

Article

Comparative Analysis of Clustering-Based Approaches for 3-D Single Tree Detection Using Airborne Fullwave Lidar Data

Sandeep Gupta *, Holger Weinacker and Barbara Koch

Department of Remote Sensing and Landscape Information System (FeLis), Faculty of Forestry, Albert-Ludwigs University, Tennenbacher str. 4, 79106 Freiburg, i.Br., Germany;

E-Mails: holger.weinacker@felis.uni-freiburg.de (H.W.); barbara.koch@felis.uni-freiburg.de (B.K.)

* Author to whom correspondence should be addressed; E-Mail: sandeep.gupta@felis.uni-freiburg.de; Tel.: +49-761-203-3698; Fax: +49-761-203-3701.

Received: 18 January 2010; in revised form: 20 February 2010 / Accepted: 27 February 2010 /

Published: 1 April 2010

Abstract: In the past, many algorithms have been applied for three-dimensional (3-D) single tree extraction using Airborne Laser Scanner (ALS) data. Clustering based algorithms are widely used in different applications but rarely being they used in the field of forestry using ALS data as an input. In this paper, a comparative qualitative study was conducted using the iterative partitioning and hierarchical clustering based mechanisms and full waveform ALS data as an input to extract the individual trees/tree crowns in their most appropriate shape. The full waveform Light Detection And Ranging (LIDAR) data was collected from the Waldkirch black forest area in the south-western part of Germany in August 2005 with density of 4–5 points/m². Both the clustering algorithms were used in their original and modified form for a comparative qualitative analysis of the results obtained in the form of individual clusters containing 3-D points for each tree/tree crown. A total of 378 trees were found in all the 1.2 ha area with height ranging from 15 m to 50.9 m. The forest contains mainly older trees with deciduous, coniferous and mixed stands. The findings showed that among the three kind of clustering methods applied (normal *k*-means, modified *k*-means and hierarchical clustering), the modified *k*-means algorithm using external seed points and scaling down the height for initialization of the clustering process was the most promising method for the extraction of clusters of individual trees/tree crowns. A 3-D reconstruction of extracted individual clusters was carried out using QHull algorithm. In this study, the result was not possible to validate quantitatively due to lack of the field inventory data.

Keywords: LIDAR; clustering; pattern; *k*-means; hierarchical; 3-D; tree crown; QHull

1. Introduction

Over the last decade, the usage of Airborne Laser Scanner (ALS) data by applying different algorithms for three-dimensional (3-D) single tree extraction has been commonly exploited in the field of forestry in order to minimize the traditional forest inventory practices which are very time, manpower and cost consuming. There has been a multifold increase in the demand for single tree related information for more precise estimation of biophysical parameters, forest management and environmental planning practices. This was the main motivation factor behind research work to test the ALS data for the extraction of pattern of single tree crowns using clustering based methodologies. The full waveform digitized data from latest laser scanners are capable of giving more information as compare to the traditional ones. The multiple return echoes of the full waveform laser data with high point density scattered from the tree crown is helpful for the purposes of single tree crown delineation. Clustering, in pattern recognition, is the process of partitioning a set of pattern vectors into subsets called 'clusters'. For example, if the pattern vectors are pairs of real numbers which can be viewed as points, then clustering consist of finding subsets of points that are 'closed' to each other with supported space partitioning measures (for example, Euclidean distance matrix) in *n*-dimension. There are different types of clustering algorithms which are useful in image segmentation and partitioning the clusters, for example, *k*-means, fuzzy C-means and hierarchical tree clustering. Because of the high point density full waveform LIght Detection And Ranging (LIDAR) data provide a good platform to implement the clustering mechanisms via partitioning the data into individual clusters of single tree/tree crown cover.

In the past 10 years, the demand for high quality LIDAR data with more information has tremendously increased for various applications. Work has been carried out in the past using LIDAR data for 3-D vegetation and related information extraction using various methods [1-4]. For single tree delineation, the majority of the existed algorithms are Digital Surface Model (DSM) based [5-7]. Apart from individual tree detection, methods for reconstructing the tree crowns were also provided [8]. With the pre-knowledge about the locations and the crown sizes of each tree, they extracted the raw points belonging to the tree, tree heights, and the average radius of the crowns at different heights were derived [3]. A new method for 3-D single tree crown contour extraction at different height levels using hierarchical morphological processes has been presented [3]. It was showed that the new full waveform LIDAR data significantly improve the detection rate of single trees using a 3D segmentation technique based on the normalized cut segmentation method [9]. Very little work has been reported using clustering based approaches for 3-D single tree extraction using airborne LIDAR data [10-15]. First and last pulse data with an overall density of 30 points m^{-2} and the *k*-means method to extract single tree in the Swiss National Park were used [10]. The authors in [10] used local maxima derived from smoothed DSM as seed points. In contrast to the modified algorithm presented below, instead of scaling-down the z-coordinates, they scaled-up z-value by 3. For this they argued that they did it to accommodate the aspect ratio of pine tree crowns, which ranged from 3 to 6, based on the field data. However, here it is important to note that by scaling down the height value (z-coordinates) of the normalized raw points as

well as seed points ‘height-wise’ in the space helps in minimizing the intra-cluster variance or the squared error function which is the ultimate objective of the k -means method. The closer the points will be, the more precise will be the cluster formation with regard to actual tree/tree crown and its shape. This fact was applied in this paper while modifying the algorithm. In another study conducted for deciduous-coniferous classification using single leaf-on full waveform LIDAR data, branches of 27 coniferous and 38 deciduous trees were derived by calculating the mean silhouette values repeatedly for different k values using simple k -means clustering for improved visualization [14]. The method lacks the efficiency for finding the suitable value of k with respect to different tree types, tree age and forest conditions. In their study the result is not validated using any field data. In another approach, supervised classification strategy using linear discriminant analysis, random forest algorithm and support vector machines for tree species classification was tested using first, single and last echoes ALS data [15]. In their study, unsupervised k -means clustering and k -means clustering in combination with the unsupervised random forest algorithm was used for species classification. However, their result showed that accuracies were lower in case of unsupervised one than for supervised methods applied for overall species classification. This shows that supervised methods are more promising which was found true during the investigation after a comparative qualitative analysis of the output.

Multi-tier single tree crown extraction within single clustering process was also applied in a test site to assess its feasibility which is lacking in all the previous studies conducted through different clustering based approaches [10-15]. In the study area, different tree types, multistoried trees, tree density, tree crown density, gap variation, presented a big challenge to extract individual trees and associated information. The multiple return pulses from single emitted laser pulse with additional information like amplitude and pulse width besides the 3-D coordinates provides relevant and quality information for the user community. However, these parameters were not used in the present study.

A lot of research work has been carried out for single tree/tree crown extraction using airborne LIDAR data with one or a combination of method(s). But, very little research work has been carried out using raw airborne LIDAR data and clustering based mechanisms. Different clustering algorithms are widely used for many applications but this attempt is unique. For the first time such a comparative analysis of different clustering algorithms was qualitatively analyzed using leaf-on raw full waveform LIDAR data for single tree/tree crown extraction and the outcomes has been presented in this paper. The main objective of this paper is to present a comparative qualitative analysis of the different clustering algorithms for the extraction of clusters of individual trees/tree crowns using airborne full waveform LIDAR data. The agglomerative hierarchical tree algorithm (bottom-up) and most commonly used k -means algorithm, an iterative partitioning based top-down approach, were used and presented in this paper. The 3-D reconstruction of the extracted individual cluster points representing individual trees/tree crowns shape into convex polytope was also carried out using the QHull algorithm. However, in this study, the result was not possible to validate quantitatively due to the lack of field inventory data.

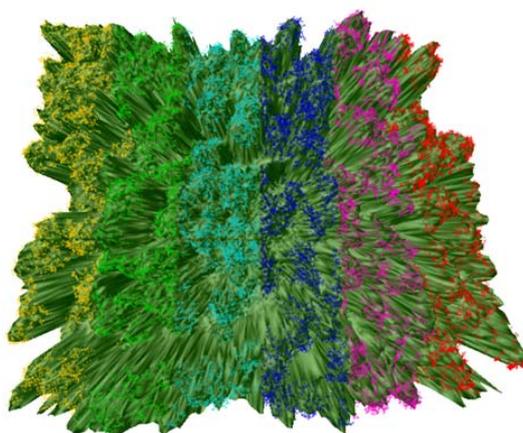
2. Materials and Methodology

2.1. LIDAR Data

For the proposed feature extraction strategy, airborne full waveform LIDAR data with multiple return echoes was acquired during August 2005 in the Waldkirch forest area of the

Baden-Württemberg region of Germany (Figure 1). The LiteMapper-5600 LIDAR system, which uses the RIEGL LMS-Q560 laser scanner with a scan angle range of ± 30 degrees, was used during the data acquisition. The density of the data is 4–5 points m^{-2} and the flying height of the aircraft was 400 m above mean ground level. More information about the LIDAR RIEGL LMS-Q560 has been presented by other authors [16,17].

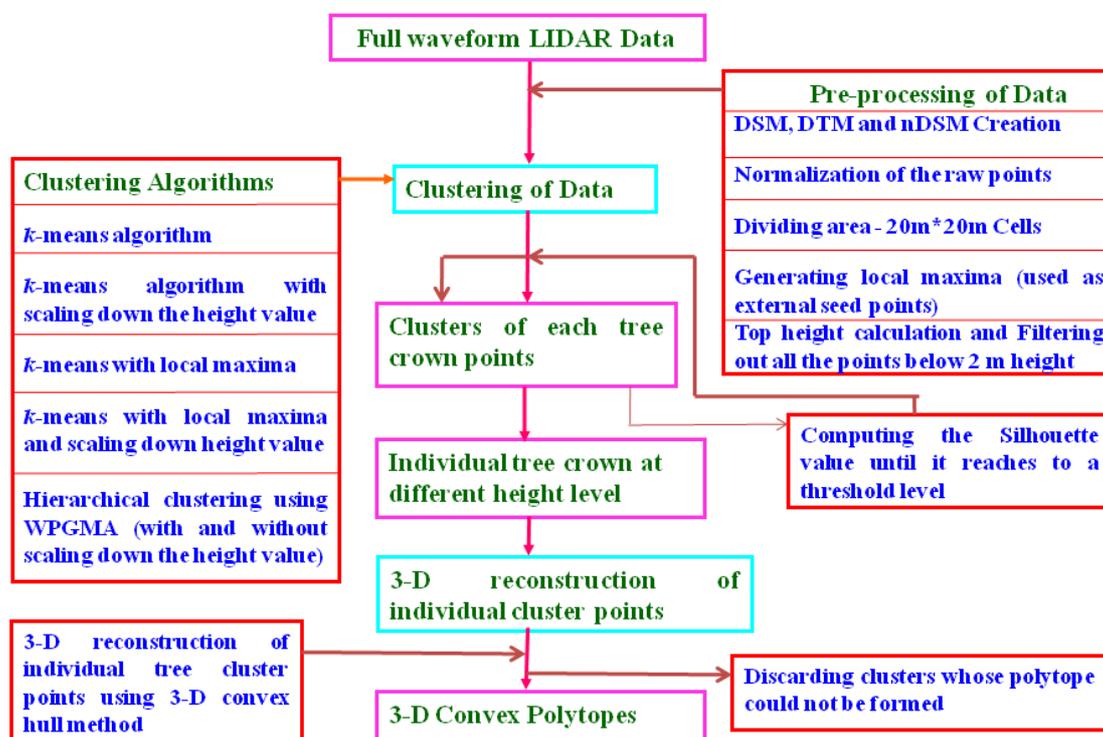
Figure 1. LIDAR raw data overlaid on DSM.



