Remote Sensing of Environment 115 (2011) 3329–3342

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Classifying individual tree genera using stepwise cluster analysis based on height and intensity metrics derived from airborne laser scanner data

Sooyoung Kim *, Thomas Hinckley, David Briggs

School of Forest Resources, University of Washington, P.O. Box 352100, Seattle, WA 98195-2100, USA

ARTICLE INFO

Article history:
Received 23 June 2010
Received in revised form 22 July 2011
Accepted 23 July 2011
Available online 4 September 2011

Keywords:
Stepwise cluster analysis
LIDAR data
Leaf-on and leaf-off conditions
Tree species classification

ABSTRACT

This paper evaluates the ability of small footprint, multiple return and pulsed airborne scanner data to classify tree genera hierarchically using stepwise cluster analysis. Leaf-on and leaf-off airborne scanner datasets obtained in the Washington Park Arboretum, Seattle, Washington, USA were used for tree genera classification. Parameters derived from structure and intensity data from the leaf-on and leaf-off laser scanning datasets were compared to ground truth data. Relative height percentiles and simple crown shapes using the ratio of a crown length to width were computed for the structure variables. Selected structure variables from the leaf-on dataset had higher classification rate (74.9%) than those from the leaf-off dataset (50.2%) for distinguishing deciduous from coniferous genera using linear discriminant functions.

Unsupervised stepwise cluster analysis was conducted to find groupings of similar genera at consecutive steps using *k*-medoid algorithm. The three stepwise cluster analyses using different seasonal laser scanning datasets resulted in different outcomes, which imply that genera might be grouped differently depending on the timing of the data collection. When combining leaf-on and leaf-off LIDAR datasets, the cluster analysis could separate the deciduous genera from evergreen coniferous genera and could make further separations between evergreen coniferous genera. When using the leaf-on LIDAR dataset only, the cluster analysis did not separate deciduous from evergreen genera. The overall results indicate the importance of the timing of laser scanner data acquisition for tree genera separation and suggest that the potential of combining two LIDAR datasets for improved classification.

1. Introduction

Airborne laser scanning and profiling are a rapidly growing technology for use in forest inventory (Boudreau et al., 2008; Næsset, 2002; Nelson et al., 2009, 2003), forest monitoring (Hopkinson et al., 2008; Næsset & Gobakken, 2005; Solberg et al., 2006) and ecological applications (Bradbury et al., 2005; Graf et al., 2009; Nelson et al., 2005; Tickle et al., 2006). Recently, forest stand types or tree species classification have been studied using laser scanner datasets (Brantberg, 2007; Brandtberg et al., 2003; Brennan & Webster, 2006; Donoghue et al., 2007; Holmgren & Persson, 2004; Kim et al., 2009a, 2009b; Moffiet et al., 2005; Ørka et al., 2009). Historically, LIDAR technology has been used to capture and detect the 3-dimensional structure of objects. Given the capability of LIDAR sensors to provide 3-D information of vegetation structure, this technology has shown great potential to estimate several biophysical and structural properties over a wide range of forest types (Lefsky et al., 1999; Morsdorf et al., 2006; Næsset, 2002; Riaño et al., 2004).

Most commercial LIDAR systems used for topographic mapping use lasers that emit energy in the near infrared range of the electromagnetic spectrum (often 1064 nm). Green vegetation reflects this wavelength well (Swain and Davis, 1978) and there are species differences in both the visible and infrared spectra (Roberts et al., 2004); therefore, LIDAR intensity data should contain information relating to forest type and condition. Because spectral reflectance changes depending on the time of a year for deciduous species (Gates, 1980), acquiring LIDAR datasets in leaf-on and leaf-off conditions could provide additional information useful for species differentiation. Most laser scanning data include an intensity value which is a relative measure of the signal strength associated with each return, a measure of the amount of energy reflected from a target. In the past, LIDAR intensity data have not been used as extensively as the three-dimensional structure data represented by laser returns. Recently, several studies have classified forest stand types or tree species using height and intensity information from airborne laser scanning (Brantberg, 2007; Brandtberg et al., 2003; Brennan & Webster, 2006; Holmgren & Persson, 2004; Kim et al., 2009b; Morsdorf et al., 2010; Ørka et al., 2009). In addition, LIDAR is increasingly being used to estimate leaf area and its distribution (Korhonen et al., 2011; Richardson et al., 2009).

Growing interest in the LIDAR intensity data leads to the study of these data as affected by scan angles, flight altitude or laser path length (Coren & Sterzai, 2006; Donoghue et al., 2007; Hasegawa, 2006). When a narrow scan angle was used, intensity data varied little; however, as scan angle increased, variation increased. Forest stand types or forest

* Corresponding author at: 6921 94th Ave. SE., Mercer Island, WA 98040, USA. Tel.: +1 2067553367.
E-mail address: kisoyo@u.washington.edu (S. Kim).

0034-4257/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2011.07.016

species classification using LIDAR intensity data may be affected by variations in laser path length due to differences in topography or canopy height. However, Höfle and Pfeifer (2007) found little variation in intensity data over a range of different heights and dates of acquisition when data were acquired from homogeneous reflecting areas. Kim et al. (2009a) used raw intensity data without additional radiometric calibration because the topographic range of study elevations (15 to 55 msl) and slopes (<30%) were small. However, since they compared intensity values from two different LIDAR datasets, which had different range values, they normalized these datasets by randomly extracting multiple man-made samples whose conditions were considered to be constant between the two LIDAR datasets. They then compared intensity values for each sample using box-plots that showed that the variability across the samples was constant for each LIDAR dataset even though the ranges of intensity values were different. Finally, they computed the ratio of the leaf-off to leaf-on medians for each sample and used the average ratio of 16.4 to scale the leaf-on intensity medians. Their data suggested that some species and species groups might be differentiated using LIDAR intensity data. Here we explore this further using the original datasets from Kim et al. (2009a). Understanding that some tree species have similar characteristics and those similarities would be discernable from LIDAR structural and intensity data, it follows that some type of hierarchical classification scheme might be used to identify species.

In this study, we evaluate the use of cluster analysis, one of the unsupervised classification methods, to classify individual trees using the *k*-medoid algorithm. Instead of using one-step cluster analysis, we use a stepwise cluster analysis, based on statistical criteria, to find hierarchical relationships between species — if sufficiently uniform and powerful, we should be able to separate or group species based upon similarities or differences derived from LIDAR data. We also investigate the potential of the laser scanning system to cluster tree genera or morphological-types using structure and intensity information derived from LIDAR datasets. Our specific objectives are: 1) to compare structure variables derived from leaf-on and leaf-off LIDAR datasets for 15 genera, 2) to conduct cluster analysis for the genera using structure and intensity variables from different LIDAR datasets and 3) to investigate what factors would affect the results of the clustering analysis.

2. Study area and datasets

2.1. Study area

The study area is the Washington Park Arboretum located in Seattle, Washington (47° 37.723′N 122° 17.732′W). The area covers 93 ha and the topographic range is 15 to 55 m above sea level with less than 30% slope for the majority of the site. Because our study aims to investigate various tree genera at the individual tree levels, this was the most suitable study site as it contained a large number of different woody plant species and genera, most of which had attained mature stature and often had crowns that did not significantly overlap, and thus could be easily detected and measured.

2.2. Laser data

In this study, we used the same intensity datasets computed from the leaf-on and leaf-off LIDAR data described in Kim, McGaughey, et al. (2009a). The following summarizes the protocol for data acquisition and the description of the laser scanner used. Leaf-on data were acquired on 30 August, 2004 using the Optech ALTM 30/70 laser scanner system. Average flying altitude was 1200 m above the ground level (a.g.l) configured to acquire data using a narrow scan angle of <11° either side of NADIR and with a point density up to 5/m^2. Scan pulse frequency was 71 kHz and a single flight line was used. Leaf-off data were acquired on 15 March 2005 using an Optech ALTM 3100. Average flying altitude was 900 m a.g.l. configured to acquire data using a narrow scan angle of <10° either side of NADIR and with a point density up to 10/m^2. Scan pulse frequency was 100 kHz and flight line was 50%. Both systems used a 1064 nm laser and beam divergence of 0.31 mrad with footprint size of 0.372 m with leaf-on data and 0.279 with leaf-off data. The leaf-off dataset did not capture all trees in full leaf-off conditions due to widely varying phenology across the diverse range of species within the arboretum and an unusually early bud break in 2005.

It was necessary to normalize the two LIDAR datasets because the features that were assumed spectrally invariant had different ranges of intensity values (Kim et al., 2009a, 2009b). Before normalizing two LIDAR datasets, the mean intensity values per genera were between 1.2 and 4.5 for the leaf-on data while the mean intensity values per genera were between 7.0 and 55.0 for the leaf-off data. After normalizing two LIDAR datasets by multiplying a scaling factor of 16.4 to the leaf-on LIDAR datasets (Kim et al., 2009a), the mean intensity values per genera for the leaf-on data became between 26.0 and 53.0.

A digital terrain model (DTM) was created with 1 by 1-m resolution using FUSION/LDV software (McGaughey et al., 2004; McGaughey & Carson, 2003). The method for creating the LIDAR-based DTM is described by Andersen et al. (2006).

2.3. Selection of tree genera and species

Kim et al. (2009a) used 15 tree genera with a total of 223 individual trees but did not investigate individual tree species within a genus. For this study, we separated the 223 individual trees by species and their respective genus. Table 1 describes the genera used in this study and their characteristic leaf structures and traits. Table 2 lists the genera, individual species, classification as to deciduous or evergreen, number of trees, and notes as to whether or not deciduous individuals were past bud break and flowering or developing leaves when the leaf-off data were acquired. Flowering or partial leaf formation could influence classification of individuals that were in various states of leaf or flower emergence.

Table 1
Tree species used in this study categorized by leaf structures.

