Spring 5-19-2016

Exploring Data Mining Techniques for Tree Species Classification Using Co-Registered LiDAR and Hyperspectral Data

Julia K. Marrs
CUNY Hunter College

Exploring Data Mining Techniques for Tree Species Classification Using Co-Registered LiDAR and Hyperspectral Data

by

Julia Marrs

Submitted in partial fulfillment
of the requirements for the degree of
Master of Arts in Geography
Hunter College of the City of New York

2016

Thesis Sponsor:

05/19/2016	Dr. Wenge Ni-Meister
Date	First Reader
05/19/2016	Dr. Carsten Kessler
Date	Second Reader

Abstract

The use of LiDAR techniques for recording and analyzing tree and forest structural variables shows strong promise for improving established hyperspectral-based tree species classifications, but previous multi-sensoral projects have often been limited by error resulting from seasonal or flight path differences. NASA Goddard's LiDAR, Hyperspectral, and Thermal imager is now providing co-registered data on experimental forests in the United States, which are associated with established ground truths from existing forest plots. Free, user-friendly data mining applications like the Orange Data Mining Extension for Python have recently simplified the process of combining datasets, handling variable redundancy and noise, and reducing dimensionality in remotely sensed datasets. Data mining methods are used here to achieve a final tree species classification accuracy of 68% on Howland Experimental Forest, a mixed coniferous-deciduous forest with ten dominant tree species. This accuracy is higher than that produced using LiDAR or hyperspectral datasets separately, suggesting that combined spectral and structural data have a greater richness of information than either dataset alone. This work was performed on data aggregated above the individual tree level, thus the high classification accuracy achieved is encouraging given that many researchers predict shifting environmental conditions will necessitate future work at such a scale. Overall, the data mining methodology described here shows promise for integrating and analyzing remotely sensed datasets, and opens the possibility of addressing large-scale forestry questions like deforestation and carbon sequestration on a species-specific level.

1. Introduction

1a. History and Use of Light Detection and Ranging

Management and inventory of forest resources has been an essential concern of Geographic Information Systems since the advent of the discipline, beginning with the early Canadian Geographic Information System (CGIS) developed in the 1960s to monitor one of North America's most prolific natural resources (Foresman 1998). Since that time, the photogrammetric and other remote sensing techniques used to monitor global forests have undergone a technological revolution. Myriad satellite and airborne systems are employed by institutions and governments worldwide to monitor natural resources on a large scale. One of the most recent additions to the field of remote sensing has been the increasing use of single-wavelength laser light pulses to calculate very precise point elevations of features on and above Earth's surface. Initially conceived as a method for creating highly accurate, high-resolution topographic maps, Light Detection and Ranging (LiDAR) technologies are increasingly being employed to collect data on structural features of tree canopies and branching patterns, forest structure and succession, and even estimates of tree physiological metrics such as leaf area index (LAI), the ecological metric of total broadleaf surface area exposed to sunlight in a given area of forest (Korhonen *et al.* 2011). The use of LiDAR sensors, usually on airborne platforms such as small airplanes, has proved to be a boon to commercial forest resource monitoring and valuation. The ability to accurately assess the height and, in many cases, diameter at breast height (DBH) of each tree in a forest with a single flyover has greatly simplified the valuation of forests grown for timber (Schardt *et al.* 2002).

However, whether a forest is assessed for conservation or commercial purposes, one key element of forest systems has remained difficult to quantify; individual LiDAR data points have little to say about the species identity of a given tree, and few studies have so far used LiDAR sensors with sufficient point density to overcome this barrier. Species is a crucial attribute of trees, and one that is increasing in global relevance as climates shift and extreme weather events become more common. Changing weather patterns will reshape the ranges of species worldwide, and the ability to monitor the changes in community dynamics of trees and other plants, which play a fundamental role in overall ecosystem functioning and identity, will be key in understanding trends in terrestrial biomes and in creating effective strategies for conservation, resilience, and human livelihoods (Wulder *et al.* 2008).

Airborne laser scanning is an active remote sensing method used to collect LiDAR data. The sensor emits pulses of light in the near-infrared range at a high frequency, ranging from 50 – 300 kHz (or thousand pulses per second) depending on the instrument. Pulses are emitted from a laser mounted on an airborne platform, and may be directed along the flight path of the airplane or satellite in profiling LiDAR systems, or may move to scan a swath along a fixed angle beneath the platform in scanning LiDAR systems (Lim *et al.* 2003). Each laser pulse is emitted downward, and will reflect off any opaque surface in its direct path; in open terrain, the closest target may be the ground. In forested areas, tree tops, branches, or foliage may present a nearer target, though some laser pulses will still reach the forest understory or ground through gaps in the tree canopy. Depending on the number of targets off of which a single laser pulse is reflected on its way to the ground, a pulse of light may be reflected back to its source as a single point or as multiple returns of varying intensity. Using a rotating mirror, the airborne sensor collects returning laser pulses and records the round-trip travel time and intensity of each return. Using the known speed of light, round-trip travel time of each return can be converted into highly precise data on the elevation of the surface off which the laser pulse was reflected. Airborne laser scanners are also equipped with an onboard global positioning system (GPS) unit in order to account for the distance traveled by the airplane or satellite since the emission of the original laser pulse and as a means of tracking the flight path of the recording session (Campbell and Wynne, 2011).

The way in which this data on laser returns is recorded and stored depends on the model of the system and the intended method of future data analysis. Data may be stored onboard the sensor as discrete values potentially representing multiple returns from the same original pulse. Depending on the system, only the first and last return may be stored, or some or all of the intermediate returns may also be saved. Other systems record the entire waveform of a laser pulse, local maxima of which would correspond to the discrete returns recorded by the other system type. One benefit of these full-waveform systems is that the full duration of the reflection is recorded, and can be used to calculate the overall intensity of the signal. Recent studies have asserted that there is useful information to be gleaned from variation in waveform intensity and width beyond that provided by discrete amplitude values, and that this information could be useful in discriminating among tree species (Heinzel and Koch 2011). However, most available ALS systems are still constrained by a tradeoff between high pulse density, discrete-return systems that can discern detailed structural differences among individual trees and the ability to collect and record full-waveform data.

1b. Deriving Structural Information from LiDAR Point Clouds

Before processing, discrete-return LiDAR is stored as a point cloud representing elevations of individual returns throughout the flight path. Various methods including single or variable elevation thresholds, segmentation algorithms, and other computational techniques can be used to separate out vegetation and canopy points from ground returns. In forest sites, understory, shrub layers, and accumulated snowfall may also be separated out into another category. As early as 1985, LiDAR sensors were being used to trace the height profile of tree canopies and differentiate between ground and tree returns (Schreier *et al.* 1985). Subsequent work focused mainly on commercial applications including the recording of accurate tree heights for use in calculating timber value, a problem that continues to challenge researchers using LiDAR for commercial forestry purposes. Although LiDAR data can provide elevation data with a very high precision, the chance that even high-density laser pulses will hit each tree exactly at its highest point during a whole-forest flyover is slim (Brandtberg 2007). Furthermore, LiDAR data collected during leaf-on conditions typically record leaf canopy height more accurately than stem height. Several models have been proposed for minimizing the error between these two metrics (Magnussen *et al.* 1999).

Counter intuitively, though small-footprint scanning laser systems had been in use for gathering data on the structure of individual trees, overall canopy structure is more easily determined using large-footprint laser scanning. The advent of the National Aeronautics and Space Administration (NASA)’s Scanning LiDAR Imager of Canopies by Echo Recovery (SLICER) sensor allowed for three-dimensional canopy modeling by gathering laser echoes from within an area larger than a tree crown, thus allowing for the identification of gaps in the tree canopy. This method proved more useful than attempting classifications on spectral data from NASA’s Landsat and Airborne Visible / Infrared Imaging Spectrometer (AVIRIS) satellite sensors in distinguishing canopy structural characteristics (Lefsky *et al.* 2001). Though SLICER data proved useful in examining vegetative strata and canopy properties, it was discovered that internal tree geometry affected models’ ability to predict gaps, implying that important data on branching patterns also exists within LiDAR point clouds (Ni-Meister *et al.* 2001).

The overall branching pattern, or bifurcation ratio, of growing trees has been shown to be generally characteristic of a plant species (Whitney 1976, Borchert and Slade 1981). Though plants show phenotypic plasticity in branch and leaf arrangement in response to limited light or other competitive stress (Canham 1988), trees of the same species growing in the same area (such as a forest stand) could be assumed to have a similar and identifiable branch arrangement. The development of species-specific branching patterns in heterogeneous forest stands has even been proposed as an adaptive mechanism for successful species co-existence (Ishii *et al*. 2003). It should thus be possible to take advantage of high-density LiDAR data to examine branching patterns and other architectural data on a single-tree or tree stand basis, and to relate this to the species of the individual or the predominant species in a stand. This technique has been used to differentiate between species in conifer forests with success (*e.g.* Donoghue *et al.* 2007), but methods for the optimal use of LiDAR on deciduous and mixed forests are still developing. Though no system yet exists with sufficient point density to map the exact branching details of each tree in a forest, it is possible to analyze the tree height information in a LiDAR point cloud and to calculate a variety of summary metrics that serve as a reasonable proxy for the variability of branch heights and angles within a given area. Some early LiDAR studies attempted this kind of analysis by calculating standard deviation, skewness (asymmetry of distribution), and kurtosis (peakedness of distribution) of height values, as well as segmenting the point cloud into three or more layers (*e.g.* Holmgren and Persson 2004). However, the utility of such metrics in creating accurate species-level classifications was limited, possibly because of the lower point-density afforded by sensors at the time. Researchers also noted the potential for error from differences in reflectance due to different bark colors of trees sampled in leaf-off conditions (Brandtberg *et al.* 2003) and the limitations of LiDAR's utility in accurately detecting smaller trees in dense understory (Maltamo *et al.* 2004). Though current allometric models of the relationship between tree height and diameter do not predict an exact linear relationship (Pretzsch 2013), it is still clear that larger trees have exponentially more biomass than shorter trees, due to the geometric relationship between height and volume of a three-dimensional structure like the central stem of a tree. Thus, the tendency of LiDAR data to omit information on small trees may be negligible for commercial assessments of overall timber volume, but it naturally presents a problem for accurate species classifications using LiDAR data alone.

1c. Discrete-Return LiDAR and Summary Metrics

As the use of LiDAR data for analyzing vegetative structure became more common, processing methods for summarizing LiDAR height values eventually proliferated into a variety of metrics, many of them specific to one article or team. In response, a set of recommendations and standards for LiDAR data collection and processing, including definitions of common metrics, was published (Evans *et al.* 2009). These are the standards used and referenced in the following work (see Methods section for further details and definitions). A number of studies performed since this time have generally taken advantage of the structural information summarized in these standard metrics to attempt classifications of tree species distributions in forest sites. Though the final accuracy of some of these classifications has been limited by the point density of the LiDAR data, many groups have succeeded in distinguishing between at least major taxonomic groups (*i.e.* conifers vs. deciduous trees) and even among individual tree species. Recent studies carried out with airborne laser scanners capable of recording six or more returns per square meter have been able to achieve considerable success using only summary metrics on point elevations. Korpela *et al.* (2009) were able to classify the distributions of three Scandinavian tree species with accuracies up to 93%. Other groups have reported similar accuracies in separating out spruce trees from birches (Ørka *et al.* 2009) and in distinguishing among coniferous, deciduous, and mixed-forest stands (van Aardt *et al.* 2008), with an emphasis on the power of point density deciles in explaining between-group variability. Such metrics give information on the vertical distribution of branches and foliage in a single tree crown or within a tree stand, depending on the resolution of the final processed raster. Even rasters with a relatively large pixel size have proved useful in sites with single-species or known mixed-species plots with historical management regimes. Analysis of vegetation strata, with a focus on density metrics, was successfully used to classify even multi-layered canopies among managed tree plots (Morsdorf *et al.* 2010).

Though classifications of relatively homogenous forests have achieved success, some limitations in the utility of summary LiDAR metrics have been found when attempting to extend similar classifications to a larger number of tree species (a phenomenon documented and investigated by Alonzo *et al.* 2014). Classifications performed on Scandinavian forests with only a few tree species have achieved greater total accuracy than those attempted on tropical forests with both higher diversity and a more heterogeneous canopy and sub-canopy structure (Gillespie *et al.* 2004). This challenge has also been confronted in non-tropical forests that are in an intermediate stage of forest succession; small trees in the forest sub-canopy may be primarily composed of a less dominant species that is being suppressed or outcompeted. This poses a double problem in that such trees may be both less common and more difficult for segmentation algorithms to find in the point cloud, meaning that errors in individual tree detection and uneven distribution in the absolute number of trees in each species category may be simultaneously detrimental to classification accuracy (Ørka *et al.* 2009). Furthermore, debate continues over the optimal resolution of LiDAR data, at least in comparison to individual crown size. While some researchers warn against trying to characterize species distribution data on anything other than the individual tree level (Yu *et al.* 2010), others have asserted that there is unavoidable within-species variability due to an individual-tree effect that explains up 65% of intraspecies variability (Hovi *et al.* 2016). The latter researchers therefore highlight the necessity of applying classification methods to aggregated tree groups for large-scale forest inventory or classification.

