:1 line.

for the TRV response. When the Euclidean distance metric was used using the 10 most correlated predictors performed better than the predictors selected by VSSA. Values of k that minimize deviance from the validation dataset estimates of TRV were also calculated (Table 3).

Model validation and agreement analysis

The validation dataset was used to compare aggregations of imputed pixels that had their centroid within the forest manager's stand boundaries with predictions from the fully independent conventional stand inventory data. Correspondence between the imputed and inventory values was very high for TRV ($R^2 = 0.94$, RMSE = 4.3 per cent) and MTH ($R^2 = 0.97$, RMSE = 1.3 per cent). Imputed and inventory values were slightly less well correlated for basal area ($R^2 = 0.69$, RMSE = 4.5 per cent) and stand density ($R^2 = 0.70$, RMSE = −1.8 per cent) (Figure 2). The calculated RMSE and MD confirm a strong agreement between both measurement techniques (Table 2). The MD statistics reveal a slight underestimation by the k-NN approach compared with the traditional approach for all response variables except stand density where there was a slight overestimation.

Estimates of TRV for the majority (76 per cent) of stands in the validation dataset were within 10 per cent of each other using both estimation techniques (Figure 3). The MTH values for most (87 per cent) stands were within 5 per cent agreement and all stands were within 10 per cent agreement using both estimation methods. There was greater inconsistency for the basal area and stand density response with a small number of stands showing a discrepancy of >20 per cent but for the majority there was correspondence within 10 per cent.

There was strong agreement between the k-NN estimate and the independent conventional estimate of product mix at the study area level (Figure 4). There was some discrepancy in the

Table 3 The RMSD on TRV and optimal k values calculated for different combinations of variable selection and distance metric

Variable Selection	Distance metric	RMSD on TRV	Optimal k
VSSA	Random forests	51.1	2
Correlated	Random forests	62.4	17
VSSA	Euclidean	93.5	1
Correlated	Euclidean	72.4	1