Angiosperm genera (all were deciduous)			Gymnosperm genera		
Leaf structure	Stem features	Genera	Leaf structure	Leaf traits	Genera
Opposite simple leaves	No thorns	*Acer*	Clustered needles	Evergreen	*Pinus*
Alternate compound leaves	No thorns	*Sorbus*	Clustered needles	Deciduous	*Larix*
Alternate simple leaves	Thorns	*Prunus* *Malus*	Single needles on woody pegs	Evergreen	*Picea*
Spirally arranged, lobed leaves	No thorns	*Quercus*	Flat, single needles	Evergreen	*Pseudotsuga* *Tsuga* *Sequoia*
Simple leaves, toothed or pointed	No thorns	*Betula*	Scale-like leaves	Evergreen	*Thuja*
Alternate simple single or doubly serrated leaves	No thorns	*Ulmus*			
Simple, smooth edged leaves	No thorns	*Magnolia*			

Table 2
Lists of the genera, individual species, classification as to deciduous or evergreen, number of trees, and conditions as to whether or not deciduous individuals were past bud break and flowering or developing leaves when the leaf-off data were acquired.

Genus	Species	Deciduousness	Conditions	Number of trees	
Broadleaved/Angiosperms				Species	Genus
Betula	*B. alleghaniensis*	Deciduous	Leaf-off	5	20
	B. nigra		Leaf-off	5	
	B. platyphylla		Leaf-off	5	
	B. utilis		Leaf-off	5	
Acer	*A. macrophyllum*	Deciduous	Leaf-off	11	11
Ulmus	*U. americana*	Deciduous	Leaf-off	10	10
Magnolia	*M. grandiflora*[a]	Deciduous	Leaf-on/flowering	4	19
	M. dawsoniana		Leaf-off/flowering	4	
	M. denudata		Leaf-off /flowering	4	
	M. fraseri		Leaf-on/flowering	4	
	M. acuminate		Leaf-on	3	
Malus	*M. coronaria*[b]	Deciduous	Leaf-off	2	10
	M. domestica		Leaf-on	3	
	M. florentina		Leaf-on	2	
	M. fusca		Leaf-on	3	
Prunus	*P. serotina*	Deciduous	Leaf-off	2	11
	P. domestica[b]		Leaf-on	3	
	P. incise		Leaf-off/flowering	3	
	P. sargentii		Leaf-off/flowering	3	
Quercus	*Q. garryana*	Deciduous	Leaf-off	1	19
	Q. arizonica		Leaf-off	5	
	Q. bicolor		Leaf-off	5	
	Q. alba		Leaf-off	4	
	Q. rubra		Leaf-off	4	
Sorbus	*S. americana*	Deciduous	Leaf-off	3	11
	S. anglica		Leaf-off	3	
	S. commixta		Leaf-off	3	
	S. hybrid		Leaf-off	2	
Coniferous/Gymnosperms					
Thuja	*T. plicata*	Evergreen	Leaf-on	19	19
Pseudotsuga	*P. menziesii*	Evergreen	Leaf-on	12	12
Larix	*L. deciduas*	Deciduous	Leaf-off	5	21
	L. kaempferi[b]		Leaf-on	4	
	L. laricina[b]		Leaf-on	4	
	L. occidentalis[b]		Leaf-on	3	
	L. gmelinii		Leaf-off	5	
Pinus	*P. monticola*	Evergreen	Leaf-on	3	21
	P. armandii		Leaf-on	5	
	P. densiflora		Leaf-on	5	
	P. strobus		Leaf-on	4	
	P. ponderosa		Leaf-on	4	
Sequoia	*S. sempervirens*	Evergreen	Leaf-on	10	10
Picea	*P. abies*	Evergreen	Leaf-on	4	15
	P. engelmannii		Leaf-on	4	
	P. koyamai		Leaf-on	3	
	P. montigena		Leaf-on	4	
Tsuga	*T. heterophylla*	Evergreen	Leaf-on	14	14

[a] Evergreen broadleaf.
[b] Bud-break already occurred.

Sample trees comprise 15 genera, 8 broadleaved and 7 coniferous. Six genera were only represented by a single species: *Acer macrophyllum* (big leaf maple), *Ulmus americana* (American elm), *Pseudotsuga menziesii* (Douglas-fir), *Sequoia sempervirens* (coastal redwood), *Thuja plicata* (western red cedar) and *Tsuga heterophylla* (western hemlock) and nine genera were composed of multiple species.

2.4. Field measurement

Field work was designed to select and georeference individual trees that could be used to ground truth data for the analysis. This field work was carried out between April and July 2005. We located three reference points that roughly form an equilateral triangle with 30–100 m sides using a Trimble Pro XR/XRS GPS system. Individual tree locations were recorded from at least two of the triangle points to confirm accurate tree locations. We used a laser rangefinder and compass to shoot foresights and backsights (horizontal and vertical distances and azimuth) along each side of the triangle. For the most part, isolated individual trees in open areas were selected to simplify their identification and measurement in the LIDAR point cloud. Total heights, crown base heights and crown diameters which were computed as the mean of the N–S and E–W directions were measured on 345 trees in the field. We used the custom trilateration program with the GPS locations and the distances and azimuths to the triangle points to obtain final point locations and the local magnetic declination. We processed the distance and azimuth shots to other points of interest using the corrected triangle locations and local magnetic declination. Next, the locations of individual trees were overlaid over the orthophoto of the Arboretum (see Fig. 1). Finally, we identified 223 usable individual trees by eliminating trees with severely overlapped crowns or trees that could not be clearly identified. Even though tree stems were not correctly positioned

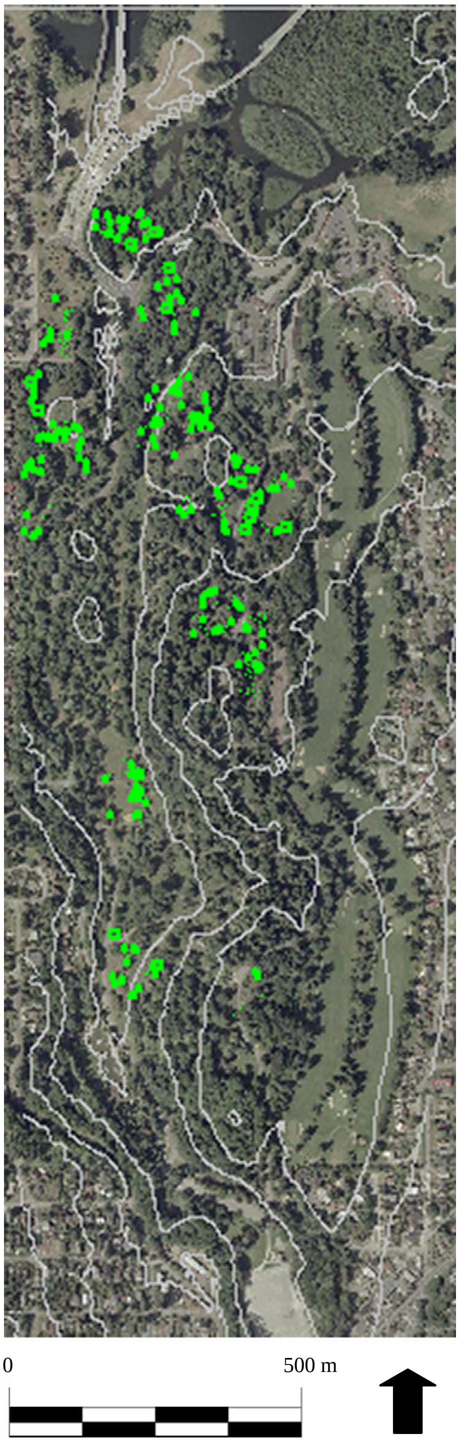

Fig. 1. The locations of individual tree crowns are marked as green dots over the orthophoto of the Washington Park Arboretum, Seattle, WA.

relative to the point cloud data, field-measured tree height and crown diameter helped the detection of trees when the x–y positioning errors are within 2 m.

2.5. Study variables

2.5.1. Intensity variables

Intensity variables derived from leaf-on and leaf-off laser scanning datasets for isolated individual tree crowns from Kim et al. (2009a) were used in this study. The following summarizes how we identified individual tree crowns in this study.

First, individual trees were detected with the aid of field measured stem locations using x–y position of the stem, tree height and crown diameter using FUSION/LDV software which displayed the LIDAR return data near the approximate tree location. Non-ground laser points were obtained by omitting laser returns representing the ground surface and those less than 1 m above the ground surface from the two data subsets.

Next, the laser points within the individual tree crowns were isolated within a cylinder defined by the field-measured location and crown diameter for each tree. Crown base height was calculated using 0.5 m height layers (Holmgren & Persson, 2004). Each layer that contained less than 1% of the total number of non-ground laser points within an individual tree was set to zero and the others to one. The crown base height was then set as the distance from the ground to the lowest laser data point above the highest 0-layer found. Using laser points within each crown, variables were computed to analyze intensity data for each tree. All variables were derived using laser returns that were located above the crown base height.

Mean intensity values were computed using returns representing the entire crown, upper crown and crown surface within each tree crown using isolated laser returns. Variables 1–9 defined in Table 3 were computed from both the leaf-on and leaf-off laser scanning data.

Table 3
Variable definition and nomenclature. Each variable was measured for both the leaf-on and leaf off datasets so in reality there are 32 variables in total.

Variable	Symbol
(1) Mean intensity values for the entire crown using all returns	entire_all
(2) Mean intensity values for the entire crown using first returns	entire_1
(3) Mean intensity values for the upper crown using all returns	upper_all
(4) Mean intensity values for the upper crown using first returns	upper_1
(5) Mean intensity values for the crown surface using all returns	surface_all
(6) Mean intensity values for the crown surface using first returns	surface_1
(7) Coefficient of variation of all return intensity for the entire crown	cv_all
(8) Coefficient of variation of first return intensity for the entire crown	cv_1
(9) Proportion of first returns	prop_1
(10) Relative 10th height percentile	rel10th
(11) Relative 50th height percentile	rel50th
(12) Relative 90th height percentile	rel90th
(13) Relative standard deviation of height	relstdev
(14) Length to width ratio at upper 10% of a crown length	ratio_10%
(15) Length to width ratio at upper 25% of a crown length	ratio_25%
(16) Length to width ratio at upper 33% of a crown length	ratio_33%

2.5.2. Structure variables

Relative height percentiles were defined as the height percentile of laser returns divided by the maximum laser return height within individual tree crowns. Height percentiles were used as promising variables to estimate forest parameters (Garcia et al., 2010; Holmgren & Persson, 2004) and relative 90th height percentiles have been shown to be related to canopy shapes (Holmgren & Persson, 2004). Since LIDAR beams exhibit differential penetration into forest canopies, information on canopy structure can be inferred. Hence, several distribution measurements related to the canopy structure were generated. Relative 10th, 50th, and 90th height percentiles and relative standard deviation of height were computed for each tree crown in both leaf-on and leaf-off datasets using Interactive Data Language (IDL) from Research Systems, Inc.

