Remote Sensing of Environment 115 (2011) 573–585

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Extracting LiDAR indices to characterise multilayered forest structure using mixture distribution functions

Dominik Jaskierniak [a,*], Patrick N.J. Lane [b,c], Andrew Robinson [d], Arko Lucieer [a]

[a] School of Geography and Environmental Studies, University of Tasmania, Tasmania, Australia
[b] Department of Forest and Ecosystem Science, University of Melbourne, Victoria, Australia
[c] Cooperative Research Centre for Forestry, Sandy Bay, Tasmania, Australia
[d] Department of Mathematics and Statistics, University of Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 31 March 2010
Received in revised form 8 October 2010
Accepted 8 October 2010

Keywords:
LiDAR
Mixture model
Theoretical distribution function
Forest structure
Forest hydrology

ABSTRACT

Discrete Light Detection and Ranging (LiDAR) data is used to stratify a multilayered eucalyptus forest and characterise the structure of the vertical profile. We present a methodology that may prove useful for a very broad range of forest management applications, particularly for timber inventory evaluation and forest growth modelling. In this study, we use LiDAR data to stratify a multilayered eucalyptus forest and characterise the structure of specific vegetation layers for forest hydrology research, as vegetation dynamics influence a catchment's streamflow yield. A forest stand's crown height, density, depth, and closure, influence aerodynamic properties of the forest structure and the amount of transpiring leaf area, which in turn determine evapotranspiration rates. We present a methodology that produces canopy profile indices of understorey and overstorey vegetation using mixture models with a wide range of theoretical distribution functions. Mixture models provide a mechanism to summarise complex canopy attributes into a short list of parameters that can be empirically analysed against stand characteristics.

Few studies have explored theoretical distribution functions to represent the vertical profile of vegetation structure in LiDAR data. All prior studies have focused on a Weibull distribution function, which is unimodal. In a complex native forest ecosystem, the form of the distribution of LiDAR points may be highly variable between forest types and age classes. We compared 44 probability distributions within a two component mixture model to determine the most suitable bimodal distributions for representing LiDAR density estimates of Mountain Ash forests in south-eastern Australia. An elimination procedure identified eleven candidate distributions for representing the eucalyptus component of the mixture model.

We demonstrate the methodology on a sample of plots to predict overstorey stand volumes and basal area, and understorey basal area of 18-, 37-, and 70-year old Mountain Ash forest with variable density classes. The 70-year old forest has been subjected to a range of treatments including: thinning of the eucalyptus layer with two distinct retention rates, removal of the understorey, and clear felling of patches that have 37 year old regenerating forest. We demonstrate that the methodology has clear potential, as observed versus predicted values of eucalyptus basal area and stand volume were highly correlated, with bootstrap based r^2 ranging from 0.61 to 0.89 and 0.67 to 0.88 respectively. Non-eucalyptus basal area r^2 ranged from 0.5 to 0.91.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

1.1. Vertically stratifying multilayered forests with discrete LiDAR data

Light Detection and Ranging (LiDAR) data are facilitating extraordinary advances in improving our understanding of the Earth's biomass by directly measuring the three-dimensional biophysical properties of the vegetation profile. The resulting representation of vertical structure of vegetation and topographic features over the terrain provides insight into the functional characteristics and processes of the land surface. Most LiDAR systems have a multi-echo capability and may capture between two and five returns for every laser pulse by penetrating beyond the first reflective surfaces of the canopy. The ability of discrete return sensors to capture a few echoes per pulse is particularly useful for forest industry applications, which require broad-area information on stand characteristics for timber inventory evaluation and forest growth modelling. For this particular purpose, mean tree height, basal area, and stand volume have been the most important forest mensuration parameters of interest (Naesset et al., 2004).

[*] Corresponding author.
E-mail addresses: dominikj@utas.edu.au (D. Jaskierniak), patrickl@unimelb.edu.au (P.N.J. Lane), a.robinson@ms.unimelb.edu.au (A. Robinson), Arko.Lucieer@utas.edu.au (A. Lucieer).

0034-4257/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2010.10.003

As well as characterising dominant forest stand attributes, LiDAR data may be used to categorise single-storey and multi-storey forest types, which has proven useful for mapping understorey fire behaviour (Zimble et al., 2003). Quantiles of height distribution in LiDAR forest data can be used to predict the vertical structure of forests (Magnussen & Boudewyn, 1998; Maltamo et al., 2005; Naesset, 1997a,b; Naesset et al., 2004). Also, Canopy Height Models (CHM), such as mean canopy height, when derived from LiDAR data, are very accurate at characterising stand attributes because they are directly measured rather than indirectly calculated.

However, LiDAR indices based on discrete statistics such as percentiles and CHM may be improved further by classifying the LiDAR data into vegetation layers to determine vegetation specific statistics. In particular, in vertically heterogeneous multilayered forests it is necessary to stratify the vegetation to address the problem of inter-stand variation in the ratio of LiDAR hits represented in the dominant canopy and the hits in the understorey.

A range of methods has been used to stratify the vegetation profile and develop layer-specific indices. Zimble et al. (2003) used height variance in LiDAR data to determine differences between single-storey and multi-storey forest types, but the method did not stratify each layer. Riano et al. (2003) on the other hand discriminated overstorey and understorey vegetation hits using a cluster analysis technique based on a minimum Euclidean distance method. The crown base of the overstorey was then defined as the 1st percentile of the overstorey layer.

A canopy volume method using volumetric pixels (voxels) was adapted by Holmgren and Persson (2004) to separate the vegetation profile into overstorey and understorey layers. With the horizontal extent of each voxel being the sample plot size, and each voxel element being 0.5 m tall, they were able to assign a value of 0 or 1 to each element according to the relative frequency of *z* values occurring within the corresponding voxel. By assigning zero to each element that contained less than 1% of the total returns in a given voxel, the authors were able to define the base of the crown as the highest voxel element with a value of zero in a given column.

Barilotti et al. (2008) use polynomial regression functions applied to frequency histograms of vegetation profile data to identify base of the crown of dominant trees, by interpreting the local frequency minimum of the linear regression function as the vegetation layer threshold. Maltamo et al. (2005) determined the existence and number of understorey trees by examining the cumulative distributions of the canopy height density, computed as the proportion of hits above different height quantiles. The authors applied a histogram threshold method, developed by Lloyd (1982), to the cumulative distributions to cluster similar data vectors into groups as a means to define a threshold of the dominant tree layer and understorey trees. Although the procedure determined whether the height distribution of hits is multimodal, the accuracy of the results was largely dependent on the density of the dominant tree layer.

Donoghue et al. (2007) used near-infrared intensity of LiDAR hits to differentiate forest species common to different forest layers, as some species reflect light more intensely than others. Distinguishing vegetation layers based on intensity of hits is complicated because intensity values are dependent on variation in laser path length, orientation of the target relative to sensor, laser beam divergence which alters the footprint size, and the attenuation of the signal by the atmosphere. As a result, this approach needs calibration of the intensity values with configurations of the LiDAR system.

A promising method for separating LiDAR hits of different vegetation layers involves fitting of probability distribution models to the density profile of LiDAR data. To date, only unimodal distributions of the Weibull distribution function have been applied to derive LiDAR indices (Coops et al., 2007; Dean et al., 2009; Maltamo et al., 2004).

Coops et al. (2007) recognised that distribution functions provide a mechanism to summarise complex canopy attributes into a short list of parameters that can be empirically analysed against stand characteristics. They found Weibull parameter β, which varies the spread of the distribution, was significantly correlated ($P<0.05$) to mean tree diameter at breast height (DBH), DBH, and stem density ($r^2=0.92$, $r^2=0.77$, $r^2=0.65$). The authors empirically identified a relationship between crown depth and Weibull parameter α, which provides for the scaling and positioning of the distribution.

Dean et al. (2009) estimated height to the base of crown and the height to the median of canopy using truncated Weibull functions. The height to the canopy median was defined as height at the median of the distribution, whereas the height to the base of the live crown was defined as the height where the upper tail asymptotes to zero. Ground-based estimates and LiDAR-based indices of crown median and crown base differed by 0.3 m and 0.6 m respectively. Maltamo et al. (2004) found parameters from the Weibull distribution function may be used to identify suppressed trees in multilayered spruce forests. By applying Weibull distribution functions to estimated tree height distributions obtained from LiDAR data, the authors used Weibull parameters to predict heights of small suppressed trees not identified in the point cloud data. The use of the method reduced RMSE values from 25% to 16% for stand volume estimates, and 75% to 49.2% for the number of stems.