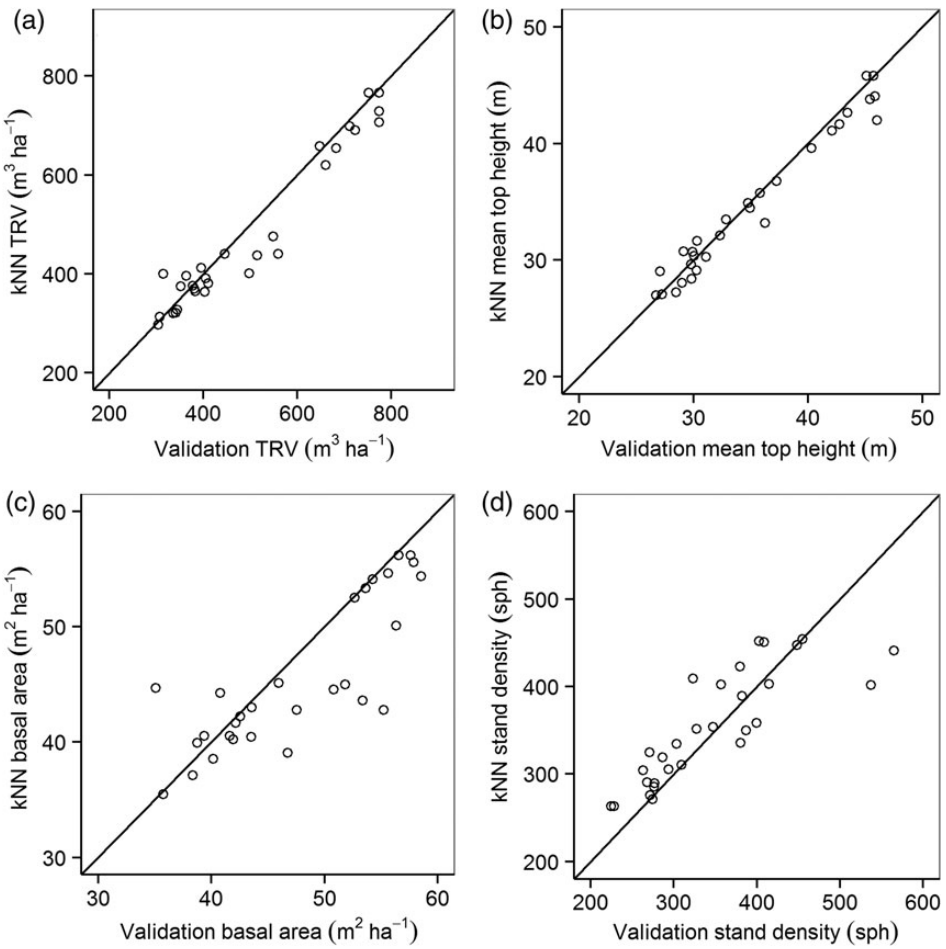

Figure 2 Relationship between k-NN estimates and conventional inventory estimates of (a) Total recoverable volume (TRV), (b) Mean top height, (c) Basal area, and (d) Stand density for stands in the validation dataset. The solid line within each panel shows the 1 : 1 line.

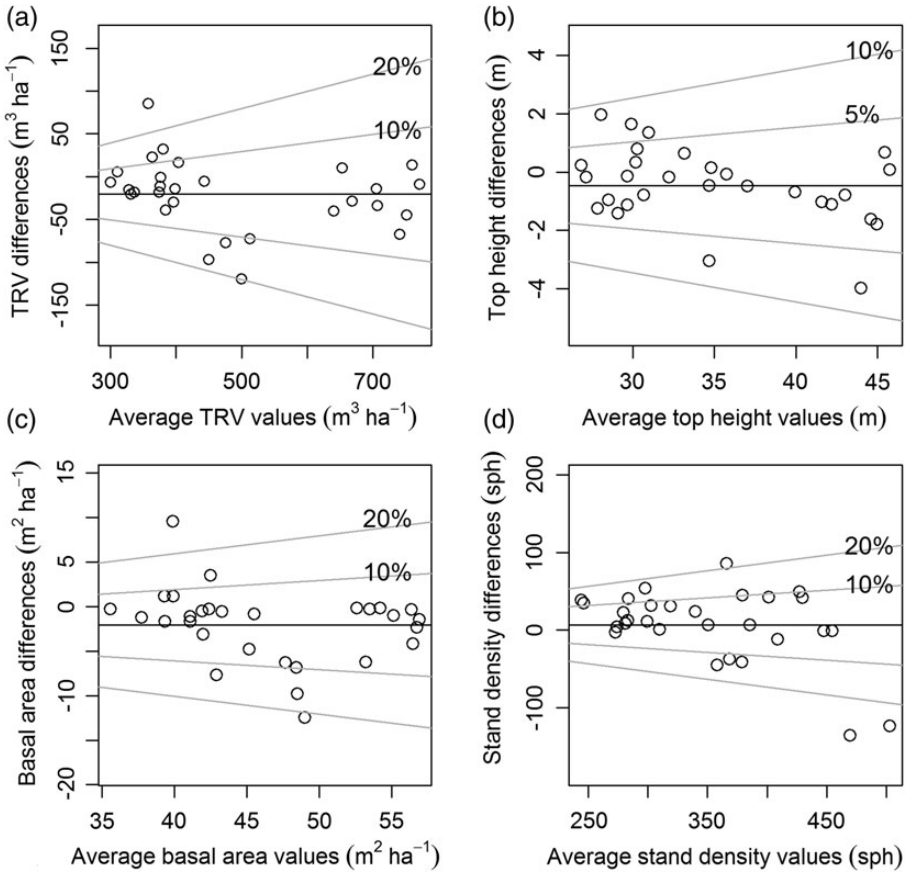

Figure 3 Bland – Altman graphs with the grey lines showing various tolerance limits for (a) Total recoverable volume (TRV), (b) Top Height, (c) Basal Area, and (d) Stand Density. Each datum represents a single stand in the validation dataset and the solid black line represents the mean difference between the two estimation techniques.