1d. Full-Waveform LiDAR

In an attempt to address the potential shortcomings of discrete-return LiDAR, some studies have incorporated the additional data offered by full-waveform LiDAR readings. Though acquiring such datasets may necessitate the purchase of another sensor that can record such readings, full-waveform measurements offer up the possibility of calculating total intensity values of each return. Such measurements are favored by some researchers because full-waveform readings measure total backscatter, thus providing information on all canopy and sub-canopy levels, as well as potentially on smaller tree components like cones or flowers, and even on tree-dwelling epiphytes or bromeliads that contribute to forest biodiversity (Korpela *et al.* 2010). Though this type of dataset may represent an untapped well of information, others have claimed that recording every echo will bury information useful to species classification inside unnecessary noise. Similarly, some worry that full-waveform recordings reduce the comparability of LiDAR datasets by introducing a seasonal effect related to budding and flowering, even in leaf-off data (Kim *et al.* 2011).

Despite these potential confounding factors, return intensity information derived from full-waveform measurements has been used in combination with a targeted set of summary metrics from discrete-return data to produce accurate classification results (Heinzel and Koch 2011). Other researchers have incorporated intensity measurement as estimates of biomass and single-tree DBH values, thereby linking carbon budget estimates with species information (Yao *et al.* 2012). Full-waveform data have also been used to perform more robust analysis of internal structural features of tree canopies than point elevation alone can facilitate, for example by calculating co-occurrence matrices representing density of LiDAR points as 3-D voxels within each tree (Li *et al.* 2013).

1e. Optical and Spectral Remote Sensing

Data on the differences in the intensity and wavelengths of light reflected off various structures on or near Earth's surface have historically constituted the bulk of remotely sensed information. Sensors attached to airplanes or satellites are able to passively record the reflectance of sunlight off of surfaces below their flight path and to separate this reflectance data into distinct ranges of wavelengths or bands. This allows researchers to manipulate the visual display of such data and to examine differential reflectance patterns recorded similar or adjacent material in more detail or in different ways than photographs can provide. Optical remote sensors, for example, may use panchromatic sensors record data on brightness of all light reflected in the visible spectrum to create black and white images that may represent more spatial detail than an equivalent color image could reproduce. Multispectral sensors take advantage of digital image displays, which represent images as combinations of three primary colors. By recording in more than three channels, multispectral datasets contain more bands than available colors on, for example, a computer monitor. These data can therefore be visualized in a non-standard way with the goal of emphasizing characteristics not normally visible, for example by choosing band combinations that provide information on relative water concentration across an area of vegetation. While multispectral datasets typically contain fewer than ten bands, hyperspectral datasets are recorded in dozens or hundreds of distinct channels, each producing bands that may be separated by as little as one or two nanometers. Both multispectral and hyperspectral datasets typically contain reflectance data from visible and infrared wavelengths, thereby further improving the capability of these datasets to represent details of non-visible characteristics of the terrain (Campbell and Wynne 2011).

Multispectral and hyperspectral datasets have widely been used for characterizing and classifying individual tree species. The basis for these classifications arises from the small differences in light reflectance off of leaves with distinctive pigment concentrations characteristic of one species (*e.g.* Blackburn 2007). Particularly detailed distinctions can be made with hyperspectral datasets, but multispectral data has also been used with success, particularly when it is combined with structural information from LiDAR measurements of the same forest (Holmgren *et al*. 2008). Fused LiDAR and remotely sensed optical datasets (*e.g.* from WorldView-2) have also been used to link carbon budget parameters and species ranges, thus opening the possibility for species-level carbon budgeting, with obvious important implications for valuing and maintaining forest resources (Karna *et al.* 2015). Combined LiDAR and optical datasets have also been used to create predictive models of North American tree species distribution and relative abundance (van Ewijk *et al.* 2014). Hyperspectral data have also been used on their own for tree species classifications. Because of the high dimensionality of hyperspectral datasets and the inherent likelihood of collinearity among readings produced by similar wavelengths, most articles present an analysis of a reduced set of hyperspectral variables produced by dimensionality reduction methods like principal components analysis or independent components analysis. Even if all bands are kept for analysis, the use of spectral libraries or the identification of a few pure pixels existing in the dataset are needed as ground truths (*e.g.* Plourde *et al*. 2007).

1f. Fused Hyperspectral and LiDAR Datasets

The level of detail contained in spectral datasets means that the explanatory power they offer may be quite high. Because of the wealth of information contained in hyperspectral datasets in particular, some analyses have found that the addition of other remotely sensed data like LiDAR measurements provides little improvement to classification accuracies. In mixed forest types, researchers have found that, while LiDAR metrics other than absolute height are able to explain a significant portion of variability on their own, they provide little benefit when added into a hyperspectral-based classification (Dalponte *et al.* 2012). Similar results have been found in studies using vegetative indices derived from hyperspectral datasets, even when LiDAR data are of a very high point density (12 points/m^2) that should contain robust structural information (Ghosh *et al.* 2014). However, when testing their results by resampling their data, this same study found that the optimal set of spectral and LiDAR metrics for distinguishing among tree species was different for each scale they tested, and the authors end by leaving open the question of how best to account for this interaction between scale and “best” classifiers. Similarly, some studies that claim that LiDAR data have limited utility in forest classifications that employ unsupervised classification methods to identify classes of forest types, rather than distributions of individual species. However, the ambiguous definition of a “forest type” leaves open the possibility that a classifier may not take advantage of an optimal set of metrics for the tree species that were actually present in the forest site in question (Hill and Thomson 2005).

On the other hand, some recent studies have found that the incorporation of LiDAR data is able to provide a major improvement. The calculation from LiDAR point clouds of volumetric canopy profiles designed to capture species- and growth stage-specific structural information has been shown to produce high classification accuracies. Additionally, the combination of this LiDAR data with hyperspectral datasets has yielded an improved species-level tree classification in comparison to classifications produced by either dataset alone (Jones *et al.* 2010). Datasets using discrete-return LiDAR data collected at a relatively low point density have leveraged the existing data in order to examine the relationship between canopy height and more basic structural characteristics like canopy cover, rather than species-level differences. Such studies have found that the use of LiDAR data to remove the influence of canopy gaps on hyperspectral-based classification improves the ability to find a relationship between reflectance and leaf pigment concentrations (Blackburn 2002), implying that fused datasets have the potential to greatly improve classifications of most of the world's forest area.

Combined LiDAR and hyperspectral datasets have also been used for ecological and vegetation-related surveys other than forest tree classification. For example, predictive modeling studies on invasive plant species distributions and ranges have relied on LiDAR mostly for analyzing ground features for habitat suitability. Those that have incorporated LiDAR to look at invasive shrubs and other low vegetation, however, have tapped into LiDAR's potential dual use in modeling both species distributions and underlying habitat features (Andrew and Ustin 2009). Combined LiDAR and hyperspectral datasets have also been used in evaluating the risk of forest fires based on vegetative properties, with a large improvement in accuracy achieved with either dataset alone (Koetz *et al.* 2008).

1g. Analytical Challenges and Data Mining Methods

The development of remote sensing technologies has given researchers the opportunity to work with unprecedented volumes of information on large areas of forest or other terrain. However, the high dimensionality of large datasets, and of hyperspectral datasets in particular, introduces new challenges into the process of analyzing and utilizing these data. The Hughes phenomenon describes the problem of decreasing predictive power of additional variables that contain information on a fixed number of known classes. In the case of tree species prediction, hyperspectral data on a forest for which the researcher has information on only a small number of ground truth areas might be more redundant than insightful (Dalponte *et al.* 2009). For this reason, machine learning and data mining techniques for dimensionality reduction and pattern finding are often employed in species classification studies, as well as for predictive models of species distributions or habitat suitability. Data mining methods are designed to take advantage of cases where a few known cases or ground truths are being used to characterize a larger area or dataset. For this reason, they are recommended above, for example, linear regression models when attempting to find explanatory patterns in data without preexisting rules or assumptions (Franklin 2009).

Decision trees split a dataset at nodes that represent whichever value of a variable is determined to best partition cases into groups with high internal similarity. The dataset is first divided in two at some optimal separation point, and each group is further subdivided into ever-smaller categories (Olden *et al.* 2008). Such trees have been widely used in data mining studies, but concerns that standard decision trees may overfit data and be less generalizable to the wider problem of interest has lead to the use of other data mining techniques as well. The random forest method avoids overfitting by constructing multiple decision trees on random subsets of one dataset (Cutler *et al.* 2007, Prasad *et al.* 2006). This method has been favored when creating species-specific classifications, due to the option of incorporating categorical data directly into the classification (Yu *et al.* 2011).

In addition to decision trees, support vector machines and k-nearest neighbors techniques also rely on the identification or assembly of similar groups of data points. Support vector machines iterate through a dataset to assign cases to categories determined by training data, eventually creating an optimal divider among categories, or hyperplane, with one fewer degree of dimensionality as the original dataset (Vapnik 1982). Many previous hyperspectral-based studies have preferred support vector machines because of their ability to handle problems in which there is no single clear solution (ill-posed problems) and to operate well on datasets with a limited number of ground truths (Mountrakis *et al.* 2011). The k-nearest neighbors method identifies cases that are either spatially or informationally proximal to training cases, and weights the known identity of the neighbor more heavily when determining a classification for unknown cases (Dudani 1976).

Other data mining methods rely on the development of rules for classification based on training data. The CN2 Rules algorithm is designed to induce rules from training data. Its main distinguishing feature is its ability to create rules that it can apply to unknown data that fit one category well, but imperfectly, rather than excluding all imperfect matches (Clark and Nibblet 1989). Neural networks consist of a system of individual learners designed to communicate like neurons in living organisms. Each “neuron” in a neural network learner tests a case against rules learned from training data and passes information to the next. These rules can shift and change as the network handles more data, similar to learning in the brain (Haykin 2004). The naïve Bayes learner is an algorithm that employs Bayesian probabilistic rules for making predictions about the value of one variable based on known information on other, related variables (Zhang 2004). The naïve Bayes learner assumes independence between these variables, and thus may not be well suited to a dataset such as the one used here, where values in one hyperspectral band are likely to be very similar to those in a band with a similar wavelength range (Rennie *et al.* 2003).

Ensemble data mining methods are used to combine several of the above described methods. Random forests, for example, are in fact an ensemble method applied to classification trees; a similar principle can be used to implement several different classifiers at once with the goal of minimizing misclassifications by comparing the overlapping predictions of each algorithm (Polikar 2006). The use of data mining methods to refine a set of variables for use in further classification and prediction work has been documented in several recent studies using remote sensing data for tree species classification (Holmgren and Persson 2004, Næsset 2007, Morsdorf *et al.* 2010, Kim *et al.* 2010, Vauhkonen *et al.* 2010) and shows strong promise for use in future work.

1h. Addressing a Gap

Though significant work has been done using LiDAR and spectral data collected on the same area, few studies have used co-registered hyperspectral and LiDAR data for plant species classifications. One exception is a study reporting classification accuracies up to 89% when using co-registered data to map the distribution of a single sagebrush species (Mundt *et al.* 2006). The use of datasets collected concurrently on the same flyover presents obvious benefits for avoiding error due to seasonality, tree growth between data collection campaigns separated by years, resampling to normalize pixel size, and differences in flight paths. For this reason, NASA Goddard created the LiDAR, Hyperspectral, and Thermal Imager (G-LiHT), an airborne sensor that came online in 2012. In the article detailing the specifications and goals of this imager (Cook *et al.* 2013), the authors stated:

> "The complimentary nature of LiDAR, optical and thermal data provide an analytical framework for the development of new algorithms to map plant species composition, plant functional types, biodiversity, biomass and carbon stocks and plant growth."

This mission statement underscores the suitability of data collected by G-LiHT for use in tree species classification. The G-LiHT data currently available includes flyovers of several experimental forests in the Northeastern United States. Some recent articles (*e.g.* Morsdorf *et al.* 2010) have recommended the use of data from experimental forests as ground truths or training classes because of the possibility for directly connecting classification results to preexisting ecological research. A field campaign undertaken in 2009 surveyed trees in the same experimental forests as the flyovers, generating species-level information that can be used in exactly this way. Thus, it is clear that there are existing datasets that are ideally suited to respond to a gap in the literature. In doing so, this study seeks to investigate efficient and novel methods for monitoring plant species ranges, for inventorying natural resources, and for tracking the effects of shifting climate patterns on forest health and composition. This work will help to assess how refined LiDAR data can help to improve hyperspectral-based classifications, and will compare several data mining techniques in order to investigate the suitability of the available methods to generating species-level tree classifications using co-registered remotely sensed data of different types. The overall purpose of this study was to explore the use data mining techniques to refine a list of available LiDAR metrics into a smaller subset of those with the most explanatory power, and to combine this subset with hyperspectral-based classifications in order to optimize the contribution of structural information to a fused dataset classification.

