Characterization and classification of vegetation canopy structure and distribution within the Great Smoky Mountains National Park using LiDAR

Jitendra Kumar*, Jon Weiner†, William W. Hargrove‡, Steven P. Norman‡,
Forrest M. Hoffman* and Doug Newcomb§
*Oak Ridge National Laboratory, Oak Ridge, TN, USA Email: jkumar@climatemodeling.org
†University of California Berkeley, Berkeley, CA, USA
‡USDA Forest Service, Southern Research Station, Asheville, NC, USA
§U.S. Fish and Wildlife Service, Raleigh, NC, USA

***Abstract*—Vegetation canopy structure is a critically important habitat characteristic for many threatened and endangered birds and other animal species, and it is key information needed by forest and wildlife managers for monitoring and managing forest resources, conservation planning and fostering biodiversity. Advances in Light Detection and Ranging (LiDAR) technologies have enabled remote sensing-based studies of vegetation canopies by capturing three-dimensional structures, yielding information not available in two-dimensional images of the landscape provided by traditional multi-spectral remote sensing platforms. However, the large volume data sets produced by airborne LiDAR instruments pose a significant computational challenge, requiring algorithms to identify and analyze patterns of interest buried within LiDAR point clouds in a computationally efficient manner, utilizing state-of-art computing infrastructure. We developed and applied a computationally efficient approach to analyze a large volume of LiDAR data and characterized the vegetation canopy structures for 139,859 hectares (540 sq. miles) in the Great Smoky Mountains National Park. This study helps improve our understanding of the distribution of vegetation and animal habitats in this extremely diverse ecosystem.**

I. Introduction

Forest ecosystems are a complex mosaic of diverse plant and tree species, the location and distribution of which are driven by a number of gradients like climate (ex. temperature, precipitation regimes), topography (ex. elevation, slope, aspect), geology (ex. soil types, textures, depth), hydrology (ex. drainage, moisture availability) etc. Diverse combinations of these gradients support diverse composition and distribution of vegetation which in turn supports an array of wildlife. Understanding the vegetation canopy structure is critical to understand, monitor and manage the complex forest ecosystems like those in the Great Smoky Mountain National Park (GSMNP). Vegetation canopies not only help understand the vegetation, but are also a critically important habitat characteristics of many threatened and endangered animal and bird species for which the GSMNP is home.

Remote sensing has been widely used to monitor regional to global forest ecosystems and for mapping of vegetation types. However, traditional remote sensing methods for vegetation classification often use light reflectance from the top layer of vegetation. Advances in Light Detection and Ranging (LiDAR) technologies have enabled remote sensing-based studies of vegetation canopies by providing a three-dimensional representation of vegetation structure throughout the canopy. While the application of LiDAR for study of forest ecosystems is becoming more common, the richness of these data sets are generally under-utilized due to the large volumes of the data produced by these instruments and lack of computational resources and analysis algorithms. Most of the LiDAR studies focus on the development of high resolution Digital Elevation Models, canopy heights and occasionally understory density [1], [2]. While LiDAR dervied metrics have proven to be useful for an array of applications [1]–[5], three-dimensional information provided by the LiDAR are left unutilized.

The objective of this study is to develop methods to realize the potentials of rich LiDAR data set to map and characterize the three-dimensional structure and distribution of vegetation canopies. We develop and apply data analytic techniques to identify the ecologically important and understandable structural types by mining the large and complex volumes of LiDAR data.

II. Materials

A. Study area

The geographic area for this study was the Great Smoky Mountains National Park (GSMNP), which in part covers the Great Smoky Mountains and the Blue Ridge Mountains, encompassing 816 sq. miles across Tennessee and North Carolina in the United States. Results presented here focus primarily on the Tennessee side of the GSMNP (approximately 540 sq. miles). The GSMNP covers complex topography with elevations ranging from 876–6,643 feet above mean sea level. The GSMNP is ecologically rich and diverse, consisting of about 1,600 species of flowering plants, including 100 native tree species and over 100 native shrub species [6]. The distribution of vegetation in the park is strongly influenced by topography, moisture and other environmental gradients [7].

B. Airborne LiDAR data

Airborne LiDAR for 1,400 sq. km (540 sq. miles) for the Tennessee portion of the GSMNP and the Foothills Parkway was acquired by The Center for Remote Sensing and Mapping Science at the University of Georgia and Photo Science, Inc. under a U.S. Geological Survey (USGS)-funded program [8]. While details of data acquisition and processing are described by [8], we briefly summarize the data here.

A total of 1,658 flight miles of data were collected during the period of February–April 2011. Four multiple discrete returns per pulse were collected at a rate of 20.2 Hz by the LiDAR instruments employed for the data collection. There was overlap of 40–50% between adjacent flight lines for a nominal flying height of 1,981.2 m above ground level. Scan angles were $\pm 16^{\circ}$ for a swath width of 1,134.7 m. Data were calibrated and LiDAR points categorized as Unclassified, Ground, Noise or Overlap. Data sets were split up into 1,500 m $\times$ 1,500 m adjacent and non-overlapping tiles (Figure 1). The tiled data sets, consisting of 724 tiles in "*las*" format (94 GB total size), were obtained from the Great Smoky Mountains National Park Service.