1.2. LiDAR indices using mixture distribution functions

Mixture models are often used in forest management to quantify merchantable timber by characterising the irregular diameter frequency distributions of mixed-species or uneven-aged forest stands (Liu et al., 2002; Zhang et al., 2001; Zhang & Liu, 2006). The present study distinguishes itself from this typical use of mixture models in forest inventory analysis by applying mixture models to LiDAR height distributions in order to estimate plot level stand characteristics. This study generalises the unimodal distribution approach applied by Coops (2007), Dean (2009), and Maltamo (2004) by using mixture models with a range of theoretical distribution functions to develop LiDAR indices that are useful for a broad range of forest management purposes, including forest hydrological research. Forest structure regulates evapotranspiration rates through its influence on the wind profile, which partially determines the vapour pressure deficit at the transpiring leaf surface (Monteith, 1965). For this reason, LiDAR indices relating to crown height, density, depth, and closure of both understorey and overstorey layers, are of interest for quantifying forest aerodynamic properties that influence evapotranspiration rates. Canopy profile attributes such as crown density, depth, and closure are also strongly related to Leaf Area Index (LAI), which is an important predictor of evapotranspiration (Vertessy et al., 2001). LiDAR indices that can predict forest productivity are important for forest hydrological research as forest growth rates may be used to predict forest water use (Raison et al., 2001).

In order to produce hydrologically related canopy profile indices, the two main objectives of this paper are:

- to develop a methodology the uses mixture models with a wide range of theoretical distribution functions as a means to provide a generalised approach for characterising the structure of specific layers of multilayered forests from LiDAR data, and
- to empirically evaluate the LiDAR derived canopy profile indices of understorey and overstorey vegetation for their capacity to predict vegetation specific plot level basal area and stand volumes in multilayered forests.

2. Methodology

2.1. Study site and field measurement description

The forested catchments used for this study were long-term research sites established in Melbourne's water catchment to investigate the impacts of land cover disturbance on the water resource. The 1939 bushfire in Victoria, Australia burnt much of Melbourne's water catchments and the regeneration process resulted in changes to the rainfall–runoff relationship as the dense regrowth forest consumed more water than the pre-disturbance mature forest (Kuczera, 1987). Permanent growth plots were established in a set of treated catchments in the early 1970s to investigate the impacts of forest density and age on forest water use. Treatments included: thinning of the eucalypt layer with two distinct retention rates, removal of the understorey, and clear felling of patches that have 37 year old regenerating forest.

For this study, the permanent plots were revisited in the summer of 2008/09 for measurements of diameter at breast height over bark (DBHOB) of all eucalyptus trees, and non-eucalyptus trees greater than 10 cm in DBHOB. In 2007, LiDAR data acquisition took place from a fixed wing aircraft, and Table 1 provides the flight details and sensor configurations. The plot level analysis of LiDAR data involved 18-, 37-, and 70-year old Mountain Ash forest with variable density classes.

Field measurements were taken for the original plots as well as 'extended' plots, which were adjusted to allow for changing forest conditions since the plots were established. Differential GPS measurements were collected at permanent pegs located at each plot using a dual frequency surveying grade Leica GPS1200 receiver. The DGPS accuracy was compromised by dense forest conditions and steep terrain, and for this reason the plot locations were visually adjusted in GIS by 0–5 m to manually correspond the pattern of tree tops in LiDAR data with measured tree locations allocated into 5 by 5 m sub-plots. The sub-plots also allowed the construction of consistent 15 by 20 m plots for all catchments to evaluate differences in mixture model outcomes when plot size is constant over the range of forest types. Table 2 provides summary statistics and treatment effects for the six catchment sites in the study area.

2.2. Generation of height above the ground

The height of the point cloud at an intercepted surface was measured relative to sea level which consequently needed to be converted into a height above a ground surface to yield point clouds that represent vegetation height. The procedure involved producing an accurate digital terrain model (DTM) representing the bare ground surface. A DTM for each catchment was produced with a thin-plate spline interpolator using Topo to Raster, an ArcGIS interpolation tool based on the ANUDEM algorithm (Hutchinson, 1989, 2005). The interpolation used classified ground hits separated from vegetation hits by the LiDAR contractor using TerraScan Software. The ground-classification procedure used an iterative procedure to build a triangulated model of the ground surface (Axelsson, 1999; Kraus & Pfeifer, 1999). The DTM height was subtracted from the remaining vegetation hits to obtain vegetation height for each LiDAR point.

2.3. Preparing LiDAR data for plot-based analysis

To use probability distributions to capture density of LiDAR points across the vertical profile, we needed to ensure that overlapping flight paths did not distort the density of the vegetation point cloud. The density of the vertical profile is otherwise distorted for all plots and grids partially represented by one flight path and partially by overlapping flight paths. To address this problem, a point density map was generated to identify strips that had overlapping flight paths, which were delineated and intersected down the middle to define the boundary used to adjoin adjacent flight paths. The GPS timestamp of the dataset was then used to group the point cloud into representative flight paths and the overlapping edges were removed.

Table 1
Flight details and sensor configurations for the LiDAR data acquisition.

LiDAR system configurations	
Date of flight	August 26th 2007
Sensor type	Optech ALTM3100
Flight altitude (m)	800
Airspeed (km/h)	220
Wavelength (Hz)	69
Pulse repetition rate (kHz)	100
Laser beam divergence (mrad)	0.3
Scan angle (degrees)	28
Mean footprint size (m)	0.16
Pulses per square metre	4
Maximum returned signals	4

The resulting point cloud consisted of all four LiDAR returns and represented the vegetation density from the ground level up. The vegetation at the field sites predominantly consisted of a ground, understorey, and overstorey layer. As field measurements did not include the ground layer shrubs, all points with a height value less than 3 m were removed. The removal of these points was necessary as the methodology that follows only implements bimodal distributions, which do not fit the complete vegetation profile adequately. Alternatively, a multimodal mixture modelling exercise would be necessary. Such an extension is beyond the scope of this study, due to challenges addressed in the discussion.

2.4. Generation of mixture models to estimate vertical profile density

To process the computationally intensive technique outlined below, the University of Melbourne's servers running Red Hat Enterprise Linux 5.3 (64bit) with open source software R, version 2.8.1 was available (R Development Core Team, 2009). Four SunFire ×4600M2 servers were used, each of which had 64 GB of memory and 8 CPUs × 4 cores (32 cores) with a CPU speed of 2.3 GHz.

Using plot-based LiDAR data for each of the plot sizes, mixture distributions were applied to estimate the density of LiDAR points across the vertical profile of the vegetation structure as a means to develop a robust predictor of basal area and stand volume. In a complex native forest ecosystem, the form of the distribution of LiDAR points may be highly variable between forest types and age classes. To accommodate for such complexity in the density distributions, Generalized Additive Models for Location, Scale and Shape (GAMLSS) are used. GAMLSS are semi-parametric regression type models fitted with a parametric distribution assumption for the response variable, and may include non-parametric smoothing functions, hence "semi-parametric", to model the parameters of the distribution (Stasinopoulos et al., 2008). The GAMLSS framework has been implemented using a series of packages available as part of the open source R software (R Development Core Team, 2009).

The GAMLSS method is suitable for handling complexity in the forest structure as there are 44 different continuous distribution functions available to capture the variable density estimate of LiDAR points across the vertical profile (Stasinopoulos et al., 2008). The number of parameters represented in the GAMLSS distributions varies from one to four, with almost all distributions represented by a location (μ) and scale (σ) parameter and some distributions represented by one or two shape parameters (υ and τ) to represent skewness and kurtosis in the response variable data. For this reason, the form of the distribution assumed for the response variable y, $f(y_i|\mu_i, \sigma_i, \upsilon_i, \tau_i)$, can be very general.

To create mixture models with GAMLSS distributions the R package *gamlss.mx* uses the expectation minimization (EM) algorithm (Rigby & Stasinopoulos, 2008). A mixture model of GAMLSS distributions has the form

$$f_y(y|\psi) = \sum_{k=1}^{K} \pi_k f_k(y|\theta_k) \tag{1}$$

where f_Y depends on parameters $\psi=(\theta, \pi)$ where $\theta=(\theta_1, \theta_2,\ldots, \theta_K)$ and $\pi^T=(\pi_1, \pi_2,\ldots, \pi_K)$; $f_k(y|\theta_k)$ is the probability function of y for component k; and $0\le\pi_k\le 1$ is the prior probability of component k,

Table 2
Summary statistics of the extended plots located in six 1939 regenerating forest catchments exposed to a range of silvicultural treatments.

Catchment	Treatment	Original Plot Size (m)	2009 Extended Plot Size (m)	Number of plots	Eucalyptus tree count per hectare			Eucalyptus basal area her hectare			Non-Eucalyptus basal area per hectare		
					Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Black Spur 1	Patch cut 40% (1972)	10×20	15×20	76	0	118	266	0	36	123	0	9.1	51.4
Black Spur 2	Thinned 40% (1972)	40×40	40×40	7	75	96	131	34	48	59	1.9	4.7	10.9
Black Spur 3	Thinned 60% (1972)	40×40	40×40	7	50	76	112	36	48	63	0.7	3.1	5.6
Ettercon 2	Understorey removed (1972)	10×40	15×40	21	0	133	217	0	43	67	0	2.9	28.1
Ettercon 3	No treatment	10×40	15×40	40	17	140	317	15	51	84	1.9	7.6	18
Myrtle 2	1984 regeneration after clear fell	5×20 & 10×20	15×20	21	167	458	969	16	36	63	0	4.1	14

for $k = 1,2,\ldots, K$ (Rigby & Stasinopoulos, 2008). In the present study, the K value is set at two, because the vegetation's vertical profile above 3 m was predominantly well represented with a bimodal distribution function.