Methodology to isolate and separate trees consists of ignoring the laser returns positioned at the edge (Kim et al., 2009a). Hence, variables based on x, y, and z coordinates of laser returns at the upper crown would be more reliable than those at the lower crown. Because most of the upper crowns were not overlapped in this dataset, crown parameters taken at the 90th percentile were likely robust. Trees with conical shapes are likely to have longer crown lengths versus crown widths as compared to trees with round shapes at the upper portion of the crown. Three different portions of an upper crown were used:

upper 10%, upper 25% and upper 33% of a crown length. For each portion of an upper crown, a crown length and a crown width were computed. First, radius was computed as the maximum horizontal distance using x and y coordinates of laser returns from the tree center to the farthest laser returns for each 45 degree sector. Next, crown width was computed by averaging the computed sector radii. A crown length was computed by measuring the distance from the z-value of the lowest positioned laser return and the z-value of the highest positioned laser return for each crown. Finally, the length to width ratio at each portion of an upper crown for each individual tree was computed. Variables 10–16 defined in Table 3 were derived from both the leaf-on and leaf-off datasets using isolated laser returns within each crown.

3. Computation and analysis

Mean values for each structure variable for each genus were compared for leaf-on and leaf-off datasets, respectively. To compare significant differences between genera that were deciduous and those that were evergreen, Student's two sample *t*-test was conducted for each variable in both LIDAR datasets. However, we anticipated that our ability to use structural and intensity data to distinguish between trees of different species, genera, or leaf-life traits would be complex and require several different statistical approaches to thoroughly evaluate. We chose the following analyses, Principal Component (PCA), Linear discriminant (LDA), and Cluster, to further explore this. The methodological details of each of these analyses are provided below.

3.1. Principal component analysis

Because we dealt with sixteen variables (Table 3), each with a leaf-on and leaf-off variant, the number of variables needs to be reduced in order to simplify subsequent analysis while retaining as much information as possible (Everitt & Dunn, 2001). For this purpose, principal component analysis (PCA) was employed using the R package (Kaufman & Rousseeuw, 1990). The size of the subset of original variables to be retained was determined by the number of components (Jolliffe, 2002). Jolliffe (2002) suggested retaining only components extracted from a correlation matrix whose associated Eigen values were greater than 0.7. The variable in the component with the greatest absolute coefficient value was chosen, provided that the variable was not already chosen to represent a different component.

3.2. Linear discriminant analysis

A simple tree classification test for broadleaved and coniferous genera was performed on selected variables using a linear discriminant function. In this study, the linear discriminant function was derived from just n-1 members of the sample and then used to classify the member not included. The process was carried out n times, leaving out each sample member in turn (Everitt & Dunn, 2001). The percentage of the correctly classified rate was then calculated; (1-misclassification rate) and is called the classification rate.

As a result of conducting a principal component analysis using variables 10–16 in Table 3 for both the leaf-on and leaf-off datasets, two crown length to width ratio variables were chosen based upon the criterion in Section 3.1; variable #15 (ratio_25%), and variable #16 (ratio_33%). When we conducted linear discriminant functions for the combined leaf-on and leaf-off datasets, we used the leaf-on and leaf-off versions of these variables, i.e. four variables in total, to test our ability to classify broadleaved and coniferous genera based upon these structure metrics.

3.3. Cluster analysis

For many clustering problems, one is usually interested in characterization of the clusters by means of typical or representative objects. In the method used in the program PAM (*P*artitioning *A*round *M*edoids) (Kaufman & Rousseeuw, 1990), the representative object of a cluster is its medoid, defined as that object of the cluster for which the average dissimilarity (typically Manhattan distance or the distance between two points measured along axes at right angles) between all the objects of the cluster is minimal. Because the objective is to find *k* such objects, the method is called the *k*-medoid. After finding a set of *k* representative objects, the *k* clusters are constructed by assigning each object of the dataset to the nearest representative object (Kaufman & Rousseeuw, 1990).

One of the simplest unsupervised learning algorithms to solve the well known clustering problem is *k*-means (MacQueen, 1967) which defines *k* centroids, one for each cluster by computing Euclidean distances. The advantages of using the *k*-medoid versus the k-means method are that *k*-medoid (1) minimizes the sum of dissimilarities instead of the sum of squared Euclidean distances and (2) is more robust with respect to outliers. The program PAM operates using the dissimilarity matrix of the given dataset. When it is presented with an ${}_nx_p$ data matrix where n indicates the number of samples and p indicates the number of variables, this program first computes a dissimilarity matrix. Next, it computes k representative objects or k-*medoids*, which together determine a clustering. Finally, each object is then assigned to the cluster corresponding to the nearest medoid. That is, object i is put into cluster vi when medoid m_{vi} is nearer to that object than any other medoid m_w:

$$_d(i, m_{vi}) \le _d(i, m_w) \text{ for all } w = 1, \ldots, k \tag{1}$$

The k representative objects should minimize the sum of the dissimilarities of all objects to their nearest medoid:

$$\text{Objective function} = \ \min \sum_{i=1}^{n} d(i, m_{vi}) \tag{2}$$

The algorithm sequentially selects k centrally located objects to be used as initial medoids. If the objective function can be reduced by interchanging (swapping) a selected object with an unselected object, then the swap is carried out. This is continued until the objective function no longer decreases.

3.3.1. Validation of cluster analysis (Silhouettes)

Rousseeuw (1987) proposed a graphical display to judge the quality of the clustering obtained and to determine the number of clusters best representing the given datasets. Each cluster is represented by a *silhouette* which is based on the comparison of its tightness and separation. The average silhouette width provides an evaluation of clustering validity and may be used to select an 'appropriate' number of clusters. To construct silhouettes, we need the collection of all proximities between objects as follows. Take any object i in the data set, and denote by A the cluster to which it has been assigned. When cluster A contains other objects apart from i, compute a (i), which is an average dissimilarity of i from all other objects of A. Next, consider any cluster C which is different from cluster A, and compute d (i, C) which is an average dissimilarity of i to all objects of cluster C. After computing d (i, C) for all clusters $C \neq A$, select the smallest of the d (i, C) and denote it by

$$b(i) = \min_{C \neq A} d(i, C) \tag{3}$$

The cluster B for which this minimum is attained (that is, $d(i, B) = b(i)$) is called the neighbor of the object. We now define:

$$-1 \leq s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \leq 1 \tag{4}$$

Which can be written as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i), \\ 0, & \text{if } a(i) = b(i), \\ b(i)/a(i) - 1, & \text{if } b(i) > a(i), \end{cases} \tag{5}$$

It is clear that we have,

$$-1 \leq s(i) \leq 1 \tag{6}$$

When $s(i)$ is close to $+1$, the 'within' dissimilarity, $a(i)$, is much smaller than the smallest 'between' dissimilarity, $b(i)$. In this case, i is considered to be 'well-clustered'. When $s(i)$ is close to -1, then $a(i)$ is much larger than $b(i)$, which implies that i lies on average much closer to B than to A. In this case, this object, i, is considered to have been misclassified. The average silhouette width defined as the average of the $s(i)$ for all objects, i, belonging to that cluster can distinguish 'good clusters' with large silhouette width from 'weak clusters' with small silhouette width. Rousseeuw (1987) suggested that the appropriate k can be determined by selecting that value of k for which the overall average silhouette width for the entire plot, $\bar{s}(k)$, with $k = 2, \ldots, n$ where n denotes the number of objects (for further details see Rousseeuw, 1987). In our study, $k = 2, \ldots, m$, where m denotes the number of genera, since the objective is to cluster genera into groups.

3.4. Stepwise cluster analysis

As one of the clustering methods, hierarchical clustering techniques proceed using either a series of successive merges or a series of successive divisions. Agglomerative hierarchical clustering techniques produce partitions by a series of successive fusions of the individual objects. With fusion methods, when an agglomerative algorithm has placed an individual in one group, it cannot subsequently appear in another group. Since agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, the chosen division should be based on the purpose of getting the best fitting number of clusters (Everitt & Dunn, 2001). In this study, we were more interested in clustering trees by genera than in clustering individual trees. Hence, to seek hierarchy among tree genera, we developed a modified approach to the typical hierarchical clustering techniques which is detailed below.

3.4.1. Process of stepwise cluster analysis

We conducted stepwise cluster analyses for the three datasets: leaf-on, leaf-off and combined leaf-on and leaf-off. For each dataset, we conducted stepwise cluster analyses using selected variables derived from intensity and structure information. Stepwise cluster analysis requires the following three steps for each dataset:

- *Step 1*: Conduct principal component analysis to determine the number of components to be used and to select corresponding variables using the criteria outlined in Section 3.1.
- *Step 2*: Conduct a cluster analysis using PAM with the subset of variables from step 1 and determine the most appropriate number of clusters by means of maximal average silhouette width.
- *Step 3*: Determine if more than 80% of individual trees within each genus are assigned to a single cluster. If so, include the remaining individual trees of that genus in the same cluster. Otherwise, delete the genus from the next step of the cluster analysis.
- *Step 4*: If the maximal overall average silhouette exceeds 0.5, the threshold suggested by Kaufman and Rousseeuw (1990) for deciding a reasonable structure has been achieved and then return to step 1 with the step 3's clusters. Otherwise, stop and consider step 3's clusters as final.