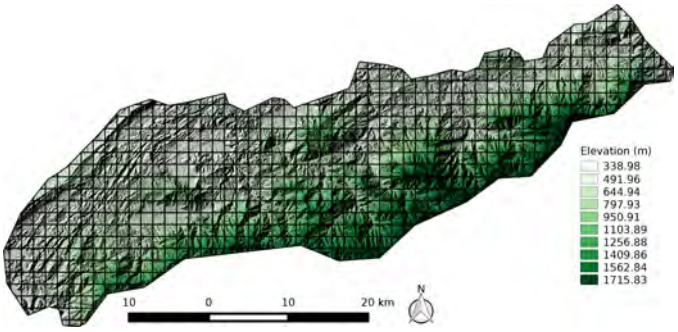


Fig. 1. LiDAR tiles for TN side of Great Smoky Mountains National Park. The underlying color image is a 1.5 m resolution digital elevation map for the region.

C. Digital elevation model (DEM)

The LiDAR point cloud was processed by [8] using ESRI ArcGIS software to create a 1.5 m resolution bare-earth digital elevation model raster (DEM) (Figure 1). This DEM was used as the bare Earth topography in the analysis presented here.

III. Methods

A computationally efficient *Python*-based workflow (Figure 2) was developed to process and analyze the LiDAR point cloud data sets.

A. Topographic detrending of LiDAR point cloud

The LiDAR point cloud data set for the GSMNP was based on a vertical datum (NAVD88 – Geoid09). Raw LiDAR point cloud elevations contain the imprints of the underlying topography (Figure 3(a)). A topographic detrending was required in order to convert the elevations from an absolute datum to an above ground level (AGL) elevation (Figure 3(b)). Thus, for every point in the point cloud data set, the corresponding ground level elevation was identified using the 1.5 m DEM (described in Section II-C), and the AGL elevation was calculated (*elevation from datum* $-$ *ground level elevation*). All further analysis was done on the detrended point cloud.

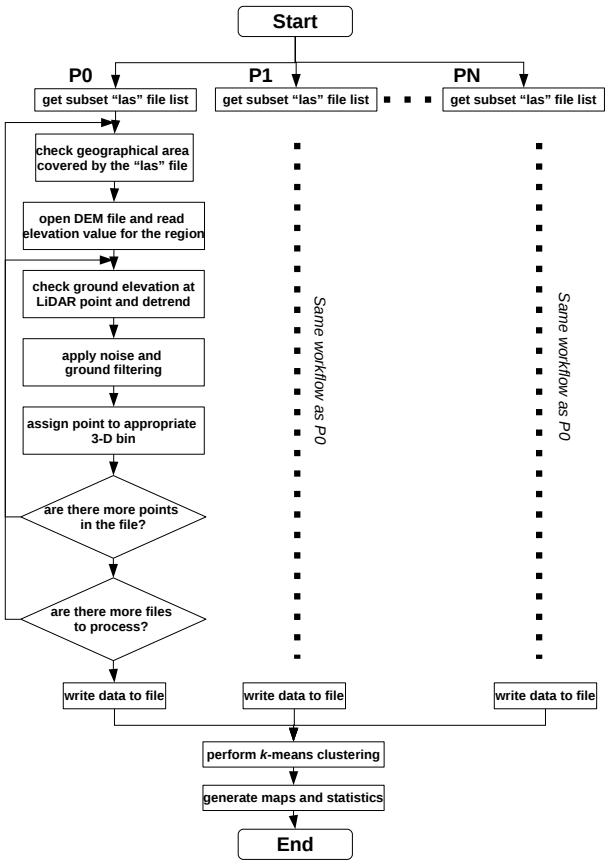


Fig. 2. Computational workflow for the analysis carried out using n processes in embarrassingly parallel fashion.

B. Vertical canopy structure

The topographically detrended LiDAR point cloud was processed to generate the vertical canopy structure of vegetation in the full study area. A horizontal grid of 30 m $\times$ 30 m resolution was used, to match the resolution of LANDSAT, NLCD and other existing vegetation mapping products for the GSMNP [9] and to enable comparison and further analysis. Employing a 30 m $\times$ 30 m resolution also ensured sufficient LiDAR point density to construct a three-dimensional vegetation canopy structure. A 1 m vertical resolution was used to identify vegetation height from the ground surface to a maximum height of 75 m. The number of LiDAR points in each vertical 1 m bin (at each 30 m $\times$ 30 m cell in the horizontal grid) was identified to construct a vertical density profile (Figure 3(d)). Normalized density profiles were created by computing the percent of total points (at that cell) in 1 m vertical bins.

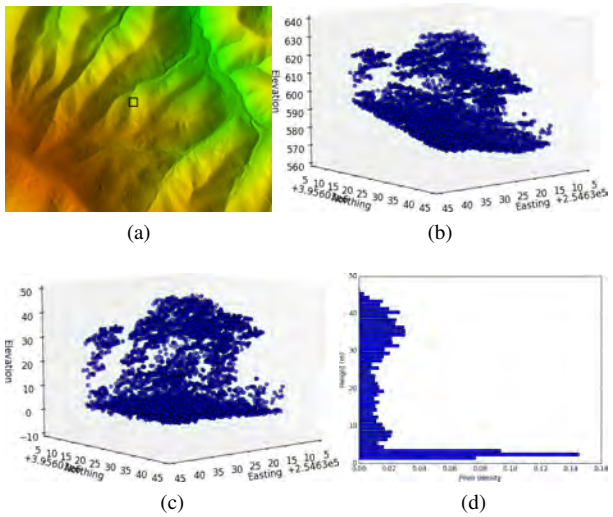


Fig. 3. (a) 3-D LiDAR point cloud at 30 m × 30 m region (black square) in a typical GSMNP cove forest. (b) The raw LiDAR point cloud (3,985 points), showing the imprints of the underlying cove topography. (c) LiDAR point cloud after topographic detrending and filtering (3,936 points) that converted the elevations to above ground level elevation. (d) Distribution of LiDAR point density along the vertical profiles in a cove forest dominated by tall trees and a dense understory.