The Akaike Information Criterion (AIC) value was used as a goodness of fit measure to identify the most suitable distributions in the bimodal density estimates (Akaike, 1974). The AIC value offers a relative measure of the information that is lost when a distribution function is used to describe the data, and has the form $AIC = 2p - 2\ln(L)$, where p is the number of parameters and L is the maximised value of the likelihood function for the estimated model.

The GAMLSS mixture models can use any combination of distribution functions, so we needed to reduce the 1936 combinations of possible bimodal distributions to a manageable amount. We placed emphasis on exhausting likely candidate distributions of the dominant vegetation layer in the second component of the mixture model to result in the evaluation of 390 bimodal distributions on each plot in the study.

The procedure began by using the normal distribution function in the first component (understorey) as it had proven reliable at converging, whilst evaluating each of the 44 available GAMLSS distributions in the second component (overstorey). The five second component (overstorey) distribution functions that proved most successful at converging with the lowest AIC value were then used to represent the first component (understorey), whilst evaluating each of the 44 available GAMLSS distributions in the second component (overstorey). Using the same performance criteria, the seven most successful second component (overstorey) distribution functions not yet assigned in the first component (understorey) were assigned as the first component (understorey) and coupled with 18 s component (overstorey) distributions most successful at addressing the performance criteria.

To reduce computational time in extrapolating 390 bimodal distributions over each catchment, four catchment specific mixture models were selected with the highest convergence rate and lowest 90th percentile in AIC values. We used the mixture model with the lowest plot-specific AIC value out of the four catchment specific mixture models to identify the optimal plot-specific bimodal distribution for generating LiDAR indices in the predictive models.

Allowing each plot's LiDAR data to determine the distribution function of each vegetation layer addresses the: (a) variation in each layer's canopy profile structure; (b) variation in ratio of LiDAR hits between layers; and (c) variation in the transition area between layers. The interaction of distribution functions influences how well a particular distribution function performs in conjunction with others. For example, Fig. 1 shows the Gumbel (GU) distribution in the first component of a bimodal curve behaves very differently depending on whether an inverse Gaussian (IG) or logistic (LO) distribution is used in the second component. In this particular example, it is evident that the lower tail of the logistic distribution in the overstorey is more compatible with the Gumbel distribution in the understorey and hence provides an overall better fit. In a mixture modelling procedure, the selection of each distribution as well as the interaction of the distributions will affect the final bimodal density estimate. For this reason, the resulting bimodal curves within a particular forest type can be highly variable and may need to be accommodated with a range of candidate distribution functions.

2.5. Generation of LiDAR indices

Using the GAMLSS package, the different distribution functions have corresponding parameters with identical physical interpretations as they represent the same forest structural attributes. For example, the location parameter, μ, and probability density estimate parameter, ρ, are both comparable between different bimodal distributions as they always represent the canopy mode and proportional density of a particular vegetation layer. As a result, it is not necessary to find one bimodal distribution to represent all sample plots when generating LiDAR indices for a study. On the other hand, the scaling parameter σ, and shape parameters υ and τ, have physical interpretations specific to a distribution function so may not be compared between different mixture models.

Table 3 lists plot level LiDAR indices generated with the following three methods: indices produced with no stratification of vegetation hits, indices vertically stratified with a eucalyptus and non-eucalyptus component, and indices horizontally and vertically stratified by calculating plot level vegetation specific averages using sub-plot grids. All indices have been calculated using all four LiDAR returns and indices with an astricts (*) have also been calculated using only the first return LiDAR hits.

Plot level indices with no vertical stratification of vegetation layers included height percentiles, the number of ground hits, and standard deviation of hits greater than 3 m. Height percentiles provided a height value for the proportion of data below a given percentile. The 99th percentile defined a measure of the maximum height in the plot, whereas the rest of the percentiles provided an indication of the variation in density across the vegetation profile. The number of ground points was inversely proportional to the total vegetation density. Standard deviation of all points greater than 3 m provided an indication of the clumpiness of the canopy profile.

Parameters extracted from the bimodal distribution functions include the canopy mode of each vegetation layer, represented by the location parameter, μ. The probability density estimate parameters, ρ, represent the proportion of hits in the eucalyptus and non-eucalyptus layer relative to each other. To vertically stratify the LiDAR hits, the canopy base height of the eucalyptus layer was calculated by determining the height percentile that separates the two strata. For this purpose, the following percentile needed to be solved:

$$(1-\rho_n) * 100 \tag{2}$$

where ρ_n is the probability density estimate representing the second component of the bimodal distribution function. The number of hits intercepted by the eucalyptus layer reflects the density of the layer and was calculated by determining the total count of LiDAR values

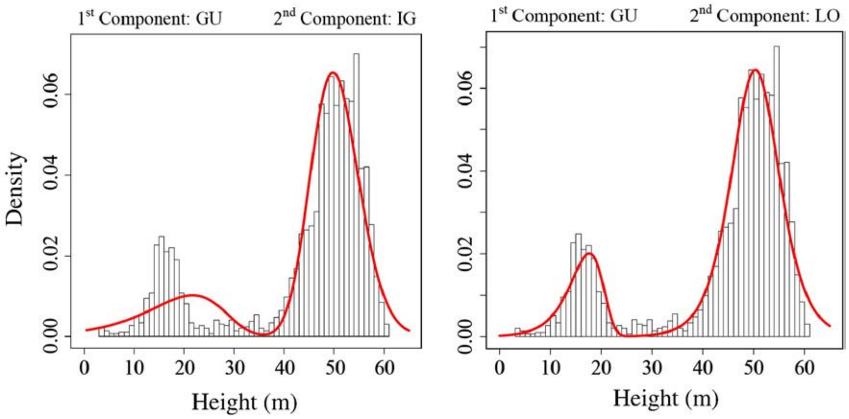

Fig. 1. Example of how the interaction of distribution functions determines the fit of each component of a mixture *model.*

greater than the eucalyptus canopy base height. The stratified LiDAR points were used to determine a statistical summary of the minimum, maximum, range, variance, mean, and percentiles for each vegetation layer.

The adjusted probability density estimate index accounted for the number of ground points and non-ground points less than 3 m in height to determine a probability density estimate of each vegetation layer relative to the total count of LiDAR hits within a given plot. The overall proportion of LiDAR points intercepted by each vegetation layer was calculated by dividing the number of hits in the vegetation layer by the sum of ground and all non-ground hits.

Finally, for each vertically stratified vegetation layer, a set of canopy profile indices was adjusted to correct for the within-plot variation in the horizontal heterogeneity of the canopy profile. To produce vertically and horizontally stratified indices, 5 m sub-plots were used to spatially average for each vegetation layer the mean, minimum, maximum and range statistics over each plot.

Table 3
List of plot level LiDAR indices generate for each plot.

Plot level LiDAR indices	Symbol
Non stratified indices	No prefix
1. Number of ground points (numeric)	*Gnd*
2. 99th, 95th, 90th, 80th...20th, 10th, 5th, 1st percentile (m)	*P99,…, P1*
3. Standard deviation of hits>3 m (m)	*SD*
Vertically stratified indices (Vegetation specific)	Prefix: *Euc_* or *Non_*
4. Canopy mode using μ parameter (m)	*Mu*
6. Probability density estimate parameter, ρ, using hits>3 m (%)	*Den>3*
7. *Number of hits (numeric)	*F1_Hit* or *Hit*
8. Probability density estimate parameter, ρ, corrected with all hits (%)	*Den*
9. *Minimum height (m)	*F1_Min* or *Min*
10. *Maximum height (m)	*F1_Max* or *Max*
11. *Height range (m)	*F1_Rg* or *Rg*
12. *Height variance (m)	*F1_Var* or *Var*
13. *Mean height (m)	*F1_Avg* or *Avg*
14. *99th, 95th, 90th, 80th...20th, 10th, 5th, 1st percentile (m)	*F1_P99,…, P1* or *P99,…, P1*
Vertically and horizontally stratified indices (Vegetation specific)	Prefix: *Euc_* or *Non_*
14. Mean of sub-plot minimum height (m)	*Avg_Min*
15. Mean of sub-plot maximum height (m)	*Avg_Max*
16. Mean of sub-plot range height (m)	*Avg_Rg*
17. Mean of sub-plot mean height (m)	*Avg_Avg*

2.6. Regression analysis of LiDAR indices against field measured forest characteristics

We used a total of 104 LiDAR indices as candidate predictor variables. These indices were derived from a combination of methods using summary statistics, percentile extraction methods, and mixture models. Non-eucalypt basal area, eucalyptus basal area, and eucalyptus stand volume were the response variables.