2. Methods

2a. G-LiHT Specifications

The G-LiHT imager is composed of several off-the-shelf remote sensing products that were selected for their compact size, high resolution, and compatibility. These components include the following: an RT4041 (Oxford Technical Solutions, Oxfordshire, UK) GPS/ Inertial Navigation System (INS), a VQ-480 (Riegl USA, Orlando, FL) scanning LiDAR sensor, a LD321-A40 (Riegl USA, Orlando, FL) profiling LiDAR sensor, a Hyperspec^{TM} VNIR Concentric Imaging Spectrometer (Headwall Photonics, Fitchburg, MA), a ruggedized RA1000m/D digital fine gain imaging camera (Adimec, Stoneham, MA), an Ocean Optics USB 4000-VIS-NIR spectrometer (Dunedin, Fl, USA) for measuring downwelling radiance, a Gobi-384 thermal imaging camera (Xenics, Leuven, Belgium), and an onboard PC for data storage during flyovers (Cook *et al*. 2013). For further details on instrument calibration and attachment to the Cessna 206, NASA UC-12B (King Air), or Piper Cherokee aircraft used for flyovers, see the G-LiHT White Paper (Cook and Corp 2012).

2b. Remotely Sensed Data

Data from flyovers conducted in June 2012 can be found at the G-LiHT data archive at *ftp://fusionftp.gsfc.nasa.gov/G-LiHT.* The scanning LiDAR sensor produces point clouds that have been processed into standard metrics (as described in Evans et al. 2009). Definitions for each of these metrics can be found in Table 1. Data from the profiling LiDAR sensor were used to create a canopy height model available for each site. All LiDAR data are available on the G-LiHT data archive in geotiff format, with data aggregated to 13m^2 pixels.

Table 1: List of LiDAR Metrics and Abbreviations

Name and Description of Metric	Units	Abbreviation
Mean Absolute Deviation = mean(|height – mean height|) of tree returns	meters	AAD
Canopy Relief Ratio = (mean-min:max-min) of tree returns	unitless	CRR
Density deciles (10% increments) of tree returns	fraction	D0 – D9
Fraction of first returns intercepted by tree	fraction	FCover
Fraction of all returns classified as tree	fraction	Fract_All
Interquartile range (P75 - P25) of tree returns	meters	IQR
Kurtosis of tree return heights	meters	Kurt
Median Absolute Deviation = median(|height - median height|) of tree returns	meters	MAD
Mean of tree return heights	meters	Mean
Height percentiles (10% increments) of tree returns	meters	P10 – P100
Rugosity (Standard deviation of gridded canopy height model values)	meters	Rug
Quadratic mean of tree return heights	meters	QMean
Skewness of tree return heights	meters	Skew
Standard deviation of tree return heights	meters	StDev
Vertical Distribution Ratio = [P100 - P50] / P100	unitless	VDR

Hyperspectral data for the Howland Experimental Forest site can also be found on the G-LiHT data archive. Available data include at-sensor reflectance data covering a spectrum between 418 – 918 nm, with an approximately 4.5-nm interval between bands for a total of 114 individual bands. A total of 44 different vegetative indices calculated from these reflectance measurements are also available. Also recorded are data on radiance along each swath, as well as ancillary data on flight path, atmospheric conditions, potential errors in data collection, and other data acquisition conditions. Reflectance and vegetative index data are available for each swath as well as in a mosaicked version, which is the version used in the following analysis. Table 2 presents a list of vegetative indices used in further analysis.

Many vegetative indices are designed to emphasize the concentration or exact reflectance of plant leaf pigments (Agapiou et al. 2012, Verrelst *et al.* 2008). The Anthocyanin Reflectance Indices 1 and 2 (Gitelson *et al.* 2001) and Carotenoid Reflectance Index 2 (Gitelson *et al*. 2002) look particularly at carotenoid and anthocyanin pigments related to plant stress and senescence. The Photochemical Reflectance Index measures xanthophyll pigment content in order to estimate photosynthetic activity and efficiency (Gamon *et al.* 1992). Other indices compare the concentrations of leaf pigments: the Red Green Ratio Index calculates the ratio of anthocyanin concentration to chlorophyll concentration to determine the source of leaf redness, as a proxy for plant type and life stage (Gamon and Surfus 1999). Similarly, the Structure Insensitive Pigment Index compares the ratio of carotenoid concentration to chlorophyll *a* concentration as another proxy for plant life stage (Peñuelas *et al.* 1995).

Chlorophyll is the key photosynthetic pigment, and is unsurprisingly the subject of numerous vegetative indices. Leaf chlorophyll content is estimated by several indices, mostly by performing calculations on bands in the red or infrared ranges of the electromagnetic spectrum (referred to as the red edge), since these are the wavelengths most dramatically absorbed by green foliage, thus showing the strongest spectral signature in vegetated areas (Filella and Peñuelas 1994). The Red Edge Inflection Point quantifies the exact wavelength at which this effect appears in a specific area or species of vegetation (Broge and Leblanc 2000). Total chlorophyll content is estimated in the Gitelson and Merzlyak 1 and 2 indices by calculating a ratio of reflectance values for wavelengths in the near infrared and red ranges (Gitelson and Merzlyak 1997), and in the Greenness Index by comparing two wavelengths on the red edge of the visible spectrum (Zarco-Tejada *et al.* 2005). The derivative of values recorded in a single red band is used to calculate chlorophyll content in the Datt 2 index (Datt 1999). A ratio involving three red bands is used to characterize chlorophyll content in the MERIS Terrestrial Chlorophyll Index (Dash and Curran 2004). The Vogelmann index uses as similar calculation on red wavelength values to examine both chlorophyll content and water content (Vogelmann 1993). Vegetation cover and density, which are often used in estimates of total photosynthetic activity in an area, can be estimated via the Renormalized Difference Vegetation Index, which compares near infrared wavelength reflectance values to those in the visible spectrum (Roujean and Breon 1995).

Table 2: Select List of Hyperspectral Vegetative Indices and Abbreviations
List of full names and abbreviations of vegetative indices used in later analysis. For further justification, see methods section and Table 5.

Name of Vegetative Index	Abbreviation
Anthocyanin Reflectance Index 1	ARI1
Anthocyanin Reflectance Index 2	ARI2
Carotenoid Reflectance Index 2	CRI2
Datt 2	DATT2
Gitelson and Merzlyak 1	GM1
Gitelson and Merzlyak 2	GM2
Greenness Index	GI
MERIS Terrestrial Chlorophyll Index	MTCI
Photochemical Reflectance Index	PRI
Renormalized Difference Vegetation Index	RDVI
Red Edge Inflection Point	REIP
Red Green Ratio Index	RGRI
Structure Insensitive Pigment Index	SIPI
Vogelmann	VOG

During analysis, two datasets were removed because of incomplete data. The rugosity file for Howland Experimental Forest contained no data, so the matching rugosity file from the Penobscot site was removed in the interest of an equal comparison between sites. There were also two missing values in the MRESR dataset. The neural network data mining method is incompatible with missing data values, so this index was also removed.

2c. Field Campaign

Data on tree species and locations were collected in NASA-funded field campaigns to Penobscot Experimental Forest and Howland Experimental Forest in Maine, USA in 2009. These experimental forests are predominantly evergreen forests. Howland Experimental Forest is a 558-acre forest with a centroid at 45°12'00" N, 68°44'00" W. Penobscot Experimental Forest is an approximately 3,900-acre forest with a centroid at 44°85'20" N, 68.62'00" W. Data were collected in forest plots of 50 m × 200 m, each of which is divided into 16 subplots of approximately 25 m x 25 m, arranged as shown in Figure 1. During the summer 2009 field campaign, data on the species, diameter at breast height, and estimates of aboveground biomass (AGB) calculated using the formula described by Jenkins *et al.* 2003 were recorded for each tree in these plots (Montesano *et al.* 2013). However, exact coordinates were not recorded for individual trees, meaning that they can be located with only the precision of the subplot in which they reside.

Figure 1: Layout of Subplots Within Whole Plots

L1	L2	L3	L4	L5	L6	L7	L8
R1	R2	R3	R4	R5	R6	R7	R8

2d. Data Preparation

In order to account for this spatial limitation, the dominant species in each subplot was determined using Excel pivot tables. The species with the greatest number of individual trees was chosen as the dominant species. In cases a tie between two or more species, the tie was resolved by picking the dominant species for the whole plot to which the subplot belonged. A list of tree species abbreviations is given in Table 3.

Table 3: Tree Species Abbreviations

	Howland Experimental Forest	**Penobscot Experimental Forest**
Dominant Tree Species: Abbreviation, Common Name, and Latin Name	ABBA = balsam fir (*Abies balsamea*)	ABBA = balsam fir (*Abies balsamea*)
	ACRU = red maple (*Acer rubrum*)	ACRU = red maple (*Acer rubrum*)
	FAGR = American beech (*Fagus grandifolia*)	ACSA = silver maple (*Acer saccharinum*)
	FRAM = white ash (*Fraxinus Americana*)	ACSP = mountain maple (*Acer spicatum*)
	PIAB = Norway spruce (*Picea abies*)	BEAL = yellow birch (*Betula alleghaniensis*)
	PIMA = black spruce (*Picea mariana*)	BEPA = paper birch (*Betula papyrifera*)
	PIRU = red spruce (*Picea rubens*)	BEPO = gray birch (*Betula populifolia*)
	PIST = eastern white pine (*Pinus strobus*)	FAGR = American beech (*Fagus grandifolia*)
	THOC = northern white-cedar (*Thuja occidentalis*)	PIRE = red pine (*Pinus resinosa*)
	TSCA = eastern hemlock (*Tsuga Canadensis*)	PIRU = red spruce (*Picea rubens*)
		PIST = eastern white pine (*Pinus strobus*)
		POGR = bigtooth aspen (*Populus grandidentata*)
		POTR = quaking aspen (*Populus tremuloides*)
		THOC = northern white-cedar (*Thuja occidentalis*)
		TSCA = eastern hemlock (*Tsuga Canadensis*)
Additional Tree Species: Abbreviation, Common Name, and Latin Name	ACPE = striped maple (*Acer pensylvanicum*)	FRAM = white ash (*Fraxinus Americana*)
	BEAL = yellow birch (*Betula alleghaniensis*)	FRPE = green ash (*Fraxinus pennsylvanica*)
	BEPA = paper birch (*Betula papyrifera*)	OSVI = eastern hophornbeam (*Ostrya virginiana*)
	BEPO = gray birch (*Betula populifolia*)	POBA = balsam poplar (*Populus balsamifera*)
	LALA = tamarack (*Larix laricina*)	QURU = northern red oak (*Quercus rubra*)
	POGR = bigtooth aspen	TIAM = American basswood

	(*Populus grandidentata*)	(*Tilia americana*)
	POTR = quaking aspen (*Populus tremuloides*)	ULAM = American elm (*Ulmus americana*)

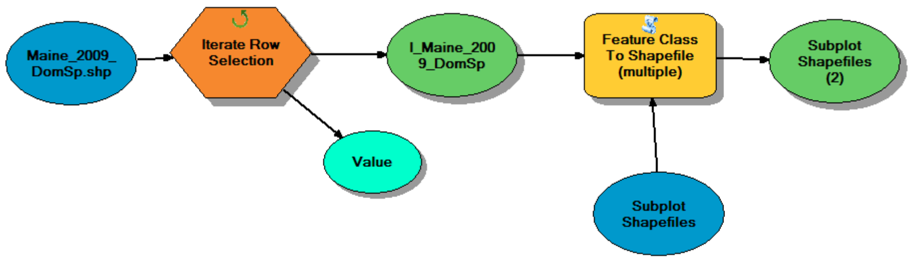
In four cases, none of the species involved in the tie was the dominant species in the corresponding whole plot. These four ties were resolved as discussed in Table 4.

Table 4: Justification For Resolving Dominant Species Ties in Four Subplots
In this table, H indicates a plot in Howland Experimental Forest and P indicates a plot in Penobscot Experimental Forest.

Subplot	Classification and Justification
H02R2	TSCA chosen – dominant for several other subplots in plot H02
P10R7	ACSA chosen – neighboring subplot P10R8 is dominantly ASCA
P14L2	POGR chosen – dominant in other subplots in plot P14, unlike tied option ACRU
P14R5	PIST chosen – dominant for several other subplots in plot P14

A spreadsheet of dominant species information was then joined to the original shapefile of plot locations using the Table Join tool in ArcGIS. The original shapefile for each experimental forest site represented plots as multipolygons, each composed of its 16 constituent subplots. In order to create separate subplot shapefiles for use in further analysis, a custom tool was created using the ArcGIS ModelBuilder. This tool was designed to iterate through the original plot shapefile and create a new shapefile from each row (representing a subplot) using the Feature Class to Shapefile (Multiple) tool (Figure 2).

