LIDAR-BASED SPECIES CLASSIFICATION USING MULTIVARIATE CLUSTER ANALYSIS

Sooyoung Kim
School of Forest Resources, College of the Environment,
University of Washington, Seattle, WA, 98195-2100
kisoyo@u.washington.com

ABSTRACT

Understanding that various tree species have characteristics similar to each other, it follows that some type of hierarchical classification scheme could be used to identify species using LIDAR data. Cluster analysis, one of the unsupervised classification methods, was conducted for all individual trees using the *k*-medoid algorithm. Instead of using one-step cluster analysis, a stepwise cluster analysis was developed based on the statistical criteria to test hierarchical relationships between species. Two seasonal LIDAR datasets collected at the Washington Park Arboretum in Seattle, Washington were used for this study. Parameters derived from structure and intensity measurements using two LIDAR datasets were used for the stepwise clustering analysis. This paper shows that a variety of tree species can be naturally clustered with a hierarchy using LIDAR-derived structure and intensity measurements. Stepwise cluster analysis showed that the species with similar characteristics seem to be clustered into a single group while the species with different characteristics are likely to be clustered into different groups based on the reliable statistical criteria. The clustering results using different seasonal datasets revealed that using both seasonal datasets clustered species more reasonably than using either one of the datasets. When using only leaf-on data, the structure of clusters was not reasonably formed even at the first step of cluster analysis. It should be noted that the clustering results would vary depending not only on the variables used but also on the selected species groups or the number of individual trees.

INTRODUCTION

Recently, forest stand types or tree species classification have been studied using laser scanner datasets (Brandtberg et al., 2003 and 2007; Brennan & Webster, 2006; Donoghue et al., 2007; Holmgren and Persson, 2004; Kim et al., 2009a and 2009b; Moffiet et al., 2005; Ørka et al., 2009).

Most laser scanning data includes an intensity value which is a relative measure of the return signal strength associated with each return; a measure of the amount of energy reflected from a target. Several authors report efforts to distinguish tree species using positions of laser points within individual tree crowns as well as intensity data (Brandtberg et al., 2003 & 2007; Brennan & Webster, 2006; Holmgren and Persson, 2004; Ørka et al., 2009).

Kim et al. (2009a) normalized intensity data from the leaf-on and leaf-off laser scanning datasets based on numerous man-made features collected from two LIDAR datasets. They found that normalized intensity data can be used for tree species classification. Kim et al. (2009b) found that using both intensity and height data derived from laser scanning data improved classification of deciduous and coniferous species groups compared with using either intensity or height data alone. These previous studies used, discriminant functions, one of the supervised classification methods, for classification.

Understanding that some tree species have characteristics similar to each other, it follows that some type of hierarchical classification scheme can be used to identify species using LIDAR data. We report on 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 used a stepwise cluster analysis, based on statistical criteria, to find hierarchical relationships between species. If the variables derived from the cluster analysis represent characteristics of individual tree species well, the resulting clusters would be reliable and one could reasonably assume that closely related species will be assigned to the same cluster while less closely related species will be assigned to other clusters.

In this study, stepwise cluster analysis was used to test if the previously derived intensity and height metrics are reliable to classify various species and to test the potential of the laser scanning data for hierarchical cluster analysis using the given samples and datasets.

STUDY AREA

The study area is the Washington Park Arboretum located in Seattle, Washington (47° 37.723N 122° 17.732W, figure1). The area covers 93 hectares and a topographic range is 15 to 55 m above sea level with less than 30% of slope for the majority of the site.

DATA

This study was based on laser scanning data and field data collected by Kim et al. (2009a). For thorough descriptions of the field data and the species selection and field measurement, the reader is referred to Kim et al. (2009a).

Laser Data Acquisition

Laser scanner data were acquired under leaf-on and leaf-off conditions. Leaf-on data were acquired on 30 August, 2004 using the Optech ALTM 30/70 laser scanner system (Kim et al., 2009a). 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 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 use a 1064 nm laser and beam divergence of 0.31mrad 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 leaf-off conditions due to widely varying phenology across the wide range of species within the arboretum and unusually early bud break in 2005. Table 1 lists the genera, individual species, classification as to deciduous or non-deciduous, 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 this state.

Raw intensity data were used without additional radiometric calibration (Coren and Sterzai, 2006; Donoghue et al. 2007; Hasegawa, 2006; Kim et al., 2009a) because a topographic range of this study site is not significant and scan angles are narrow (<11 ^{o} off-nadir) for both datasets.