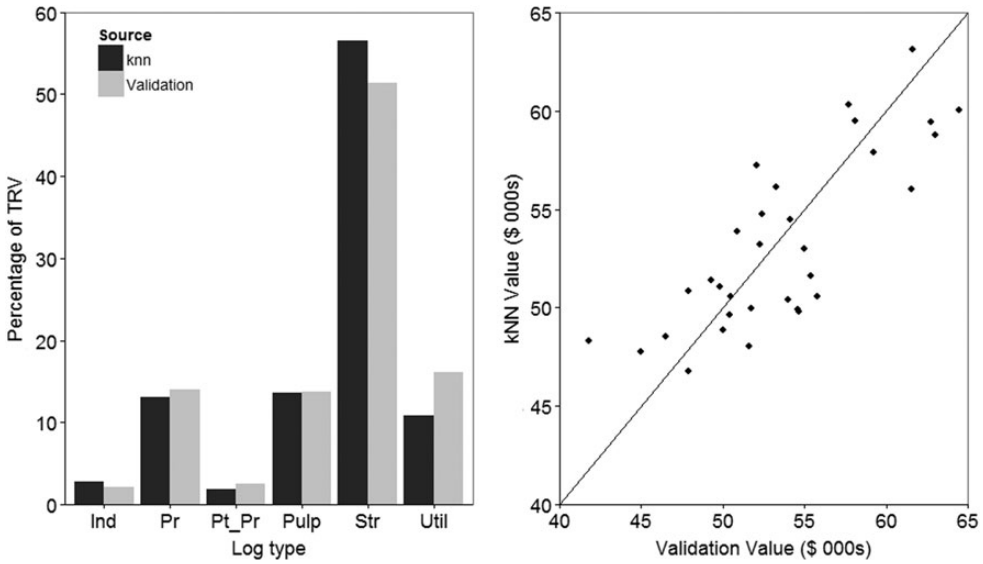

Figure 4 The bar chart in the left hand panel shows average product mix for stands in the validation dataset estimated using *k*-NN and LiDAR metrics (*k*-NN) and conventional inventory (validation). The log types refer to groupings of the forest manager's log sorts (Ind = industrial saw log, Pr = pruned saw log, Pt_Pr = part-pruned saw log, pulp = pulp log, Str = structural saw log, Utl = utility saw log) The right hand panel shows the relationship between the value (in $NZ) estimated using *k*-NN and a conventional inventory. Values shown are at the stand level. The 1 : 1 line is shown on the plot.

percentage of structural (Str) and utility (Util) grade products estimated by the two approaches. In most instances the values produced by both methods were comparable at the stand level although there were some outliers. In some stands, the imputed value was considerably higher than the validation dataset value due to the prediction of, high-value, pruned saw log volumes in un-pruned stands. This occurred as the reference cells providing the product mix were obtained from pruned stands whereas in fact the target stands were in un-pruned stands. This may be a cause for concern in a production setting and would need to be resolved.

Spatial correlation

Spatial correlation was examined by visual assessment of the variance of the differences between imputation errors for pairs of reference plots graphed against the distance between the plots, for all possible pairs of reference plots. A trend in which variation increases with distance to a plateau was taken as evidence of spatial correlation. Evidence for the presence of such a trend was inconclusive. The effective range of spatial correlation, if any, was found to be several hundred metres. Most of the reference plots were located on a 400 m grid and the spatial correlation had an important influence only when additional plots were placed off the grid.

As the exact magnitude and range of spatial correlation was uncertain, it was considered useful to see what effect the assumed spatial correlation had on estimates of sampling error. A comparison of the relative sampling error calculated both with, and without, recognition of spatial correlation for TRV revealed, as expected, that sampling error was lower when spatial correlation was ignored. Across all stands ignoring spatial correlation would lead to an under estimation of the standard error for TRV of 15 per cent; i.e. a probable limit of error of 10 per cent would increase to 11.5 per cent. The effect of incorporating spatial correlation into the estimates of sampling errors was sensitive to the response variable as well as the value of k. The equivalent increases in the standard errors for other responses were 19 per cent for basal area, 4 per cent for stand density and 12 per cent for MTH. The effects of incorporating spatial correlation into estimates of sampling error are relatively consistent across many stands. The effects of spatial correlation were incorporated into the results presented.

Sampling error estimates

Estimates of all response variables, with associated sampling error, were produced for all stands in the study area using k-NN estimation and LiDAR metrics whether or not that stand contained reference plots. The confidence intervals were narrow enough to suggest that the k-NN stand-level estimates could be useful (Figure 5). A comparison of the sampling error for k-NN estimates and for the conventional estimates for stands in the validation dataset (Figure 5) shows that for the majority (72 per cent) of stands the k-NN estimate of sampling error was smaller than the conventional estimate currently used by the forest manager. The median confidence interval for the k-NN estimates (29.13 m^3 ha^{-1}) was smaller than the independent conventional confidence interval (37.89 m^3 ha^{-1}).