Figure 2: ModelBuilder Tool for Exporting Rows as Individual Shapefiles



The resulting subplot shapefiles were then combined with the information contained in the LiDAR and hyperspectral geotiffs. The LiDAR geotiffs were directly downloaded from the G-LiHT data archive. The hyperspectral geotiffs were created using the Save File As > TIFF/GeoTIFF function in ENVI Classic (version 5.2) to save each band as a separate geotiff.

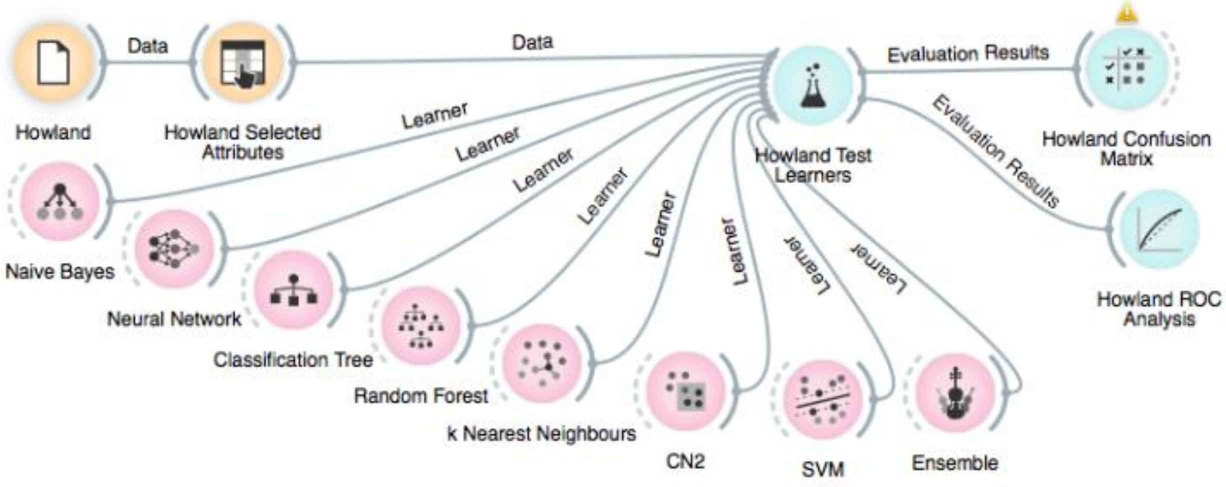
Further data processing steps were performed using the ArcPy package for Python (Python Software Foundation). The different pixel size of the LiDAR metric geotiffs and the hyperspectral geotiffs necessitated the use of two different methods for obtaining the mean value for each subplot. For the LiDAR data, a set of points representing the centroid of each subplot was created using the Feature To Point tool in ArcGIS. These centroids could then be used as points for interpolation of pixels in the underlying LiDAR geotiff layers containing data on LiDAR metrics. Using the Extract MultiValues To Points tool while iterating through the list of geotiffs, a new column containing the interpolated value for each subplot centroid was added to the shapefile. The Extract MultiValues To Points tool uses a bilinear interpolation method (ESRI *n.d.*), which calculates a mean value using the value of the pixel underlying the centroid as well as the values of the four pixels bordering it in a "T" or plus shape. This method was identified as generating the most representative mean values for each shapefile by visual comparison with other interpolation methods, such as the Zonal Statistics method described below. Using the ArcPy ListValues and InsertCursor methods, the attribute tables of all shapefiles were exported as comma separated value tables for use in data mining work. All Python code can be found in Appendix 1.

For the hyperspectral geotiffs of both vegetative indices and reflectance bands, the Zonal Statistics as Table function (ESRI 2011) was used to create an average of all the pixels the majority of whose area overlaps with the subplot shapefile. A custom Python script was used to run this function on all hyperspectral geotiffs, thereby generating an ESRI info table with the mean value for each subplot. Using the Feature Class to Table (Multiple) function in ArcCatalog, these info tables were saved as dbase tables, which were then opened in Microsoft Excel and concatenated, resulting in tables containing the mean value for each subplot of each vegetative index or reflectance band. These two tables were saved as comma separated value tables for use in data mining work.

2e. Data Mining Methods, Accuracy, and Validation

All data mining work was carried out using the Orange data mining toolbox for Python (Demsar *et al.* 2013). In addition to being compatible with Python for batch processing and the creation of custom tools, Orange offers a visual programming interface that can be downloaded at *http://orange.biolab.si/*. Initially, available data mining methods including naïve Bayes learner, k-nearest neighbors, neural network, classification tree, random forest, support vector machine, and CN2 rules methods, along with an ensemble method for combining multiple methods, were tested simultaneously to determine which produced the best classifications. Further information on algorithm implementation in Orange can be found at the Docs link on the above cited Orange webpage. A sample workflow is reproduced in Figure 3.

Figure 3: Sample Orange Workflow for Comparing Data Mining Methods



All the above listed data mining methods assess the performance of the classifications they produce by calculating the validity of the results the classification produces. Validity is assessed using several metrics to quantify potential sources of error. In addition to correctly classifying cases in a given category and rightfully excluding other cases from this category, it is also possible for error to be introduced in two ways: commission error (also referred to as Type I error or the false positive rate) quantifies the percentage of cases incorrectly classified in each possible category, while omission error (also known as Type II error or the false negative rate) quantifies the percentage of cases left out of the category in which they should have been included. In remote sensing datasets, known ground truth measurements are used to construct classification rules. These same cases are “blindly” classified according to these rules and the resulting differences are used to produce assessments of classifier performance and measures of classification accuracy. Overall accuracy or classification accuracy is determined by summing the number of correctly classified pixels (true positives and true negatives) and dividing by the total number of pixels to produce a ratio or percentage.

Similarly, classifications can be assessed for their accuracy for certain applications. User’s accuracy is calculated by dividing the number of correctly classified pixels in each category by the total number of pixels assigned to that category, and summing across all categories, thus summarizing the probability that the user of a classification will obtain a valid result. Producer’s accuracy compares the number of pixels correctly classified into each category to the original number of ground truth pixels used to characterize that category, thus summarizing the probability that the producer of a classification was able to train the classifier effectively for future applications. These metrics are calculated using a confusion matrix summarizing how each pixel was classified. These confusion matrices can also be used to calculate a summary statistic known as Cohen’s Kappa coefficient (Cohen 1960), which represents the degree of overall agreement between ground truth pixels and the classification being summarized. This metric is preferred to overall accuracy when comparing among studies or classification methods because it takes into account both user’s and producer’s accuracy (Congalton and Green 2009).

In addition to assessing the performance of a single classification method for any given dataset, different data mining methods used on the same dataset may be compared using several available indicators of performance. In this case, two such indicators were used: the area under the curve of the receiver operating characteristics graph (AUC-ROC) and the Brier Score. These methods of assessment rely on error metrics that are the related to the Type I and Type II error metrics discussed above; sensitivity (or the true positive rate) is the inverse of Type I error and specificity (or the true negative rate) is the inverse of Type II error. An ROC graph plots the rates of true positives against false positives, which is to say specificity versus the rate of Type I error. This comparison creates a curved hull, the area underneath which can be calculated and compared across different classifiers. Since the area under the curve represents the probability that a random case will be classified as true positive rather than a false positive, a larger AUC-ROC score indicates better performance (Fawcett 2005). Brier curves accomplish a similar goal to ROC graphs, except that the curve it displays represents a metric of the cost of an incorrect classification across different operating parameters. As in an ROC graph, the area under this curve can be calculated, and is referred to as the Brier score (Hernandez-Orallo 2011). These two metrics are automatically calculated by Orange when comparing different data mining methods.

Three methods of resampling are also built in to the Orange visual programming interface. Resampling is a method of validation for data mining methods and can be carried out in several ways. Cross-validation resampling is performed by splitting the dataset into groups or "folds," one of which is held out and compared to a classifier induced from the rest of the cases. This process is typically repeated several times. Leave-one-out resampling uses a similar technique, but holds out one case at a time instead of one group. Random sampling divides the dataset into two groups, for example by holding out 30% of cases as training data to be used for testing the remaining 70% of the cases. As with cross-validation resampling, this process is usually repeated several times, with a different random sample being held out in each repeat (Demsar *et al.* 2013). Each of these methods requires different computational time to accomplish, with the leave-one-out method being the most time-consuming.

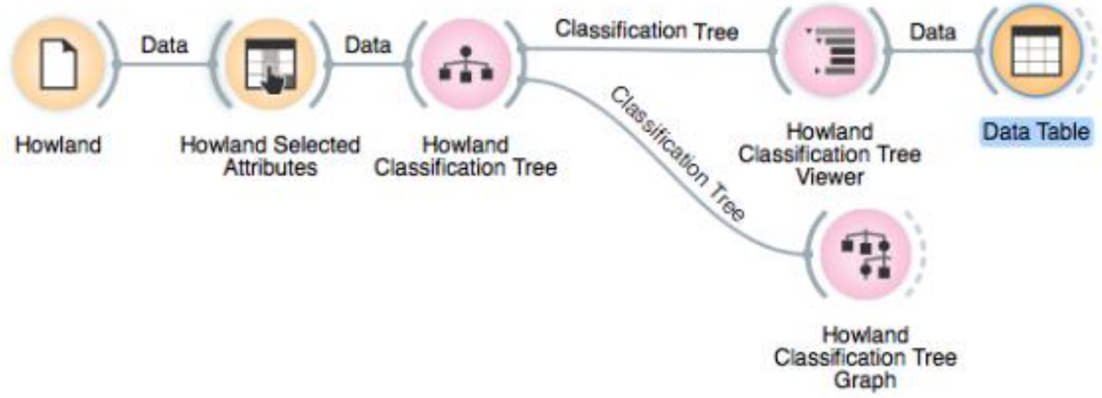
2f. Refining LiDAR Metrics and Hyperspectral Data for Further Work

All previously discussed data mining method assessments were carried out on the outputs of classifiers run on the full lists of 32 LiDAR metrics, 43 vegetative indices, and 114 reflectance bands. However, one main goal of this study was to refine this list into a subset with high explanatory power. This was made possible using the Classification Tree Viewer widget in

Orange. This widget shows a list of details of each node in the classification tree created on a dataset. Since these nodes are chosen to break a dataset into smaller categories, they were assumed to be explanatory of the variability in the dataset as a whole. This method has previously produced promising results on a full-wavelength LiDAR dataset being used for tree species classification (Heinzel and Koch 2011). A similar method for variable reduction using the results of random forest-generated trees produced good results in a study using LiDAR data for forest inventory (Vauhkonen *et al.* 2010).

To apply this method here, a classification tree was run on each dataset, and the resulting list of variables used as nodes in the first five levels of each tree was kept for further work. This was done for the LiDAR metrics at each forest site and individually for the vegetative indices and for the reflectance bands at the Howland site. For the Howland site, another simplified list was produced from a combined input of all LiDAR and hyperspectral data. A sample workflow of the classification tree step can be found in Figure 4.

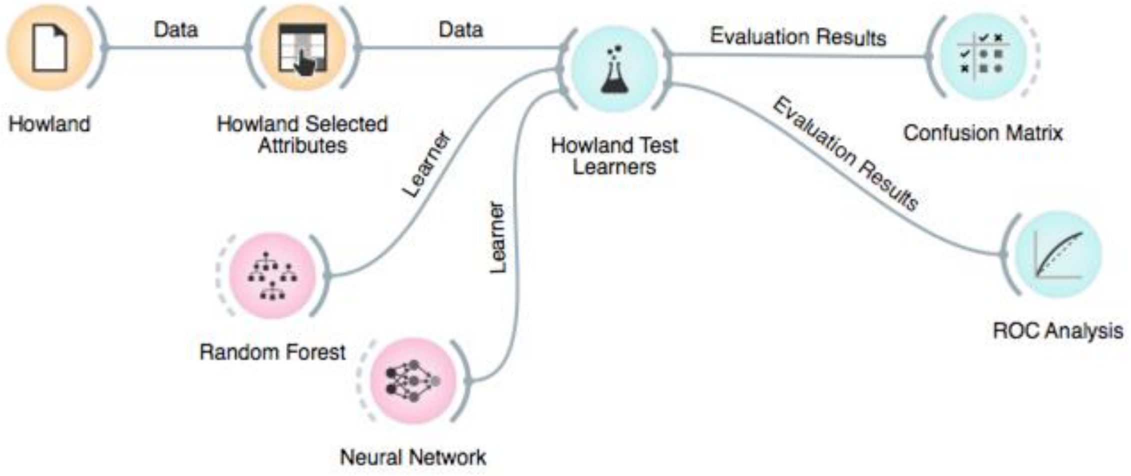
Figure 4: Sample Classification Tree Workflow


Data
Data
Classification Tree
Data
Howland
Howland Selected Attributes
Howland Classification Tree
Classification Tree
Howland Classification Tree Viewer
Data Table
Howland Classification Tree Graph

The results of the initial test of all data mining classifiers and resampling methods were used to determine the best settings for each dataset. These simplified and optimized settings and the refined list of metrics, indices, or bands were used to rerun the data mining procedure. A sample simplified workflow can be found in Figure 5. In the case of the LiDAR metrics, a shared, or generic, set of metrics repeated across the simplified lists from the two forest sites, was generated and used as input for another set of classifiers. Classifier performance was again assessed by comparing AUC-ROC, Brier scores and classification accuracy generated by Orange. In all cases, the resampling method that produced the best results in the initial comparison was used for this assessment.