Intensity data from the leaf-on laser scanning system multiplied by a scaling factor (16.43949) was used to directly compare with leaf-off intensity data (Kim et al., 2009a). The digital terrain model (DTM) described by Kim et al. (2009a) was used in this study with 1- by 1- m resolution using FUSION/LDV software (McGaughey and Carson, 2003; McGaughey et al., 2004).

Field Measurement

The purpose of the field work was to select and georeference various tree species that could be used as ground data for the analysis. The field work was carried out in the period of April to July 2005.

Seven coniferous and eight broadleaved species were used for the analysis. Seven coniferous species are western red cedar (*Thuja plicata*), Douglas-fir (*Pseudotsuga mensiesii*), larch (*Larix*), pine (*Pinus*), western hemlock (*Tsuga heterophylla*), redwood (*Sequoia sempervirens*. Eight broadleaved species are bigleaf maple (*Acer macrophyllum*), birch (*Betula)*, elm (*Ulmus*), oak (*Quercus), Prunus, Magnolia, Malus, Sorbus.* The locations of selected individual trees are overlaid over the orthophoto of the Arboretum.

Tree heights, crown base heights and average crown diameters computing the mean value of the two perpendicular directions (N-S and E-W) measured in Kim et al. (2009a) were used for the analysis in this study. In total, 345 trees were collected. The post-processing of collected GPS points for individual tree locations are also described in Kim et al. (2009a). After post-processing to eliminate where with severely overlapped crowns or trees that could not be clearly identified in the office, 223 individual trees were selected for the analysis.

Variables

Intensity metrics derived from leaf-on and leaf-off laser scanning datasets using isolated individual tree crowns by Kim et al.(2009a) were used for the analysis in this study. 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. The following nine variables

were computed from each of the leaf-on and leaf-off laser scanning data: (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), and (9) proportion of first returns (prop_1).

Height metrics composed of vertical distributions of laser returns and upper crown shapes computed by Kim et al. (2009b) were used in this study. Four vertical distributions of laser returns using isolated individual tree crowns are the relative 90^{th} height percentile, relative median height percentile, relative 10^{th} height percentile and relative standard deviation of height. Three variables regarding upper crown shapes are upper 10%, upper quarter (25 %) and upper one-third (33.3%) of the crown length.

COMPUTATION AND ANALYSIS

Cluster Analysis

In the method used in the program PAM (*P*artitioning *A*round *M*edoids) in the R statistical package, the representative object of a cluster is its medoid, which we defined as that object of the cluster for which the average dissimilarity (typically Manhattan distance which is defined as the distance between two points measured along axes at right angles) to all the objects of the cluster is minimal. As the objective is to find *k* such objects, we call this the *k*-medoid method. After finding a set of *k* representative objects, the *k* clusters are constructed by assigning each object of the data set to the nearest representative object (Kaufman and Rousseeuw, 1990).

One of the simplest unsupervised learning algorithms that solve the well known clustering problem is *k*-means (MacQueen, 1967) which defines *k* centroids, one for each cluster by computing Euclidean distances. Advantages of *k*-medoid method are that it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances employed in *k*-mean method and that it is more robust with respect to outliers. By construction, the *k*-medoid method tries to find "spherical" clusters, that is, clusters that are roughly ball-shaped. It is therefore not suited to discover drawn-out clusters. The *k* representative objects should minimize the sum of the dissimilarities (distance) of all objects to their nearest medoid. Basically, dissimilarities are nonnegative numbers $d(i, j)$ that are small (close to zero) when *i* and *j* are "near" to each other and are large when *i* and *j* are far apart. PAM operates using the dissimilarity matrix of the given dataset. When it is presented with an $_n$ x $_p$ data matrix where *n* indicates the number of samples and *p* indicates the number of variables, PAM first computes a dissimilarity matrix. The algorithm computes *k* representative objects, called *medoids*, which together determine a clustering. The number of clusters, *k*, is an argument of the function. 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:

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

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

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

The algorithm proceeds in two steps:

a. Build-step

This step sequentially selects *k* centrally located objects to be used as initial medoids.

b. Swap-step

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.

Validation of cluster analysis (Silhouettes). There are questions about the validity of cluster analysis. For example, how many clusters best represent the given datasets and if the quality of clusters is high, i.e. the 'within' dissimilarities are small when compared to the 'between' dissimilarities. To solve these problems, Rousseeuw (1987) proposed a new graphical display for partitioning techniques. Each cluster is represented by a so-called, *silhouette*, which is based on the comparison of its tightness and separation. This silhouette shows which objects lie

well within their cluster, and which ones are merely somewhere in between clusters. The entire clustering is displayed by combining the silhouettes into a single plot, allowing an appreciation of the relative quality of the clusters and an overview of the data configuration. The average silhouette width provides an evaluation of clustering validity and might be used to select an 'appropriate' number of clusters. In order to construct silhouettes, we need the partition we have obtained and the collection of all proximities between objects. 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*, then we can compute

$a(i)$ = average dissimilarity of i to all other objects of A.