Assessing model error

Multiple models were tested and all provided the same conclusion about the magnitude and significance of the stand-level effect. However, a random-slopes model was the most informative

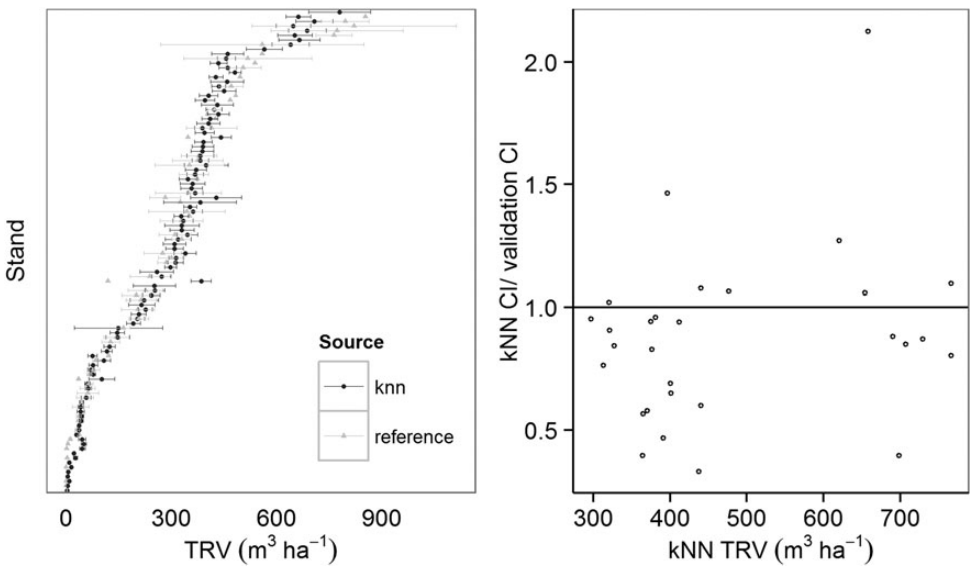

Figure 5 The left panel shows estimates of total recoverable volume, by stand, and where these could be calculated the error bars show the 95% confidence intervals. The black diamonds represent k-NN estimates where $k = 2$, the grey triangular points represent the average of the reference plot values in a given stand and the grey error bars show the 95% confidence interval (CI) calculated by treating the reference plots in a stand as a sub-population of the study area. The right panel shows the relationship between total recoverable volume estimated by k-NN and the k-NN estimate of 95% confidence intervals expressed as a fraction of the conventional estimate for the stands in the validation dataset. The solid horizontal line is drawn at a value of 1 to indicate where the 95% CI estimated by k-NN and conventional estimation are equivalent.

in that the slopes were directly interpretable as a proportional error. The random slopes model is $(imputed)_{ij} = [(\beta + b_i)\ observed]_{ij} + \in$ here i represents stands, j represents plots within stands, β 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 control heteroscedasticity. Significance tests relied on a likelihood ratio test of nested models with the base model omitting the random effects. The between-stand variation, calculated in this way, had a standard deviation of 6 per cent of the stand mean. This was not significantly different to zero in a statistical sense ($P = 0.09$) and could be ignored. However, it does provide a best estimate of the potential magnitude of stand-level bias; one in which for 95 per cent of stands the absolute model error would be <12 per cent of the mean. This is in addition to the estimated sampling error but, unlike sampling error, cannot be calculated at a stand-level.

Discussion

Findings from this paper demonstrate the successful implementation of k-NN estimation as a means of integrating aerial LiDAR scanning data into a forest information system. This case study used for illustration was the first use of this approach with data of this type in New Zealand. The results suggest that the technique can be used to provide accurate predictions when compared with a validation dataset derived from independent stand assessments in the study area. Estimates of the error associated with stand level k-NN estimates were calculated in a manner that accounted for spatial correlation and were encouragingly small.

A variable selection algorithm was implemented and used to select important predictor variables and discard those that are unimportant based on model prediction error. The algorithm can be used to select the important variables from the numerous LiDAR metrics in an unsupervised manner. It was observed that using the variables selected by the algorithm produced better predictions than predictors selected using alternative techniques. The quality of prediction was assessed both in terms of the model performance statistics and by comparing the stand-level model outputs with stand estimates from the validation dataset.