C. Noise and ground filtering of LiDAR data

Raw LiDAR point cloud data often contain anomalous and noisy returns used as elevation data, and identification and elimination of these points is essential prior to analysis (Figure 3(c)) [1], [10], [11]. Anomalously high elevation values are often caused by atmospheric aerosols, dust and smoke, birds and insects, or other unknowns, while very low elevations are caused by low-lying vegetation (like shrubs, grasses) or ground litter and complex slopes and topographic relief. The following noise-removal steps were performed on the entire LiDAR point cloud:

1) *Anomalous high elevation points*: The tallest tree recorded within the GSMNP is a 190 foot (57.9 m) high tulip-tree (*Liriodendron tulipifera*) [12] [13]. A maximum height of 75 m (higher than the tallest possible tree) was used in the analysis, and any LiDAR points with AGL elevation higher than 75Am were filtered out as noise.
2) *Anomalous low elevation points*: Topographic features (like steep slopes, and high reliefs, etc.) and surface litter often cause anomalous reflections that are recorded by the sensors as negative elevations, especially after topographic detrending due to errors in the DEM. Thus, the points with negative elevations were filtered out.
3) *Low height vegetation*: Some areas of the GSMNP are dominated by low height vegetation (shrubs and grasses) and are often lower than one meter in height. Grid cells where 95% or more LiDAR return points were within 1 m from the ground surface were identified and classified as low lying vegetation.
4) *Anomalously high number of returns at the same elevation*: Due to errors and noise in the point cloud data, some locations contain an anomalously large fraction of total returns at a cell recorded at the same elevation. Thus, a correction was performed if more than 30% of total returns at a cell were recorded from the same 1 m vertical bin, by applying a smoothing and replacing the anomalous return counts in the vertical bin with a simple average of return counts in bins directly above and below. This simple correction allowed identification and removal of anomalous spikes in vertical canopy structures. The scheme was manually spot checked at a number of locations and was determined to be adequate and not to introduce any artifact in the data.

While we developed a number of steps to remove noise and applied corrections to the data, inaccuracies remain in the data and are carried throughout the rest of the analysis, requiring careful ecological interpretation of the results to identify and filter useful signals from noise.

D. Classification and spatial distribution of vegetation canopy structure

A key objective of this study was to understand the spatial distribution, pattern and dominance of various vegetation types and canopy structures in the study area. Multi-variate clustering techniques have been widely used in Earth science for delineation of ecoregions that are relatively homogeneous with respect to a collection of observable environmental and climate characteristics [14]–[17].

Here, we used a k-means algorithm to cluster the gridded vertical canopy structure data set (Section III-B) into groups containing locations with similar vertical canopy structures (Figure 3d). The k-means algorithm groups data (X_1, X_2, …, X_n) with n records into a desired number of clusters, k, equalizing the full multi-dimensional variance across clusters [18]. The number of clusters, k, is supplied as an input and remains fixed. The k-means algorithm starts with initial centroid vectors (C_1, C_2, …, C_n) and calculates the Euclidean distance of each pixel (X_i, $1 \leq i \leq n$) to every centroid (C_j, $1 \leq j \leq k$), assigning it to the closet existing centroid. The centroid vector is recalculated as the vector mean of all dimensions of each pixel assigned to that centroid. This classification and re-calculation process is iteratively repeated until fewer than some small fixed proportion of observations changes their cluster assignment between iterations. We assumed convergence was achieved when fewer than 0.05% of the observations changed cluster assignments.

In [16], we developed a parallel version of the k-means algorithm to accelerate convergence, handle empty cluster cases, and obtain initial centroids through a scalable implementation of the triangular equality based acceleration method [19]. [20] extended this to a fully distributed and highly scalable parallel version of the k-means algorithm for analysis of very large data sets, which was used in this study.

E. Computational workflow

A computationally efficient workflow was developed in *Python* for processing and analyzing the massive LiDAR point cloud. To exploit the inherent parallelism in analysis of LiDAR point cloud data, an embarrassingly parallel scheme was implemented to allow processing of each "*las*" file in a different process on a multi-core machine. The "*laspy*" [21] Python module was used for processing LiDAR point cloud data sets in "*las*" format. The *Geo-spatial Data Abstraction Library 2.0.0* [22] was used for analyzing geospatial data sets (e.g., DEM), which allowed for efficient access to elevation data sets for desired geographical regions within the parallel workflow. Figure 2 shows a schematic of the analysis workflow implemented for mapping vegetation canopy structure and distribution using LiDAR.

IV. Results and Discussion

A. Unique Vegetation Canopy Structures

A gridded data set of vertical canopy structure (Section III-B) was classified using a k-means clustering algorithm (Section III-D) to identify patterns of vegetation and to create clusters of unique vegetation canopy structures. While large volumes of LiDAR data are typically difficult to understand, our classification method enables derivation of higher order products that can be easily analyzed and understood by forest and wildlife managers. Data sets were classified at various levels of division (k = 5, 10, 15, 20, 25, 30, 50, 75, 100). While at lower levels of division (small k), different canopy structures of interest may be lumped together, higher levels of division may define clusters with insignificant differences in canopy structures. Various approaches for determining the optimal level of division for k-means clustering have been developed and reported in the literature [23], [24]. However, most of these methods are not effective for data sets like LiDAR that contain significant amounts of errors and noise. Thus, we classified and analyzed the data sets at various level of divisions. k=30 was selected for subsequent analysis because it appeared to have an optimal signal to noise ratio while allowing sufficient resolution to distinguish different vegetation canopy structures. A geospatial map of 30 vegetation classes was developed (Figure 4(a)), with each class defined by a nominal vertical canopy structure (Figure 4(b)).