The large number of candidate predictors and the inherent collinearity between percentiles were of particular concern when developing predictive models with standard regression techniques such as least-squares and stepwise selection. For prediction purposes in high-dimensional predictor spaces, models generated by shrinkage regression techniques may be more accurate than standard regression techniques (e.g. Hastie et al. (2001) and citations therein). Shrinkage regression techniques, such as ridge regression, are model-fitting methods that use penalties or constraints that shrink parameter estimates to avoid over-fitting. For example, ridge regression minimises the residual sums of squares with a penalty on the sum of the squares of the regression coefficient estimates (Hoerl & Kennard, 1970). The coefficients that estimate $\hat{\beta}j$ are those that minimise the ridge regression objective function:

$$O_{RR} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (3)$$

where λ controls the amount of shrinkage, and is usually selected by cross-validation, y_i is the response variable, x_{ij} are the p predictor variables, and β_j are the $p+1$ unknown parameters.

Although ridge regression benefits from a lower variance of parameter estimates and increased prediction accuracy, the procedure keeps all parameters in the model, which makes it undesirable for seeking a parsimonious solution that consists only of the most dominant explanatory variables. For this reason, we applied a pre-screening step that involved selecting a list of 2, 3, 4, and 5 predictor variables that had the highest absolute conditional correlation with the response variable. That is, after the first variable had been identified, we chose the next variable that had maximum absolute correlation with ordinary least-squares fit of the already chosen parameters against the nominated response variable. For each candidate list of predictor variables, a family of competing models with the same number of parameters but different parameter shrinkage levels was generated. To identify the best model from each family of competing models with the same predictor variables, the Generalized Cross Validation (GCV) (Golub et al., 1979) procedure was applied, defined as:

$$GCV = \frac{\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\sum_{j=1}^{p}x_{ij}\hat{\beta}_j\right)^2}{n(1-p/n)^2} \tag{4}$$

where p is the number of parameters in the model, N is the number of observations, $\hat{\beta}_j$ is the estimate of the j^{th} parameter and the other symbols as are above.

From each family of models with a predetermined number of predictor variables, the models with the most optimal level of shrinkage, based on the GCV values, were compared to find the overall model that offers the best predictive accuracy. For this purpose the Prediction Squared Error (PSE) is a metric of prediction accuracy used to identify the smallest difference between the observed values and those predicted by the model, defined as:

$$PSE = \sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\sum_{j=1}^{p}x_{ij}\hat{\beta}_j\right)^2 \tag{5}$$

The quality of any regression model's coefficient estimates are over-optimistic in a procedure that determines the quality of the model's estimate using the same data to fit and then assess the model. To address this concern, we adopted the “0.632+” bootstrap method (Efron & Tibshirani, 1997) to correct for misleading estimates of future observation variances. The 0.632+ bootstrap is comparable to cross-validation but is more efficient, making it desirable when observations are few. The procedure involved randomly selecting 1000 subsets from the original dataset with replacement, with each sample having the same number of observations as the original dataset. The 0.632+ bootstrap mimics cross validation, as for each observation i, all bootstrap samples that do not contain observation i are used to predict the value of observation i and measure the error. Our bootstrap operation included the initial screening of the variables, so any uncertainty created in the model step by that screening was included in the bootstrap estimates of PSE.

3. Results

3.1. Identifying the best fitting mixture models

The first step in identifying the most suitable bimodal distribution function for each plot required the following iterative procedure. We used the normal distribution function in the first component (understorey) of the mixture model whilst testing all available distribution functions in the second component (overstorey). The five best performing second component distribution functions are listed in the first column of Table 4. In the second step, these five distributions were used in the first component and coupled to each of the available distribution functions assigned to the second component. The second column of Table 4 provides the seven best performing second component distributions in the second step, based on the criteria outlined in the methodology. In the third step, these seven distributions were represented in the first component and coupled to eighteen best performing second component distribution functions identified in the second step. The eighteen best performing distribution functions were the normal distribution and all the distribution functions listed in Table 4.

To identify the best catchment specific bimodal distributions, all candidate mixture models were evaluated and only those mixture models that converged for all plots in the given catchment were considered. For these mixture models, the 90th percentile of the catchment's plot AIC values was used as a performance criterion to determine the four best performing mixture models. Table 5 lists the four most successful mixture models for each plot size in each catchment. A count of the number of plots that performed the best for each of the four most successful mixture models is also provided for both the original plot size and the extended plot size.

The results show that the optimal bimodal distribution function varied between catchments, and the Gumbel function (GU) was often successful at representing the overstorey Mountain Ash forest type. Within any given catchment, there was also variation between plots of a given plot size and there was often one predominantly successful bimodal curve. These results provide evidence that the use of a single distribution function will not adequately capture the heterogeneity of Mountain Ash forest structure, and the deployment of a range of distributions needs consideration.

In Mountain Ash forests, the elimination procedure identified eleven likely candidate distributions for representing the eucalyptus component of the mixture model. The eleven distribution functions included all distributions in first two columns of Table 4 excluding the Zero Adjusted Inverse Gaussian (ZAIG) and Generalised Gamma (GG) function. It is worth noting that eight of the

Table 4
Best performing distribution functions for plot-based LiDAR evaluated in this study.

	Best performing distribution functions	
First component in second step	First component in third step	Second component in third step[a]
Weibull (WEI)	Zero adjusted inverse Gaussian (ZAIG)	Log normal (LogNo)
Gumbel (GU)	Weibull (WEI3)	Generalised t (GT)
Gamma (GA)	Generalised Gamma (GG)	Generalised Inverse Gaussian (GIG)
Inverse Gaussian (IG)	T Family (TF)	Generalised Beta type 1 (GB1)
Skew T type 4 (ST4)	Logistic (LO)	Box-Cox t (BCT)
	Reverse Gumbel (RG)	
	Normal family (NOF)	

[a] Distribution functions listed in first and second rows are also evaluated as second component distribution functions in the third step.

Table 5
The four best performing distribution functions for each plot extent in each catchment and the number of plots that performed the best for a given mixture model in a given catchment. Empty records imply that the plot size is the same as the original plot size for the given catchment.

Catchment	Original plot size		Extended plot size		15×20 plot size
	Mixture model	Number of plots	Mixture model	Number of plots	
Black Spur 1	f (GA, GU)	53	f (RG, GU)	60	–
	f (GA, WEI)	16	f (GU, GA)	2	–
	f (GA, IG)	7	f (WEI, IG)	11	–
	f (GA, WEI3)	0	f (WEI, RG)	3	–
Black Spur 2	f (IG, WEI)	7	–	–	f (IG, GU)
	f (GA, TF)	0	–	–	f (ST4, GA)
	f (ST4, IG)	0	–	–	f (RG, ST4)
	f (TF, ST4)	0	–	–	f (WEI, IG)
Black Spur 3	f (LO, ST4)	5	–	–	f (GU, RG)
	f (ST4, IG)	2	–	–	f (IG, GU)
	f (ST4, ZAIG)	0	–	–	f (GU, IG)
	f (ZAIG, WEI)	0	–	–	f (GU, NO)
Ettercon 2	f (ZAIG, ST4)	11	f (GA, GU)	18	f (NO, ST4)
	f (ST4, GU)	9	f (GU, WEI3)	1	f (GA, GU)
	f (ST4, LO)	1	f (TF, IG)	1	f (WEI, IG)
	f (RG, GU)	0	f (IG, NO)	1	f (WEI, WEI)
Ettercon 3	f (WEI, ST4)	20	f (RG, GU)	32	f (ST4, GU)
	f (NO, GU)	17	f (GU, LO)	4	f (WEI, WEI)
	f (GU, NO)	3	f (GU, NO)	2	f (WEI, LO)
	f (ST4, LO)	0	f (WEI, RG)	2	f (GU, GU)
Myrtle 2	f (ZAIG, TF)	13	f (RG, GU)	7	–
	f (WEI, NO)	6	f (IG, WEI)	9	–
	f (WEI, RG)	2	f (WEI, RG)	3	–
	f (NO, NOF)	1	f (GU, GA)	2	–