The VSSA algorithm and random forest distance metric were compared with more simplistic modelling approaches. This analysis revealed that using random forest for neighbour selection provided considerably lower RMSD values than Euclidean distance. This result is consistent with previous studies (Hudak *et al*., 2008a,b; Hudak *et al*., 2014). The reason for the greater accuracy provided by random forest is unclear although, as noted previously (Hudak *et al*., 2014), the bootstrapping nature of random forest may be the best explanation for its effectiveness. When using the random forest proximity metric the predictor variables selected by VSSA produced a RMSD that was substantially (18 per cent) lower than that produced using the 10 predictors most correlated with TRV. A differing effect was observed when using the Euclidean distance metric. The RMSD based on the variables selected by VSSA was 23 per cent higher than that produced using the 10 most highly correlated predictors. The cause of this occurrence is unknown. However, the predictors that were most correlated with TRV were all height percentiles whereas other types of predictors, such as intensity kurtosis and intensity percentiles, were selected by VSSA. It is logical that the inherent flexibility in the random forest distance metric can accommodate a more varied data structure in the predictors.

Model validation and agreement analysis showed that the k-NN stand estimates did not vary significantly, or systematically, from the conventional inventories in the validation dataset. Furthermore, rasters detailing the response show detailed intra-stand distribution of stand dimensions that were previously unavailable. This is of considerable value to forest managers for managing silvicultural operations, harvest planning and matching production to market conditions. The k-NN stand estimates varied minimally from the conventional inventory estimates used by the forest manager. Wherever there was a notable discrepancy between k-NN and conventional estimates of TRV, stands were visited to investigate the cause. This revealed that several stands had experienced wind damage at some time between the conventional measurement and LiDAR acquisition date. These are visible as a cluster of points below the diagonal line in panel a) of Figure 2. This is an encouraging finding as it highlights the improved information available through integrating remotely sensed data into forest assessment.

Log-product volume estimates produced from k-NN were consistent with those from the conventional stand assessments. In a minority of cases the imputed product mix estimate for a stand differed significantly from the conventional estimate. There are several possible causes for this and these issues would be addressed in a production environment. Silvicultural treatments that have an effect on log-product mix can be accounted for in model development. These could be included in the proximity calculation used to select donors. This would require accurate stand records, and knowledge of the silvicultural treatments that affect log-product mix. Alternatively, a statistic that serves as an index of log-product mix (e.g. plot value) could be included as a response during predictor variable selection. This process should result in the selection of predictors that ensure appropriate donors are selected for improved log-product estimates. Management differences could be accounted for during sampling design by splitting the area based on a silvicultural treatment that is known to result in the production, or otherwise, of a product of interest (e.g. pruning produces pruned saw logs). Separating the imputation in this way would ensure that reference plots could only be selected from stands that had received appropriate management. Challenges remain for the imputation of log-product mixes but this study illustrates a valid proof of concept and suggests that the approach can be extended to produce log-product volume estimates suitable for use in a commercial environment.

The sampling errors for imputed response variable estimates were calculated using a method (McRoberts *et al*., 2007) that accounts for the correlation between target pixels that share the same reference pixel(s), and for correlation between reference pixels that are in close proximity. Analysis showed that failing to account for these correlations would result in an underestimate of sampling error by 15 per cent for TRV, 12 per cent for MTH, 19 per cent for basal area and 4 per cent for stand density. The short range of spatial correlation observed is consistent with the findings of previous work (McRoberts *et al*., 2007; Magnussen *et al*., 2009). It is tempting to conclude that spatial correlation could be ignored in a production inventory system if plot spacing is kept ~400 m. While this might in fact be the case, it would be premature to arrive at this conclusion on the basis of limited evidence.

It was possible to impute values with useful and consistent precision for every stand in the study area using *k*-NN estimates based on LiDAR metrics; even for stands containing no reference plots. The sampling errors for *k*-NN estimates were smaller in most cases when compared with the conventional stand assessments. For a commercial forestry application this result is particularly encouraging as it indicates the potential cost savings, or precision benefits, that incorporating LiDAR data in a *k*-NN framework can offer. Precise stand estimates covering 102 stands across 4000 ha were produced using only 213 ground plots. There is no reason to suggest that the comparable results would not be produced over a larger area, and for many more stands, if a more extensive LiDAR dataset was available. In contrast, the conventional assessments in the validation dataset contained ∼440 plots and provided stand estimates for only 29 separate stands.