This is the average length of all lines within cluster *A*. Next, consider any cluster *C* which is different from *A*, and compute, $d(i, C)$ = average dissimilarity of i to all objects of C.

This is the average length of all lines going from *i* to *C*. After computing $d(i, C)$ for all clusters $C \neq A$, select the smallest of those numbers and denote it by, $b(i) = \underset{C \neq A}{\text{minimum}} d(i, C)$

The cluster *B* for which this minimum is attained (that is, $d(i, B) = b(i)$) we call the neighbor of the object. Cluster *B* is the closest (on average) to object *i*, when *A* itself is discarded. The number $s(i)$ is obtained by combining $a(i)$ and $b(i)$ as follows: $-1 \leq s(i) = \dfrac{b(i) - a(i)}{\max\{a(i), b(i)\}} \leq 1$

When $s(i)$ is at its largest (that is, $s(i)$ close to 1) this implies that 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) pointed out that the silhouettes should look best for a 'natural' value of *k*, the number of clusters. He 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 species, since the objective is to cluster species groups.

Stepwise Cluster Analysis

Some tree species are more closely related than other species in terms of genetics or structural characteristics and so it is likely that there are natural groupings of species. Also, it is possible that tree species have different relationships depending on the criteria being evaluated. Generally, trees are classified as either broadleaved and coniferous for many forestry applications because this division is critical in a variety of ecosystem management plans. Both the broadleaved and coniferous groups contain evergreen and deciduous species; many temperate broadleaved species are deciduous but some are not and most coniferous species are evergreen but *larix* is a notable deciduous conifer. However, classification as to evergreen or deciduous is common practice and it is possible that tree species can also be classified based on leaf structures (Petrides and Petrides, 1992).

As one of the clustering methods, hierarchical clustering techniques proceed using either a series of successive merges or a series of successive divisions. Hierarchical methods result in a nested sequence of clusters which can be graphically represented with a tree, called a *dendrogram* (Kaufman and Rousseeuw, 1990). Agglomerative hierarchical clustering techniques produce partitions by a series of successive fusions of the individual objects. With such methods, fusions, once made, are irreversible, so that when an agglomerative algorithm has placed two individuals in the same groups they cannot subsequently appear in different groups. Since all agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, which division to choose should be decided for the purpose of getting the best fitting number of clusters (Everitt and Dunn, 2001). They also pointed out that determining the appropriate number of groups, that is, the appropriate partition, is not straightforward. When hierarchical clustering techniques are used in practice, the investigator is often interested in only one or two partitions rather than the complete hierarchy. In this research, we are more interested in clustering species groups than clustering individual trees. Therefore, for the purpose of seeking hierarchy among tree species, instead of using hierarchical clustering techniques, a modified approach was developed. As a first step to conduct stepwise cluster analysis, variables need to be reduced to simplify later analysis while retaining as much information as possible (Everitt and Dunn, 2001). For this purpose, principal component analysis (PCA) was conducted using the R package.

Process of stepwise cluster analysis. Starting with conducting principal component analysis using all datasets, stepwise cluster analysis iterates the following three steps.

Step 1: Conduct principal component analysis. Determine the number of components to be used and derive the corresponding variables.

Step 2: Conduct cluster analysis using PAM. Determine the most appropriate number of clusters by means of maximal average silhouette width.

Step 3: Either redistribute individual trees within one species into one cluster or delete the species.

After the number of clusters is decided at step 2, examine the individual trees of each species for assignment to a cluster. In this study, clustering species is considered to be a important objective than clustering individual trees. Ideally, individual trees of the same species will group into a single cluster. However, it is possible that some individuals of the same species may have different characteristics depending on age, growth conditions, and competition with neighborhood trees and become members of different clusters. Conversely, trees of closely related species within the same *genus* may be so similar that they cannot be separated and all become members of the same cluster. However, 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 produce confusing classifications that would not occur at times while truly deciduous before bud break/flowering or after foliage has matured. See Table 1 for cases where this occurred in this study. Therefore, we need a rule to determine if a certain species can be considered to be clustered. One approach would be to require that a certain minimum percentage of individual trees within one species must be in one cluster, in order for that species to be considered as clustered. If the minimum required percentage is not achieved the species is considered to be not clustered. After testing different percentages to construct good clusters, a range of 70 - 90 % was selected as the criterion for a species to be considered to be clustered. If a species meets this criterion and is clustered, all individual trees within that species are redistributed into the cluster where the majority of individual trees of the species were assigned. If a certain species failed criterion, that is, less than 70 – 90 % of the individuals for that species were assigned to a single cluster, that species was excluded from the next step.