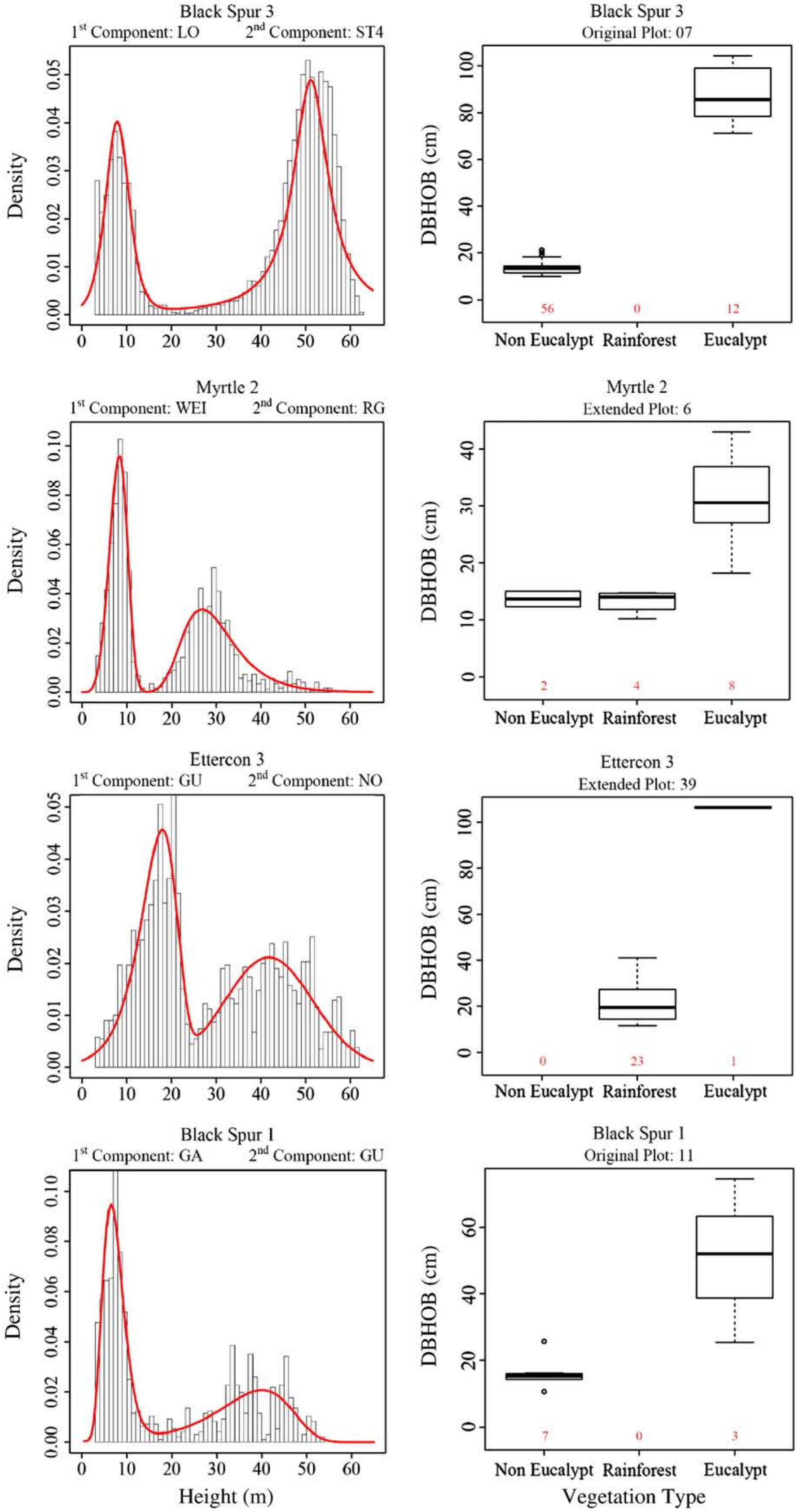

Fig. 2. Bimodal curves represented with four different second component distribution functions fitted to the plot-based LiDAR data. Box plots provide a summary of each plot's forest inventory.

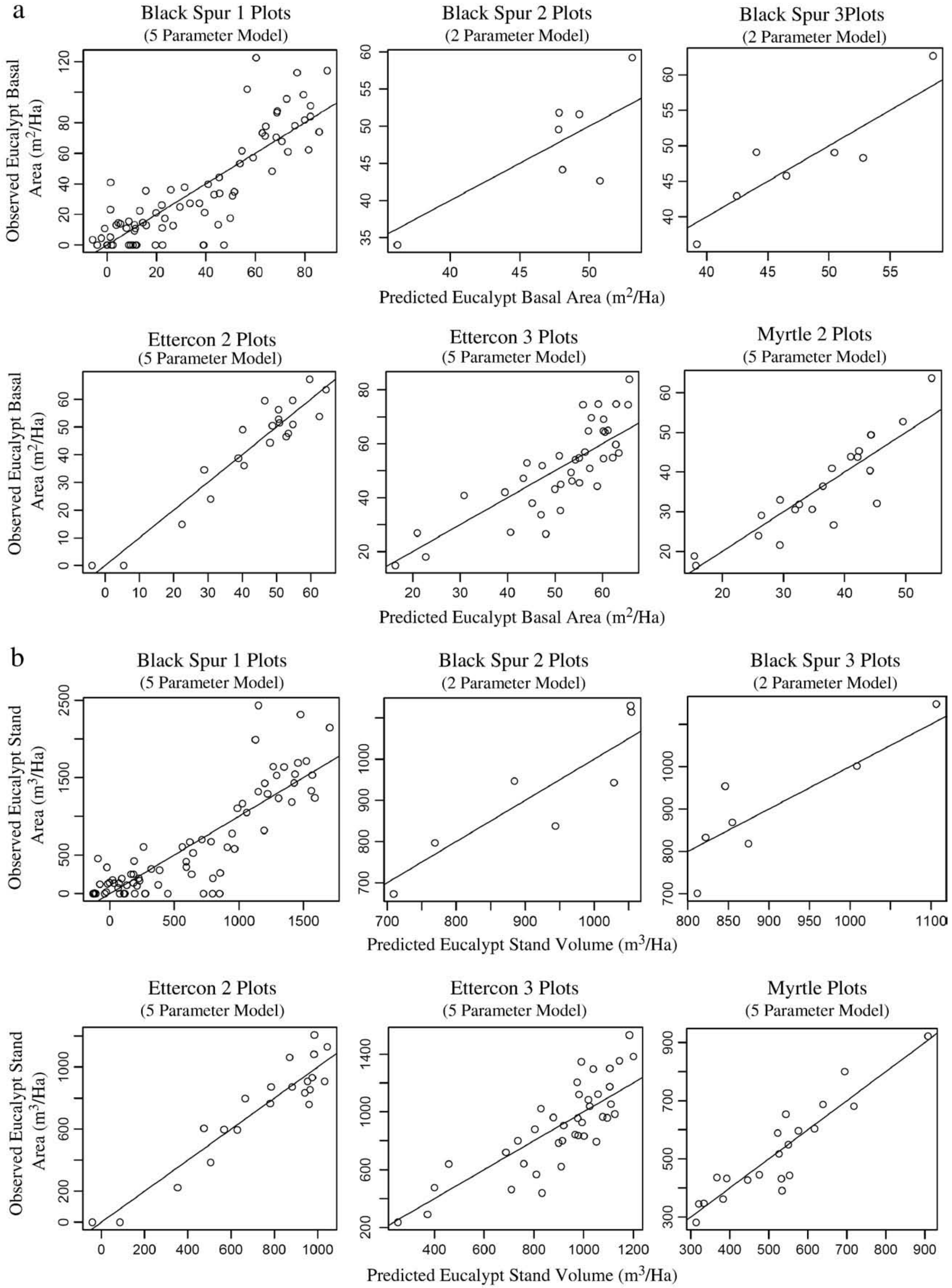

Fig. 3. a: Scatter plots of predicted versus observed eucalyptus basal area values using ridge regression modelling. b: Scatter plots of predicted versus observed eucalyptus stand volume values using ridge regression modelling.

eleven distribution functions represent two-parameter models, which are a great deal more computationally efficient to fit than the computation-intensive distribution functions that have three or four parameters.

For each of the plots, mixture model curves were superimposed over frequency histograms of LiDAR hits along the vertical vegetation profile in order to visually evaluate the effectiveness of the mixture model at representing the distribution of LiDAR hits. Fig. 2 provides one randomly selected plot for four out of the eleven most effective second component distribution functions.

3.2. Ridge regression predictions

3.2.1. Eucalyptus vegetation layer

Using ridge regression analysis for each of the catchments, Fig. 3 provides scatter plots of predicted versus observed values for eucalyptus; (a) basal area and (b) stand volume. The small number of observations at Black Spur 2 and 3 meant the model was limited to only two parameters to avoid over-fitting noise and erroneous associations between the predictor and response variables. The results for the patch cut treatment catchment, Black Spur 1, show overestimations of predicted basal area in plots with no eucalyptus trees. These plots are mainly located on the edge of cleared patches where eucalyptus regeneration is suppressed due to shading from retained trees. These plots have no eucalyptus trees but have overhanging trees external to the plot, which misrepresent the presence of eucalypts.

In Black Spur 1 there is also a tendency to underestimate basal area in plots with observed basal area greater than 80 m^2/ha. In dense Mountain Ash forest stands the intense competition results in suppressed trees having highly irregular canopy structure. The suppressed trees can contribute a substantial amount of basal area to the plot measurements with a disproportionately reduced crown structure. Such circumstances inevitably result in underestimated basal area predictions using LiDAR data.

Table 6 provides the RMSE, relative RMSE, and R^2 results of the ridge regression, as well as lists the LiDAR indices used to make predictions of: (a) eucalyptus basal area and (b) stand volume. Relative RMSE is calculated by dividing the RMSE value by the mean of the field observation values. A significant portion of the predictor variables used in the final models include vegetation specific indices generated with the mixture modelling methodology. The mixture model index, *Euc_Hits*, which represents the total hits intercepted by the eucalyptus layer, and the percentile extraction method index, *P50*, which represents the 50th percentile of all hits, are notably the most consistent predictive LiDAR indices. For comparison with models that use traditional predictor variables, Table 7 provides the results of predictive models generated using predictor variables that do not require mixture modelling. It is evident that by using mixture model LiDAR indices, basal area and stand volume predictions were respectively improved by 4–20% and 4–16%.