Canopy structure classes (Figure 4(b)), identified well the range of vegetation present in the GSMNP from tall and dense tree canopies with very low understory vegetation (unimodal profiles like 10 and 13), to tree canopies with understory vegetation (represented by bi-modal profiles 5, 14, 17, etc.) to low height shrub dominated vegetation (profiles 1, 4, 16, etc.). While the classification method was able to identify unique canopy structures, it also identified the areas with outliers or noisy data in unique clusters (like 3 and 11), making it easy to eliminate them from further analysis. Noise and errors along the boundaries of the point cloud data tiles were identified (potentially introduced by processing of the data [8]) and filtered out in our analysis, imprints of which are visible in the final map product (Figure 4(a)).

B. Translating Canopy Structures into Vegetation Types

We used *Mapcurves* [25] to identify the best "translation table" between LiDAR clusters and vegetation types defined by [9] (Table I). Although *Mapcurves* identifies the single vegetation type having the best fit in terms of spatial overlap, each LiDAR cluster is likely to overlap with many other vegetation types; however, Table I shows only the single vegetation category [9] exhibiting the largest spatial co-registration.

Indeed, the inherently different natures of vegetation type or composition and the above-ground vertical biomass distribution might act to minimize any agreement between these two maps. A number of different forest compositions might show similar vertical structure distributions, despite substantial differences in species composition. Conversely, a single forest type might, throughout its successional development, sequentially adopt a series of substantially distinct vertical profiles. Moreover, the wide discrepancy in inherent resolution of these two maps might further complicate their direct comparison. While the vegetation type map consists of generalized descriptive polygons, the analytical LiDAR map was coarsened to 30 horizontal meters.

Despite these differences, a number of consistencies emerge from the comparison of these two maps. Successional vegetation types, including grasses, are restricted to LiDAR categories 0 and 3, although this is somewhat artificial, since cluster 0 was defined *a priori* as low-stature vegetation less than 1 m tall, and cluster 3 is anomalous, accounting for little area in the map. The Spruce-fir type predominates within a single profile cluster, number 21, and typifies these short stature, high elevation forests. Similarly, profile cluster 27 solely predominates the *Ericaceous* shrub type. The Yellow pine type has the majority of overlap with three profile clusters (15, 24, and 29), which seem to differ in their degree of canopy height, perhaps reflecting separate phases of successional development.

The two profile types predominantly associated with Montane cove types reflect the tallest forests growing on the most fertile sites, but the Montane Oak-Hickory type has four otherwise similar vertical profile forms. The Northern/acid hardwood type dominates five different profile types, which may differ in the degree of understory, possibly Rhododendron, that is present. As might be expected, the Chestnut oak vegetation type, which accounts for much of the area in the map (43%) is manifested across 12 different profile types, perhaps reflecting differences in both forest age and compositional differences.

C. Validation case studies

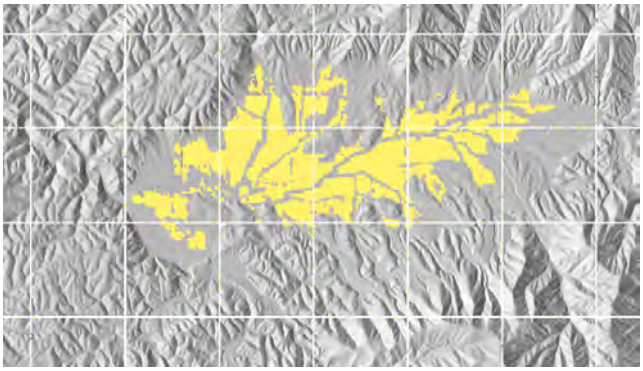
We conducted a number of case studies to verify the LiDAR based canopy structures against best available maps of vegetation in the GSMNP [9], which were available at the same spatial resolution of 30 m. While the map resolution is the same, LiDAR-derived canopy structures developed here represent aggregation from significantly higher resolution source data compared to vegetation maps that classify the region in traditional vegetation classes.

TABLE I
Mapcurves BASED TRANSLATION OF LiDAR DATA DERIVED 30 UNIQUE VEGETATION CANOPY STRUCTURES TO TRADITIONAL VEGETATION CLASSES FOR THE GREAT SMOKY MOUNTAIN NATIONAL PARK

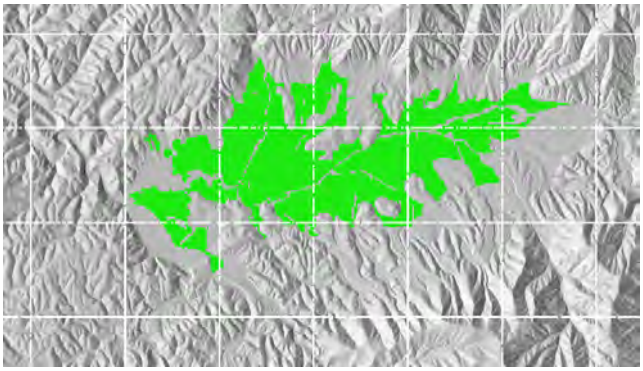
Cluster	Dominant Vegetation Type
0	Successional or modified vegetation
1	Chestnut Oak Forest
2	Chestnut Oak Forest
~~3~~	~~Successional or Modified Vegetation~~
4	Chestnut Oak Forest
5	Northern Hardwood/acid Hardwood Forest
6	Chestnut Oak Forest
7	Yellow Pine Forest
8	Northern Hardwood/acid Hardwood Forest
9	Chestnut Oak Forest
10	Montane Cove Forest
~~11~~	~~Chestnut Oak Forest~~
12	Northern Hardwood/acid Hardwood Forest
13	Montane Oak-hickory Forest
14	Northern Hardwood/acid Hardwood Forest
15	Yellow Pine Forest
16	Chestnut Oak Forest
17	Montane Cove Forest
18	Montane Oak-Hickory Forest
19	Chestnut Oak Forest
20	Montane Oak-Hickory Forest
21	Spruce-Fir Forest
22	Northern Hardwood/Acid Hardwood Forest
23	Chestnut Oak Forest
24	Yellow Pine Forest
25	Montane Oak-Hickory Forest
26	Chestnut Oak Forest
27	Ericaceous Shrubs (Heath Bald Type)
28	Chestnut Oak Forest
29	Yellow Pine Forest
30	Chestnut Oak Forest