The next step is to repeat three cluster analysis steps described above with the newly assigned clusters. This stepwise cluster analysis is continued with the reconstructed clusters until the maximal overall average silhouette width is under 0.5, This value isthe threshold suggested by Kaufman and Rousseeuw (1990)for deciding that a reasonable structure has been achieved who suggest a subjective interpretation of the Silhouette Coefficient (SC) as the maximal average silhouette width for the entire data set.

RESULTS

Stepwise Cluster Analysis using all datasets

The analysis used 223 individual sample trees. As a result of Principal Component Analysis, eleven variables were derived, including three intensity variables and eight height variables. The first two components account for 56.0 % of the variance of the combined leaf-on and leaf-off datasets. Two variables, coefficient of variation using all returns (cv_all) and mean intensity values for an upper portion of a crown using all returns (upper_all) in leaf-off data were selected based on the greatest absolute coefficient value on each component. After testing various numbers of clusters, two clusters were suggested based on the maximal average silhouette width. The average silhouette width using two clusters had the highest value, 0.615, compared to that using other numbers of clusters, considered to be a reasonable structure by Kaufman and Rousseeuw(1990). Table 1 shows the result of using two clusters indicated by the number of individual trees and the percentage assigned to each group as well as the total number of individuals and the percentage for each species. All individual bigleaf maple, elm and oak trees were assigned to *Group 2* while all individual Douglas-fir, pine, spruce and western hemlock trees were assigned to *Group 1*. Although all individual trees were not assigned to a single group, birch and *Sorbus* were redistributed to *Group 1* while cedar and redwood were redistributed to *Group 2* according to the clustering criterion described in section 4.2. Individual trees of *Magnolia*, *Malus*, *Prunus* and larch were assigned to both groups, and therefore, these species were defined not to be clustered into any groups according to the clustering criterion and consequently were not used at the next step.

The result of redistributing individual trees within the same species into a single cluster by deleting species which failed criterion is shown in Table 2. *Cluster 1* was composed of broadleaved species which had no foliage at the time of leaf-off data acquisition in March. *Cluster 2* was composed of evergreen coniferous species.

For *Cluster 1* and *Cluster 2*, cluster analysis was repeated using the derived variables from the first step of the cluster analysis above.

Table 1. The result of cluster analysis using all datasets indicated by the number of trees and the percentage assigned to each group as well as the total number of trees and the percentage for each species

Species	Group1		Group 2		Total
	Number of trees	Percentage (%)	Number of trees	Percentage (%)	Number of trees (%)
Birch	18	90	2	10	20 (100)
Bigleaf maple	11	100	0	0	11 (100)
Elm	10	100	0	0	10 (100)
Magnolia	11	58	8	42	19 (100)
Malus	2	20	8	80	10 (100)
Prunus	5	45	6	55	11(100)
Oak	19	100	0	0	19 (100)
Sorbus	10	91	1	9	11 (100)
Cedar	2	11	17	89	19 (100)
Douglas-fir	0	0	12	100	12 (100)
Larch	10	48	11	52	21 (100)
Pine	0	0	21	100	21 (100)
Redwood	2	20	8	80	10 (100)
Spruce	0	0	15	100	15 (100)
Western Hemlock	0	0	14	100	14 (100)

Table 2. The result of redistributing individual trees within the same species into a single cluster (*Cluster 1* or *Cluster 2*) after deleting the species which failed to the clustering criterion

	Cluster 1	Cluster 2
Species	Birch	Cedar
	Bigleaf maple	Douglas-fir
	Elm	Pine
	Oak	Redwood
	Sorbus	Spruce
		Western Hemlock

Clustering result for Cluster 1. As a result of PCA, eight variables were selected, including two intensity variables and six height variables. The first four components accounted for 53.2% variability of the given datasets and the variables selected from the components were all leaf-off variables. The maximal average silhouette width was 0.45 indicating that cluster analysis did not produce a good structure.