The rationale for the 80% criterion in step 3 can be understood by considering that it is possible for some individual trees of the same genus to fall into different clusters because species within the same genus, especially those that are deciduous, may have differences in phenology such that variation in the timing of flowering and leafing out may lead them to be placed in different clusters more reflective of phenological versus evolutionary differences. Table 2 illustrates cases when flowering and foliage development were already underway in spite of the data being collected when there should have been no leaves or flowers. Consequently, a procedure was needed to determine if a particular genus can be considered as clustered or not. One approach would be to require that a certain minimum percentage of individual trees belonging to that genus be in one cluster, in order for that genus to be considered clustered. If the minimum required percentage is not achieved, the genus is considered as not clustered. After testing different percentages to construct good clusters, of the value of 80% was selected as the criterion for a genus to be considered clustered. If a genus meets this criterion and is clustered, the remaining trees of that genus are also included in that cluster. If a certain genus fails that criterion, that is, fewer than 80% of the individual trees of that genus were assigned to a single cluster, that genus was excluded from the next step. Table 4 presents a subjective interpretation of the Silhouette Coefficient (SC) as the maximal average silhouette width for the entire data set (Kaufman & Rousseeuw, 1990).

4. Results

4.1. Analysis of structure variables

4.1.1. Height percentiles of laser returns

Mean values of the four height variables, variables 10–13 defined in Table 3, for each genus are shown in Fig. 2 with (a) leaf-on and (b) leaf-off datasets. *Pinus* showed the highest values for the three relative height percentiles among coniferous genera in both leaf-on and leaf-off datasets while *Pseudotsuga* showed the lowest values.

The results of the two-sample t-tests for deciduous and coniferous genera for each variable are shown in Table 5. For leaf-off data, the two leaf-trait groupings of genera (deciduous versus evergreen) did not show significant differences for any of the height variables ($p > 0.05$) while for leaf-on data, these were significant differences for the three height percentiles ($p < 0.01$).

For the analysis of height metrics, all three relative height percentiles in the leaf-on dataset, variables 10–12 in Table 3, showed significant differences at a $p < 0.01$ level using Student's t-test while none of the height percentiles in leaf-off dataset showed significant differences between deciduous and coniferous genera. This result implies that the leaf-on dataset discriminates deciduous from evergreen genera better than the leaf-off dataset in terms of height metrics. This result was counter-intuitive and will be discussed later.

4.1.2. Length to width ratios within upper portions of a crown

The results for length to width ratios for the upper portions of a crown, variables 14–16 in Table 3, are shown in Fig. 3 with (a) leaf-on

Table 4
Subjective interpretation of the Silhouette Coefficient (SC), defined as the maximal average silhouette width for the entire dataset (Kaufman & Rousseeuw, 1990).

SC	Proposed interpretation
0.71–1.00	A strong structure has been found
0.51–0.70	A reasonable structure has been found
0.26–0.50	The structure is weak and could be artificial; try additional methods on this dataset
≤0.25	No substantial structure has been found

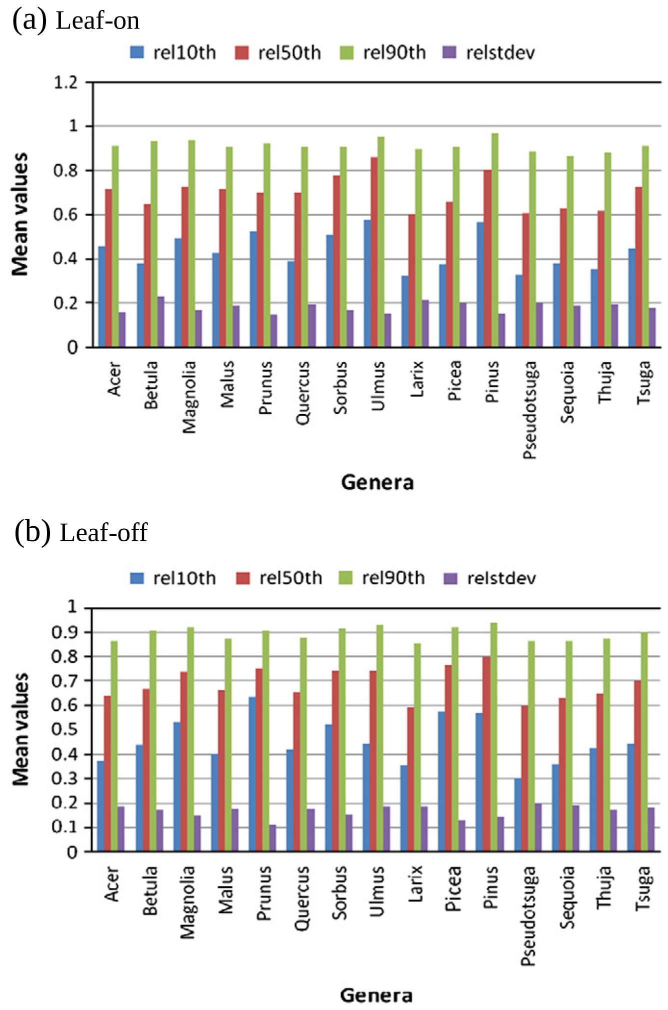

Fig. 2. Mean values for four height variables by genus in (a) leaf-on and (b) leaf-off datasets (rel10th, rel50th and rel90th—relative 10th, 50th, and 90th height percentile; relstdev—relative standard deviations of heights).

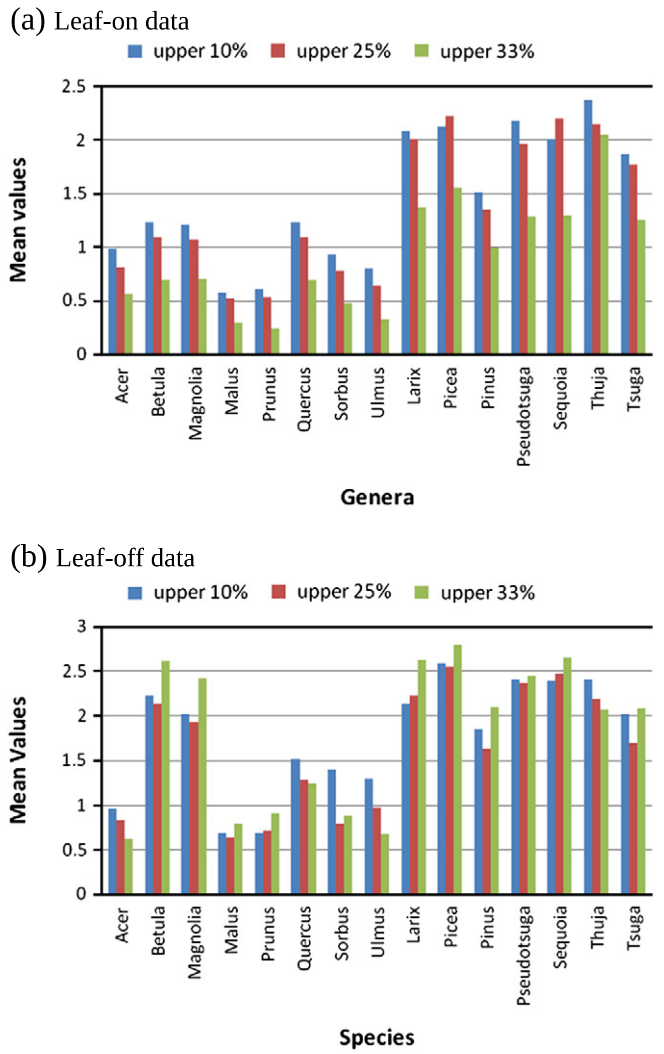

Fig. 3. Length to width ratio by genus at the upper 10%, 25% and 33% of crown length in (a) leaf-on and (b) leaf-off datasets. The greater the number or ratio, the more conical the tree.

and (b) leaf-off datasets. For the leaf-on dataset, all coniferous genera had greater ratios (i.e., their crowns were long relative to width or more conical) than all deciduous genera for all three height percentiles. However, in the leaf-off dataset, not all of the coniferous genera showed greater crown length/width ratios than all of the broadleaved genera. For example, *Betula* showed greater crown length/width ratios than *Pinus* and *Tsuga* for all three height percentiles. Also, *Magnolia* showed greater crown length/width ratios than *Pinus* for all three height percentiles. It is likely that *Magnolia* had comparatively high length to width ratio values because individual trees in the *Magnolia* genus had very different leaf phenologies in leaf-off datasets. For example, *M. grandiflora* is evergreen whereas other magnolia species are deciduous and the deciduous *Magnolia* species were in different stages of leaf and flower emergence ranging from almost complete flower emergence to no flowers or leaves. *Pinus* showed the lowest crown length/width ratios among coniferous genera in both leaf-on and leaf-off datasets. Amongst the coniferous genera, *Pinus* has least conical shapes partly as a result of many of the study specimens having reached maximum heights and, therefore, the tendency for the crown to become wider over time. In both datasets, broadleaved and coniferous genera showed significant differences based upon *t*-test (p <0.001) with higher mean crown length/width ratios for coniferous genera (See Table 5).

Table 5
The *p*-values using Student's *t*-test for structure variables for broadleaved and coniferous genera with the leaf-on and leaf-off datasets.

Structural variable	Variables	*P* values (using Student's *t*-test)	
		Leaf-on	Leaf-off
Height	Relative 10th height percentile	0.006	0.337
	Relative 50th height percentile	0.001	0.159
	Relative 90th height percentile	0.001	0.285
	Relative standard deviation of height	0.483	0.373
Crown shape	Length to width ratio at upper 10% of a crown length	<0.001	<0.001
	Length to width ratio at upper 50% of a crown length	<0.001	<0.001
	Length to width ratio at upper 90% of a crown length	<0.001	<0.001

4.1.3. Classification of broadleaved and coniferous genera from structure variables

The results of classification percentages using linear discriminant functions for deciduous and evergreen genera showed that the leaf-on dataset had greater correct classification (74.9%) than the leaf-off dataset (50.2%). When combining two LIDAR datasets, correct classification was 74.0%, slightly less than achieved with the leaf-on dataset alone. The leaf-on dataset was much more effective than the leaf-off dataset and slightly better than the combined dataset in classifying individual trees into the correct genus.