2.2. Methodology Flow Chart

A flow chart showing the overview of methodology adopted in the whole process to fulfill the aims and objectives of the study is presented in Figure 2. The figure shows the pre-processing, main process and different clustering algorithms applied during the processing steps and the output of the process.

Figure 2. Methodology flow chart.

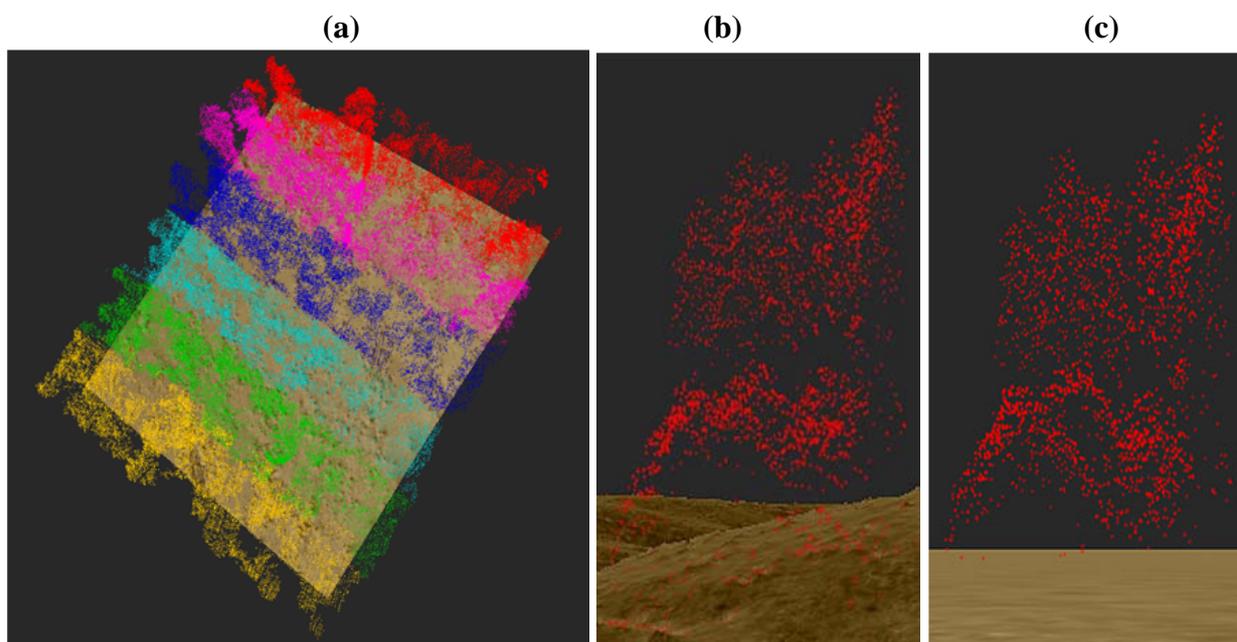


2.3. Generation of DSM, DTM and nDSM and Normalization of Raw LIDAR Data

A raster DSM and Digital Terrain Model (DTM) were generated from the full waveform LIDAR raw point clouds by TreesVis—a software for LIDAR data processing and visualization [18]. Normalized DSM (nDSM) was generated by subtracting the gray values of DTM from DSM in TreesVis. DSM, DTM and nDSM were used as an input at various preprocessing steps.

Normalization of raw LIDAR point clouds was carried out in order to get the absolute height above ground of each point and to eliminate the influence of the terrain. Raw points were projected above the DTM [Figure 3a and Figure 3b)]. The height difference between a raw point and its correspondent terrain is marked as the normalized height of the point [Figure 3c]. The height of each normalized point represents the absolute height of the respective point.

Figure 3. LIDAR raw and normalized points. (a) LIDAR raw points of the whole test area projected above DTM. (b) LIDAR raw points above DTM in a closed view. (c) Normalized points above zero height in a closed view.



2.4. Study Area and Its Subdivision

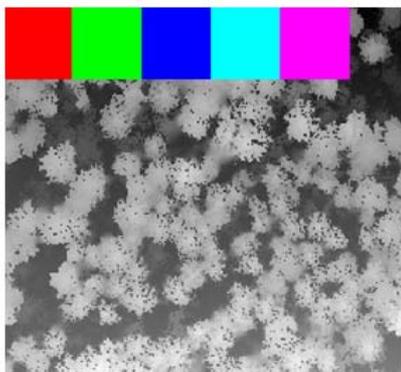
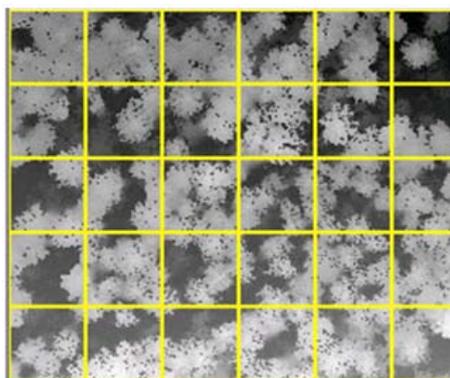
The normalized LIDAR data was used for the investigation of the study area. The coordinates of the upper left corner of the study area in Gauss-Krüger are 3424043 (Easting), 5328613 (Northing) and the unit of the coordinate values are in meters. The total forest area selected is 1.2 ha. The forest mainly contains mature and old trees. The dominant tree types are Scots pine (*Pinus sylvestris*), oak (*Quercus sp.*), European beech (*Fagus sylvatica*), silver birch (*Betula pendula*), Norway spruce (*Picea abies*), hornbeam (*Carpinus betulus*) and few minor species.

The tree top height ranged between the 37 and 52 meters (Table 1). The top height is the peak height of the 100 strongest rises that defined trees with the largest diameter in one hectare [19]. Because trees with the largest diameter are usually the tallest trees, the top height represents the height of trees in the upper canopy level which can be modeled with ALS data.

Table 1. Top height (in meters) of 30 cells of 20 m × 20 m size each.

TopHeight [m]					
Cell1- 39.79	Cell2- 42.78	Cell3- 44.89	Cell4- 49.73	Cell5- 47.47	Cell6- 38.17
Cell7- 39.35	Cell8- 40.41	Cell9- 46.01	Cell10- 51.57	Cell11- 50.63	Cell12- 41.64
Cell13- 39.31	Cell14- 41.23	Cell15- 44.58	Cell16- 43.72	Cell17- 39.54	Cell18- 39.32
Cell19- 37.27	Cell20- 39.96	Cell21- 41.02	Cell22- 39.30	Cell23- 39.39	Cell24- 39.38
Cell25- 37.92	Cell26- 43.23	Cell27- 41.98	Cell28- 40.5	Cell29- 38.45	Cell30- 37.14

There is a large variation in terrain, tree type, tree density, tree crown density of the forest area. The total study area was therefore divided into 30 grid cells of 20 m × 20 m area each to obtain more reliable information, reduce the computation run time and simplify the complexity (Figures 4 and 5).

Figure 4. Subdivision of study area (in progress) into a 20 m × 20 m grid.**Figure 5.** Divided cells (30) in yellow color overlaid on the DSM.

With an area of each cell of 400 m², the four highest trees in each grid were found. Mean height of those trees were calculated as a top height in each cell using the following equation:

$$\text{TopHeight} = (\text{Max}_1 + \text{Max}_2 + \dots + \text{Max}_n)/n \quad (1)$$

Where, TopHeight is defined as the mean peak height among all the high rise trees in each grid. Max₁, Max₂, ..., Max_n are the maximum heights of all the selected highest trees present in each grid and *n* is the number of high rise trees present in each grid.

2.5. Clustering

Iterative partitioning and hierarchical tree based approaches for extracting the clusters of individual trees/tree crowns were used in the present study. A short description is given below.