Clustering result for Cluster 2. As a result of PCA, eight variables were selected, including three intensity variables and five height variables. The first four components account for 53.0 % variability of the given datasets and they were composed of intensity and height variables in both leaf-on and leaf-off datasets: (1) relative 10^{th} height percentile in leaf-off data, (2) length to width ratio within the upper 10 % of a crown in leaf-on data, and coefficient of variation of intensity using all returns in (3) leaf-off data and (4) leaf-on data. The average silhouette width was largest, 0.56, with four clusters and the second largest was 0.55 with two clusters. Since the difference between silhouette widths was not large enough, individual objects assigned to clusters were examined. With two clusters, all individuals have silhouette width, (s (i)), greater than zero while three objects had s (i) less than zero with four clusters. Therefore, two clusters are suggested to be the most natural number of clusters. Table 3 presents the result of using two clusters indicated by the number of individual trees and the percentage assigned to each group as well as the total number of individuals and the percentage for each. All individuals within western hemlock were assigned to *Group 1*. The majority of cedar and pine were assigned to *Group 2* while the majority of redwood and spruce were assigned to *Group 1*. Douglas-fir was evenly assigned to the both groups. The result of redistributing individual trees within the same species into a single cluster without Douglas-fir is shown in Table 4.

For the cluster analysis with *Cluster 2-1*, six variables were selected, including intensity and height variables in both datasets. The first five principal components account for 57.7% variability of the given datasets. Three clusters were suggested with the maximal average silhouette width, 0.61. Table 5 presents the result of cluster analysis using three clusters indicated by the number of individual trees and the percentage assigned to each group as well as the total number of individuals and the percentage for each species. Most of the individuals within western hemlock were assigned into *Group 1*. However, the majority of redwood and spruce were also assigned to *Group 1*. Therefore, it is hard to say that these three species were clustered into separate groups at this step.

For the cluster analysis using *Cluster 2-2*, six variables were derived, including intensity and height variables in both datasets. The first five principal components account for 57.7% variability of the given datasets. Two clusters were suggested with the maximal average silhouette width, 0.57. Table 6 presents the result of cluster analysis using two clusters indicated by the number of individual trees and the percentage assigned to each group with the total number of individuals and the percentage for each species. Individuals within cedar and pine were assigned to the both groups and therefore, they were not clustered into any groups.

The diagram of stepwise cluster analysis. The overall stepwise cluster analysis using both leaf-on and leaf-off datasets was summarized with diagrams and is shown in Figure 1. At the first step of cluster analysis, deciduous broadleaved species and evergreen coniferous species were well divided into separate groups. The left side bubble diagram within *Cluster 1* was composed of deciduous broadleaf species while the right side bubble diagram within *Cluster 2* was composed of evergreen coniferous species. Four species, *Magnolia*, *Malus*, *Prunus* and larch were not clustered into any group. Evergreen coniferous species in *Cluster 1* were divided into two separate groups again at the next step of cluster analysis. Cedar and pine were clustered into one group while redwood, spruce and western hemlock were clustered into the other group. The overall silhouette widths were larger than 0.50 at every step which suggests that the separations between clusters are acceptable

Table 3. The result of cluster analysis using *Cluster 2* indicated by the number of individuals and the percentage assigned to each group as well as the total number of individuals and the percentage for each species

Cluster 2	Group 1		Group 2		Total
	Number of trees	Percentage (%)	Number of trees	Percentage (%)	Number of trees (%)
Cedar	3	16	16	84	19 (100)
Douglas-fir	6	50	6	50	12 (100)
Pine	4	19	17	81	21 (100)
Redwood	8	80	2	20	10 (100)
Spruce	12	80	3	20	15 (100)
Western Hemlock	14	100	0	0	14 (100)

Table 4. The result of redistributing individuals within the same species into a single cluster (*Cluster 2-1* or *Cluster 2-2*) without Douglas-fir

	Cluster 2-1	Cluster 2-2
Species	Redwood	Cedar
	Spruce	Pine
	Western Hemlock	

Table 5. The result of cluster analysis with *Cluster 2-1* using three groups indicated by the number of individuals and the percentage assigned to each group as well as the total number of individuals and the percentage for each species

Cluster 2-1	Group 1		Group 2		Group3		Total
	Number of trees	Percent (%)	Number of trees	Percent (%)	Number of trees	Percent (%)	Number of trees (%)
Redwood	5	50	3	30	2	20	10 (100)
Spruce	8	53	6	40	1	7	15 (100)
Western Hemlock	11	79	3	21	0	0	14 (100)

Table 6. The result of cluster analysis with *Cluster 2-2* using two clusters indicated by the number of individuals and the percentage assigned to each group with the total number of individuals and the percentage for each species

Cluster 2-2	Group 1		Group 2		Total
	Number of trees	Percentage (%)	Number of trees	Percentage (%)	Number of trees (%)
Cedar	12	63	7	37	19 (100)
Pine	15	71	6	29	21 (100)