4.2. Stepwise cluster analysis using both leaf-on and leaf-off datasets

Principal component analysis (Step 1 in Section 3.4.1) was conducted using all intensity and structure variables for 223 individual trees (see Table 3 for definitions of the variables). Variables 2, 10, 13, 14, 15, and 16 for the leaf-on dataset and variables 3, 7, 10, 12, and 13 for the leaf-off dataset were chosen using the criteria in Section 3.1 for the next step in classification. After testing different numbers of clusters with the subset of variables from Step 1, two clusters (maximal average silhouette width = 0.615) were chosen.

Table 6a presents the number and percentage of trees assigned to each cluster. All individual *Acer*, *Ulmus* and *Quercus* trees were assigned to cluster 1 while all individual *Pseudotsuga*, *Pinus*, *Picea* and *Tsuga* trees were assigned to cluster 2. For the most part, clustering by deciduous and evergreen genera made sense — some of the deciduous genera (e.g., *Betula*, *Larix*, *Sorbus*) were already partially leafed-out and one of the *Magnolia* was evergreen, which was correctly placed into a leaf-on group. Using the Step 3 80% criterion, the small percentages of *Betula* and *Sorbus* trees were redistributed to cluster 1 whereas *Thuja* and *Sequoia* were redistributed to cluster 2. Large percentages of *Magnolia*, *Prunus* and *Larix* trees were assigned to both clusters; trees of these genera were not considered clustered and were dropped. Table 6b presents the final clustering from this first pass through the stepwise cluster analysis. Cluster 1 is composed of broadleaved genera which had no foliage at the time of leaf-off data acquisition; *Acer*, *Betula*, *Quercus*, *Sorbus* and *Ulmus*. Cluster 2 is composed of five evergreen conifers, *Pseudotsuga*, *Picea*, *Pinus*, *Sequoia*, *Thuja*, and *Tsuga,* and one deciduous broadleaf, *Malus*.

The stepwise cluster analysis process was repeated to identify possible clusters within the Table 6b clusters; these new clusters, if any, will be referred to as sub-clusters.

4.2.1. Sub-clustering result for Table 6b cluster 1

As a result of the Step 1 in PCA, eight variables were selected using the criteria described in Section 3. These variables (see Table 3 for definitions of variables) are 10, 11, 13, 15, 16 for the leaf-on dataset and variables 2, 7, 12 for the leaf-off dataset. The maximal average *silhouette* width was 0.45 indicating that this cluster analysis did not produce a good sub-cluster structure.

4.2.2. Sub-clustering result for Table 6b cluster 2

As a result of the Step 1 in PCA, variables (see Table 3 for definitions of variables) 7, 10, 13, and 14 from the leaf-on dataset and variables 3, 7, 10, and 16 from the leaf-off dataset were selected using the criteria described in Section 3.1. The average *silhouette* width was 0.56 with four sub-clusters and 0.55 with two sub-clusters. Since the difference between *silhouette* widths was not significant, the

Table 6a
Initial cluster analysis using combined leaf-on and leaf-off datasets indicating the number of trees and the percentage (%) assigned to each cluster as well as the total number of trees and the percentage (%) for each species.

Genus	Cluster 1		Cluster 2		Total
	Species	Number of trees (%)	Species	Number of trees	Number of trees (%)
Betula	*B. alleghaniensis* *B. nigra* *B. platyphylla* *B. utilis*	18 (90)	*B. nigra*[a]	2 (10)	20 (100)
Acer	*A. macrophyllum*	11 (100)		0 (0)	11 (100)
Ulmus	*U. americana*	10 (100)		0 (0)	10 (100)
Magnolia	*M. dawsoniana* *M. denudata* *M. acuminata*	11 (58)	*M. grandiflora*[b] *M. fraseri*	8 (42)	19 (100)
Malus	*M. coronaria*[a]	2 (20)	*M. domestica* *M. florentina* *M. fusca*	8 (80)	10 (100)
Prunus	*P. serotina* *P. domestica*[a]	5 (45)	*P. incise* *P. sargentii*	6 (55)	11 (100)
Quercus	*Q. garryana* *Q. arizonica* *Q. bicolor* *Q. alba* *Q. rubra*	19 (100)		0 (0)	19 (100)
Sorbus	*S. americana* *S. anglica* *S. commixta*	10 (91)	*S. hybrid*[a]	1 (9)	11 (100)
Thuja	*T. plicata*	2 (11)	*T. plicata*	17 (89)	19 (100)
Pseudotsuga		0 (0)	*P. menziesii*	12 (100)	12 (100)
Larix	*L. deciduas* *L. gmelinii*	10 (48)	*L. kaempferi*[a] *L. laricina*[a] *L. occidentalis*[a]	11 (52)	21 (100)
Pinus		0 (0)	*P. monticola* *P. armandii* *P. densiflora* *P. strobus* *P. ponderosa*	21 (100)	21 (100)
Sequoia	*S. sempervirens*	2 (20)	*S. sempervirens*	8 (80)	10 (100)
Picea		0 (0)	*P. abies* *P. engelmanni* *P. koyamai* *P. montigena*	15 (100)	15 (100)
Tsuga		0 (0)	*T. heterophylla*	14 (100)	14 (100)

[a] Bud-break already occurred.
[b] Evergreen broadleaf.

Table 6b
Final cluster analysis using combined leaf-on and leaf-off datasets after applying the 80% rule.

Genus	Cluster 1		Cluster 2		Total
	Species	Number of trees (%)	Species	Number of trees	Number of trees (%)
Betula	*B. alleghaniensis*	20 (100)			20 (100)
	B. nigra				
	B. platyphylla				
	B. utilis				
	B. nigra[a]				
Acer	*A. macrophyllum*	11 (100)			11 (100)
Ulmus	*U. americana*	10 (100)			10 (100)
Malus			*M. domestica*	10 (100)	10 (100)
			M. florentina		
			M. fusca		
			M. coronaria[a]		
Quercus	*Q. garryana*	19 (100)			19 (100)
	Q. arizonica				
	Q. bicolor				
	Q. alba				
	Q. rubra				
Sorbus	*S. americana*	11 (100)			11 (100)
	S. anglica				
	S. commixta				
	S. hybrid[a]				
Thuja			*T. plicata*	19 (100)	19 (100)
Pseudotsuga			*P. menziesii*	12 (100)	12 (100)
Pinus			*P. monticola*	21 (100)	21 (100)
			P. armandii		
			P. densiflora		
			P. strobes		
			P. ponderosa		
Sequoia			*S. sempervirens*	10 (100)	10 (100)
Picea			*P. abies*	15 (100)	15 (100)
			P. engelmanni		
			P. koyamai		
			P. montigena		
Tsuga			*T. heterophylla*	14 (100)	14 (100)

[a] Bud-break already occurred.

assignment of individual trees to two versus four sub-clusters was compared. With two sub-clusters, all individual trees have *silhouette* width, $s(i)$, greater than zero while three individual trees had $s(i)$ less than zero with four sub-clusters. Hence, two sub-clusters are suggested to be the most natural number of sub-clusters.

Table 7a presents the results using two sub-clusters within the Table 6b cluster 2. All individual trees within *Tsuga* were assigned to sub-cluster 2–1. The majority of *Malus* (80%), *Thuja* (84%) and *Pinus* trees (81%) were assigned to sub-cluster 2–2 while the majority of *Sequoia* (80%) and *Picea* (80%) were assigned to sub-cluster 1. Half of the *Pseudotsuga* trees were assigned to each sub-cluster so this species failed the 80% rule and was dropped as not clustered. Table 7b shows the final sub-clusters; sub-cluster 2–1 is composed of *Picea*, *Sequoia* and *Tsuga* while sub-cluster 2–2 is composed of *Malus*, *Pinus* and *Thuja*.

4.2.3. Sub-clustering result for Table 7b sub-clusters 2–1 and 2–2

The stepwise cluster analysis process was repeated to identify possible clusters within the Table 7b sub-clusters; these new sub-clusters, if any, will be referred to as sub-sub-clusters. For the cluster analysis using sub-cluster 2–1, variables 3, 8, and 10 from the leaf-on dataset and variables 2, 11, and 15 from the leaf-off dataset were

Table 7a
Initial cluster analysis of Table 6b cluster 2 using combined leaf-on and leaf-off datasets indicating the number of trees and the percentage (%) assigned to each sub-cluster as well as the total number of trees and the percentage (%) for each species.

Cluster 2	Sub-cluster 2–1		Sub-cluster 2–2		Total
Genus	Species	Number of trees (%)	Species	Number of trees	Number of trees (%)
Malus	*M. coronaria*[a]	2 (20)	*M. domestica*	8 (80)	10 (100)
			M. florentina		
			M. fusca		
Thuja	*T. plicata*	3 (16)	*T. plicata*	16 (84)	19 (100)
Pseudotsuga	*P. menziesii*	6 (50)	*P. menziesii*	6 (50)	12 (100)
Pinus	*P. ponderosa*	4 (19)	*P. monticola*	17 (81)	21 (100)
			P. armandii		
			P. densiflora		
			P. strobus		
Sequoia	*S. sempervirens*	8 (80)	*S. sempervirens*	2 (20)	10 (100)
Picea	*P. abies*	12 (80)	*P. koyamai*	3 (20)	15 (100)
	P.engelmanni				
	P. montigena				
Tsuga	*T. heterophylla*	14 (100)		0 (0)	14 (100)

[a] Bud-break already occurred.

Table 7b
Final cluster analysis of Table 6b Cluster 2 after applying the 80% rule.

Cluster 2	Sub-cluster 2–1		Sub-cluster 2–2		Total
Genus	Species	Number of trees (%)	Species	Number of trees	Number of trees (%)
Malus			*M. domestica*	10 (100)	10 (100)
			M. florentina		
			M. fusca		
			M. coronaria[a]		
Thuja			*T. plicata*	19 (100)	19 (100)
Pinus	*P. ponderosa*	4 (19)	*P. monticola*	17 (81)	21 (100)
			P. armandii		
			P. densiflora		
			P. strobus		
Sequoia	*S. sempervirens*	10 (100)			10 (100)
Picea	*P. abies*	15 (100)			15 (100)
	P.engelmanni				
	P. montigena				
	P. koyamai				
Tsuga	*T. heterophylla*	14 (100)			14 (100)

[a] Bud-break already occurred.