2.5.1. Iterative Partitioning

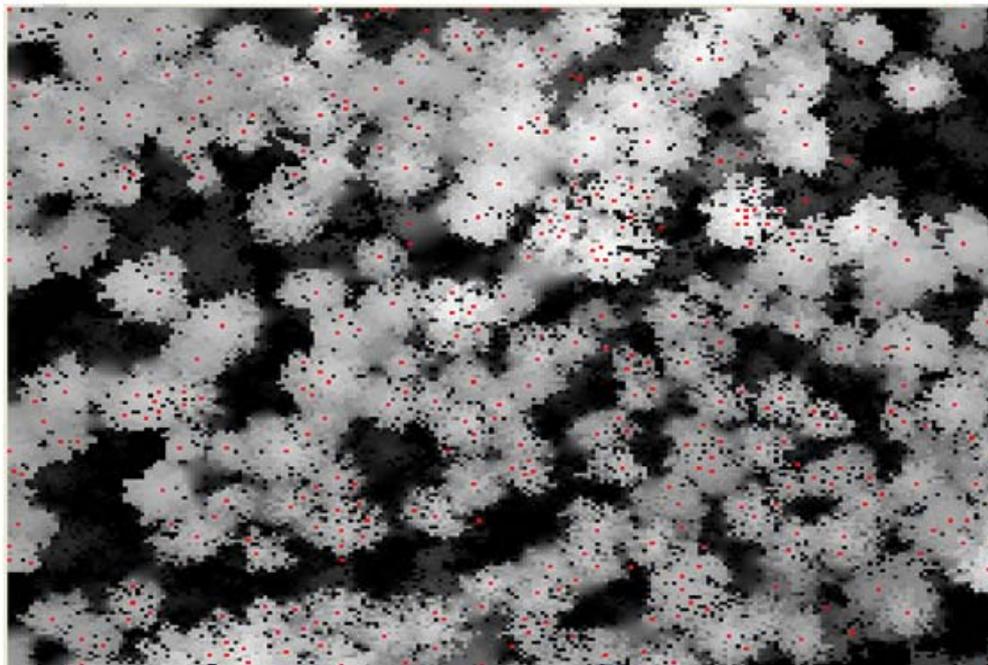
The k -means algorithm was chosen because it is an iterative hill-climbing method and is a staple of clustering methods. In the first approach, the k clusters were chosen using the default random seed points. In the second approach, k clusters were chosen using local maxima (extracted as all points from the nDSM image having a gray value larger than the gray value of all its neighbors and returned as an output, Figure 6) as external seed points. The selection of random seed points in the first approach was kept equal to the external seed points obtained through second approach. This was done in order to maintain an equal number of clusters, each representing an individual tree/tree crown, and carrying out a comparative analysis about the approximate shape of the trees/tree crowns. Additionally, in both the approaches quality of output (tree clusters) was analyzed by scaling down the height value. In this, the value of z-coordinates of raw normalized LIDAR points as well as external seed points is reduced before initialization of the k -means process. The logic behind scaling down the height value of the points is that it helps in minimizing the intra-cluster variance, or, the squared error function (S_e) which is the ultimate objective of the k -means method. This is also important in viewing the natural dataset used where pattern recognition is not a simple task. In k -means algorithm, the sum of absolute differences in Euclidian distance between each point and its closest center is minimized. The squared error function (S_e) is given by the following equation:

$$S_e = \sum_{i=1}^k \sum_{j=1}^n |x_j^i - c_i|^2 \quad (2)$$

where $|x_j^{(i)} - c_i|^2$ is a chosen distance measure between a data point $x_j^{(i)}$ and the mean vector or cluster centre c_i , is an indicator of the distance of the n data points from their respective cluster centers. The value of S_e depends on how the samples are grouped into clusters and the number of k clusters. At every iteration, new cluster memberships and new cluster centers are computed, and it is guaranteed that the goodness measure S_e can only decrease with each iteration. In the optimal partitioning, S_e is minimized. Since, here S_e directly depends on distance measure which is computed in Euclidean 3-D space, so z-factor is directly linked with the S_e minimization process. The more 'closer' the points are 'height-wise', the more sound is grouping of neighboring points resulting in optimal partitioning of the sample data in to a desired k clusters.

The k -means can converge to a local optimum, in this case, a partition of points in which moving any single point to a different cluster increases the total sum of distances. Fortunately, this problem was solved by providing local maxima as seed point in the second approach and scaling down the height value to a threshold level. Thus, in the k -means algorithm, the number of seed points chosen has the direct relationship over the number of tree clusters to be generated (quantity).

Figure 6. Extracted local maxima overlaid on DSM (red colored points).



2.5.2. Hierarchical Tree Method

Apart from clustering through k -means there exist several methods, one of which is constructing clusters from the hierarchical cluster tree. It starts with each sample as a singleton (a new cluster consisting of the one point furthest from its centroid) cluster and iteratively merge clusters that are most similar according to chosen similarity or distance measures. The clusters are generated as the maximum number of clusters to form from the hierarchical tree in a matrix of size $(m-1)$ -by-3 as an output, where m is the number of observations in the original data. The leaf nodes (objects at the bottom of the hierarchical cluster tree) in the cluster hierarchy are the objects in the original data set, numbered from 1 to m .

For the creation of hierarchical cluster tree, the weighted average distance algorithm or also known as **w**eighted **p**air-**g**roup **m**ethod using arithmetic averages (WPGMA), one type of agglomerative or bottom-up algorithm [20] was used. This algorithm is based on different ways of measuring the distance between two clusters of objects. The distance is computed between objects in the data matrix using the standardized Euclidean distance. Each coordinate in the sum of squares is inverse weighted by the sample variances of that coordinate. The weighted average distance algorithm compute the distance by measuring the distance between two clusters of objects by weighting the member most recently admitted to a cluster equal with all previous members. Thus, the size of the respective clusters (*i.e.*, the number of objects contained in them) is used as a weight in this algorithm.

2.6. 3-D Reconstruction of Individual Tree Clusters

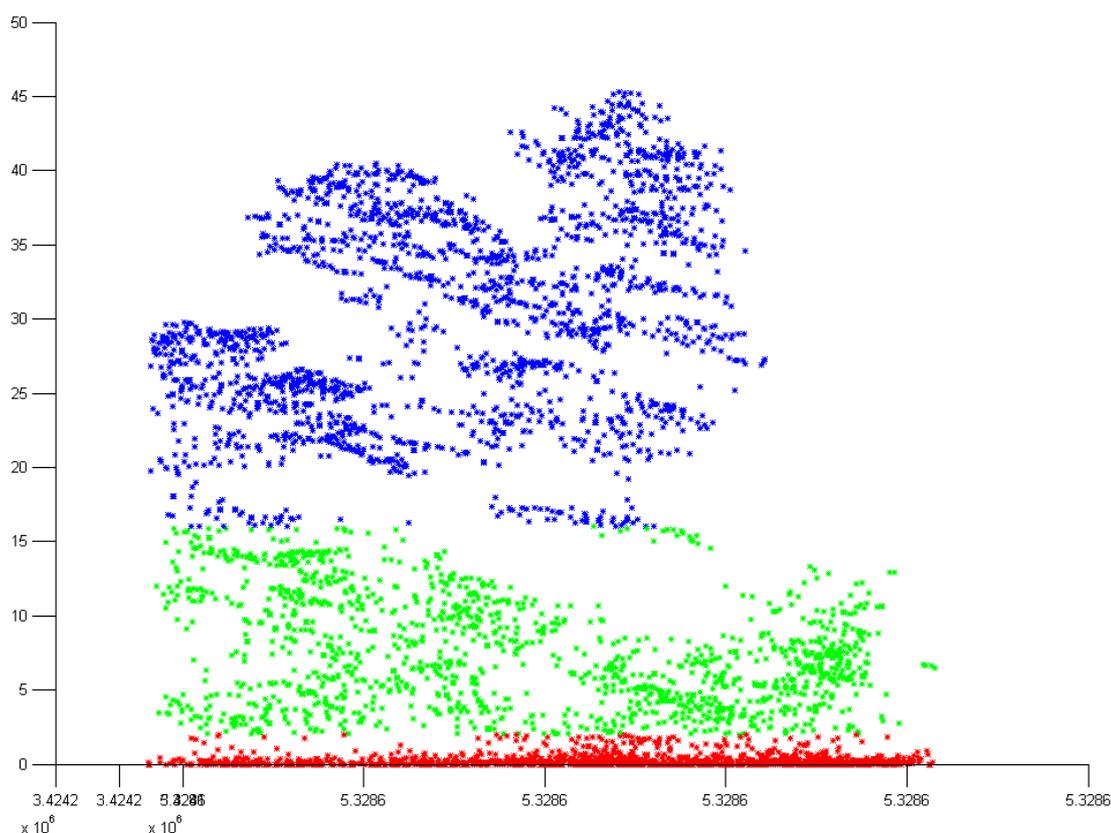
The 3-D reconstruction of clusters representing the individual tree crown was carried out using the QHull algorithm [21]. QHull is a general dimension code for computing convex hulls using Quickhull

algorithm. QHull may be used for n-dimensions. A detail description about the convex hull is also given [22]. Finally, each tree crown is shown as a 3-D convex polytope with triangular surface.

3. Results and Discussion

Figure 7, which corresponds to a sample dataset (Cell 6 with top height 38.17 m), shows the distribution of normalized 3-D LIDAR points between 0 and 2 m (red), between 2 and 16 m (green) and above 16 m (blue) height, respectively. The normalized raw LIDAR data was filtered below a certain threshold height value so as to minimize the disturbances during the clustering process. The return laser hits from lower and ground vegetation interferes in clustering process, so the points below 2 m were filtered out before the processing. While applying *k*-means algorithm, following four different criteria to assess the point distribution of individual trees in 3-D space through manual visualization was considered.

Figure 7. Cell 6 - distribution of normalized 3-D LIDAR points, projected into a freely chosen vertical plane, at 3 different height levels (0–2 m, 2–16 m, and above 16 m) shown in 3 different colors (red, green and blue, respectively). The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.



3.1. Normal *k*-means (*N k*-means)

In this case, the seeds for initialization of the process are randomly selected by the algorithm to minimize the error from the point in a cluster to its centroid. However, the number of random seed points was kept equal to the number of external seed points obtained via modified *k*-means approach.

This was done in order to maintain an equal number of clusters and carrying out a comparative qualitative analysis about the approximate shape of the trees/tree crowns.

3.1.1. Without Scaling Down the Height Value

In the given example (Cell 6) as displayed in Figure 7, clustering was performed on the LIDAR data selected from different height levels. In the first dataset, the normalized data points were split into two height classes (from 2–16 m and above 16 m height) after visual inspection. In the second dataset the normalized data points selected which were only at above 16 m height. This was done because the returned laser hits were from dense forest unlike the case of open forest where there is enough gap to separate the LIDAR points of individual tree. Since the top height for Cell 6 was 38.17 m, hence, it was assumed that there were very less chances of getting the low height trees at first-tier. But, for a better partitioning, the data in Cell 6 were split at height value of 16 meters after visual check. The *k*-means algorithm was carried out on both the datasets and the result after partitioning has been presented in Figure 8a) and Figure 8b), respectively. The black circular dots and the rectangles shown in the Figure 8a) and Figure 8b) are cluster centroids. It is clear from the Figure 8 that the normal *k*-means failed to separate the clusters of points from natural forest conditions in both the datasets.

Figure 8. Result after running N *k*-means without scaling down the height value on two different datasets of Cell 6 at different height levels. (a) Cell 6—clusters from N *k*-means in 2 height classes (between 2 and 16 m and above 16 m). (b) Cell 6—clusters from N *k*-means above 16 m height. The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.