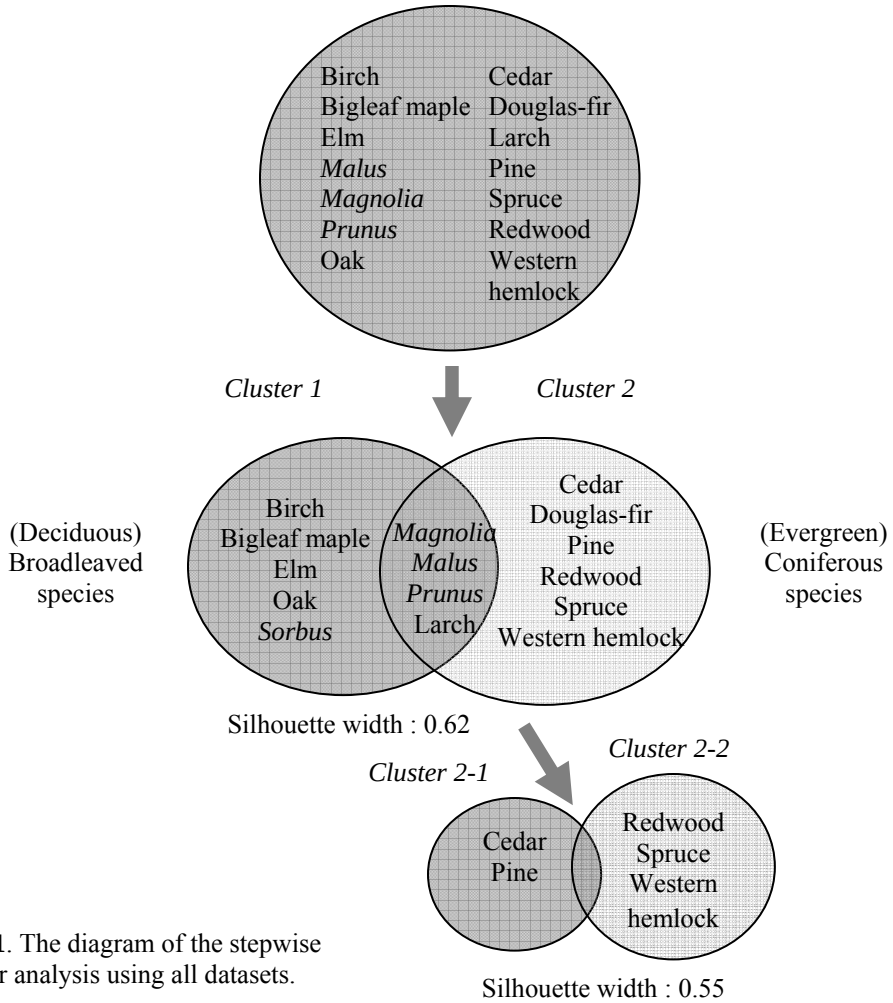


Figure 1. The diagram of the stepwise cluster analysis using all datasets.

Stepwise Cluster Analysis using Leaf-on Data

As a result of PCA, seven variables were selected, including two intensity variables and five height variables. As a result of PAM, the maximal average silhouette width was less than 0.5. Therefore, natural clustering was not found with only leaf-on variables.

Stepwise Cluster Analysis using Leaf-off Data

The variables based on leaf-off data were used and PCA was conducted. ass Two clusters were suggested as the most natural clustering with an average silhouette width, 0.62. The result of cluster analysis using two clusters with the number of individual trees is shown in Table 7. The clustering result looks similar to the result in Table 2 using both leaf-on and leaf-off datasets. Except *Magnolia* and *Prunus*, all species were more clearly clustered into either *Group 1* or *Group 2* than the clustering result using both datasets. All individuals within birch and redwood were assigned to *Group 1* and *Group 2*, respectively while clustering result using both datasets showed there were outliers within these species (see Table 2). All individuals within *Malus* and the majority of *Sorbus* were assigned to *Group 2*, which is different from the clustering result using both datasets where *Sorbus* was clustered into *Group 1* and *Malus* was not clustered into any groups (see Table 2). The majority of individuals within larch were assigned to *Group 1* which is also different from the clustering result using both datasets where larches failed to the clustering criterion (see Table 2).

The result of redistributing individual trees within the same species into a single cluster after deleting species which failed to the clustering criterion is shown in Table 8. *Group 1-1* was composed of species which had no or little foliage at the time of March data acquisition. *Group 1-2* was composed of evergreen coniferous species and one broadleaved species, *Malus*.