Table 8
Cluster analysis of Table 7b sub-cluster 2–1 using three sub-sub-clusters indicating the number of trees and the percentage (%) assigned to each sub-sub-cluster as well as the total number of trees and the percentage (%) for each species.

Sub-cluster 2–1	Sub-cluster 2–1–1		Sub-cluster 2–1–2		Sub-cluster 2–1–3		Total
	Species	# trees (%)		# trees (%)		# trees (%)	# trees (%)
Sequoia	*S. sempervirens*	5 (50)	*S. sempervirens*	3 (30)	*S. sempervirens*	2 (20)	10 (100)
Picea	*P. abies* *P. engelmannii*	8 (53)	*P. montigena* *P. koyamai*	6 (40)	*P. koyamai*	1 (7)	15 (100)
Tsuga	*T. heterophylla*	11 (79)	*T. heterophylla*	3 (21)		0 (0)	14 (100)

selected using the criteria described in Section 3.1. Three clusters were indicated with the maximal average silhouette width, 0.61. Table 8 presents the result of cluster analysis using three sub-sub-clusters within sub-cluster 2–1. Fewer than 80% of individual trees of any of the genera within sub-cluster 2–1 were assigned to any of the three sub-sub-clusters. Hence, the genera in sub-cluster 2–1 could not be further clustered.

For the cluster analysis using sub-cluster 2–2, variables 2, 10, 16 from the leaf-on dataset and variables 1, 4, 14 from the leaf-off dataset were selected because each of these variables had the greatest absolute coefficient value in the component whose associated Eigen value was greater than 0.7. Two clusters were suggested with the maximal average silhouette width, 0.57. Table 9 presents the result of cluster analysis using two sub-sub-clusters within sub-cluster 2–2. Individual trees assigned to a single group were less than 80% for all genera within sub-cluster 2–2; therefore, these genera could not be further clustered.

4.2.4. Venn diagram of stepwise cluster analysis using both leaf-on and leaf-off dataset

Fig. 4 summarizes the overall stepwise cluster analysis using both leaf-on and leaf-off datasets using Venn diagrams. The overlapping area of the Cluster 1 and Cluster 2 comprises *Magnolia*, *Prunus* and *Larix*. The exclusive set of Cluster 1 comprises deciduous broadleaved genera; *Acer*, *Betula*, *Quercus*, *Sorbus* and *Ulmus*. The exclusive set of Cluster 2 comprises *Malus*, *Picea*, *Pseudotsuga*, *Pinus*, *Sequoia*, *Thuja*, and *Tsuga*.

The genera in Cluster 2 were further divided into two sub-clusters (Cluster 2–1 and Cluster 2–2). *Malus*, *Thuja* and *Pinus* compose Cluster 2–1 and *Sequoia*, *Picea* and *Tsuga* compose Cluster 2–2. There was no overlap of the Venn diagrams of Cluster 2–1 and Cluster 2–2. The overlapping area of the Cluster 1 and Cluster 2 comprises *Pseudotsuga*. The *silhouette* width was greater than 0.5 at every step in each Venn diagram, which suggests that the separations between clusters are acceptable as indicated.

4.3. Stepwise cluster analysis using leaf-on data

As a result of PCA using all variables from the leaf-on dataset, variables 3, 7 and variables 10, 13, 14, 15, and 16 were selected using the criteria described in Section 3.1. As a result of PAM, the maximal average silhouette width was 0.46 (<0.5). Hence, natural clustering was not suggested with only leaf-on variables.

Table 9
Cluster analysis of Table 7b sub-cluster 2–2 using two sub-sub-clusters indicating the number of trees and the percentage (%) assigned to each sub-sub-cluster with the total number of trees and the percentage (%) for each species.

Sub-cluster 2–2	Sub-sub-cluster 2–2–1		Sub-sub-cluster 2–2–2		Total
Genus	Species	Number of trees (%)	Species	Number of trees (%)	Number of trees (%)
Malus	*M. domestica* *M. fusca*	6 (60)	*M. coronaria M. florentina*	4 (40)	10 (100)
Thuja	*T. plicata*	12 (63)	*T. plicata*	7 (37)	19 (100)
Pinus	*P. armandii* *P. densiflora* *P. strobus*	15 (71)	*P. ponderosa* *P. monticola*	6 (29)	21 (100)

4.4. Stepwise cluster analysis using leaf-off data

As a result of PCA using all variables from the leaf-off dataset, variables 10, 11, 13, 14, and 16 in Table 3 were selected using the criteria described in Section 3.1. Two clusters were suggested as the most natural clustering with an average silhouette width, 0.62. Table 10a presents the result of cluster analysis using two clusters. These clusters are similar to those in Table 6a which used both leaf-on and leaf-off datasets. Except for *Magnolia* and *Prunus*, which failed the 80% rule, all genera were more clearly clustered into either cluster 1-off or cluster 2-off than were the Table 6a clustering results using both datasets. For example, all individual *Betula* and *Sequoia* trees were assigned to cluster 1-off and cluster 2-off, respectively while clustering result using both datasets showed some individual trees within these genera in different

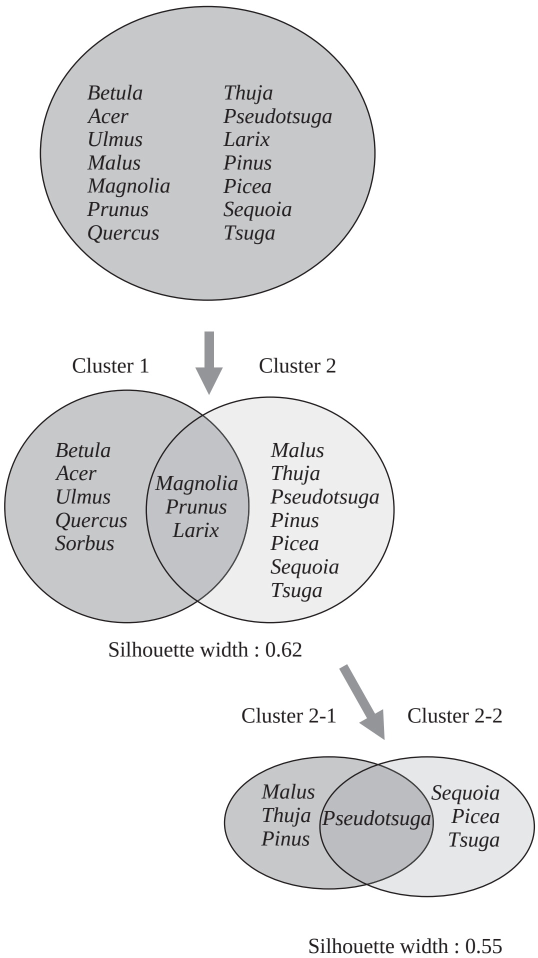

Fig. 4. Venn diagram of the stepwise cluster analysis using combined leaf-on and leaf-off datasets.

sub-clusters (see Table 6a). The majority of individual *Larix* trees were assigned to cluster 1-off (86%) which is also different from the clustering result using both datasets where *Larix* failed the 80% criterion. Table 10b presents the final clustering using only the leaf-off dataset. Cluster 1 is composed of genera which had no or little foliage at the time of data acquisition in March; *Acer*, *Betula*, *Larix*, *Quercus*, *Sorbus* and *Ulmus*. Cluster 2 is composed of six evergreen coniferous genera, *Pseudotsuga*, *Picea*, *Pinus*, *Sequoia*, *Thuja* and *Tsuga* and one deciduous broadleaved genus, *Malus*.

4.4.1. Sub-clustering result with cluster 1-off

As a result of PCA, variables 11, 12, 14, and 16 were selected using the criteria described in Section 3.1. The maximal average *silhouette* width was 0.36 (<0.5). Hence, natural clustering is not suggested.

4.4.2. Clustering result with cluster 2-off

As a result of PCA, variables 10, 13, 14, and 15 were selected using the criteria described in Section 3.1. The maximal average *silhouette* width was 0.31 (<0.5). Hence, natural clustering is not suggested.

4.4.3. Venn diagram of the stepwise cluster analysis using leaf-off dataset

Fig. 5 summarizes the overall stepwise cluster analysis using leaf-off dataset using Venn diagrams. The cluster analysis did not continue after the first clustering. The exclusive set of cluster 1-off comprises one deciduous coniferous genus, *Larix*, and five deciduous broad-leaved genera, *Acer*, *Betula*, *Quercus*, *Sorbus* and *Ulmus*. The exclusive set of cluster 2-off comprises all evergreen coniferous genera and one broadleaved genus, *Malus*.

5. Discussion

5.1. Analysis of structure variables

We examined the classification rate (i.e., separating deciduous and evergreen individual trees) using structural variables derived from two separate LIDAR flights, one when leaves were on trees and the other when leaves were not. Discriminant analysis showed that the leaf-on dataset had a higher classification rate (74.9%) than the leaf-off dataset (50.2%). Interestingly, when combining structure variables from the two LIDAR datasets, the classification rate was 74.0%, which was slightly lower than that of the single leaf-on dataset (74.9%). In contrast to the results presented here for structural variables, Kim et al. (2009a) showed for the intensity data, derived from the same LIDAR flights, that the leaf-off dataset had a higher classification rate (83.4%) than that of leaf-on dataset (73.1%) for deciduous and evergreen genera using a linear discriminant function. They also found that when combining two

Table 10a
Initial cluster analysis using leaf-off (off) data indicating the number of trees and the percentage (%) assigned to each cluster with the total number of trees and the percentage (%) for each species.