The sampling error does not incorporate error due to the *k*-NN model being potentially biased at the stand level for some stands. Analysis of stands with multiple reference plots showed that this error was small (±6 per cent) and not significantly different to zero in this study area.

There are two challenging issues associated with calculating and using the sampling errors for AOIs; the size of the computation and the estimation of spatial correlation. Neither issue is insurmountable. Estimation of sampling error requires computation of the interactions between each pair of target pixels and their *k* reference pixels in the AOI. In this study, an AOI consisting of the entire study area contained 43 548 target pixels. With its five nearest neighbours ($k = 5$), computation and summarization required consideration of $(5 \times 43\,548)^2$ over 50 billion values. The size of the computational problem has two implications. Firstly, it cannot all be fitted into computer memory at the same time so the calculations must be staged adding to its complexity. Secondly, it can take considerable time to compute the sampling error for large areas of interest. Three approaches suggested by McRoberts *et al.* (2007), were used to reduce the size of the problem to a manageable level in this study. These included sub-division of the covariance matrix into manageable blocks that were processed sequentially (blocking), the use of symmetry (the upper diagonal is the same as the lower diagonal), and sub-sampling. Blocking and symmetry allowed for computation across the entire study area. Sub-sampling allowed for computation across the entire study area within the time limits that would be practical in an operational context. All three methods would be required in a production implementation involving large AOIs.

Understanding spatial correlation among the reference plots is complex but not time-consuming. It is complex because it requires multiple iterations of the process of fitting a semi-variogram. All of this can be automated, and was automated for the case study. Problems arise because the spatial data tend to be noisy so fitting a semi-variogram 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 would require more experience with the modes of failure and how best to deal with them. McRoberts *et al.* (2007) recommended 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 would be prudent and the necessity for manual intervention should be assumed.

Conclusions

In the course of this research methods for small area estimation using *k*-NN and aerial LiDAR have been refined. In particular a review of the effect of spatial correlation between reference observations was investigated and practical methods to account for it identified. The case study has shown that valid estimates of stand yields, including log-product volumes, can be produced with improved precision compared with conventional methods. This offers forest managers a substantial potential cost saving in avoided measurement plots.

Acknowledgements

The contribution of Susana Gonzalez Aracil is gratefully acknowledged for assistance with LiDAR processing. The diligent and professional contribution of all field teams involved is also acknowledged by the authors. Mike Watt and Ruth Falshaw of Scion Research are acknowledged for providing extremely useful reviews of an early draft of this document. Mark Ducey and the two anonymous reviewers provided useful comments that have improved the quality of this work.

Conflict of interest statement

None declared.

Funding

Funding for this research was provided by Future Forests Research Ltd. and Timberlands Ltd. The research providers were Interpine Forestry Ltd., Silmetra Ltd. and Scion Research.

References

Bland, J.M. and Altman, D.G. 1986 Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310.

Breiman, L. 2001 Random forests. *Mach. Learn.* **45**, 5–32.

Crookston, N.L. and Finley, A.O. 2008 yaImpute: an R package for κNN imputation. *J. Stat. Software* **23**, 1–16.

Dalponte, M., Bruzzone, L. and Gianelle, D. 2008 Estimation of tree biomass volume in Alpine forest areas using multireturn LIDAR data and support vector regression. In *Proceedings of SPIE - The international Society for Optical Engineering*, vol 7109. Image and Signal Processing for Remote Sensing XIV, Cardiff, Wales.

Falkowski, M.J., Hudak, A.T., Crookston, N.L., Gessler, P.E., Uebler, E.H. and Smith, A.M.S. 2010 Landscape-scale parameterization of a tree-level forest growth model: a *k*-nearest neighbor imputation approach incorporating LiDAR data. *Can. J. For. Res.* **40**, 184–199.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E. and Falkowski, M.J. 2008a Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sens. Environ.* **112**, 2232–2245.