Orange does not provide built-in functionality for calculating the Kappa coefficient, so the confusion matrices generated by the Confusion Matrix widget were used as inputs for a custom Python script designed to calculate Kappa (see Appendix 1) for further classifier comparison. This script was designed to take advantage of the specific functionality of NumPy arrays (see van der Walt e*t al.* 2013). This functionality allows for calculations to be performed simultaneously on the whole array or on a particular slice thereof, which is particularly useful for the conversion of data from confusion matrix to intermediate values used to calculate Cohen's Kappa.

Figure 5: Sample Simplified Orange Workflow

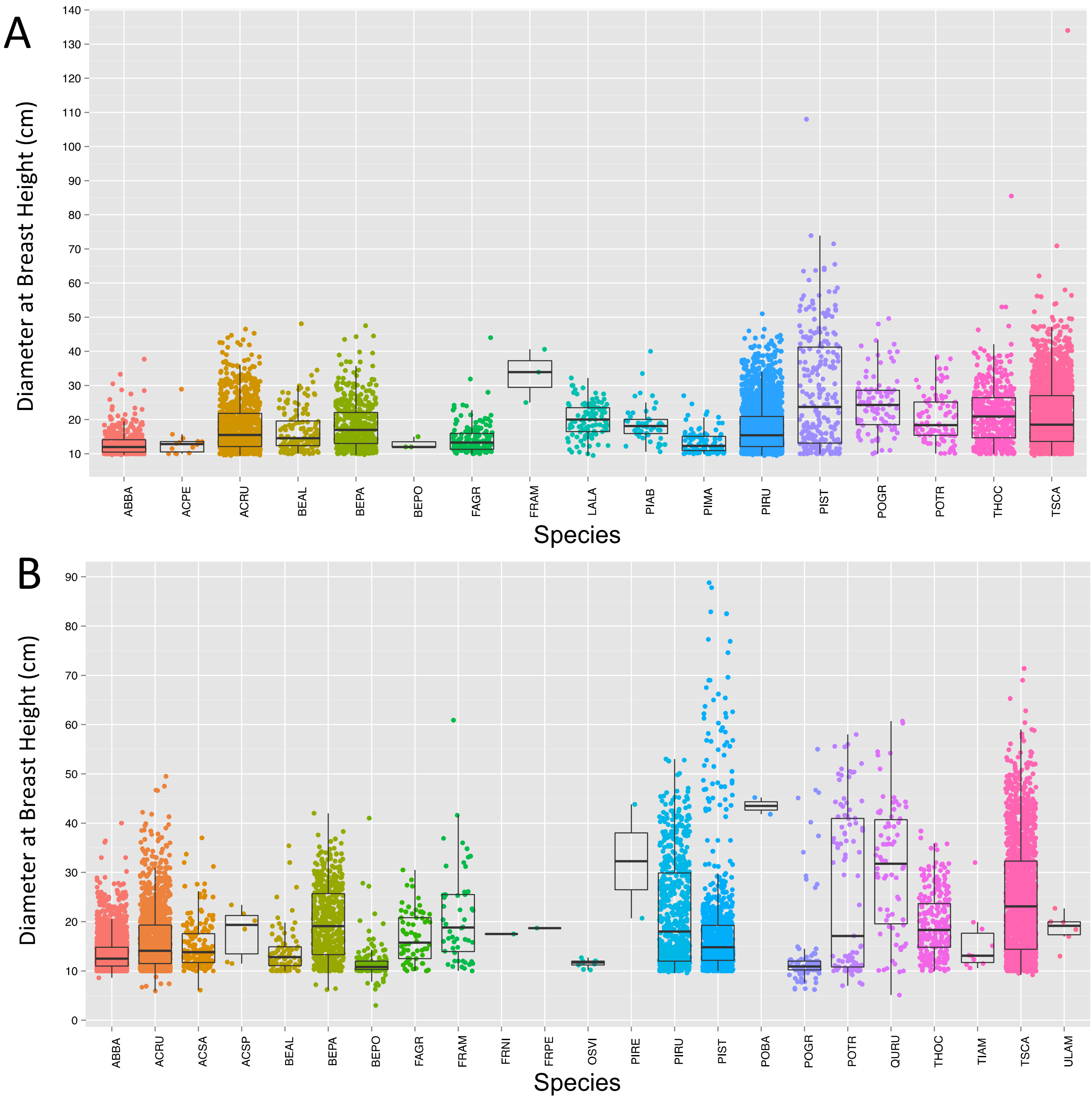


As a baseline for comparison, principal components analyses were also performed on the vegetative indices and reflectance hyperspectral files, using the Forward PC Rotation function in ENVI Classic. The resulting principal components with eigenvalues greater than one (nine total for the vegetative indices file and ten total for the reflectance file) were exported as geotiffs, processed in the same way as the original hyperspectral data, and used as inputs for data mining in Orange.

3. Results

In order to confirm a relationship between structural variables and tree species, the diameter at breast height (DBH) data collected in the field campaign were plotted for each species class. Though individual tree heights were also recorded in some cases, there were too many missing values in the Penobscot dataset for a robust analysis or comparison. The trends in DBH by species can be seen as scatterplots in Figure 6 or as boxplots in Figure 7. While some patterns can be detected between overall tree size and species classification, DBH values alone do not seem to provide sufficient data to classify tree species on their own. An unexpected frequency of 10 cm as the recorded value for DBH also suggests that this value may have been used as a default measurement for small trees, and may be skewing distributions toward lower values overall. Additionally, DBH measurements are necessarily associated with the age of individual trees as well as their species. Nonetheless, the weak trends that can be detected suggest that further examining structural information on trees in the form of the LiDAR metrics is a sound avenue of analysis.

Figure 6: Plots of Individual Tree Diameter at Breast Heights by Species
Figure shows all available data points on diameter at breast height of individual trees in Howland Experimental Forest (A) and Penobscot Experimental Forest (B).



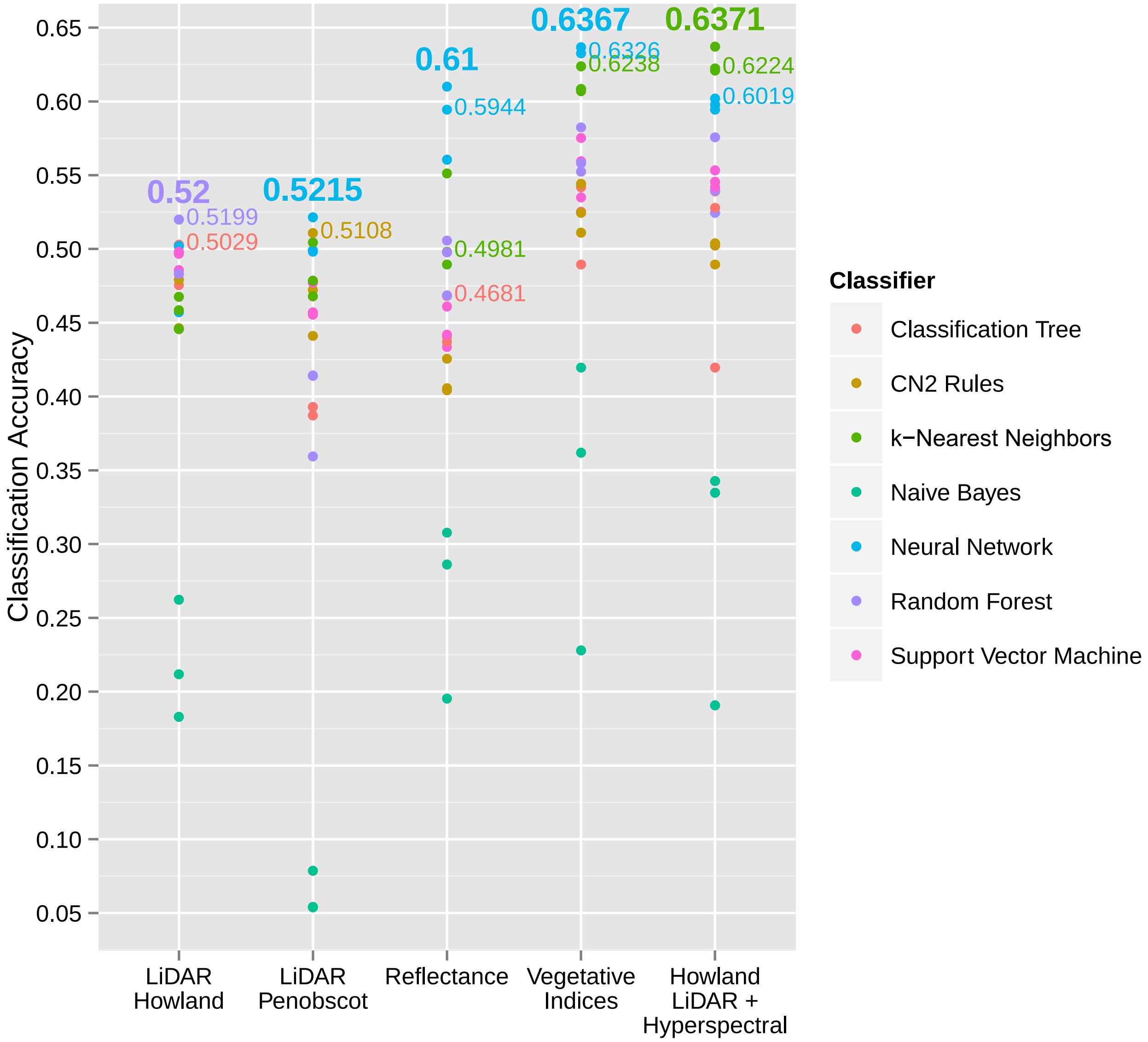
Initial comparisons of data mining methods and resampling techniques performed on all LiDAR metrics across Howland Experimental Forest and Penobscot Experimental Forest showed relatively consistent classification accuracy (CA) values across all combinations tested (Figure 7). While the data mining methods that produced the best results varied across the two forests, leave-one-out resampling consistently allowed for the highest CA values. Using the LiDAR metrics alone, the highest classification accuracies achieved were CA = 0.5200 for Howland Experimental Forest, using a random forest classifier, and CA = 0.5215 for Penobscot Experimental Forest, using a neural network classifier. Second-best methods to be included for comparison on subsequent analyses included the classification tree method for Howland and the CN2 Rules method for Penobscot.

Initial comparisons of data mining methods and resampling techniques performed on vegetative indices for Howland Experimental Forest produced higher classification accuracy values than the same protocol run on LiDAR metrics (Figure 7). In this case, cross-validation resampling produced the highest CA values. Using the vegetative indices alone, the highest classification accuracy was CA = 0.6367, achieved using a neural network classifier. The k-nearest neighbors method produced a comparable classification accuracy of 0.6238, and was also kept for inclusion in subsequent analyses.

Initial comparisons of data mining methods and resampling techniques performed on reflectance data for Howland Experimental Forest produced classification accuracy values slightly lower than those from the vegetative index comparison (Figure 7). Cross-validation resampling again produced the highest CA values in this comparison. Using only data on the hyperspectral reflectance bands, the highest classification accuracy was CA = 0.6100, again achieved using a neural network classifier. The k-nearest neighbors and random forest methods produced comparable classification accuracies of 0.4981 and 0.5057, respectively. All three of these methods were kept for inclusion in subsequent analyses.

Figure 7: Comparison of Resampling Techniques and Data Mining Methods Using Complete Lists of Metrics

Available LiDAR data were used in separate analyses of each forest. Hyperspectral data (reflectance bands and vegetative indices) were only available for Howland Experimental Forest, so those analyses, as well as an analysis of all available data together, were only conducted for that site. Repeated colors in the same column indicate the results of different resampling methods used in combination with each classifier.



Another initial comparison of data mining methods and resampling techniques was performed, this time on all available data for Howland Experimental Forest (Figure 7). Cross-validation resampling once again produced the highest CA values in this comparison, CA =

0.6371, achieved using a k-nearest neighbors classifier. The neural network method produced a comparable classification accuracy of 0.6019, and was also included in further analyses.