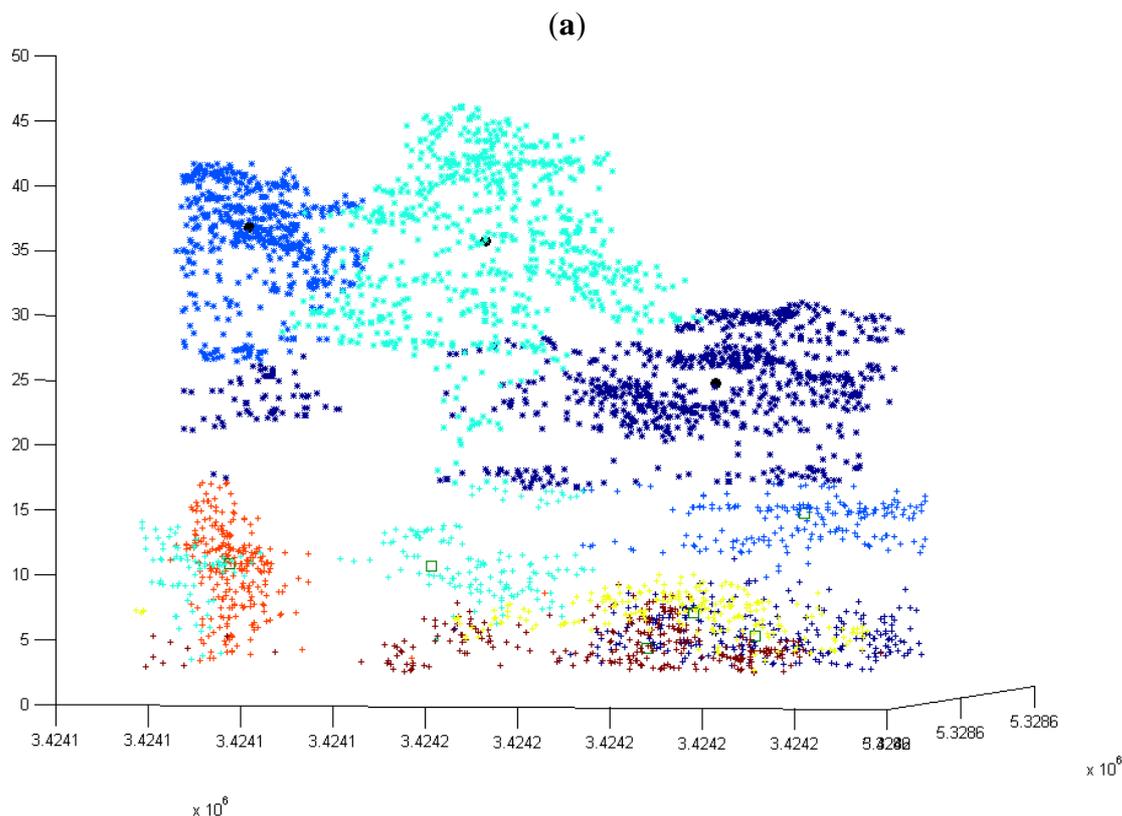
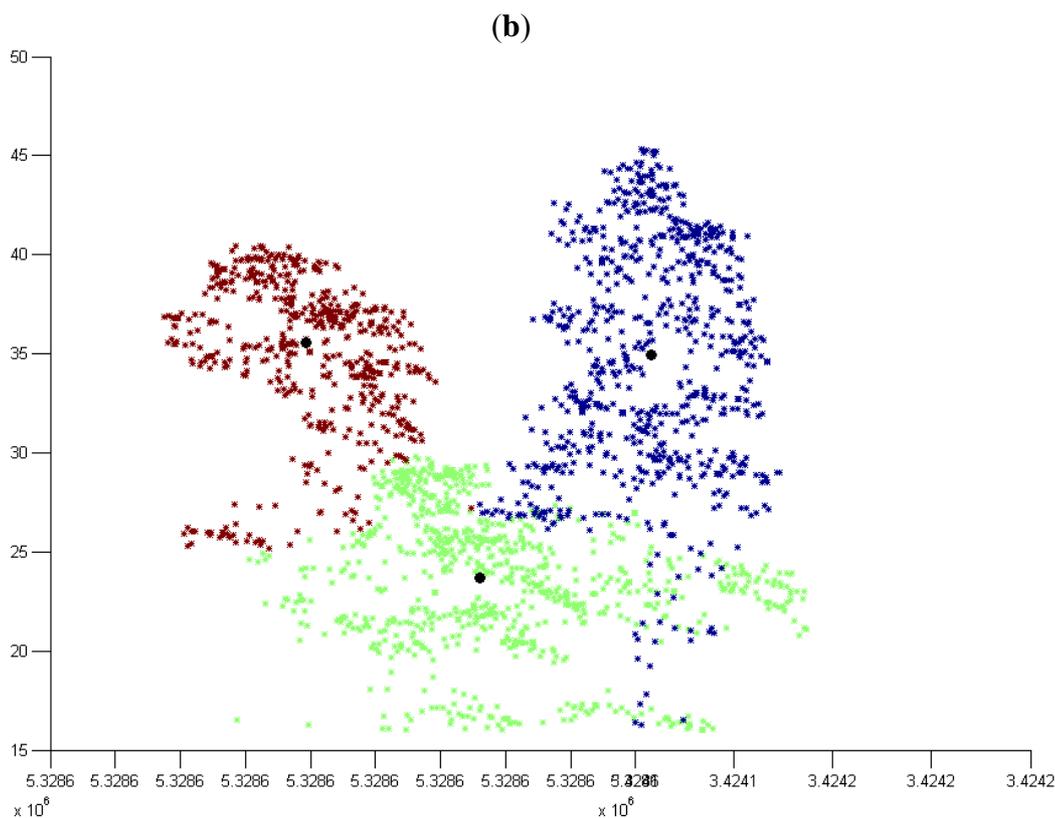


Figure 8. Cont.



3.1.2. By Scaling Down the Height Value

The height value of the both the datasets of Cell 6 (between 2 and 16 m, and above 16 m) and respective seed points was scaled down to a threshold level before initialization of the k -means algorithm. The threshold value for scaling down the height was chosen after a trial and error approach. The value was kept constant for all the grid cells. By scaling down the height value, points came close to each other in z -direction. In this way, the basic idea behind the k -means algorithm in its modified form was fully utilized for better partitioning of the data and cluster formation. Once the algorithm was completed, the height value of the clustered data was scaled-up to its original. Figure 9a shows the improvement in pattern classification of the first dataset in its upper layer, but no considerable improvement was obtained in the lower layer. In the second dataset, where there are points only above 16 m height, a more accurately classified result was obtained when compared to the upper layer of the first dataset [Figure 9b].

As an example, a cluster of an individual tree crown above 16 m is shown in Figure 10a. This tree cluster has been extracted by applying a N k -means approach without scaling down the height value. The corresponding 3-D convex hull is shown in Figure 10b. Similarly, a cluster containing the point cloud of an individual tree crown above 16 m by applying N k -means approach with scaling down the height value and the corresponding 3-D convex hull has been shown in Figure 10c and Figure 10d, respectively. It is apparent from Figure 10b and Figure 10d that the shape of the 3-D polytope is considerably different. It supports the view that the scaling down the height value of the data as well as seed points before initialization influences the clustering. The shape of the extracted clusters of the same tree crown can be visualized from Figure 10a and Figure 10c.

Figure 9. Result after running N *k*-means by scaling down the height value on two datasets of Cell 6 at different height levels. **(a)** Cell 6 – tree clusters from N *k*-means in two height classes (between 2 and 16 m and above 16 m). **(b)** Cell 6 – tree clusters from N *k*-means above 16 m height. The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.