Table 7. The result of cluster analysis using leaf-off data indicated by the number of trees and the percentage assigned to each group with the total number of trees and the percentage for each species

Species	Group1		Group 2		Total
	Number of trees	Percent (%)	Number of trees	Percent (%)	Number of trees (%)
Birch	20	100	0	0	20 (100)
Bigleaf maple	11	100	0	0	11 (100)
Elm	10	100	0	0	10 (100)
Magnolia	11	58	8	42	19 (100)
Malus	0	0	10	100	10 (100)
Prunus	4	36	7	64	11(100)
Oak	19	100	0	0	19 (100)
Sorbus	10	91	1	9	11 (100)
Cedar	3	16	16	84	19 (100)
Douglas-fir	0	0	12	100	12 (100)
Larch	18	86	3	14	21 (100)
Pine	0	58	21	42	21 (100)
Redwood	0	0	10	100	10 (100)
Spruce	0	36	15	64	15 (100)
Western Hemlcok	0	0	14	100	14 (100)

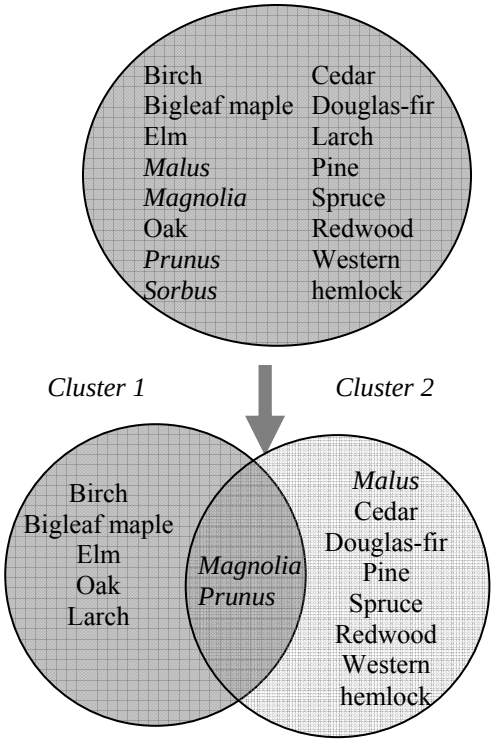
Table 8. The result of redistributing individuals within the same species into a single cluster (*Cluster 1 or Cluster 2*) after deleting species which failed the clustering criterion in leaf-off data

	Cluster 1	Cluster 2
Species	Birch	*Malus*
	Bigleaf maple	Cedar
	Elm	Douglas-fir
	Oak	Pine
	Sorbus	Redwood
	Larch	Spruce
		Western Hemlcok

Clustering result for Cluster 1. As a result of PCA, four variables were selected, including only height variables. The maximal average silhouette width was 0.36 (< 0.5), so natural clustering was not found at this level.

Clustering result for Cluster 2. As a result of PCA, four variables were selected, including only height variables. The maximal average silhouette width was 0.31 (< 0.5), so natural clustering was not found at this level.

The diagram of the stepwise cluster analysis. The overall stepwise cluster analysis using leaf-off data was summarized with diagrams and is shown in Figure 2. The cluster analysis was performed using only one step. Except *Magnolia* and *Prunus*, all species were well divided into two groups. The left side bubble diagram within *Cluster 1* was composed of deciduous species including one deciduous coniferous species, larch, and deciduous broadleaved species while the right side bubble diagram within *Cluster 2* was composed of evergreen coniferous species with one broadleaved species, *Malus* which had foliage at the time of leaf-off data acquisition.



Silhouette width : 0.62

DISCUSSIONS

This study showed that a variety of tree species could be clustered naturally with a hierarchy using height and intensity measurements derived from laser scanning data. Stepwise cluster analysis showed that species with similar characteristics would be clustered into the same group while species with different characteristics would be clustered into different groups based on reliable statistical criteria.

The three stepwise cluster analyses conducted using different seasonal laser scanning datasets showed different results. This implies that tree species might be grouped differently depending on the timing of the data collection. The diagrams generated by the stepwise cluster analysis using all variables based on both leaf-on and leaf-off datasets showed reasonable relationships between species groups at each step, implying that the derived variables described the characteristics of species appropriately. For example, at the first step of stepwise cluster analysis, broadleaved species were mostly separated from coniferous species. This result implies that two clusters are probably the most natural number of clusters when dealing with both broadleaved species and coniferous species. At the next step of the stepwise cluster analysis using coniferous species, a leaf structure was probably the critical factor to divide these species. For example, cedar and pine which have scale-like needles and clustered needles, respectively, were separated from the species with single needles such as spruce, redwood and western hemlock.

This finding is supported by the result that pine and cedar showed lower intensity values than the latter three species in Kim et al. (2009a).