Table 7
RMSE, and R^2 of ridge regression models using only predictor variables that do not require mixture modelling (i.e. rows 1, 2, 3, 10, and 15 in Table 3).

Catchment	Basal area (m^2/ha)			Stand volume (m^3/ha)		
	R^2	RMSE	Number predictor variables	R^2	RMSE	Number predictor variables
Black Spur 1	0.68	19.2	4	0.7	365	4
Black Spur 2	0.53	6.4	2	0.75	91	2
Black Spur 3	0.77	4.2	2	0.74	73.8	2
Ettercon 2	0.69	10.4	4	0.72	179	4
Ettercon 3	0.57	10.7	5	0.59	194	5
Myrtle 2	0.74	6.1	4	0.8	72.3	4

3.2.2. Non-eucalyptus vegetation layer

Fig. 4 provides scatter plots of non-eucalyptus stand basal area predictions for each catchment and Table 8 provides the RMSE, relative RMSE, and R^2 results of the ridge regressions, as well as the list of LiDAR indices used to make predictions of non-eucalyptus basal area. Using LiDAR data to model basal area of understorey vegetation is more challenging than modelling overstorey vegetation as the number of hits intercepted by the understorey is a function of the overstorey density. Furthermore, modelling basal area of understorey vegetation is confounded by measured trees leaning out of the plot and unmeasured trees leaning in, as well as unmeasured ferns and non-eucalyptus trees with DBHOB less than 10 cm rightfully in the plot and contributing substantially to the understorey profile.

4. Discussion

A generalised methodology has been presented for representing the vertical forest structure of a broad range of forest types. We have demonstrated that canopy attributes captured by LiDAR data may be summarise into a short list of parameters for empirical analyses against field measured stand characteristics using mixture modelling methods. To evaluate the robustness of the methodology, mixture models of each sample plot were visually assessed to determine how

Table 6
RMSE, relative RMSE, and R^2 of the ridge regression model, as well as the list of predictor variables used in the final model to predict eucalyptus: (a) basal area, and (b) stand volume, for each catchment and all catchments lumped together. Predictor variables with an astricts symbol (*) were developed by stratifying the vegetation layers using mixture models.

Catchment	R^2	RMSE (m^2/ha)	Relative RMSE	Predictor variables used in final model				
(a)								
Black Spur 1	0.72	18	0.5	P50	F1_Euc_Hits*	SD	F1_Euc_P30*	Non_Euc_P70*
Black Spur 2	0.61	5.1	0.10	Den>3*	P50			
Black Spur 3	0.81	3.5	0.07	F1_Euc_Hits*	Euc_Max*			
Ettercon 2	0.89	6.2	0.14	Euc_Hits*	F1_Euc_P10*	Gnd	SD	Euc_Avg_Avg*
Ettercon 3	0.66	9.6	0.18	P50	F1_Euc_Rg*	Euc_Hits*	F1_Euc_Hits*	F1_Euc_P50*
Myrtle 2	0.84	4.7	0.13	F1_Euc_P10*	Den>3*	Euc_P99*	SD	F1_Euc_Hits*
Catchment	R^2	RMSE (m^3/ha)	Relative RMSE	Predictor variables used in model				
(b)								
Black Spur 1	0.76	324.5	0.53	P50	Non_Euc_P70*	Euc_Avg_Avg*	SD	Gnd
Black Spur 2	0.81	72.9	0.08	P50	Euc_Avg_Rg*			
Black Spur 3	0.78	67.3	0.07	F1_Euc_Hits*	Den*			
Ettercon 2	0.88	117.0	0.16	F1_Euc_P60*	Euc_Hits*	Gnd	Den>3*	Euc_Avg_Avg*
Ettercon 3	0.67	176.0	0.19	P50	F1_Euc_Hits*	Euc_Hits*	F1_Euc_Min*	F1_Euc_P50*
Myrtle 2	0.85	63.7	0.12	Euc_Avg_Avg*	F1_Euc_Var*	SD	Non_mu*	Gnd

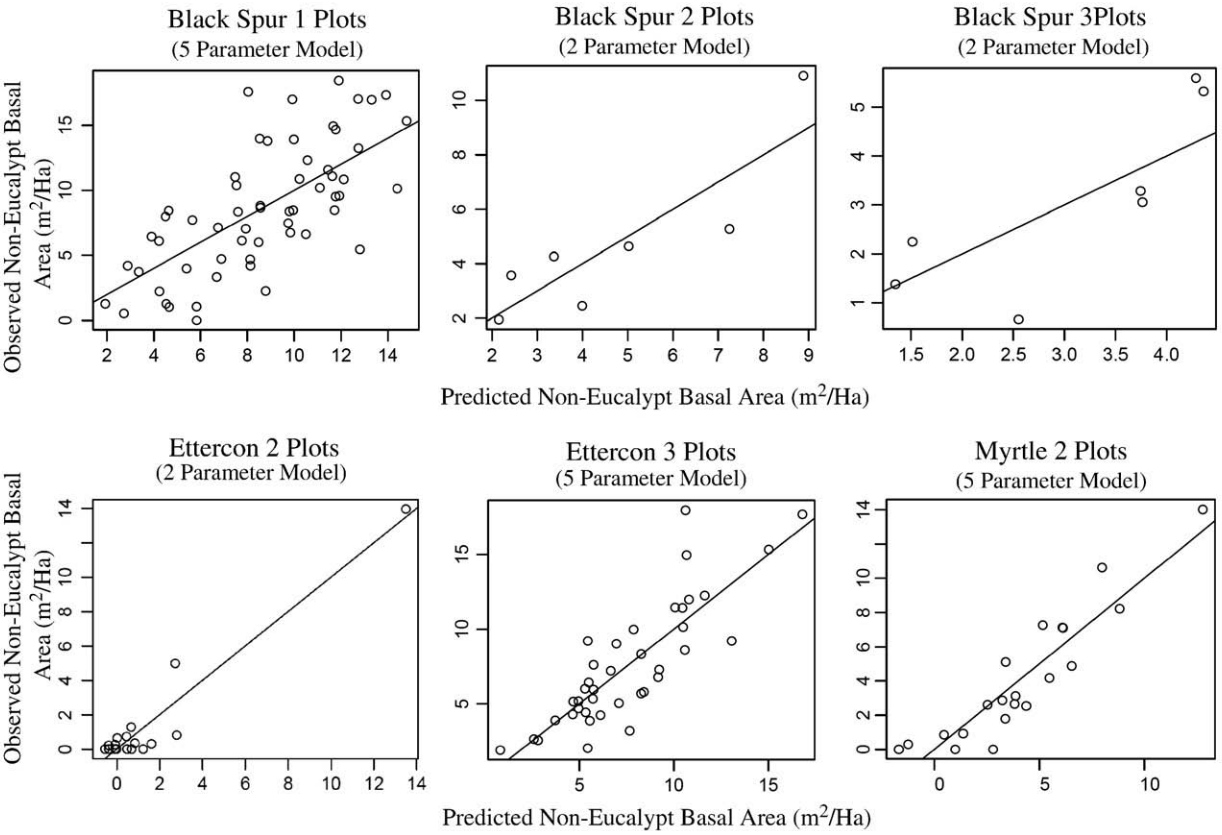

Fig. 4. Scatter plots of predicted versus observed values of non-eucalyptus basal area using ridge regression modelling.

well each component represented the correct vegetation layer. Fig. 5 illustrates the following four types of erroneous fits identified in the mixture modelling procedure:

(a) Distorted distribution functions in the young regrowth forest of the Myrtle 2 catchment when the plot contained old growth stags unmeasured in the field. It was possible to correct these plots by identifying and removing LiDAR data above an expected maximum young regrowth height across the whole catchment.

(b) Bimodal distribution functions that did not effectively identify the appropriate vegetation layer. This predominantly occurred for plots along streams where overstorey vegetation consisted of rainforest vegetation in the absence of eucalyptus trees. The second component of the mixture model was assigned to represent the eucalyptus layer under such circumstances as no conditions or constraints were applied in the modelling exercise to correct such anomalies. Plots that were within a 20 m buffer of a stream were removed from the regression analysis as it may be assumed that eucalyptus trees are not present along the riparian strip when extrapolating the regression over the catchment. Hill slope plots with no eucalyptus trees were used in the regression and weakened the final results, which was most evident in the patch cut silvicultural treatment catchment Black Spur 1.

(c) Bimodal distribution functions that attempted to represent vegetation profile containing more than two vegetation layers. This error was not corrected and resulted in an underestimate or overestimate of the eucalyptus canopy density depending on how the bimodal distribution captured the three or more vegetation layers.

(d) Distribution functions that integrate vertically overlapping rainforest and eucalyptus trees into the second component. This error was not corrected and resulted in an over-estimated density of the eucalyptus canopy.

The complexity in vertical heterogeneity of multilayered forests may be addressed by generalising the mixture modelling procedure with multimodal distributions. Developing distribution curves with variable modes based on the site specific vegetation profile requires a procedure that can determine how many modes best represent the vertical profile, and identify which particular vegetation layer each component of the mixture model represents. To determine how many modes best represent the vegetation profile it may be possible to fit

Table 8
RMSE, relative RMSE, and R^2 of ridge regression used to predict non-eucalyptus basal area and the list of predictor variables in the final model.