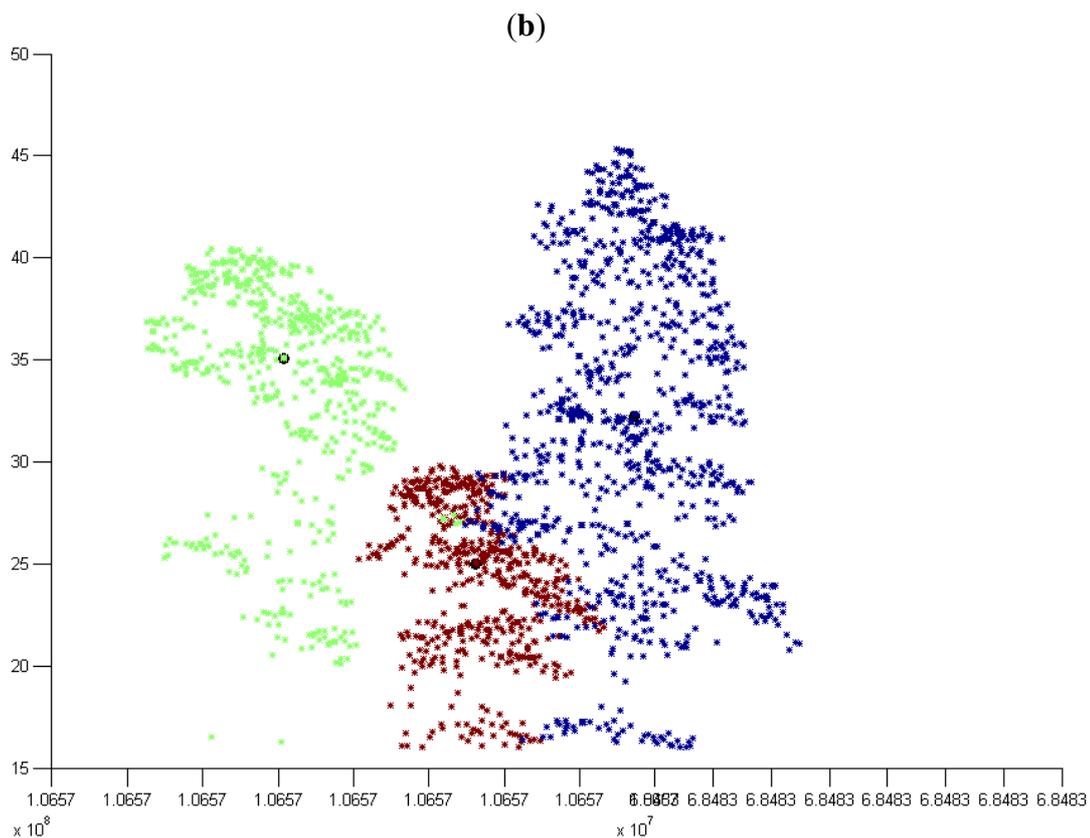
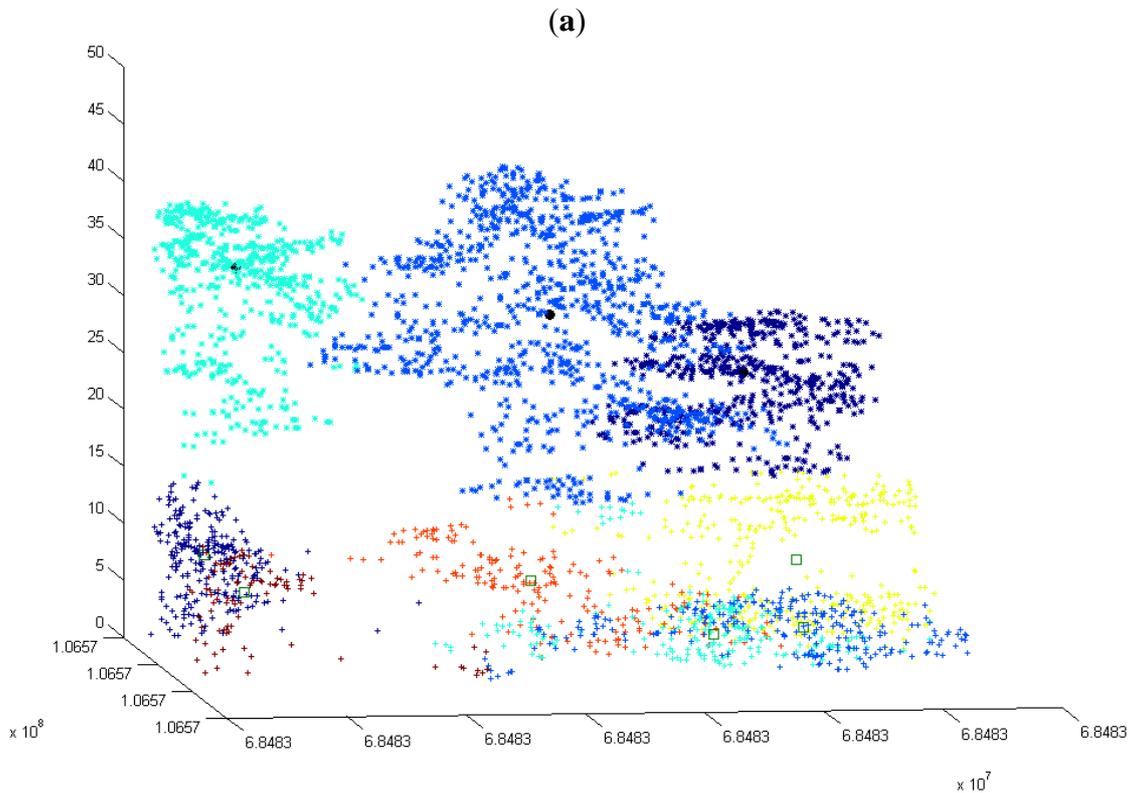
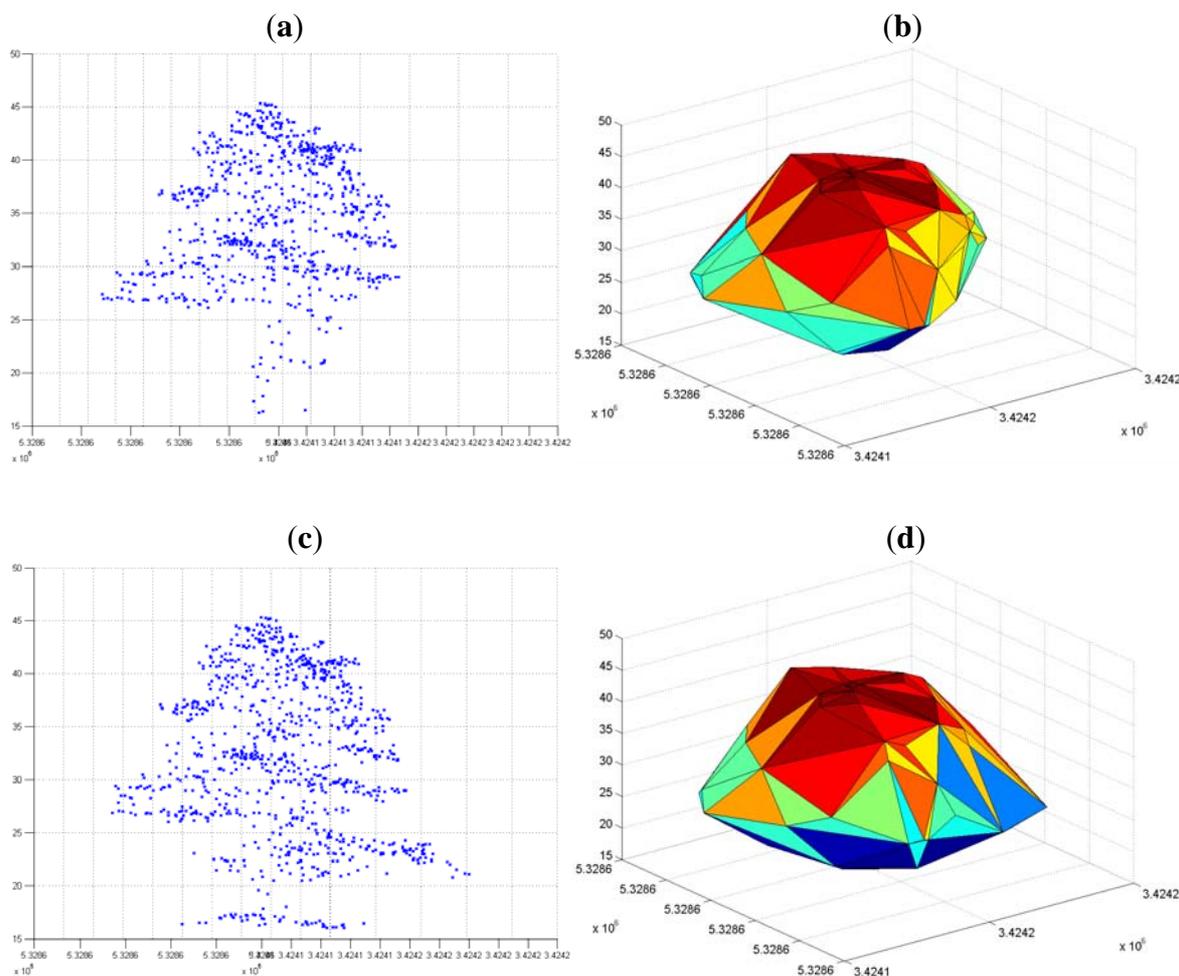


Figure 10. Cluster of an individual tree from Cell 6 after running N k -means on the dataset above 16 m height and respective convex polytope, projected into a freely chosen vertical plane. **(a)** Cell 6—an individual tree cluster by applying N k -means without scaling down the height value. **(b)** 3-D convex polytope reconstructed from an individual tree cluster as shown in **(a)**. **(c)** Cell 6—an individual tree cluster by applying N k -means after scaling down the height value. **(d)** 3-D convex polytope reconstructed from an individual tree cluster as shown in **(c)**. The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.



3.2. Modified k -means (M k -means) with External Seed Points

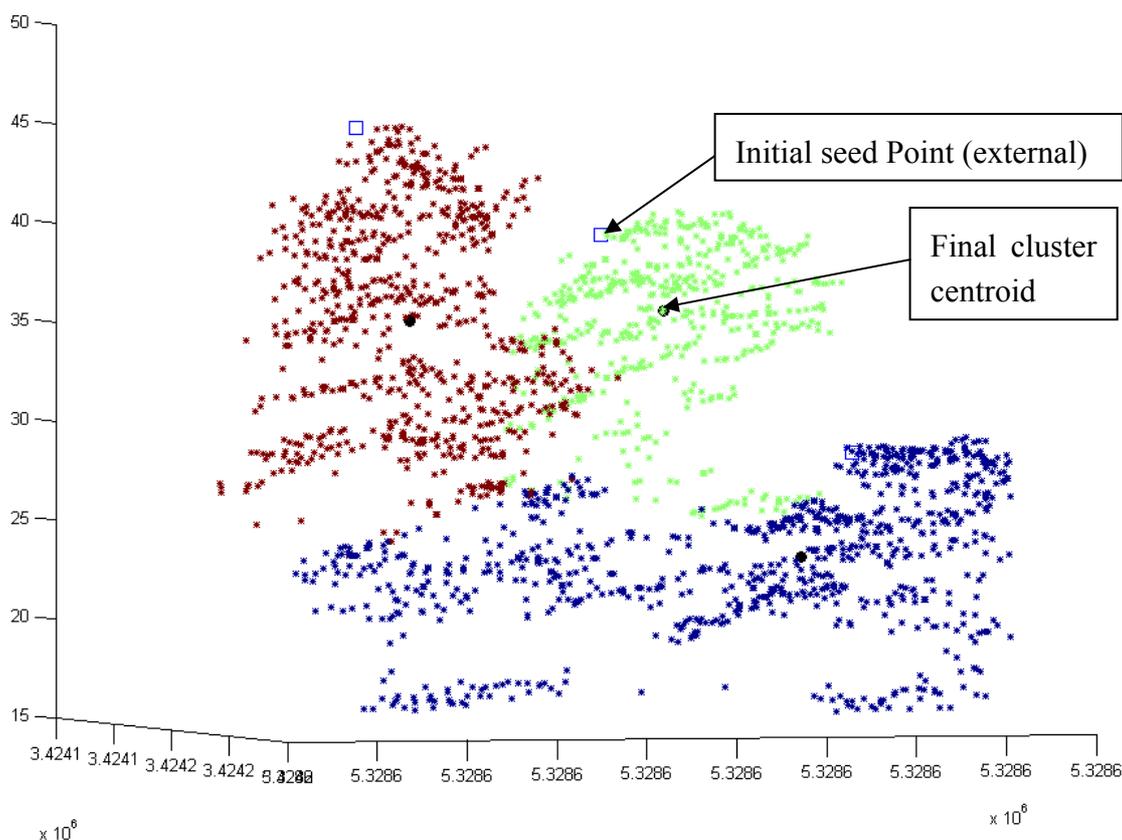
In this case, instead of random seed points, the local maxima derived from DSM were used as external seed points in the algorithm to minimize the error from the point in a cluster to its centroid.