Cades Cove, located in a valley surrounded by mountains, is one of the most popular destinations in the GSMNP. Cades Cove consists of woodlots interspersed within old-fields that are mowed and burned to mimic a 19^{th} century agrarian settlement [26]. Figure 5(a) shows the area of low height (less than 1 m tall) vegetation class identified by our study which shows very good correspondence to the "Successional or modified vegetation" types (Figure 5(b)) as mapped by vegetation map of the GSMNP [9].

In contrast to the low height vegetation in Cades Cove, forests in the mountain coves of the GSMNP are dominated by tall trees with dense canopies, especially on North-facing slopes. The Great Smoky Mountain Institute at Tremont (GSMIT) is surrounded by "Montane Cove" and "Hemlock" forests with tall and dense canopies (Figure 6(b)). Strong correspondence and spatial overlap with vegetation canopy classes 10 and 13 was identified for the region (Figure 6(a)).



(a)



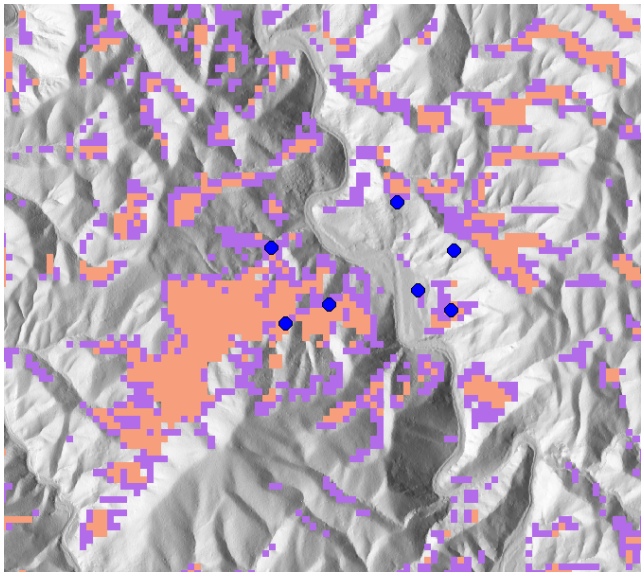
(b)

Fig. 5. Vegetation in Cades Cove valley in the GSMNP. (a) Low height (less than 1 m tall) vegetation class identified by LiDAR derived canopy structure product. (b) "Successional or modified vegetation" mapped by [9].

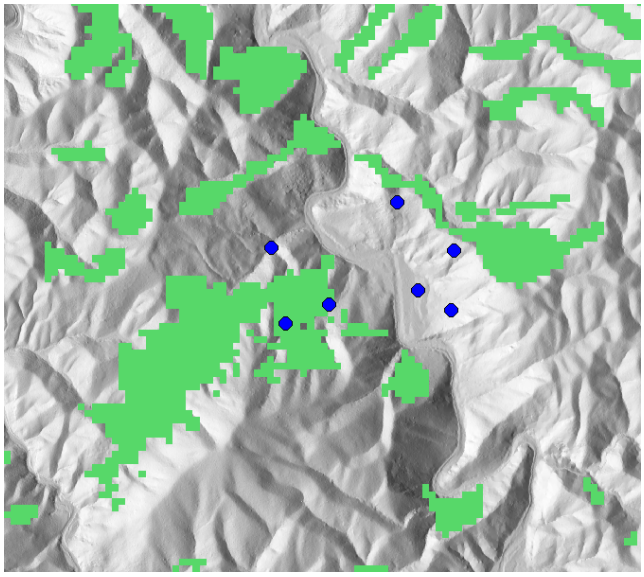
Vegetation canopy classes 10 and 13 represent some of the tallest vegetation in the GSMNP, having canopy heights of up to 50 m from the ground, high density/biomass in tree canopies with relatively low understory growth due to competition for light and nutrients. The canopy profiles at a number of "Citizen Science" phenology plots (marked by blue circles in Figure 6) maintained by GSMIT were studied to match the known tree types at the sites.

D. Computational performance

The compute and data intensive steps involved in processing, filtering/correcting and gridding of the LiDAR point cloud data (Section III-A,III-B,III-C) were conducted with a Python based workflow (Section III-E) developed for this research. The workflow was tested and optimized for parallel performance on low to moderate core-count Linux-based platforms. We were able to process individual LiDAR "*las*" files within 15 seconds on average on Intel Xeon 2.40 GHz processors. Each production run processed 98 GB LiDAR "*las*" data sets and were performed using 8 processes with an average turn around time of 22 minutes. The parallel k-means clustering tool developed by [20] was used to classify the resulting gridded canopy structure data sets (Section III-D).



(a)



(b)

Fig. 6. Vegetation around The Great Smoky Mountain Institute at Tremont (a) Tall canopy structures classes 10 (orange) and 13 (violet) derived using LiDAR (b) "Montane Cove" forest mapped by [9]. Blue circles shows the locations of phenology plots maintained by the Great Smoky Mountain Institute at Tremont (GSMIT) Citizen Science program.