Genus	Cluster 1-off		Cluster 2-off		Total
	Species	Number of trees (%)	Species	Number of trees	Number of trees (%)
Betula	*B. alleghaniensis*	20 (100)		0 (0)	20 (100)
	B. platyphylla				
	B. utilis				
	B. nigra[a]				
Acer	*A. macrophyllum*	11 (100)		0 (0)	11 (100)
Ulmus	*U. americana*	10 (100)		0 (0)	10 (100)
Magnolia	*M. dawsoniana*	11 (58)	*M. grandiflora*[b]	8 (42)	19 (100)
	M. denudata		*M. fraseri*		
	M. acuminata				
Malus		0 (0)	*M. coronaria*[a]	10 (100)	10 (100)
			M. domestica		
			M. florentina		
			M. fusca		
Prunus	*P. serotina*	4 (36)	*P. incisa*	7 (64)	11 (100)
			P. sargentii		
	P. domestica[a]				
Quercus	*Q. garryana*	19 (100)		0 (0)	19 (100)
	Q. arizonica				
	Q. bicolor				
	Q. alba				
	Q. Rubra				
Sorbus	*S. americana*	10 (91)	*S. hybrid*[a]	1 (9)	11 (100)
	S. anglica				
	S. commixta				
Thuja	*T. plicata*	3 (16)	*T. plicata*	16 (84)	19 (100)
Pseudotsuga		0 (0)	*P. menziesii*	12 (100)	12 (100)
Larix	*L. deciduas*	18 (86)	*L. occidentalis*[a]	3 (14)	21 (100)
	L. gmelinii				
	L. kaempferi[a]				
	L. laricina[a]				
Pinus		0 (0)	*P. monticola*	21 (100)	21 (100)
			P. armandii		
			P. densiflora		
			P. strobus		
			P. ponderosa		
Sequoia	*S. sempervirens*	0 (0)	*S. sempervirens*	10 (100)	10 (100)
Picea		0 (0)	*P. abies*	15 (100)	15 (100)
			P. engelmanni		
			P. koyamai		
			P. montigena		
Tsuga		0 (0)	*T. heterophylla*	14 (100)	14 (100)

[a] Bud-break already occurred.
[b] Evergreen broadleaf.

Table 10b
Final cluster analysis using leaf-off (off) after applying the 80% rule.

Genus	Cluster 1-off		Cluster 2-off		Total
	Species	Number of trees (%)	Species	Number of trees	Number of trees (%)
Betula	*B. alleghaniensis* *B. platyphylla* *B. utilis* *B. nigra*[a]	20 (100)			20 (100)
Acer	*A. macrophyllum*	11 (100)			11 (100)
Ulmus	*U. americana*	10 (100)			10 (100)
Malus			*M. coronaria*[a] *M. domestica* *M. florentina* *M. fusca*	10 (100)	10 (100)
Quercus	*Q. garryana* *Q. arizonica* *Q. bicolor* *Q. alba* *Q. Rubra*	19 (100)			19 (100)
Sorbus	*S. americana* *S. anglica* *S. commixta* *S. hybrid*[a]	11 (100)			11 (100)
Thuja			*T. plicata*	19 (100)	19 (100)
Pseudotsuga			*P. menziesii*	12 (100)	12 (100)
Larix	*L. deciduas* *L. gmelinii* *L. kaempferi*[a] *L. laricina*[a] *L. occidentalis*[a]	21 (100)			21 (100)
Pinus			*P. monticola* *P. armandii* *P. densiflora* *P. strobus* *P. ponderosa*	21 (100)	21 (100)
Sequoia	*S. sempervirens*		*S. sempervirens*	10 (100)	10 (100)
Picea			*P. abies* *P. engelmanni* *P. koyamai* *P. montigena*	15 (100)	15 (100)
Tsuga			*T. heterophylla*	14 (100)	14 (100)

[a] Bud-break already occurred.

LIDAR datasets, the classification rate improved to 90.6%. Hence, the results in the current study using structural variables implied that the leaf-on dataset was a more promising tool than the leaf-off dataset in terms of separating deciduous and evergreen genera.

These results from either structural or intensity LIDAR datasets demonstrated a degree of robustness in separating genera or individuals into deciduous or evergreen groups; however, the classifications were not perfect. A number of reasons exist for this including (1) that the two datasets were collected with different laser scanning systems with differences in the scanner used, the height of the flight, and the intensity of data points collected and (2) that there were leaves (and flowers) forming on some of the deciduous trees during the leaf-off collection period.

5.2. Stepwise cluster analysis

The stepwise cluster analysis of the leaf-on, leaf-off, and combined datasets showed different results. This implies that genera might be grouped differently depending on the timing of the data collection. The Venn diagrams generated by the unsupervised stepwise cluster analysis using variables derived from the leaf-off and combined leaf-on and leaf-off datasets demonstrated reasonable relationships between groups of genera at each step, implying that the derived variables described foliar characteristics of species appropriately. For example, at the first step of stepwise cluster analysis, most deciduous genera were separated from evergreen genera although some individual trees did not cluster with the other trees in the same genus. Since most of the deciduous genera were angiosperms and most of the evergreen genera were gymnosperms, one might also state that separate of genera fell into broad evolutionary differences.

Individual trees of the following four genera, *Acer*, *Pseudotsuga*, *Tsuga*, and *Ulmus,* were all grouped into the same clusters as a result of the first cluster analysis (see Table 6a and Table 10a). Also, individual trees of the following three genera, *Quercus*, *Picea* and *Pinus,* were all grouped into the same clusters, respectively. For these cases, individual trees within a particular genus are more similar to each other than to those in other genera.

When splitting genera into different clusters, we applied a criterion requiring that a certain minimum percentage (80%) of individual trees within one genus must be in one cluster. Because species within the same genus, especially those that are deciduous, may have differences in the timing of leafing and flowering phenology, the relatively late March leaf-off LIDAR acquisition flight, unfortunately captured this variation, thus producing confusing classifications that would not occur before bud break when these trees would have been truly deciduous. The results from cluster analysis demonstrated that species in one genera were often grouped with species of another genera and in cases where there was only one species in a given genera, these species were either not placed in separate clusters (i.e., genera) or were sometimes placed with trees with very different morphological traits in different genera (see Table 6a and Table 10a). It is clear that the morphological and intensity variables analyzed and chosen from the two LIDAR acquisitions were not adequate to consistently separate genera. The results from cluster analysis implies that the potential for differentiation of these species is generally low and only broad differentiation is possible (mostly

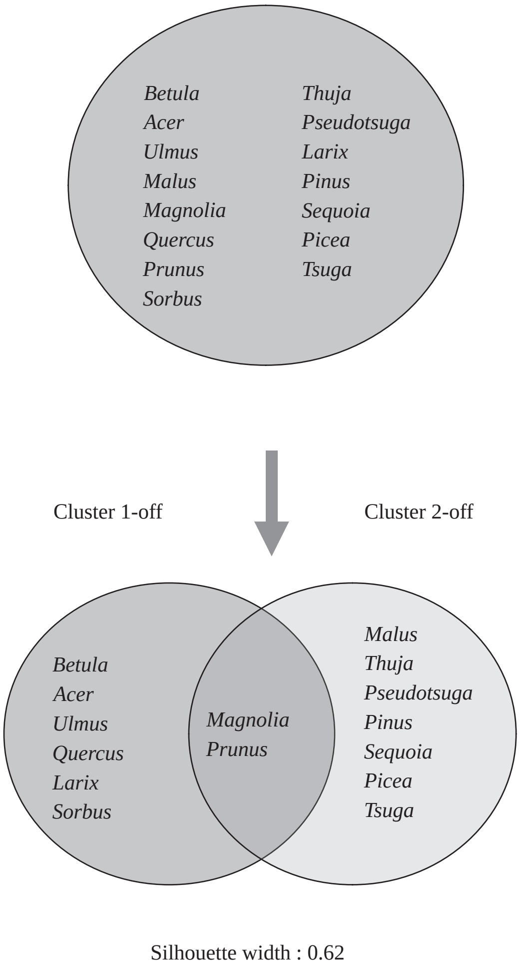

Fig. 5. Venn diagram of stepwise cluster analysis using leaf-off dataset.

between deciduous and evergreen species, with some potential to differentiate between two major classes of coniferous species), and it is best done with a combined, and thus complex, datasets (leaf-on and leaf-off datasets).

In most of the cluster analyses in this study, we obtained two clusters according to the methodology described by Kaufman and Rousseeuw (1990), which implies that two clusters are probably the most natural number of clusters when dealing with multiple genera. Since we chose genera having either deciduous or evergreen leaf traits, our classification was unable to extend beyond this.

When we conducted cluster analysis from a combined leaf-on and leaf-off datasets using cluster 2 (see Fig. 4), which comprised six coniferous genera and *Malus*, the combination of leaf reflectance and density of the leaves was probably one of the factors affecting the clustering (See Tables 7a and 7b). For example, *Picea*, *Sequoia* and *Tsuga* which have single needles, were separated from *Malus*, *Pinus* and *Thuja*, which do not have single needles. Kim et al. (2009a) found that mean intensity values between *Picea*, *Sequoia* and *Tsuga* were similar to each other and higher than those of *Pinus* and *Thuja*. The similar intensity values among the former three genera and among the latter two genera probably affected the separation of these two groups. The reason why *Malus* was clustered with the latter two genera was not investigated in this study but may be due to the stage of bud break, flowering, and new foliage development of different individuals of *Malus* at the time when the leaf-off data set was acquired.

Because the difference between mean intensity values between genera was very significant in leaf-off data compared with other variables (Kim et al, 2009a), clustering results might be mostly affected by mean intensity variables. This finding was also consistent with the result of the principal component analysis. That is, these mean intensity variables from the leaf-off dataset were always selected as a result of the principal component analysis because they had the greatest absolute coefficient values on the first few principal components which would be critical at the following cluster analyses. That is probably the reason why the stepwise cluster analysis using only leaf-off variables was similar to the result using both leaf-on and leaf-off variables. Cluster analysis using only leaf-on data was not successful probably due to the seasonal issue and because both datasets were acquired from different laser scanner systems with different flight parameters; the leaf-on data were acquired with lower point density, fewer returns per pulse, and lower scan pulse repetition frequency than the leaf-off data. LIDAR returns of leaf-on data merely represent the outer hull of the crowns, which is only particularly representative of tree species, whereas the inner structure as sampled by leaf-off data is more representative.

Usually, forest species classification is studied in forest stands where canopy overlap is greater than for our Arboretum trees which were mostly isolated individuals. Because we used the tree crown isolation method, which was especially useful in open-grown canopy, the clustering results may be different if we chose tree samples from a dense stands where overlap is common. It should be noted that the structural variables would likely be different in a close-canopy forest, because the branching structure and canopy dimensions would potentially be quite different.