3.2.1. Without Scaling Down the Height Value

It was hypothesized by the authors that due to the use of external seed points for the initializing of the k -means algorithm, it was better to perform clustering using M k -means only on datasets containing points above some threshold height (for example, 5–16 m and above 16 m in the test area). This is due to the fact that return hits were from natural forest conditions containing mainly older trees

with few at the pole or mature stages and to reduce the errors caused by low ground vegetation. Partitioning of the whole dataset into individual clusters using local maxima as external seed points was found much better (Figure 11) as compare to the result obtained through N *k*-means algorithm (Figure 8b) without scaling down the height value of the data points. However, the clustering was not as good when compared to the clusters presented in Figure 9 (b), where the height value was scaled down before clustering.

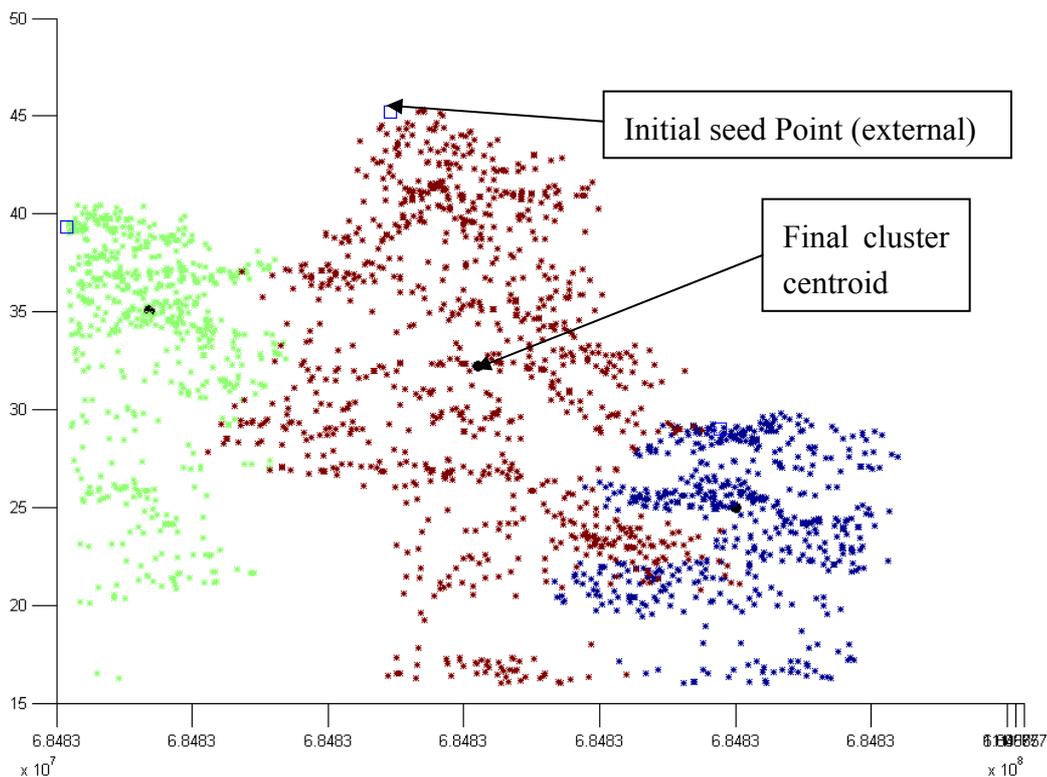
Figure 11. Result after running M *k*-means without scaling down the height value on Cell 6 datasets of height above 16 m. The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.



3.2.2. By Scaling Down the Height Value

Figure 12 shows the tree clusters (above 16 m height) obtained by running *k*-means while using local maxima as external seed points (M *k*-means) and scaling down the height value of both the LIDAR data points and seed points. This method for partitioning the data in to individual clusters representing single trees was found as the best method compared to all the above *k*-means options (Figure 8b, Figure 9b and Figure 11). The fact is *k*-means is a top-down approach. Feeding of local maxima points as external seeds, before initializing the algorithm, determined the number of clusters to be formed. Scaling down the height value brought the points closer for distance measurement and to minimize the bias during the run-time. Combining these two factors and using them in the *k*-means algorithm, generated individual clusters of each trees/tree crowns in their most appropriate form.

Figure 12. Result after running M *k*-means after scaling down the height value on Cell 6 datasets of height above 16 m. The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.



From Cell 6, two examples has been shown containing clusters of an individual tree crown above 16 m height extracted through M *k*-means algorithm and its respective reconstructed 3-D convex hulls. The result without scaling down the height value has been shown in Figure 13a and Figure 13b.

Figure 13. Cluster of an individual tree from Cell 6 after running M *k*-means without scaling down the height value on the dataset above 16 m height and respective convex polytope. (a) Cell 6—an individual tree cluster above 16 m height. (b) 3-D Convex polytope reconstructed from tree cluster as shown in (a). The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.

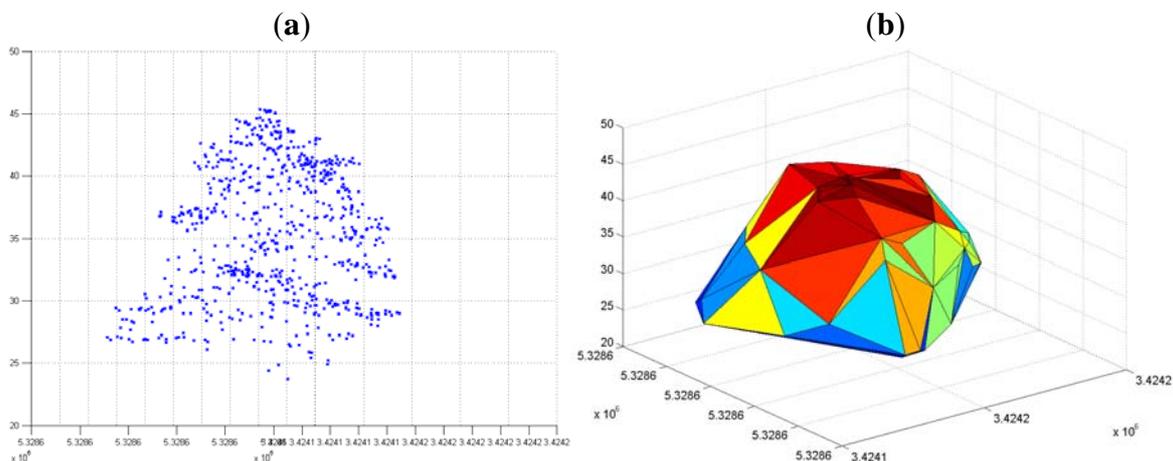
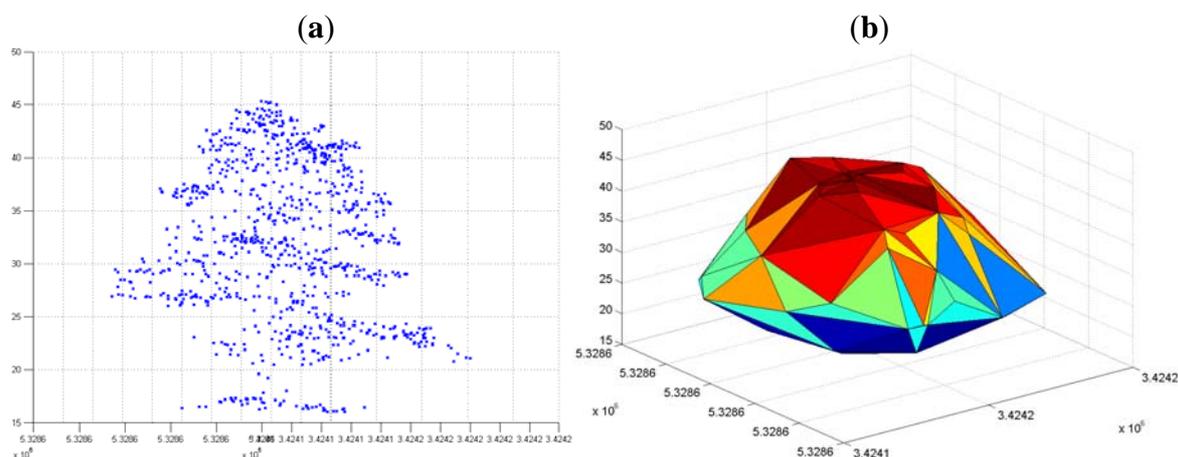


Figure 14a and Figure 14b shows the result obtained by applying M k -means after scaling down the height value on the normalized raw LIDAR points above 16 m height and reconstructing the respective convex polytope. The form of the 3-D crown cluster and the corresponding convex polytope showed that better qualitative accuracy was obtained by M k -means procedure by scaling down the height value as compared to the N k -means results [Figures 10a-10d]. Thus, after applying k -means, it was found that there is a direct relationship of the number of tree clusters to be generated (quantity) with respect to the number of seed points selected by the user before initialization of the process. Similar relationship was found after scaling down the height value of the LIDAR data points and corresponding external seed points with respect to the shape of the tree/tree crown extracted (quality) after a careful visual check.

Figure 14. Cluster of an individual tree from Cell 6 by applying M k -means after scaling down the height value on the dataset above 16 m height and respective convex polytope. (a) Cell 6—an individual tree cluster above 16 m height. (b) 3-D Convex polytope reconstructed from an individual tree cluster as shown in (a). The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.



The silhouette value for each point is a measure of how similar that point is to points in its own cluster compared to points in other clusters, and ranges from -1 to $+1$. Silhouette value and silhouette plot was used for the data distribution and validation purposes. A negative value for some clusters was obtained because the points were too far from each other and acted as an outlier. Therefore, such clusters could not be generated by the algorithm.