Catchment	R^2	RMSE (m^2/ha)	Relative RMSE	Predictor variables used in model				
Black Spur 1	0.5	3.5	0.38	F1_Euc_P1*	SD	F1_Euc_Hits *	Non_Avg*	Euc_Hits *
Black Spur 2	0.9	1.0	0.21	Den>3*	F1_Euc_Var*			
Black Spur 3	0.69	1.0	0.32	Non_Mu*	Den *			
Ettercon 2	0.91	2.1	0.72	P70	Gnd			
Ettercon 3	0.71	2.2	0.29	Den>3*	F1_Euc_Var	Gnd	F1_Euc_Hit*	P60
Myrtle 2	0.84	1.5	0.37	Den>3*	F1_Euc_Rg*	Non_Mu *	Gnd	Non_P99

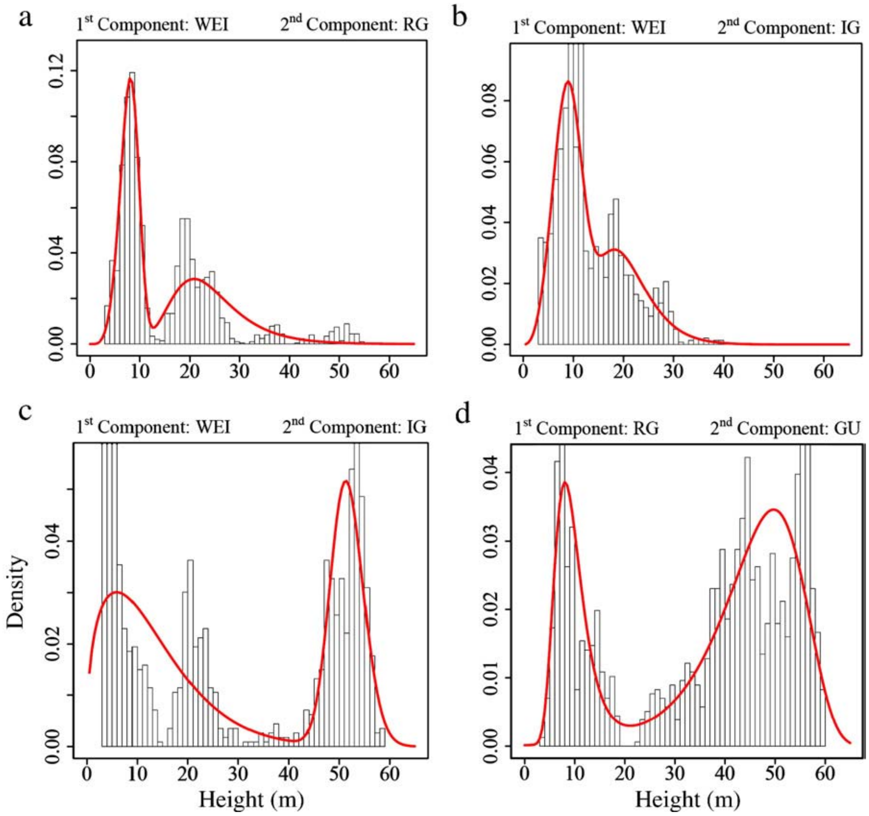

Fig. 5. Types of erroneous fits identified in the mixture models of the vegetation profile, where: (a) has old growth stags distorting the eucalyptus regrowth distribution, (b) has no eucalyptus trees but the mixture models assumes rainforest layer is eucalyptus layer, (c) has three vegetation layers that are poorly fitted with a bimodal distribution, and (d) has a rainforest layer that has been integrated into the overstorey density estimate.

non-parametric kernel smoothing methods through frequency histograms to count the local frequency maxima and use this value to represent the number of components in the mixture model. Identifying which particular vegetation layer each component of the mixture model represents is necessary to successfully identify circumstances where two distributions are more appropriate for representing one vegetation layer in order to combine them and improve the density estimates. We suggest a modelling procedure that uses conditions and constraints for each component of the distribution curve to identify particular vegetation layers in the multimodal mixture model. An understanding of the general forest structure of a particular forest type may allow for mode values and vertical distances between the modes to be used in condition statements to interpret whether two components of a distribution curve represent one or two vegetation layers.

For example, Fig. 6 shows how a Black Spur plot may be more accurately represented with a four modal curve to separate the rainforest middle storey from the eucalyptus layer which is more effectively represented by the third and fourth component of the mixture model. By establishing conditions based on the mode values or vertical distances that separate them, the density estimates of distributions represented by particular modes may be combined to represent a particular vegetation layer. Such an approach would also recognise when a vegetation layer is missing from the vegetation profile. For example, if it is expected for a eucalyptus layer of a given age to have a canopy mode no smaller than a particular height value and when there are no distributions of this characteristic, then the overstorey layer may be redefined as a rainforest layer. Using such techniques, the predictability of the regression models generated in this study may be improved as erroneous interpretations of mixture models such as those illustrated in Fig. 5 would be corrected. Further research is necessary to determine whether such techniques may identify suppressed trees or overlapping vegetation layers to further improve the predictability of basal area in targeted vegetation layers.

Fitting mixture models to diameter distribution data has been well developed (Liu et al., 2002; Zhang et al., 2001; Zhang & Liu, 2006), and multimodal mixture modelling of LiDAR data may also prove useful in forestry applications that require estimates of diameter frequency distributions. Mixture model parameters characterising the irregular diameter distributions of mixed-species or uneven-aged forest stands may be regressed against multimodal mixture models of LiDAR vegetation height distributions to identify suppressed trees, and different aged cohorts. Further research needs to be undertaken to determine how well such methods may improve timber inventory modelling.

We speculate that the results in the present study may be further improved by improving the optimisation of the mixture model selection procedure. To reduce computational time during the extrapolation procedure, the present methodology was designed to evaluate only four of the most successful mixture models at each location of a catchment. At the expense of increased computational time, an alternative approach may involve identifying the most successful mixture model for each plot out of all converged mixture models for the given plot. To extrapolate such results all successful plot-specific distributions would then need to be tested as candidates and evaluated at each grid over the catchment area. If the sample plots are highly variable with a large range of optimal multimodal distributions then the extrapolation procedure over a large area may prove computationally intensive. The results on the other hand would be more accurate and furthermore, the spatial distribution of the large variety of mixture models may themselves provide useful insight for predicting forest characteristics.

An important advantage in identifying many candidate distributions (eleven in this case) for fitting a range of vegetation profiles is that particular distributions or mixture models may represent specific

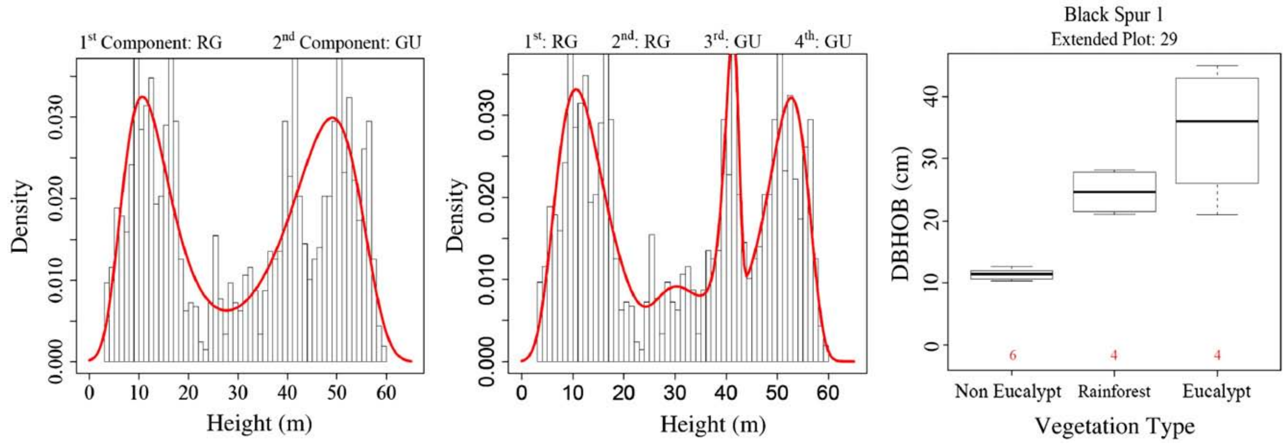

Fig. 6. An example plot that may be more accurately represented with a four modal curve to capture the density estimate of the eucalyptus vegetation profile.

ecosystems or tree species. For example, overlapping vegetation layers due to a particular species in the middle storey may be recognised with a particular skewed distribution function in a particular component of the mixture model. An empirical analysis relating ecosystem types to particular mixture models was beyond the scope of this study but may prove useful for a very broad range of forest management applications.