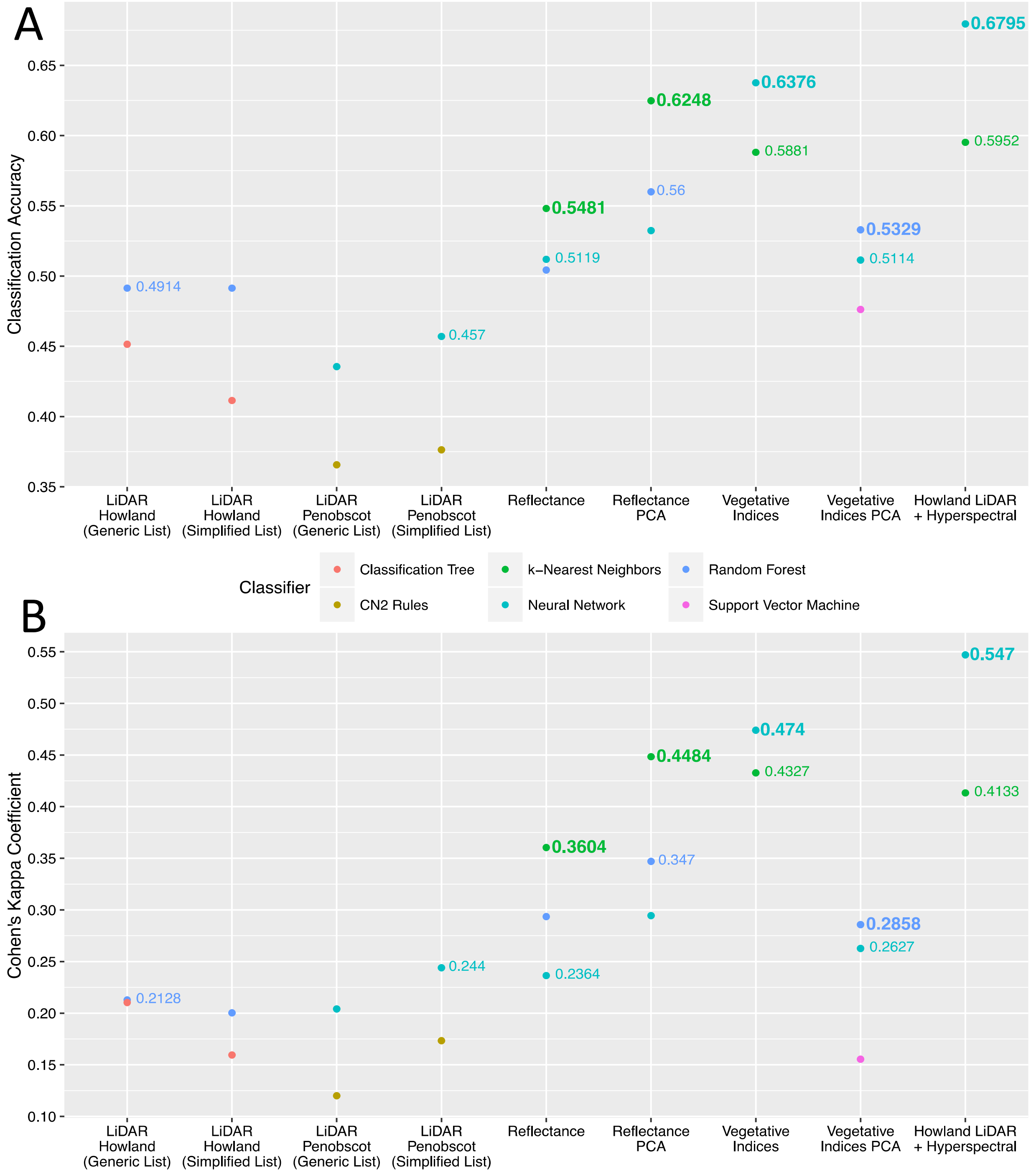
As described above, classification trees were also run on each dataset discussed above, regardless of the performance of this method during the initial comparison. These classification trees were not used to generate metrics of classifier performance, but were viewed in list format in order to identify which metrics, indices, or bands served as the nodes in the first five levels of the tree. The results of that analysis provided simplified lists of inputs for use in further analyses (Table 5). A list of thirteen LiDAR metrics from Penobscot Experimental Forest and a list of ten LiDAR metrics from Howland Experimental Forest were cross-referenced to produce a generic list of five LiDAR metrics shared between the classification trees produced on the two forests. This generic list was generated in an attempt to identify some universal or generalizable aspects of LiDAR data that may have strong explanatory power in other forests. The hyperspectral data available on the Howland site were also used to produce simplified lists of reflectance bands, vegetative indices, and a list generated from the full dataset of LiDAR and hyperspectral data together. All further results were generated using only the inputs shown in these simplified lists.

Table 5: Simplified Lists of Data Mining Inputs

	Penobscot	**Howland**			
Generic LiDAR	LiDAR Metrics	LiDAR Metrics	Reflectance Bands	Vegetative Indices	Hyperspectral + LiDAR
D9	D0	D3	B003	ARI1	B023
Fcover	D2	D5	B019	ARI2	B034
Fract_all	D9	D6	B026	CRI2	DATT2
P50	Fcover	D9	B037	DATT2	Fract_All
P100	Fract_all	Fcover	B038	GI	GM1
	Kurtosis	Fract_all	B054	GM2	GM2
	Mean	P50	B059	PRI	Mean
	P10	P60	B062	RDVI	MTCI
	P40	P90	B063	RGRI	P60
	P50	P100	B070	SIPI	P70
	P80		B072		PRI
	P100		B079		REIP
	St_Dev		B086		RGRI
			B102		SIPI
			B112		VOG
			B113		

The classifiers that produced the best results using the LiDAR metrics from both forests as inputs were run again, using only the simplified and generic lists discussed above. At this stage, final comparisons were made by adding in the Cohen’s Kappa coefficient as an additional means of assessing the performance of these optimized protocols. Figure 8 shows the results of classifications run on the simplified and generic LiDAR metrics datasets, using the two best data mining methods for each forest as determined by the initial comparison of classifiers.

Figure 8: Comparison of Resampling Techniques and Data Mining Methods Using Simplified and Generic Lists of Metrics



For the Howland site, the random forest classifier run with the simplified list showed the highest classification accuracy of 0.4914, with a Kappa coefficient of 0.2003. For the Penobscot site, the neural network classifier run with the simplified showed the highest classification accuracy of 0.4570, with a Kappa coefficient of 0.2440. All of these results represent a loss of accuracy as compared to the classifiers run using the full list of LiDAR metrics as inputs.

However, the variable reduction alone does not fully explain this decrease in accuracy, as shown in the result of the same classifiers run using the generic list of metrics shared across the simplified lists for the two forests. In this case at the Howland site, the random forest classifier run with the generic list showed the highest classification accuracy of CA = 0.4914 with a Kappa coefficient of 0.2128. Though none of these results surpasses the accuracy achieved by using the full list, all are better than with the simplified list. For the Penobscot site, the neural network classifier run with the generic list showed the highest classification accuracy of CA = 0.4355 with a Kappa coefficient of 0.2041. These results, unlike the Howland site, show another small decrease in accuracy (Figure 8).

A comparable method was used on the hyperspectral datasets, using the simplified list of metrics discussed above in combination with the two best classifiers on the vegetative index dataset and the three best classifiers for the reflectance dataset. These classifiers were run using the cross-validation resampling technique that yielded the best results in the initial assessment (Figure 8). For the vegetative indices dataset, the neural network classifier performed best, with an overall classification accuracy of 0.6376 and a Kappa score of 0.4740, a dramatic improvement in accuracy over the performance of any classifier run with the full list of metrics. There was also a small decrease in accuracy for the reflectance dataset as compared to the classifiers run on the full list of bands. The k-nearest neighbors classifier achieved a classification accuracy of 0.5481, with a Kappa score of 0.3604.

For comparison, the results of the best classifiers used on the outputs of the principal components analysis on the hyperspectral datasets are also shown in Figure 8. The PCA results show the inverse trend when compared to the results of the data mining performed with the simplified lists of hyperspectral metrics. On the vegetative indices PCA dataset, the neural network classifier produced the best result, with a classification accuracy of 0.5329 and a Kappa coefficient of 0.2858. This represents a decrease compared to both the results generated using the simplified list of indices and the original data mining results using the full list. For the reflectance dataset, the k-nearest neighbors classifier produced the best result, with a classification accuracy of 0.6248, with a Kappa coefficient of 0.4484. This represents a decrease in accuracy as compared to either previous analysis.

In a final assessment, the simplified list of inputs from the combined LiDAR and hyperspectral dataset on the Howland site were used in combination with the two best classifiers as identified by the initial assessment. These classifiers were again run using the cross-validation resampling technique that yielded the best results in the initial assessment. The better of the two methods tested in this analysis was the k-nearest neighbors classifier, which achieved a final classification accuracy of 0.6795 and a Kappa score of 0.5470, which represents an improvement over any other list of inputs or classifier discussed thus far.

4. Discussion

The results of the final assessment, using the combined LiDAR and hyperspectral dataset, outperform all of the previous assessments. Since all previous steps used LiDAR and hyperspectral data separately, these final results suggest that the combination of spectral and structural information is richer in detail than either dataset alone. This improvement is in line with other studies that have found a similar effect (*e.g.* Liu *et al.* 2013), and stands in contrast to several cases in which other authors have not found a significant improvement when incorporating LiDAR data into existing hyperspectral analyses. The fact that the incorporation of LiDAR data improved the hyperspectral-based classifications of trees Howland Experimental Forest speaks to the utility of data mining techniques in solving problems like this one. One notable element of the data mining procedure discussed here is the high performance of classifier types that are not typically favored in remote sensing work. In particular, support vector machines (SVM) and other methods that are best equipped to handle the very high dimensionality of hyperspectral data in particular are the established standard for of remote sensing work. However, when tested concurrently, the SVM method available through Orange was significantly outperformed. Some researchers have previously postulated that LiDAR datasets do not suffer as much from the issues of ill-posed problems and very high dimensionality and are therefore better suited to classification techniques that would not necessarily be optimal for other remote sensing work (Ducic *et al.* 2006), which may account for some of the differences between the methods described here and other previously published workflows.

Whatever the context-specific details of classifier choice, the capability of data mining interfaces like Orange to simplify and optimize classification workflows is clearly powerful. The variable reduction technique used here showed mixed results in this context; when comparing this technique to the principal components analysis, it appears that each technique may have its merits under different circumstances. The PCA produced the best result of any method on the reflectance dataset, but had the poorest results of any method on the vegetative indices dataset. The classification tree-based variable reduction produced the best result of any method on both the vegetative indices and LiDAR + hyperspectral datasets, but had the poorest results of any method on the reflectance dataset. Thus, it appears that variable reduction based on classification tree nodes is a technique worth trying when seeking to reduce dimensionality. it is adaptable to any dataset, and has the desirable effect of both reducing the dimensionality of a very large hyperspectral dataset into a more manageable form, and improving the outcome of classifications. Additionally, data mining software allows a comparison to PCA to be performed quickly, so that an optimal dimensionality reduction technique for the context can be easily determined. This dual benefit indicates that the technique discussed here may be useful for a variety of commercial forestry and inventory applications, even for organizations without the computing power or resources to use expensive and computation-intense programs like ENVI.

Nonetheless, there remain some limitations to the analysis as presented here, the most important of which is the necessity of using aggregated data. While this is not a constraint that will necessarily apply to all future studies, aggregation of data to a subplot level was required in this case because of the lack of data on the coordinates of individual trees within either forest. This means that some detail was necessarily lost, particularly from the field campaign dataset, which provided data on height and DBH at an individual tree level, and from the hyperspectral datasets. In most cases, 500 or more pixels were averaged together during the Zonal Statistics summarization process, meaning that a great deal of detail on differential reflectance from within individual tree crowns could not be used. This effect was much less pronounced on the LiDAR dataset, since the 13m^2 pixel size meant that individual trees could not be distinguished even before averaging to the subplot level.

This is a problem that has been confronted by numerous researchers in the past, since G-LiHT's is certainly not the only dataset to include data aggregated to different sizes or to rely on ground truth data with some limitations. Some authors have argued that attempts to identify or classify species at anything above the individual tree level will be met with difficulty (Yu *et al.* 2010), but other researchers have previously published classifications with up to 90% on tree stands (Korpela *et al.* 2010). While that level of accuracy is partially due to the fact that the latter analysis was run on a forest with low species diversity and very homogenous tree stands, their results viewed in combination with those presented here make a relatively convincing argument that the use of data at a larger scale than individual trees is, while not ideal, still quite serviceable. This effect is paralleled in the hyperspectral data, in which there was most likely a larger limitation. Because pixels were aggregated into an overall mean value for a subplot identified only by the dominant species, there was necessarily some error that hindered species classification both because of the loss of detail and because of the contaminating effect of non-dominant species' spectral signatures for which it was impossible to fully account in this classification. Nonetheless, a relatively high classification accuracy of over 67% demonstrates again that such datasets can still be used to generate reliable results, an encouraging result given that researchers have recently begun to acknowledge that most forest classification work will need to be done on forest stands for practical reasons (Hovi *et al.* 2016).