3.3. Hierarchical Tree Based Approach Using WPGMA Algorithm

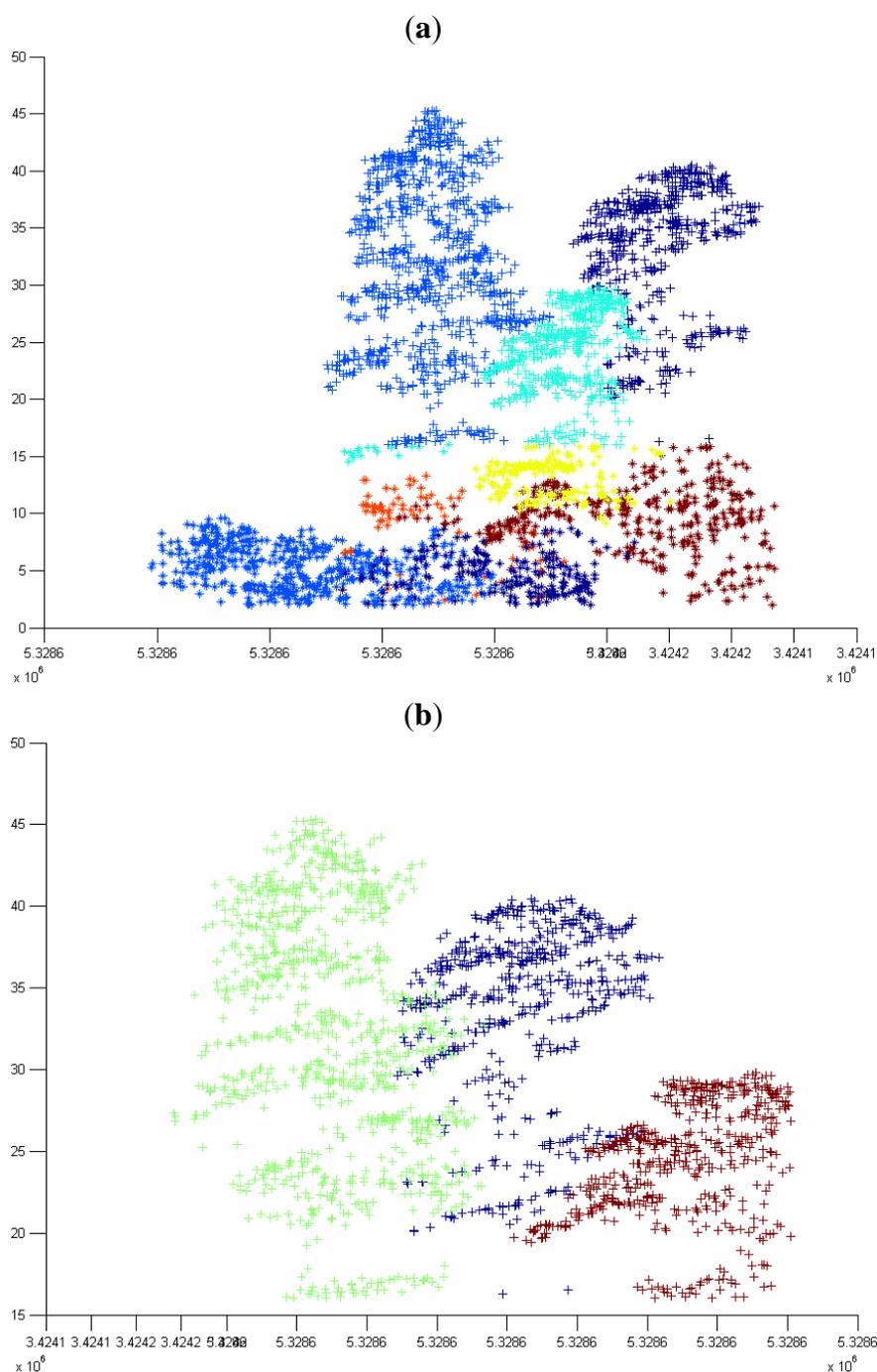
For the creation of hierarchical cluster tree, distance between every pair of objects in a dataset in Euclidian 3-D space using the weighted average distance algorithm was calculated. The size of the respective clusters (*i.e.*, the number of objects they contained) is used as a weight.

3.3.1. Without Scaling Down the Height Value

Similar to k -means, two datasets were used for hierarchical tree clustering. In one dataset, data points were split into two height classes (above 16 m and from 2 to 16 m) as shown in Figure 15a. In

the second dataset, points selected were above 16 m in height [Figure 15 (b)]. Clustering was carried out on both the datasets and result showed that clustering was better in the second dataset. However, in this case, there was no significant difference in the result as compare to the N *k*-means method applied [Figures 8a and 8b] mainly because of smaller number of clusters.

Figure 15. Result after running hierarchical tree clustering without scaling down the height value on two datasets of Cell 6 at different height levels. **(a)** Cell 6 clusters after hierarchical clustering in 2 height classes (between 2 and 16 m height and above 16 m height). **(b)** Cell 6 clusters after hierarchical clustering performed on dataset above 16 m height. The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.



3.3.2. By Scaling Down the Height Value

In this case, the hierarchical clustering was performed by scaling down the height value for both the test datasets with different height classes as mentioned in Section 3.3.1. Clusters obtained from data points separated into two height classes [Figure 16a] is slightly better partitioned as compare to without scaling down the height value [Figure 15a] as seen after visual inspection. In this case, hierarchical clustering gave good result in the data type of height above 16 m [Figure 16b] as compared to without scaling [Figure 15b and Figure 8b] and also found better than M *k*-means [Figure 12]. However, it failed where there is more numbers of clusters were formed.

So, with this method alone satisfactory result was not obtained as compare to N *k*-means and M *k*-means methods even by scaling down the height value. But, in some cases, where numbers of trees are few (3–5), it worked quite well. However, possibilities cannot be ignored to use this method in conjunction with other algorithms.

Figure 16. Result after running hierarchical tree clustering and scaling down the height value on two datasets of Cell 6 at different height levels. **(a)** Cell 6 clusters after hierarchical clustering in 2 height classes (between 2 and 16 m height and above 16 m height). **(b)** Cell 6 clusters after hierarchical clustering performed on dataset above 16 m height. The x and y coordinate values (in meters) are displayed horizontally and the z value (in meters) is displayed vertically.

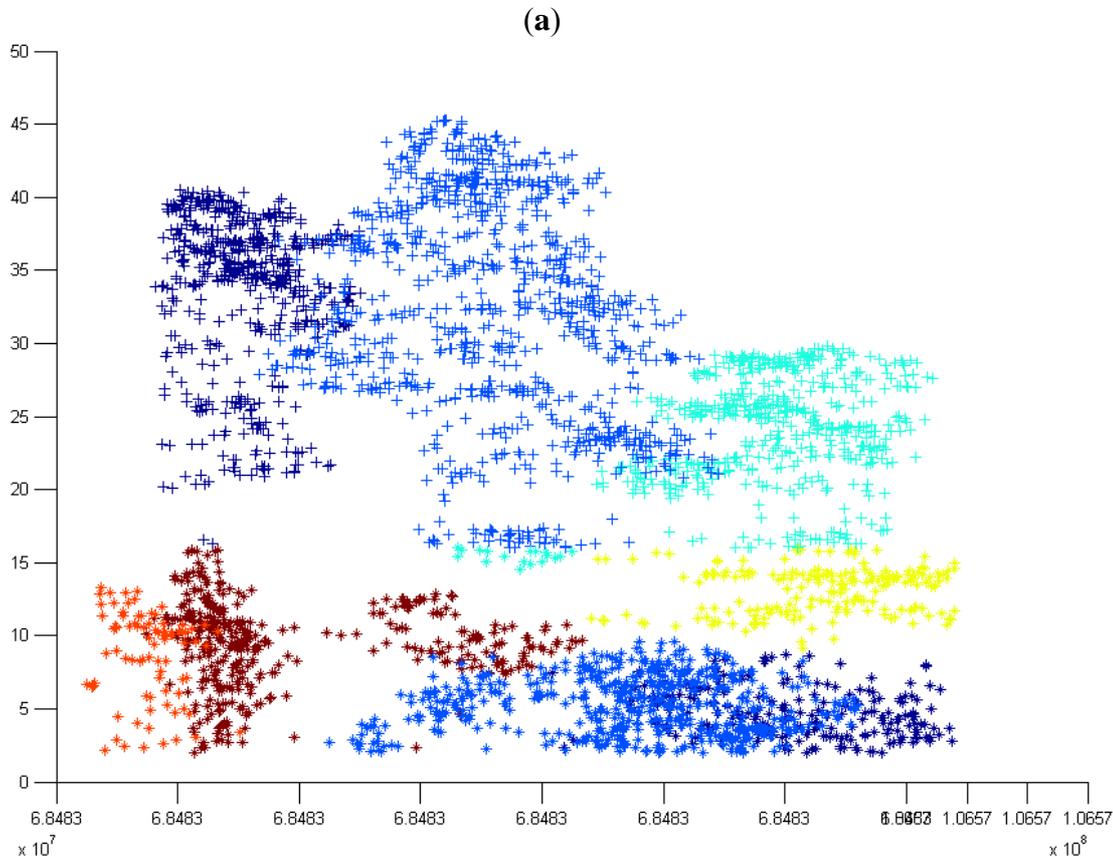
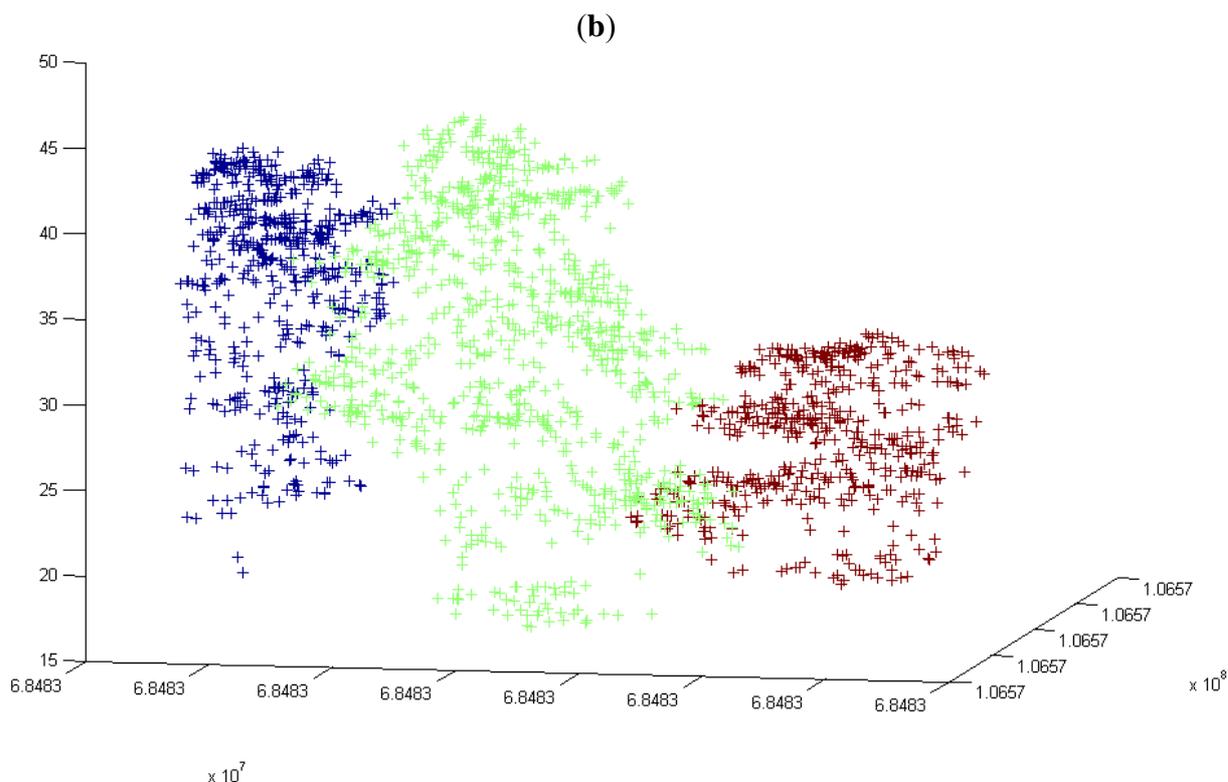


Figure 16. Cont.