Because the difference between mean intensity values between species was very significant in leaf-off data compared with other variables (Kim et al, 2009a and Kim et al., 2009b), clustering results were probably mostly affected by intensity variables. This finding was also consistent with the result of the principal component analysis. That is, these variables were always selected as the first few principal components which would be critical for the continued cluster analyses. Therefore, the stepwise cluster analysis using only leaf-off variables was similar to the result using both leaf-on and leaf-off variables. However, with leaf-off data, looking at the species assigned to the two separate groups, *Malus* was assigned to *Cluster 2* which was composed of evergreen coniferous species while larch was assigned to *Cluster 1* which was composed of deciduous broadleaved species. This implies that the clustering analysis using only leaf-off data resulted in less natural clustering results than using both datasets where these two species, *Malus* and larch, failed criterion. Also, cluster analysis was conducted using more than one step using both datasets while a single step cluster analysis was conducted using leaf-off data. Therefore, leaf-on data seems to be also useful to do clustering analysis between species groups although the clustering result using only leaf-on data implies that even two species groups, broadleaved species and coniferous species, were not separated naturally. A part of the reasons why cluster analysis using only leaf-on data is not successful is not only due to a seasonal issue but also due to other factors. Because both datasets were acquired from different laser scanner systems with different flight parameters, for example, leaf-on laser scanning data were acquired by smaller numbers of point density, smaller numbers of returns per pulse, and lower scan pulse repetition frequency than leaf-off laser scanning data, characteristics of species are probably better described using leaf-off data than using leaf-on data.

At each step of the cluster analysis, the criterion using a certain percentage of individual trees within species was applied because individual trees within the same species are not always in the same conditions. Age and competition with neighboring trees probably affect the shape of a tree. Non-native species within genus may affect different characteristics of these species. For example, pine included three individual trees within one native species, western white pine, while the rest individuals were not native species which have different needle and crown shapes. Oak included one native species, Oregon white oak, while the rest individual trees were not native species. If the number of sample trees per each species increased, the accuracy of clustering results could be improved. Especially, *Magnolia* and *Prunus* were composed of a variety of species within genus from worldwide collections and so, foliage conditions also varied, for example, some individual trees had foliage with only leaves, others had foliage with leaves and flowers and the rest had no foliage at the time of data collection in March. This is probably one reason why they failed criterion in leaf-off data.

The result of cluster analysis using PAM varies depending on the species and the variables used. Depending on the determined number of clusters, species would be clustered differently, too. Therefore, the clustering results shown in this study don't suggest any absolute separation between species. Instead, the stepwise cluster analysis introduced in this study suggests the possibility of natural clustering for various tree species based on their structural and spectral characteristics.

REFERENCES

Ahokas, E.,Yu, X.,Oksanen, J.,Hyyppäe, J., Kaartinen,H., and H. Hyyppäe, 2005. Optimisation of the scanning angle for countrywide laser scanning. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *36* Part 3/W19.

Brandtberg, T., T. Warner, R. E. Landenberger and J. B. McGraw, 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. and T. L. Webster, 2006. Object-oriented land cover classification of lidar-derived surfaces. *Canadian Journal of Remote Sensing,* 32(2), 162-172.

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

Donoghue, D.N.M. , P. J. Watt, N. J. Cox and J. Wilson, 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., and G. Dunn, 2001. *Applied multivariate data analysis*: Second edition, Oxford University Press Inc., New York.

Fisher, R.A., **1936**. The use of multiple measurements in taxonomic problems, *Annals of Eugenics,* 7, pp. 179-188.

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

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

Huberty, Carl J. and S. Olejnik, 2006. *Applied MANOVA and Discriminant Analysis*: Second Edition. Wiley Series in Probability and Statistics.

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

Kaufman, Leonard and P.J. Rousseeuw. 1990. Finding groups in data: an introduction to cluster analysis. Wiley, New York.

Kim, Sooyoung, R. McGaughey, H. - E. Andersen, and G. Schreuder, 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, Sooyoung, T. Hinckley, and D. Briggs, 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.

McGaughey, R. J. and W. W. Carson, 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.

MacQueen, J. B. 1967. Some Methods for classification and Analysis of Multivariate bservations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297.

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

Moffiet, T., K .Mengersen, C. Witte, R. King, and R. Denham, 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., and B. Allgower, 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.

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

Petrides, George A. and O. Petrides., 1992. Western trees. Peterson field guides. Houghton Mifflin company.

Popescu , S.C. and R.H. Wynne, 2004. Seeing the trees in the forest: using lidar and multispectral data fusion with local filtering and variable window size for estimating tree height. *Photogrammetric Engineering & Remote Sensing*, *70*(5): 589-604.

Roberts, D. A., S. L. Ustin, S. Ogunjemiyo, J. Greenberg, S. Z. Dobrowski, J. Chen, and T. M. Hinckley, 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 (1987) 53-65. North- Holland.

Venables, W. N. and B.D. Ripley, 1994. *Modern Applied Statistics with S-PLUS.* Springer- Verlag, New York.