Another concern that has been discussed in the literature is the level of management that the forest in question has undergone. Though most studies (*e.g.* Maltamo *et al.* 2004) have found that classification accuracies produced with LiDAR data alone tend to be lower on unmanaged forest plots than on those that are managed, some researchers have achieved improved root mean square error values when using combined LiDAR and hyperspectral data to classify unmanaged forest plots (Anderson *et al.* 2008). In this case, aggregating data to the subplot level may have created a homogenizing effect that is comparable to the more easily classified regularity of a managed forest with stands of the same species intentionally growing together or a cleared understory reducing less useful backscatter in LiDAR datasets. It is also possible, however, that a combination of structural and spectral information like the one used here is able to reveal canopy gaps or other irregularities that would otherwise make the association of spectral data with species information more difficult, as previously suggested by others, including Brennan and Webster (2006).

The improved association between species identity and structural variables when moving from DBH alone to the G-LiHT-generated structural variables further supports this idea. It has been shown that using data on aboveground biomass (for which DBH is often used as a proxy) in conjunction structural information on forest structure generated by the Laser Vegetation Imaging Sensor (LVIS) improves the ability of models to predict the size of forest carbon stocks (Ni-Meister *et al*. 2010). It now seems that the combination of these two data types may be able to simultaneously help identify tree species, thereby opening up the possibility of generating species-specific carbon estimates with a similar combined dataset. Other researchers looking to the future of remote sensing have also highlighted the utility of LiDAR data in addressing large-scale questions like deforestation and carbon sequestration in whole forests on a species-specific basis (Koch 2010, Karna *et al.* 2015). Maltamo and Packalén (2014) recommended a similar species-specific approach to forest inventory and classification, which may help to reduce error by relying on very targeted ground truth measurements. When looking to the future of multi-sensoral and fused datasets, one of the commonly cited challenges is the development or discovery of analytical methods that can properly integrate data collected by different sensors or by different projects altogether. Based on the results of this analysis, it appears that data mining methods can be used to produce simplified datasets combining information from a variety of sensors and to optimize classifiers depending on context, with high resulting accuracy even across many heterogeneously distributed temperate tree species.

5. Appendix

I. Creating Centroids, Interpolating, and Exporting Attribute Tables to Text Files

```python
import arcpy
from arcpy import env
from arcpy.sa import *
import os
arcpy.CheckOutExtension("Spatial")
env.overwriteOutput = True

# Iterate over subplot shapefiles and create centroid shapefiles

env.workspace =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots"
subplots = arcpy.ListFeatureClasses()
for fc in subplots:
    try:
        outfc = arcpy.Describe(fc).basename + "_Centroids"
        arcpy.FeatureToPoint_management(fc, outfc, "CENTROID")
    except Exception as e:
        print e
    print "Centroids Created"

# Extract values around centroids to fields in centroids shapefile attribute
table - Howland

env.workspace =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Howland_Metrics"
metrics = arcpy.ListFiles(wild_card = "*.tif")
for fc in metrics:
    try:
        centroids =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Howland_Subplots_Centroids.shp"
        ExtractMultiValuesToPoints(centroids, fc, "BILINEAR")
    except Exception as e:
        print e
    print "New Column Added"

# Create text file from attribute table

try:
    input =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Howland_Subplots_Centroids.shp"
```

```
    fieldList = arcpy.ListFields(input)
    field_names = []
    for field in fieldList:
        field_names.append(field.name)
    fields_to_keep = field_names[15:17] + field_names[19:]
    rows = arcpy.SearchCursor(input)
    out_string = ""
    file_name =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Howland.txt"
    out_file = open(file_name, 'w')
    for r in rows:
        for f in fields_to_keep:
            val = r.getValue(f)
            out_string += "\t" + str(val)
        out_string += "\n"
        out_file.write(out_string)
    print(val)
    out_file.close()
except Exception as e:
    print e
print "Text File Created"

# Create CSV file from attribute table

try:
    input =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Howland_Subplots_Centroids.shp"
    fieldList = arcpy.ListFields(input)
    field_names = []
    for field in fieldList:
        field_names.append(field.name)
    fields_to_keep = field_names[15:17] + field_names[19:]
    out_file =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Howland.csv"
    arcpy.ExportXYv_stats(input, fields_to_keep, "COMMA", out_file,
"ADD_FIELD_NAMES")
except Exception as e:
    print e
print "CSV File Created"

# Extract values around centroids to new shapefile with dominant species info
included - Penobscot

env.workspace =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Penobscot_Metrics"
metrics = arcpy.ListFiles(wild_card = "*.tif")
for fc in metrics:
    try:
        centroids =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Penobscot_Subplots_Centroids.shp"
        ExtractMultiValuesToPoints(centroids, fc, "BILINEAR")
    except Exception as e:
        print e
    print "New Column Added"

# Create text file from attribute table

try:
    input =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Penobscot_Subplots_Centroids.shp"
    fieldList = arcpy.ListFields(input)
    field_names = []
    for field in fieldList:
        field_names.append(field.name)
    fields_to_keep = field_names[15:17] + field_names[19:]
    rows = arcpy.SearchCursor(input)
    out_string = ""
    file_name =
r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_Application
s/Final_Project/Subplots/Penobscot.txt"
    out_file = open(file_name, 'w')
    for r in rows:
        for f in fields_to_keep:
            val = r.getValue(f)
            out_string += "\t" + str(val)
        out_string += "\n"
        out_file.write(out_string)
    print(val)
    out_file.close()
except Exception as e:
    print e
print "Text File Created"
```

II. Extracting Hyperspectral Subplot Averages

```
# Perform Zonal Statistics As Table function on reflectance geotiffs

import arcpy
from arcpy import env
from arcpy.sa import *
arcpy.CheckOutExtension("Spatial")
env.overwriteOutput = True
env.workspace = r"D:\Documents\Reflectance"
veg_indices = arcpy.ListFiles(wild_card = "*.tif")
for fc in veg_indices:
    try:
        subplots = r"D:\Documents\Howland_Subplots.shp"
        outfc = arcpy.Describe(fc).basename + "_Zonal_Stats"
        outZStats = ZonalStatisticsAsTable(subplots, "SUBPLOT_ID", fc, outfc,
        "DATA", "MEAN")
    except Exception as e:
        print e
    print "Zonal Stats Calculated"

# Perform Zonal Statistics As Table function on vegetative index geotiffs

import arcpy
from arcpy import env
from arcpy.sa import *
arcpy.CheckOutExtension("Spatial")
env.overwriteOutput = True
env.workspace = r"D:\Documents\Veg_Indices"
veg_indices = arcpy.ListFiles(wild_card = "*.tif")
for fc in veg_indices:
    try:
        subplots = r"D:\Documents\Howland_Subplots.shp"
        outfc = arcpy.Describe(fc).basename + "_Zonal_Stats"
        outZStats = ZonalStatisticsAsTable(subplots, "SUBPLOT_ID", fc, outfc,
        "DATA", "MEAN")
    except Exception as e:
        print e
    print "Zonal Stats Calculated"
```

III. Calculating Kappa Coefficient – Example

```
import numpy as np

# Load csv files into numpy arrays

H_RF =
np.loadtxt(r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_
Applications/Final_Project/Howland_Confusion_Matrix_RF.csv",
    dtype = None, delimiter = ',')
H_NN =
np.loadtxt(r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_
Applications/Final_Project/Howland_Confusion_Matrix_NN.csv",
    dtype = None, delimiter = ',')
P_RF =
np.loadtxt(r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_
Applications/Final_Project/Penobscot_Confusion_Matrix_RF.csv",
    dtype = None, delimiter = ',')
P_NN =
np.loadtxt(r"/Users/juliamarrs/Documents/Computer_Programming_for_Geographic_
Applications/Final_Project/Penobscot_Confusion_Matrix_NN.csv",
    dtype = None, delimiter = ',')

# For each site below, the diagonal is extracted from the original array.
# It is then trimmed of the last (grand total) value and summed.
# The grand total number of pixels and its squared value are named and
calculated.
# The last row and column are extracted from the original array and trimmed
for the last (grand total) value.
# These trimmed arrays are multiplied together and summed
# The grand total, its square, and sum of the row and column totals are used
to calculate kappa.

howland = [H_RF, H_NN]
for h in howland:
    diag = np.diagonal(h)
    int_diag = diag[0:10]
    diag_sum = np.sum(int_diag)
    gr_total = diag[10]
    gr_total_sq = gr_total**2
    last_col = h[:,10]
    col_totals = last_col[0:10]
    last_row = h[10,:]
    row_totals = last_row[0:10]
    totals_mult = col_totals*row_totals
    totals_mult_sum = np.sum(totals_mult)
    kappa = ((gr_total*diag_sum)-totals_mult_sum) / (gr_total_sq -
     totals_mult_sum)
    print kappa

penobscot = [P_RF, P_NN]
for p in penobscot:
    diag = np.diagonal(p)
    int_diag = diag[0:15]
    diag_sum = np.sum(int_diag)
    gr_total = diag[15]
    gr_total_sq = gr_total**2
    last_col = p[:,15]
    col_totals = last_col[0:15]
    last_row = p[15,:]
    row_totals = last_row[0:15]
    totals_mult = col_totals*row_totals
    totals_mult_sum = np.sum(totals_mult)
    kappa = ((gr_total*diag_sum)-totals_mult_sum) / (gr_total_sq -
     totals_mult_sum)
    print kappa
```

Bibliography

Agapiou A., D. G. Hadjimistsis, and D. D. Alexakis. 2012. “Evaluation of Broadband and Narrowband Vegetation Indices for the Identification of Archaeological Crop Marks.” *Remote Sensing* 4:3892-3919. doi:10.3390/rs4123892.

Alonzo, M., B. Bookhagen, and D. A. Roberts. 2014. “Urban tree species mapping using hyperspectral and LiDAR data fusion.” *Remote Sensing of Environment* 148:70-83. doi: 10.1016/j.rse.2014.03.018.

Anderson, J. E., L. C. Plourde, M. E. Martin, B. H. Braswell, M.-L. Smith, R. O. Dubayah, M. A. Hofton, and J. B. Blair. 2008. “Integrating waveform LiDAR with hyperspectral imagery for inventory of a northern temperate forest.” *Remote Sensing of Environment* 112 (4):1856-1870. doi: 10.1016/j.rse.2007.09.009.

Andrew, M. E., and S. L. Ustin. 2009. “Habitat suitability modeling of an invasive plant with advanced remote sensing data.” *Diversity and Distributions* 15 (4):627-640. doi: 10.1111/j.1472-4642.2009.00568.x.

Blackburn, G. A. 2002. “Remote sensing of forest pigments using airborne imaging spectrometer and LiDAR imagery.” *Remote Sensing of Environment* 82 (2-3):311-321. doi: 10.1016/ s0034-4257(02)00049-4.

Blackburn, G. A. 2007. “Hyperspectral remote sensing of plant pigments.” *Journal of Experimental Botany* 58 (4):855-867. doi: 10.1093/jxb/erl123.

Borchert R. and N. A. Slade. 1981. “Bifurcation Ratios and the Adaptive Geometry of Trees.” *Botanical Gazette* 142 (3):394-401.

Brandtberg, T., T. A. 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. doi: 10.1016/s0034-4257(03)00008-7.

Brandtberg, 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. doi: 10.1016/j.isprsjprs.2006.10.006.

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

Broge N. H. and E. Leblanc. “Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density.” *Remote Sensing of Environment* 76:156-172.

Campbell, J. B., and R. H. Wynne. 2011. *Introduction to Remote Sensing*. 5th ed. New York, USA: The Guildford Press.

Canham, C. D. 1988. "Growth and Canopy Architecture of Shade-Tolerant Trees: Response to Canopy Gaps." *Ecology* 69(3):786-795.

Clark P. and T. Nibblet. 1989. "The CN2 Induction Algorithm." *Machine Learning* 3: 261-283.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1):37-46. doi: 10.1177/001316446002000104.

Congalton, R. G., and K. Green. 2009. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. 2nd ed. Boca Raton, USA: CRC Press, Taylor and Francis Group.

Cook, B. D., and L. A. Corp. 2012. "G-LIHT: Goddard's LiDAR, Hyperspectral, and Thermal Airborne Imager." National Aeronautics and Space Administration Goddard Space Flight Center.

Cook, B. D., L. A. Corp, R. F. Nelson, E. M. Middleton, D. C. Morton, J. T. McCorkel, J. G. Masek, K. J. Ranson, L. Vuong, and P. M. Montesano. 2013. "NASA Goddard's LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager." *Remote Sensing* 5 (8):4045-4066. doi: 10.3390/rs5084045.