V. Conclusion

In this study, we developed a methodology and computational tools to analyze large volumes of LiDAR point cloud data and applied our workflow to map and characterize vegetation canopy structures and their spatial distributions for the Tennessee portion of the Great Smoky Mountain National Park (GSMNP). LiDAR data sets often suffer from noise and errors due to reflection/detection errors, conditions at the time of data collection, complex terrain and relief and heterogeneous vegetation. We developed schemes to identify and filter out noise in the data that may induce errors in the characterization of vegetation canopy structures. Cluster analysis was employed to develop a canopy structure-based classification of vegetation in the GSMNP. We found a strong correspondence between the resulting canopy classes and a map of vegetation types present in highly biodiverse complex terrain of the park. The high resolution map of vegetation canopies will provide forest and wildlife managers with critically important information for resource management and conservation planning. Species composition in the GSMNP is in a state of flux due to various environmental stressors like fires, hemlock death due to the wooly adelgid, and other factors, leading to successional changes in this critically important ecosystem. Computationally efficient tools developed in this study allow forest managers to monitor the forest using repeat LiDAR surveys, which was not previously possible because of the complexity and volume of airborne LiDAR data sets.

VI. Data Products

All the data sets produced by this study and discussed throughout this article has been archieved and available at Oak Ridge National Laboratory Distributed Active Achive Center [http://www.daac.ornl.gov] [27]. The collection contains the following key data products from this study.

- 30 unique vegetation canopy structure classes (Figure 4(a), Section IV-A)
  - Geospatial maps of vegetation canopy classes (Figure 4(a), Section IV-A). *Format: Geotiff*
  - Representative vegetation canopy structures that define the 30 unique canopy structure classes (Figure 4(b), Section IV-A). *Format: ASCII*
- *Mapcurves* [25] based reclassification of the 30 unique vegetation canopy classes to vegetationa category exhibiting largest spatial co-registration
  - Geospatial maps of reclassified 30 unique vegetation canopy classes. *Format: Geotiff*
  - Translation table from vertical canopy structure classes (Figure 4(a), Section IV-A) to vegetation type [9] *Format: ASCII*
- *Mapcurves* [25] was also applied in opposite direction with vegetation map [9] to identify vegetation canopy classes that best co-registered with any given vegetation type
  - Geospatial maps of vegetation types [9] reclassed to vegetation canopy classes. *Format: Geotiff*
  - Translation table from vegetation types [9] to vertical canopy structure classes (Figure 4(a), Section IV-A) *Format: ASCII*

The Universal Transverse Mercator (UTM) projection system Zone 17N, Datum NAD83 was used for all the geospatial data products.

Acknowledgment

This research was partially sponsored by the U.S. Department of Agriculture, U.S. Forest Service, Eastern Forest Environmental Threat Assessment Center. Additional support

was provided by the Biogeochemistry Feedbacks Scientific Focus Area (SFA), which is sponsored by the Regional and Global Climate Modeling (RGCM) Program in the Climate and Environmental Sciences Division (CESD) of the Biological and Environmental Research (BER) Program in the U. S. Department of Energy Office of Science. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

REFERENCES

[1] K. Zhao, S. Popescu, X. Meng, Y. Pang, and M. Agca, "Characterizing forest canopy structure with LiDAR composite metrics and machine learning," *Remote Sens. Environ.*, vol. 115, no. 8, pp. 1978 – 1996, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425711001118

[2] M. Lefsky, W. Cohen, S. Acker, G. Parker, T. Spies, and D. Harding, "Lidar Remote Sensing of the Canopy Structure and Biophysical Properties of Douglas-Fir Western Hemlock Forests," *Remote Sensing of Environment*, vol. 70, no. 3, pp. 339 – 361, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425799000528

[3] M. A. Lefsky, D. Harding, W. Cohen, G. Parker, and H. Shugart, "Surface Lidar Remote Sensing of Basal Area and Biomass in Deciduous Forests of Eastern Maryland, USA ," *Remote Sensing of Environment*, vol. 67, no. 1, pp. 83 – 98, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425798000716

[4] L. Étienne Boudreault, A. Bechmann, L. Tarvainen, L. Klemedtsson, I. Shendryk, and E. Dellwik, "A LiDAR method of canopy structure retrieval for wind modeling of heterogeneous forests," *Agricultural and Forest Meteorology*, vol. 201, pp. 86 – 97, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168192314002652

[5] H.-E. Andersen, R. J. McGaughey, and S. E. Reutebuch, "Estimating forest canopy fuel parameters using LIDAR data," *Remote Sensing of Environment*, vol. 94, no. 4, pp. 441 – 449, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425704003438

[6] Great Smoky Mountains National Park. [Online]. Available: http://www.nps.gov/grsm/learn/nature/plants.htm(AccessedAugust01,2015)

[7] R. H. Whittaker, "Vegetation of the Great Smoky Mountains," *Ecological Monographs*, vol. 26, no. 1, pp. pp. 1–80, 1956. [Online]. Available: http://www.jstor.org/stable/1943577

[8] T. Jordan, M. Madden, B. Yang, J. Sharma, and S. Panda, "Acquisition of LiDAR for the Tennessee Portion of Great Smoky Mountains National Park and the Foothills Parkway," Center for Remote Sensing and Mapping Science (CRMS), Department of Geography, The University of Georgia, Athens, Georgia, USA, Tech. Rep. USGS Contract # G10AC0015, 2011.