Hudak, A.T., Evans, J.S., Crookston, N.L., Falkowski, M.J., Stiegers, B.K., Taylor, R. *et al.* 2008b *Aggregating pixel-level basal area predictions derived from LiDAR to industrial forest stands in North-Central Idaho*. US Forest Service. Idaho, USA.

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

Kirkpatrick, S., Gelatt, C.D. Jr. and Vecchi, M.P. 1983 Optimization by simulated annealing. *Science* **220**, 671–680.

Laamanen, R. and Kangas, A. 2011 Large-scale forest owner's information needs in operational planning of timber harvesting - some practical views in Metsähallitus, Finnish state-owned enterprise. *Silva Fenn.* **45**, 711–727.

Latifi, H. and Koch, B. 2012 Evaluation of most similar neighbour and random forest methods for imputing forest inventory variables using data from target and auxiliary stands. *Int. J. Remote Sens.* **33**, 6668–6694.

Liaw, A. and Wiener, M. 2012 Breiman and Cutler's random forests for classification and regression 4.6–7 Ed.

Rawley, B. 2011 YTGen. 2.9.8.4 Ed.

Maclean, G.A. and Krabill, W.B. 1986 Gross-merchantable timber volume estimation using an airborne lidar system. *Can. J. Remote Sens.* **12**, 7–18.

Magnussen, S. 2013 An assessment of three variance estimators for the k-nearest neighbour technique. *Silva Fenn.* **47**, 1–19.

Magnussen, S. and Tomppo, E. 2014 The k-nearest neighbor technique with local linear regression. *Scand. J. For. Res.* **29**, 120–131.

Magnussen, S., McRoberts, R.E. and Tomppo, E.O. 2009 Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. *Remote Sens. Environ.* **113**, 476–488.

McGaughey, R.J. 2013 *FUSION/LDV: Software for LiDAR data analysis and visualisation*. 3.30 Ed., United States Department of Agriculture.

McRoberts, R.E. 2012 Estimating forest attribute parameters for small areas using nearest neighbors techniques. *For. Ecol. Manage.* **272**, 3–12.

McRoberts, R.E., Tomppo, E.O., Finley, A.O. and Heikkinen, J. 2007 Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. *Remote Sens. Environ.* **111**, 466–480.

McRoberts, R.E., Næsset, E. and Gobakken, T. 2013 Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sens. Environ.* **128**, 268–275.

Moeur, M. and Stage, A. 1995 Most similar neighbour: an improved sampling inference procedure for natural resource planning. *For. Sci.* **41**, 337–359.

Næsset, E. 1997 Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sens. Environ.* **61**, 246–253.

NZFOA. 2013 *New Zealand Plantation Forest Industry facts and figures*. N.Z.F.O. Association (ed.), .

Packalén, P., Temesgen, H. and Maltamo, M. 2012 Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. *Stratégies de sélection de variables le plus proche pour méthodes d'imputation utilisées dans voisines à distance de détection basé sur l'inventaire forestier*, **38**, 557–569.

Pebesma, E.J. 2004 Multivariate geostatistics in S: the gstat package. *Comput. Geosci.* **30**, 683–691.

Pinheiro, J., Bates, D., DebRoy, S. and Sarkar, D. and R Core Team. 2014 *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1–117.

R Core Team. 2013 *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

Rombouts, J., Ferguson, I.S. and Leech, J.W. 2010 Campaign and site effects in LiDAR prediction models for site-quality assessment of radiata pine plantations in South Australia. *Int. J. Remote Sens.* **31**, 1155–1173.

Stephens, P.R., Kimberley, M.O., Beets, P.N., Paul, T.S.H., Searles, N., Bell, A. *et al*. 2012 Airborne scanning LiDAR in a double sampling forest carbon inventory. *Remote Sens. Environ.* **117**, 348–357.

Tomppo, E. and Katila, M. 1991 *Satellite image-based national forest inventory of Finland*.IGARSS'91 Digest, pp. 1141–1144.

Watt, P. and Watt, M. 2013 Development of a national model of Pinus radiata stand volume from lidar metrics for New Zealand. *Int. J. Remote Sens.* **34**, 5892–5904.

Wallenius, T., Laamanen, R., Peuhkurinen, J., Mehtätalo, L. and Kangas, A. 2012 Analysing the agreement between an airborne laser scanning based forest inventory and a control inventory-a case study in the state owned forests in Finland. *Silva Fenn.* **46**, 111–129.