The present study has developed LiDAR indices that are highly applicable to forest hydrological studies as these indices directly relate to the vegetation characteristics that influence forest water use. For example, generating mixture model statistics over sub-plots to produce horizontally and vertically stratified vegetation structural attributed provide a measure of the canopies aerodynamic properties. The preserved canopy profile characteristics captured in the mixture models such as canopy density, depth, and closure are strongly related to the transpiring leaf area. Research needs to be undertaken to determine whether spatially represented LiDAR indices relating to forest hydrological systems may be used to explain catchment variations in stream flow.

5. Conclusion

Mixture models provide an elegant and robust method for stratifying the vegetation profile into distinct vegetation layers whilst preserving vegetation specific characteristics of the canopy profile. Unlike most previously proposed LiDAR indices in literature that categorise the vertical profile of forest structure into a finite assemblage of statistics (Hall et al., 2005; Lefsky et al., 1999; Lefsky et al., 2005; Zimble et al., 2003), mixture models can capture a more complete representation of the continuous point density estimate. Very few studies have explored theoretical distribution functions to represent the vertical profile of vegetation structure in LiDAR data. The most notable examples by Coops et al. (2007), Dean et al. (2009), and Maltamo et al. (2004) all used unimodal Weibull distributions which are restricted in their application as the vertical and horizontal forest structure around the world is so variable.

The methodology presented in this paper is working towards a generalised approach in representing the vertical forest structure with theoretical distribution functions for a very broad range of forest types. Using the GAMLSS package available with the open source software R (R Development Core Team, 2009), the form of the distributions available is very general and there are no restrictions on the number of modes available in the mixture models. The present study systematically evaluated 44 distribution functions to produce bimodal curves that estimate canopy density of Mountain Ash forests with a range of age and density classes. The results identified eleven likely candidate distributions that were successful at representing the overstorey of Mountain Ash forests and may prove useful for other forest types.

Mixture modelling is a promising method that may summarise complex canopy attributes captured by LiDAR data into a short list of parameters for empirical analyses against field measured stand characteristics. The present study has demonstrated that parameters extracted from bimodal curves are successful at predicting eucalyptus basal area and stand volume as well as basal area of non-eucalypt understorey. Using a ridge regression procedure that accounts for sources of uncertainty ignored in standard regression techniques, the study found that observed versus predicted values of eucalyptus basal area and stand volume was highly correlated with r^2 ranging from 0.61 to 0.89 and 0.67 to 0.88 respectively. Non-eucalyptus basal area r^2 ranged from 0.5 to 0.91. A critical evaluation of the mixture model density estimates for all the study plots identified circumstances under which a more complex multimodal distribution curves may be used to improve the predictive capacity of the mixture modelling methodology.

Acknowledgements

The authors would like to thank the following funding bodies that provided assistance: Melbourne Water under the Wildfire and Water Security project, the Cooperative Research Centre for Forestry project 4.1, and the Victorian Department of Sustainability and Environment. We would like to thank three anonymous reviewers for their valuable suggestions that have improved the final manuscript. We would also like to thank Jack Snodgress for his assistance with field work.

References

Akaike, H. (1974). New look at statistical-model identification. *IEEE Transactions on Automatic Control*, *AC19*, 716—723.

Axelsson, P. (1999). Processing of laser scanner data — Algorithms and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*, 138—147.

Barilotti, A., Sepic, F., & Abramo, E. (2008). Automatic detection of dominated vegetation under canopy using Airborne Laser Scanning data. *SilviLaser. Edinburgh, UK*.

Coops, N., Hilker, T., Wulder, M., St-Onge, B., Newnham, G., Siggins, A., & Trofymow, T. (2007). Estimating canopy structure of Douglas-fir forest stands from discrete-return LiDAR. *Trees*, *21*, 295—310.

Dean, T. J., Cao, Q. V., Roberts, S. D., & Evans, D. L. (2009). Measuring heights to crown base and crown median with LiDAR in a mature, even-aged loblolly pine stand. *Forest Ecology and Management*, *257*, 126—133.

Donoghue, D. N. M., Watt, P. J., Cox, N. J., & Wilson, J. (2007). Remote sensing of species mixtures in conifer plantations using LiDAR height and intensity data. *Remote Sensing of Environment*, *110*, 509—522.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation. the 0.632+ bootstrap method. *Journal of the American Statistical Association*, *2*, 548—560.

Golub, G., Heath, M., & Wahba, G. (1979). Generalised cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*.

Hall, S. A., Burke, I. C., Box, D. O., Kaufmann, M. R., & Stoker, J. M. (2005). Estimating stand structure using discrete-return LiDAR: An example from low density, fire prone ponderosa pine forests. *Forest Ecology and Management*, *208*, 189—209.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.*

Hoerl, A., & Kennard, R. (1970). Ridge regression — Applications to non-orthogonal problems. *Technometrics*, *12*, 69—82.

Holmgren, J., & Persson, A. (2004). Identifying species of individual trees using airborne laser scanner. *Remote Sensing and Environment*, *90*, 415—423.

Hutchinson, M. F. (1989). A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology*, *106*, 211—232.

Hutchinson, M. F. (2005). *ANUDEM Version 5.1.* Canberra: The Australian National University.

Kraus, K., & Pfeifer, N. (1999). Determination of terrain models in wooded areas with airborne scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*, 193—203.

Kuczera, G. (1987). Prediction of water yield reductions following a bush-fire in ash-mixed species eucalypt forest. *Journal of Hydrology*, *94*, 215—236.

Lefsky, M. A., Cohen, W. B., Acker, S. A., Parker, G. G., Spies, T. A., & Harding, D. J. (1999). LiDAR remote sensing of the canopy structure and biophysical properties of Douglas-Fir Western Hemlock forests. *Remote Sensing and Environment*, *70*, 339—361.

Lefsky, M. A., Hudak, A. T., Cohen, W. B., & Acker, S. A. (2005). Geographic variability in LiDAR predictions of forest stand structure in the Pacific Northwest. *Remote Sensing and Environment*, *95*, 532—548.

Liu, C., Zhang, L., Davis, C. J., Solomon, D. S., & Gove, J. H. (2002). A finite mixture model for characterizing the diameter distributions of mixed-species forest stands. *Forest Science*, *48*, 653—661.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Inference Theory*, *28*, 129—137.

Magnussen, S., & Boudewyn, P. (1998). Derivations of stand heights from airborne laser scanner data with canopy-based quantiles. *Canadian Journal of Forest Research*, *28*, 1016—1030.

Maltamo, M., Eerikainen, K., Pitkanen, J., Hyyppa, J., & Vehmas, M. (2004). Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions. *Remote Sensing of Environment*, *90*, 319—330.

Maltamo, M., Packalenm, P., Yu, X., Eerikainen, K., Hyyppa, J., & Pitkanen, J. (2005). Identifying and quantifying structural characteristics of heterogeneous boreal forests using laser scanner data. *Forest Ecology and Management*, *216*, 41—50.

Monteith, J. L. (1965). Evaporation and environment. *Proceedings of the 19th Symposium, Society of Experimental Biology*. London: Cambridge University Press.

Naesset, E. (1997a). Determination of mean tree heights of forest stands using airborne laser scanner data. *Remote Sensing and Environment*, *29*, 547—553.

Naesset, E. (1997b). Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing and Environment*, *61*, 547—553.

Naesset, E., Gobakken, T., Holmgren, J., Hyyppa, J., Hyyppa, H., Maltamo, M., Nilsson, M., Olsson, H., Persson, A., & Soderman, U. (2004). Laser scanning of forest resources: The Nordic experience. *Scandinavian Journal of Forest Research*, *19*, 482—499.

R Development Core Team (2009). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raison, R. J., Brown, A., & Flinn, D. (2001). *Criteria and indicators for sustainable forest management.* CABI Publishing.

Riano, D., Meier, E., Allgower, B., Chuvieco, E., & Ustin, S. (2003). Modeling airborne laser scanning data for the spatial generation of critical forest parameters in fire behavior modeling. *Remote Sensing and Environment*, *86*, 177—186.

Rigby, R.A., & Stasinopoulos, D.M. (2008). A flexible regression approach using GAMLSS in R. In. London Metropolitan University, London: STORM Research Centre

Stasinopoulos, D.M., Rigby, R.A., & Akantziliotou, C. (2008). Introduction on how to use the GAMLSS package in R. In, *London Metropolitan University, London*: STORM Research Centre.

Vertessy, R., Watson, F., & O'Sullivan, K. (2001). Factors determining relations between stand age and catchment water balance in mountain ash forests. *Forest Ecology and Management*, *143*, 13—26.

Zhang, L., Gove, J. H., Liu, C., & Leak, W. B. (2001). A finite mixture of two Weibull distributions for modeling the diameter distributions of rotated-sigmoid, uneven-aged stands. *Canadian Journal of Forest Research*, *31*, 1654—1659.

Zhang, L., & Liu, C. (2006). Fitting irregular diameter distributions of forest stands by Weibull, modified Weibull, and mixture Weibull models. *Journal of Forest Research*, *11*, 369—372.

Zimble, D. A., Evans, D. L., Carlson, G. C., Parker, R. C., Grado, S. C., & P.D., G. (2003). Characterising vertical forest structure using small-footprint airborne LiDAR. *Remote Sensing and Environment*, *87*, 171—182.