[9] M. Madden, "Overstory Vegetation at Great Smoky Mountains National Park, Tennessee and North Carolina, Reference Code: 1047498," 2014. [Online]. Available: https://irma.nps.gov/App/Reference/Profile/1047498

[10] X. Meng, L. Wang, J. L. Silván-Cárdenas, and N. Currit, "A multi-directional ground filtering algorithm for airborne LiDAR," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 1, pp. 117 – 124, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271608000956

[11] X. Meng, N. Currit, and K. Zhao, "Ground Filtering Algorithms for Airborne LiDAR Data: A Review of Critical Issues," *Remote Sens.*, vol. 2, no. 3, p. 833, 2010. [Online]. Available: http://www.mdpi.com/2072-4292/2/3/833

[12] Native Tree Society, Accessed September 08, 2015. [Online]. Available: http://www.nativetreesociety.org/bigtree/tallest_trees_by_state.htm

[13] Jim Burnett, Accessed September 08, 2015. [Online]. Available: http://www.nationalparkstraveler.com/2012/10/tallest-native-hardwood-tree-north-america-located-national-park10714

[14] W. W. Hargrove and F. M. Hoffman, "Potential of multivariate quantitative methods for delineation and visualization of ecoregions," *Environ. Manage.*, vol. 34, no. Supplement 1, pp. S39–S60, apr 2004.

[15] F. M. Hoffman, W. W. Hargrove, D. J. Erickson, and R. J. Oglesby, "Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models," *Earth Interact.*, vol. 9, no. 10, pp. 1–27, Aug. 2005.

[16] F. M. Hoffman, W. W. Hargrove, R. T. Mills, S. Mahajan, D. J. Erickson, and R. J. Oglesby, "Multivariate Spatio-Temporal Clustering (MSTC) as a data mining tool for environmental applications," in *Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software Society (iEMSs 2008)*, M. Sànchez-Marrè, J. Béjar, J. Comas, A. E. Rizzoli, and G. Guariso, Eds., Barcelona, Catalonia, Spain, Jul. 2008, pp. 1774–1781.

[17] F. M. Hoffman, J. Kumar, R. T. Mills, and W. W. Hargrove, "Representativeness-based sampling network design for the State of Alaska," *Landscape Ecol.*, vol. 28, no. 8, pp. 1567–1586, Oct. 2013.

[18] J. A. Hartigan, *Clustering Algorithms*. John Wiley & Sons, 1975.

[19] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering," in *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Jul. 1998, pp. 91–99.

[20] J. Kumar, R. T. Mills, F. M. Hoffman, and W. W. Hargrove, "Parallel k-means clustering for quantitative ecoregion delineation using large data sets," in *Proceedings of the International Conference on Computational Science (ICCS 2011)*, ser. Procedia Comput. Sci., M. Sato, S. Matsuoka, P. M. Sloot, G. D. van Albada, and J. Dongarra, Eds., vol. 4. Amsterdam: Elsevier, Jun. 2011, pp. 1602–1611.

[21] G. Brown, "*laspy 1.2.5*: Native Python ASPRS LAS read/write library." [Online]. Available: https://github.com/grantbrown/laspy

[22] H. Butler, "*GDAL 2.0.0*: Geospatial Data Abstraction Library." [Online]. Available: http://www.gdal.org/

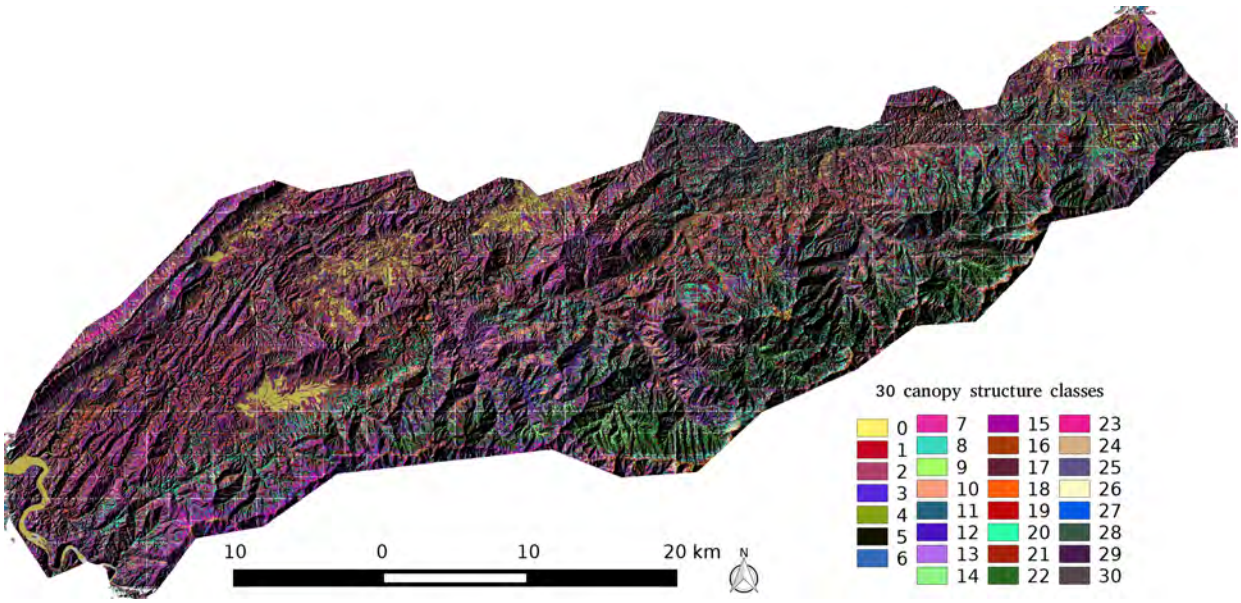
[23] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of k in k-means clustering," *Proceedings of the Institution of Mechanical Engineers,Part C: Journal of Mechanical Engineering Science*, vol. 219, no. 1, pp. 103–119, 2005. [Online]. Available: http://pic.sagepub.com/content/219/1/103.abstract