3.4. Number of Trees in All the Grids at the First Height Level

For all the cells (except Cell 6), the number of tree were extracted at first height level only. A total of 378 trees were found in all the 30 cells. The height of the extracted trees ranged from 15 m to 50.9 m. The trees were mainly older with few ones at mature or other stages. Table 2 shows the number of individual trees extracted at first tier after the clustering process.

Table 2. Number of first-tier trees in 30 cells.

Number of trees in each cell					
Cell 1-14	Cell 2-16	Cell 3-13	Cell 4-15	Cell 5-11	Cell 6-3
Cell 7-8	Cell 8-7	Cell 9-8	Cell 10-24	Cell 11-20	Cell 12-5
Cell 13-8	Cell 14-12	Cell 15-13	Cell 16-13	Cell 17-8	Cell 18-9
Cell 19-12	Cell 20-15	Cell 21-14	Cell 22-13	Cell 23-20	Cell 24-16
Cell 25-11	Cell 26-20	Cell 27-10	Cell 28-19	Cell 29-13	Cell 30-8

4. Conclusions

The general problem in clustering is how to partition a set of vectors into groups having similar values. In image analysis, vectors can be represented as LIDAR point clouds. The normal *k*-means method is numerical, unsupervised, non-deterministic and iterative. Unfortunately, the empirical speed and simplicity of the normal *k*-means algorithm come at the price of accuracy. In many natural examples, the algorithm generates arbitrarily bad clustering. However, using the external seed points for initialization of the process and scaling down the height value increases the grouping based on similarity measures. It enhances reliability of the outcome which was found true after visual checks

performed on random number of samples of different tree and vertical tree crown densities. The important findings in this research investigation are the usage of local maxima as external seed points and scaling down the height value of the LIDAR data points and seed points for partitioning the data into individual clusters representing individual trees/tree crowns in their most appropriate forms. The quality of the final solution depends largely on the initial set of clusters, and may, in practice, be much poorer than the global optimum. The clustering method found efficient in detecting the individual tree crown in their appropriate form mainly applying by modified k -means algorithms in the study area. Thus, in the k -means algorithm, the number of seed points chosen has the direct relationship over the number of tree clusters (quantity) and by scaling down the height value of the LIDAR data points and corresponding external seed points has a direct influence on the shape of the tree/tree crown (quality). The k -means uses the actual observations of objects or individuals in the data, and not just their proximities. Whereas, the user faces more problems in clustering of LIDAR points reflected back from natural objects using hierarchical tree clustering method. In general, no satisfactory result from the hierarchical clustering based approach was found as compared to the N k -means and M k -means approaches. It is widely known that LIDAR point density, forest conditions in which a tree grows, terrain type, crown cover and tree density are the main factors that determine the number and shape of trees to be extracted. Often, outliers create a problem in clustering process. Removal of outliers before initialization of the clustering process may improve the output quality. A total of 378 trees of 15 to 50.9 meters of height were found in all the 1.2 ha area with mainly older trees. At the moment, it is not possible to validate the result obtained because of the lack of the field inventory data. Further work will be done to assess the result once the field inventory data is received.

Acknowledgements

The authors acknowledge the proofreading support for the manuscript received from H. Gyde Lund, Forest Information Services, Gainesville, VA, USA. We also thank the anonymous reviewers for their corrections and helpful comments.

References and Notes

1. Hyyppä, J.; Yu, X.; Hyyppä, H.; Maltamo, M. Methods of airborne laser scanning for forest information extraction, In *Proceedings of the International Workshop on 3D Remote Sensing in Forestry*, Vienna, Austria, February 2006; pp. 63-78.
2. Persson, A.; Holmgren, J.; Söderman, U. Identification of tree species of individual trees by combining very high resolution laser data with multi-spectral images. In *Proceedings of the International Workshop on 3D Remote Sensing in Forestry*, Vienna, Austria, February 2006; pp. 91-96.
3. Wang, Y.; Weinacker, H.; Koch, B.; Krzysztof, S. Lidar point cloud based fully automatic 3d single tree modelling in forest and evaluations of the procedure. In *Proceedings of the 21st ISPRS Congress*, Vol. 37, Part B6b, Beijing, China, July 2008; pp. 45-52.
4. Vauhkonen, J.; Tokola, T.; Packalén, P.; Maltamo, M. Identification of Scandinavian commercial species of individual trees from airborne laser scanning data using alpha shape metrics. *For. Science* **2009**, *55*, 37-47.

5. Hyypä, J.; Inkinen, M. Detecting and estimating attributes for single trees using laser scanner. *Photogramm. J. Fin.* **1999**, *16*, 27-42.
6. Persson, A.; Holmgren, J.; Söderman, U. Detecting and measuring individual trees using an airborne laser scanner. *Photogramm. Eng. Remote Sens.* **2002**, *68*, 925-932.
7. Koch, B.; Heyder, U.; Weinacker, H. Detection of individual tree crowns in airborne lidar data. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 357-363.
8. Pyysalo, U.; Hyypä, H. Reconstructing tree crowns from laser scanner data for feature extraction. In *Proceedings of the ISPRS Technical Commission III Symposium for Photogrammetric Computer Vision*, Graz, Austria, September 2002; pp. 293-296.
9. Reitberger, J.; Krzystek, P.; Stilla, U. 3D segmentation and classification of single trees with full waveform LIDAR data. In *Proceedings of SilviLaser 2008*, Edinburgh, UK, September 2008; pp. 216-226.
10. Morsdorf, F.; Meier, E.; Allgöwer, B.; Nüesch, D. Clustering in airborne laser scanning raw data for segmentation of single trees. In *Proceedings of the ISPRS Working Group III/3 Workshop for 3-D Reconstruction from Airborne LaserScanner and InSAR Data*, Dresden, Germany, October 2003; pp. 27-33.
11. Morsdorf, F.; Meier, E.; Kötz, B.; Itten, K.I. Lidar based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management. *Remote Sens. Environ.* **2004**, *92*, 353-362.
12. Cici, A.; Kevin, T.; Nicholas, J.T.; Sarah, S.; Jörg, K. Extraction of vegetation for topographic mapping from full-waveform airborne laser scanning data. In *Proceedings of SilviLaser 2008*, Edinburgh, UK, September 2008; pp. 343-353.
13. Doo-Ahn, K.; Woo-Kyun, L.; Hyun-Kook, C. Estimation of effective plant area index using LiDAR data in forest of South Korea. In *Proceedings of SilviLaser 2008*, Edinburgh, UK, September 2008; pp. 237-246.
14. Ko, C.; Sohn, G.; Rimmel, T.K. Classification for deciduous and coniferous trees using airborne lidar and internal structure reconstructions. In *Proceedings of SilviLaser 2009*, College Station, TX, USA, October 2009; pp. 36-45.
15. Ørka, H.O.; Næsset, E.; Bollandsas, O.M. Comparing classification strategies for tree species recognition using airborne laser scanner data. In *Proceedings of SilviLaser 2009*, College Station, TX, USA, October 2009; pp. 46-53.
16. Rieger, P.; Ullrich, A.; Reichert, R. Laser scanners with echo digitization for full waveform analysis. In *Proceedings of the International Workshop on 3D Remote Sensing in Forestry*, Vienna, Austria, February 2006; pp. 204-210.
17. Hug, C.; Ullrich, A.; Grimm, A. LiteMapper-5600—A waveform-digitizing LIDAR terrain and vegetation mapping system. In *Proceedings of ISPRS Working Group on Laser-Scanners for Forest and Landscape Assessment—Instruments, Processing Methods and Applications*, Freiburg, Germany, October 2004; pp. 24-29.

18. Weinacker, H.; Koch, B.; Heyder, U.; Weiancker, R. Development of filtering, segmentation and modelling modules for lidar and multispectral data as a fundament of an automatic forest inventory system. In *Proceedings of ISPRS Working Group on Laser-Scanners for Forest and Landscape Assessment—Instruments, Processing Methods and Applications*, Freiburg, Germany, October 2004; pp. 50-55.
19. Burschel, P.; Huss, J. In *Grundriss des Waldbaus*, 1st ed.; Eugen Ulmer Verlag: Parey Text Series, Berlin, Germany, 1997; pp. 67-68.
20. Sneath, P.H.A.; Sokal, R.R. In *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, 2nd ed.; W.H. Freeman and Company: San Francisco, CA, USA, 1973.
21. Barber, C.B.; Dobkin, D.P.; Huhdanpaa, H.T. The Quickhull algorithm for convex hulls. *ACM Trans. Math. Software* **1996**, *22*, 469-483.
22. Berg, M.D.; Kreveld, M.V.; Overmars, M.; Schwarzkopf, O. Convex Hulls—Mixing Things. In *Computational Geometry—Algorithms and Applications*; Springer-Verlag: Berlin, Heidelberg, Germany, 1997; pp. 233-248.

© 2010 by the authors; licensee Molecular Diversity Preservation International, Basel, Switzerland. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).