6. Conclusions

Overall, our results show that LIDAR datasets can be used to cluster tree genera between deciduous and evergreen genera using unsupervised cluster analysis, and to further cluster with the evergreen coniferous genera groups. Different mean values for structural variables between genera were related not only to physical properties such as leaf structure and crown shape, but may also have been affected by system differences between the leaf-off and leaf-on LIDAR acquisitions. The two different seasonal LIDAR datasets resulted in different relative mean values for structure analysis among genera with better separation using leaf-on data than leaf-off data, albeit with two different LIDAR sensors with different settings. With the acquisition of different LIDAR datasets using the same LIDAR systems with the same conditions to control for system effects, we may produce more reliable clustering results with better differentiation between tree genera.

The stepwise cluster analysis used in this study introduces one approach to classifying tree genera based on their structural and spectral characteristics. When combining leaf-on and leaf-off LIDAR datasets, the cluster analysis was able to separate non-deciduous genera beyond the simple separation between deciduous and evergreen genera. Our results indicate that the importance of matching the timing of LIDAR data acquisition to phenologically distinct periods in order for tree genera separation. In addition, our results implied that combined LIDAR datasets was more powerful that either the leaf-on or the leaf-off dataset. Finally, our results suggested that using relatively open trees greatly aided our ability to derive distinct morphological and intensity variables. Similarly positive results from dense forest stands are likely more difficult to realize.

Acknowledgments

This study was funded by Precision Forestry Cooperative at the University of Washington, College of Forest Resources. The authors would also like to thank Robert J. McGaughey and Professors Hans-Erik Andersen and Marina Alberti at the University of Washington for their support of this project.

References

Andersen, H. E., Reutebuch, S. E., & McGaughey, R. J. (2006). A rigorous assessment of tree height measurements obtained using airborne lidar and conventional field methods. *Canadian Journal of Remote Sensing*, *32*(5), 355–366.

Boudreau, J., Nelson, R. F., Margolis, H. A., Beaudoin, A., Guindon, L., & Kimes, D. (2008). Regional aboveground forest biomass using airborne and spaceborne LIDAR in Quebec. *Remote Sensing of Environment*, *112*, 3876–3890.

Bradbury, R. B., Hill, R. A., Mason, D. C., Hinsley, S. A., Wilson, J. D., Balzter, H., Anderson, G. Q. A., Whittingham, M. J., Davenport, I. J., & Bellamy, P. E. (2005). Modelling relationships between birds and vegetation structure 768 using airborne LIDAR data: A review with case studies from agricultural and woodland environments. *Ibis*, *147*, 443–452.

Brandtberg, T., Warner, T., Landenberger, R. E., & McGraw, J. B. (2003). Detection and analysis of individual leaf-off tree crowns in small footprint, high sampling density LIDAR data from the eastern deciduous forest in North America. *Remote Sensing of Environment*, *85*(3), 290–303.

Brantberg, T. (2007). Classifying individual tree species under leaf-off and leaf-on conditions using airborne LIDAR. *ISPRS Journal of Photogrammetry and Remote Sensing*, *61*(5), 325–340.

Brennan, R., & Webster, T. L. (2006). Object-oriented land cover classification of LIDAR-derived surfaces. *Canadian Journal of Remote Sensing*, *32*(2), 162–172.

Coren, F., & Sterzai, P. (2006). Radiometric correction in laser scanning. *International Journal of Remote Sensing*, *27*(15–16), 3097–3104.

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*(4), 509–522.

Everitt, B. S., & Dunn, G. (2001). *Applied multivariate data analysis: Second edition.* New York: Oxford University Press Inc.

Garcia, M., Riaño, D., Chuvieco, E., & Danson, F. M. (2010). Estimating biomass carbon stocks for a Mediterranean forest in central Spain using LIDAR height and intensity data. *Remote Sensing of Environment*, *114*(4), 816–830.

Gates, D. M. (1980). *Biophysical Ecology.* New York: Springer-Verlag.

Graf, R. F., Mathys, L., & Bollmann, K. (2009). Habitat assessment for forest dwelling species using LIDAR remote sensing: Capercaillie in the Alps. *Forest Ecology and Management*, *257*, 160–167.

Hasegawa, H. (2006). Evaluations of LIDAR reflectance amplitude sensitivity towards land cover conditions. *Bulletin of the Geographical Survey Institute*, 53(March).

Höfle, B., & Pfeifer, N. (2007). Correction of laser scanning intensity data: Data and model-driven approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, *62*(6), 415–433.

Holmgren, J., & Persson, Å. (2004). Identifying species of individual trees using airborne laser scanner. *Remote Sensing of Environment*, *90*(4), 415–423.

Hopkinson, C., Chasmer, L., & Hall, R. J. (2008). The uncertainty in conifer plantation growth prediction from multi-temporal LIDAR datasets. *Remote Sensing of Environment*, *112*, 1168–1180.

Jolliffe, I. T. (2002). *Principal component analysis.* New York: Springer-Verlag.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York: Wiley.

Kim, S., McGaughey, R. J., Andersen,, H. -E., & Schreuder, G. (2009a). Tree species differentiation using intensity data derived from leaf-on and leaf-off airborne laser scanner data. *Remote Sensing of Environment*, *113*(8), 1575–1586.

Kim, S., Hinckley, T., & Briggs, D. (2009b). Classifying tree species using structural and spectral data from LIDAR, 2009b. *In Proceedings of the ASPRS MAPP2009 Fall Conference, 16–19 November 2009, San Antonio, Texas, CD-ROM.*

Korhonen, L., Korpela, I., Heiskanen, J., & Maltamo, M. (2011). Airborne discrete-return LIDAR data in the estimation of vertical canopy cover, angular canopy closure and leaf area index. *Remote Sensing of Environment*, *115*(4), 1065–1080.

Lefsky, M. A., Cohen, W. B., Acker, S. A., Parker, G. G., Spies, T. A., & Harding, D. (1999). Lidar remote sensing of the canopy structure and biophysical properties of Douglas-Fir western hemlock forests. *Remote Sensing of Environment*, *70*(3), 339–361.

MacQueen, J. B. (1967). *Some methods for classification and analysis of multivariate observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, *1*. (pp. 281–297) Berkeley: University of California Press.

McGaughey, R. J., & Carson, W. W. (2003). Fusing LIDAR data, photographs, and other data using 2D and 3D visualization techniques. *In Proceedings from the terrain data: applications and visualization — making the connection* (pp. 16–24).

McGaughey, R. J., Carson, W. W., Reutebuch, S. E., & Andersen, H. -E. (2004). Direct measurement of individual tree characteristics from LIDAR data. *In Proceedings from the 2004 Annual ASPRS Conference*.

Moffiet, T., Mengersen, K., Witte, C., King, R., & Denham, R. (2005). Airborne laser scanning: exploratory data analysis indicates potential variables for classification of individual trees or forest stands according to species. *ISPRS Journal of Photogrammetry and Remote Sensing*, *59*(5), 289–309.

Morsdorf, F., Frey, O., Meier, E., Itten, K., & Allgower, B. (2006). Assessment of the influence of flying height and scan angle on biophysical vegetation products derived from airborne laser scanning. *Proceedings of Workshop on 3D Remote Sensing in Forestry, Vienna, Austria* (pp. 145–150).

Morsdorf, F., Marell, A., Koetz, B., Cassagne, N., Pimont, F., Rigolot, E., & Allgower, B. (2010). Discrimination of vegetation strata in a multi-layered Mediterranean forest ecosystem using height and intensity information derived from airborne laser scanning. *Remote Sensing of Environment*, *114*(7), 1403–1415.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*, *80*, 88–99.

Næsset, E., & Gobakken, T. (2005). Estimating forest growth using canopy metrics derived from airborne laser scanner data. *Remote Sensing of Environment*, *96*, 453–465.

Nelson, R., Valenti, M. A., Short, A., & Keller, C. (2003). A multiple resource inventory of Delaware using airborne laser data. *BioScience*, *53*, 981–992.

Nelson, R., Keller, C., & Ratnaswamy, M. (2005). Locating and estimating the extent of Delmarva fox squirrel habitat using airborne LIDAR profiler. *Remote Sensing of Environment*, *96*, 292–301.

Nelson, R., Ranson, K. J., Sun, G., Kimes, D. S., Kharuk, V., & Montesano, P. (2009). Estimating Siberian timber volume using MODIS and ICESat/GLAS. *Remote Sensing of Environment*, *113*, 691–701.

Ørka, H. O., Næsset, E., & Bollandsås, O. M. (2009). Classifying species of individual trees by intensity and structure features derived from airborne laser scanner data. *Remote Sensing of Environment*, *113*, 1163–1174.

Riaño, D., Chuvieco, E., Condés, S., González-Matesanz, S., & Ustin, S. L. (2004). Generation of crown bulk density for *Pinus sylvestris* L. from lidar. *Remote Sensing of Environment*, *92*(3), 345–352.

Richardson, J. J., Moskal, L. M., & Kim, S. H. (2009). Modeling approaches to estimate effective leaf area index from aerial discrete-return LIDAR. *Agricultural and Forest Meteorology*, *149*(6–7), 1152–1160.

Roberts, D. A., Ustin, S. L., Ogunjemiyo, S., Greenberg, J., Dobrowski, S. Z., Chen, J., & Hinckley, T. M. (2004). Spectral and structural measures of Northwest forest vegetation at leaf to landscape scales. *Ecosystems*, *7*, 545–562.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65 North-Holland.

Swain, P. H., & Davis, S. M. (Eds.). (1978). *Remote sensing: The quantitative approach* (pp. 396). New York: McGraw Hill.

Solberg, S., Næsset, E., Hanssen, K. H., & Christiansen, E. (2006). Mapping defoliation during a severe insect attack on Scots pine using airborne laser scanning. *Remote Sensing of Environment*, *102*, 364–376.

Tickle, P. K., Lee, A., Lucas, R. M., Austin, J., & Witte, C. (2006). Quantifying Australian forest floristics and structure using small footprint LIDAR and large scale aerial photography. *Forest Ecology and Management*, *223*, 379–394.