[24] M.-T. Chiang and B. Mirkin, "Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads," *Journal of Classification*, vol. 27, no. 1, pp. 3–40, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00357-010-9049-5

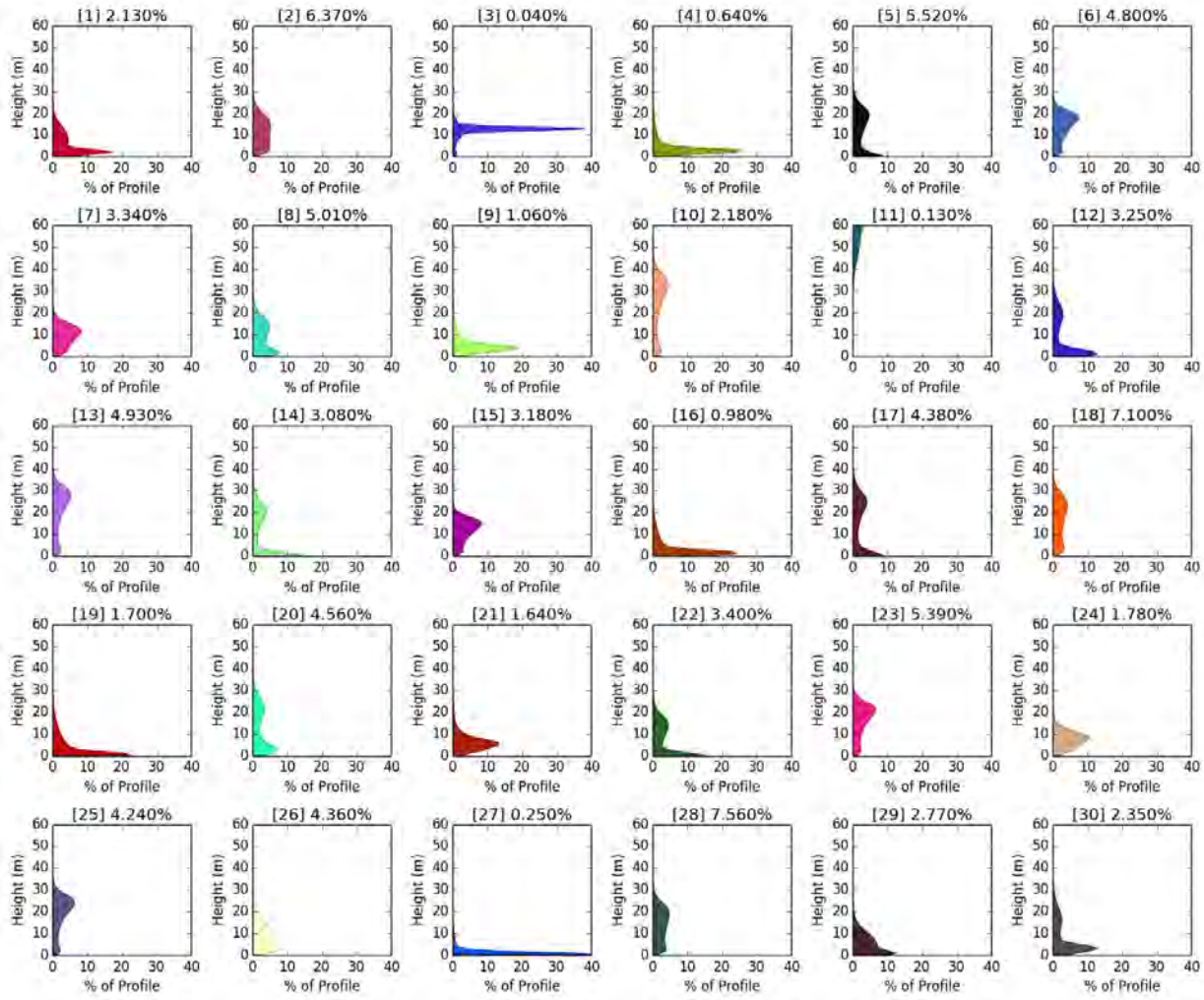
[25] W. W. Hargrove, F. M. Hoffman, and P. F. Hessburg, "Mapcurves: A quantitative method for comparing categorical maps," *J. Geograph. Syst.*, vol. 8, no. 2, pp. 187–208, Jul. 2006.

[26] J. A. Thiemann, C. R. Webster, M. A. Jenkins, P. M. Hurley, J. H. Rock, and P. S. White, "Herbaceous-layer impoverishment in a post-agricultural southern Appalachian landscape," *The American Midland Naturalist*, vol. 162, pp. 148–168, 2009.

[27] J. Kumar, J. Weiner, W. W. Hargrove, S. P. Norman, F. M. Hoffman, and D. Newcomb, "Vegetation canopy structure and distribution from LiDAR, Great Smoky Mountains, USA," 2015, ORNL DAAC, Oak Ridge, Tennessee, USA. [Online]. Available: http://dx.doi.org/10.334/ORNLDAAC/1286



(a)



(b)

Fig. 4. (a) 30 unique vegetation canopy structure classes identified by a k-means clustering algorithm for the Tennessee portion of the Great Smoky Mountains National Park. (b) Representative vegetation canopy structures that define the 30 unique canopy structure classes in (a). The percent of total area occupied by each class is described at the top of each class definition plot. Fill colors for the plots in (b) correspond to the colors for the class in the spatial map in (a).