Cutler, D. R., T. C. Edwards, Jr., K. H. Beard, A. Cutler, and K. T. Hess. 2007. "Random forests for classification in ecology." *Ecology* 88 (11):2783-2792. doi: 10.1890/07-0539.1.

Dalponte, M., L. Bruzzone, and D. Gianelle. 2012. "Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data." *Remote Sensing of Environment* 123:258-270. doi: 10.1016/j.rse.2012.03.013.

Dalponte, M., L. Bruzzone, L. Vescovo, and D. Gianelle. 2009. "The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas." *Remote Sensing of Environment* 113 (11):2345-2355. doi: 10.1016/j.rse.2009.06.013.

Dash J. and P. J. Curran. 2004. "The MERIS terrestrial chlorophyll index." International Journal of Remote Sensing 25(23):5403-5413. doi: 10.1080/0143116042000274015.

Datt, B. 1999. "Visible/near infrared reflectance and chlorophyll content in Eucalyptus leaves." *International Journal of Remote Sensing* 20(14): 2741-2759. doi:10.1080/014311699211778.

Demsar, J., T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik, and B. Zupan. 2013. “Orange: Data Mining Toolbox in Python.” *Journal of Machine Learning Research* 14:2349-2353.

Donoghue N. M. D., P. J. Watt, N. J. Cox, 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. doi:10.1016/j.rse.2007.02.032.

Dudani S. A. 1976. “The Distance-Weighted k-Nearest Neighbor Rule.” *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6(4): 325-327.

ESRI. 2011. "Zonal Statistics as Table (Spatial Analyst)." ArcGIS Resource Center Desktop 10. Accessed April 18, 2016. <http://help.arcgis.com/EN/ArcGISDesktop/10.0/Help/index.html#//009z000000w8000000.htm>.

ESRI. *n.d.* “Extract Multi Values to Points.” ArcGIS for Desktop. Accessed 18 March, 2016. < http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/extract-multi-values-to-points.htm >.

Evans, J. S., A. T. Hudak, R. Faux, and A.M. S. Smith. 2009. “Discrete Return LiDAR in Natural Resources: Recommendations for Project Planning, Data Processing, and Deliverables.” *Remote Sensing* 1 (4):776-794. doi: 10.3390/rs1040776.

Fawcett, T. 2006. “An introduction to ROC analysis.” *Pattern Recognition Letters* 27:861-874.

Filella I. and J. Peñuelas. 1994. “The red edge position and shape as indicators of plant chlorophyll content, biomass, and hydric status.” *International Journal of Remote Sensing* 15 (7):1459-1470. doi:10.1080/01431169408954177.

Foresman, T. W. 1998. *The History of Geographic Information Systems: Perspectives from the Pioneers*: Prentice Hall PTR.

Franklin, J. 2009. *Mapping Species Distributions: Spatial Inference and Prediction*. Edited by M. Usher, D. Saunders, R. Peet and A. Dobson, *Ecology, Biodiversity, and Conservation*. Cambridge, UK: Cambridge University Press.

Gamon J. A., J. Peñuelas, and C. B. Field. 1992. “A Narrow-Waveband Spectral Index That Tracks Diurnal Changes in Photosynthetic Efficiency.” *Remote Sensing of Environment* 41:35-44.

Gamon J. A. and J. S. Surfus. 1999. “Assessing leaf pigment content and activity with a reflectometer.” *New Phytologist* 143:105-177.

Ghosh, A., F. E. Fassnacht, P. K. Joshi, and B. Koch. 2014. “A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales.” *International Journal of Applied Earth Observation and Geoinformation* 26:49-63. doi: 10.1016/j.jag.2013.05.017.

Gillespie, T. W., J. Brock, and C. W. Wright. 2004. “Prospects for quantifying structure, floristic composition and species richness of tropical forests.” *International Journal of Remote Sensing* 25 (4):707-715. doi: 10.1080/01431160310001598917.

Gitelson A. A. and M N. Merzlyak. 1997. “Remote estimation of chlorophyll content in higher plant leaves.” *International Journal of Remote Sensing* 18(12):2691-2697.

Gitelson A. A., M N. Merzlyak, and O. B. Chivkunova. 2001. “Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves.” *Photochemistry and Photobiology* 74(1):38-45.

Gitelson A. A., Y. Zur, O. B. Chivkunova, and M N. Merzlyak. 2002. “Assessing Carotenoid Content in Plant Leaves with Reflectance Spectroscopy.” *Photochemistry and Photobiology* 75(3):272-281.

Haykin S. 2004. *Neural Networks: A Comprehensive Foundation*. 2nd ed. Upper Saddle River, USA: Prentice Hall.

Heinzel, J., and B. Koch. 2011. “Exploring full-waveform LiDAR parameters for tree species classification.” *International Journal of Applied Earth Observation and Geoinformation* 13 (1):152-160. doi: 10.1016/j.jag.2010.09.010.

Hernandez-Orallo, J., P.A. Flach, and C. Ferri. 2011. “Brier Curves: A New Cost-Based Visualisation of Classifier Performance.” *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*:1-8.

Hill, R. A., and A. G. Thomson. 2005. “Mapping woodland species composition and structure using airborne spectral and LiDAR data.” *International Journal of Remote Sensing* 26 (17):3763-3779. doi: 10.1080/01431160500114706.

Holmgren, J., and A. Persson. 2004. “Identifying species of individual trees using airborne laser scanner.” *Remote Sensing of Environment* 90 (4):415-423. doi: 10.1016/s0034-4257(03)00140-8.

Holmgren, J., A. Persson, and U. Soderman. 2008. “Species identification of individual trees by combining high resolution LiDAR data with multi-spectral images.” *International Journal of Remote Sensing* 29 (5):1537-1552. doi: 10.1080/01431160701736471.

Hovi, A., L. Korhonen, J. Vauhkonen, and I. Korpela. 2016. “LiDAR waveform features for tree species classification and their sensitivity to tree- and acquisition related parameters.” *Remote Sensing of Environment* 173:224-237.

Ishii H., F. E. David, D. G. Sprugel. “Comparative Crown Form and Branching Pattern of Four Coexisting Tree Species in an Old-growth *Pseudotsuga-Tsuga* Forest.” *Eurasian Journal of Forest Research* 6(2):99-109.

Jenkins, J., D. Chojnacky, L. Heath, and R. Birdsey. 2003. “National-scale biomass estimators for United States tree species.” *Forest Science* 49 (1):12-35.

Jones, T. G., N. C. Coops, and T. Sharma. 2010. “Assessing the utility of airborne hyperspectral and LiDAR data for species distribution mapping in the coastal Pacific Northwest, Canada.” *Remote Sensing of Environment* 114 (12):2841-2852. doi: 10.1016/ j.rse.2010.07.002.

Karna, Y. K., Y. A. Hussin, H. Gilani, M. C. Bronsveld, M. S. R. Murthy, F. M. Qamer, B. S. Karky, T. Bhattarai, A. Xu, and C. B. Baniya. 2015. “Integration of WorldView-2 and airborne LiDAR data for tree species level carbon stock mapping in Kayar Khola watershed, Nepal.” *International Journal of Applied Earth Observation and Geoinformation* 38:280-291. doi: 10.1016/j.jag.2015.01.011.

Kim, S., T. Hinckley, and D. Briggs. 2011. “Classifying individual tree genera using stepwise cluster analysis based on height and intensity metrics derived from airborne laser scanner data.” *Remote Sensing of Environment* 115 (12):3329-3342. doi: 10.1016/ j.rse.2011.07.016.

Koch, B. 2010. “Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment.” *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (6): 581-590. doi:10.1016/j.isprsjprs.2010.09.001

Koetz, B., F. Morsdorf, S. van der Linden, T. Curt, and B. Allgoewer. 2008. “Multi-source land cover classification for forest fire management based on imaging spectrometry and LiDAR data.” *Forest Ecology and Management* 256 (3):263-271. doi: 10.1016/j.foreco.2008.04.025.

Korhonen, L., I. Korpela, J. Heiskanen, and M. Maltamoa. 2011. “Airborne discrete-return LiDAR data in the estimation of vertical canopy cover, angular canopy closure and leaf area index.” *Remote Sensing of Environment* 115 (4):1065–1080. doi: 10.1016/ j.rse.2010.12.011.

Korpela, I., T. Tokola, H. O. Ørka, and M. Koskinen. 2009. “Small-Footprint Discrete-Return LiDAR in Tree Species Recognition.” *Proceedings of the ISPRS* 2-5.

Korpela, I., H. O. Ørka, M. Maltamo, T, Tokola, and J. Hyyppä. 2010. “Tree Species Classification Using Airborne LiDAR - Effects of Stand and Tree Parameters, Downsizing of Training Set, Intensity Normalization, and Sensor Type.” *Silva Fennica* 44 (2):319-339. doi: 10.14214/sf.156.

Lefsky, M. A., W. B. Cohen, and T. A. Spies. 2001. “An evaluation of alternate remote sensing products for forest inventory, monitoring, and mapping of Douglas-fir forests in western Oregon.” *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestière* 31 (1):78-87. doi: 10.1139/cjfr-31-1-78.

Li, J., B. Hu, and T. L. Noland. 2013. “Classification of tree species based on structural features derived from high density LiDAR data.” *Agricultural and Forest Meteorology* 171:104-114. doi: 10.1016/j.agrformet.2012.11.012.

Lim, K., P. Treitz, M. Wulder, B. St-Onge, and M. Flood. 2003. “LiDAR remote sensing of forest structure.” *Progress in Physical Geography* 27 (1):88-106. doi: 10.1191/0309133303pp360ra.

Liu L., Y. Pang, W. Fan, Z. Li, D. Zhang, and M. Li. 2013. “Fused airborne LiDAR and hyperspectral data for tree species identification in a natural temperate forest.” *Journal of Remote Sensing* 1007-4619: 679-695.

Magnussen, S., P. Eggermont, and V. N. LaRiccia. 1999. “Recovering tree heights from airborne laser scanner data.” *Forest Science* 45 (3):407-422.

Maltamo, M., K. Mustonen, J. Hyyppä, J. Pitkanen, and X. Yu. 2004. “The accuracy of estimating individual tree variables with airborne laser scanning in a boreal nature reserve.” *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestière* 34 (9):1791-1801. doi: 10.1139/x04-055.

Maltamo, M., and Packalen, P. 2014. “Species-specific management inventory in Finland.” In *Forestry applications of airborne laser scanning. Concepts and case studies*. Edited by M. Maltamo, E. Næsset, J. Vauhkonen. *Managing forest ecosystems 27*. Dordrecht, Netherlands: Springer Science + Business Media.

Montesano, P. M., B. D. Cook, G. Sun, M. Simard, R. F. Nelson, K. J. Ranson, Z. Zhang, and S. Luthcke. 2013. “Achieving accuracy requirements for forest biomass mapping: A spaceborne data fusion method for estimating forest biomass and LiDAR sampling error.” *Remote Sensing of Environment* 130:153-170.

Morsdorf, F., A. Marell, B. Koetz, N. Cassagne, F. Pimont, E. Rigolot, and B. Allgoewer. 2010. “Discrimination of vegetation strata in a multi-layered Mediterranean forest ecosystem using height and intensity information derived from airborne laser scanning.” *Remote Sensing of Environment* 114 (7):1403-1415. doi: 10.1016/j.rse.2010.01.023.

Mountrakis G., J. Im, and C. Ogole. 2011. “Support vector machines in remote sensing: A review.” *ISPRS Journal of Photogrammetry and Remote Sensing* 66(3):247-259.

Mundt, J. T., D. R. Streutker, and N. F. Glenn. 2006. “Mapping sagebrush distribution using fusion of hyperspectral and LiDAR classifications.” *Photogrammetric Engineering and Remote Sensing* 72 (1):47-54.

Næsset, E. 2007. “Airborne laser scanning as a method in operational forest inventory: Status of accuracy assessments accomplished in Scandinavia.” *Scandinavian Journal of Forest Research* 22 (5):433-442. doi: 10.1080/02827580701672147.

Ni-Meister, W., D. L. B. Jupp, and R. Dubayah. 2001. “Modeling LiDAR waveforms in heterogeneous and discrete canopies.” *IEEE Transactions on Geoscience and Remote Sensing* 39 (9):1943-1958. doi: 10.1109/36.951085.

Ni-Meister, W., S. Lee, A. H. Strahler, C. E. Woodcock, C. Schaaf, T. Yao, K. J. Ranson, G. Sun, and J. B. Blair. 2010. “Assessing general relationships between aboveground biomass and vegetation structure parameters for improved carbon estimate from LiDAR remote sensing.” *Journal of Geophysical Research* 115 (G00E11):1-12. doi:10.1029/ 2009JG000936.

Olden, J. D., J. J. Lawler, and N. L. Poff. 2008. “Machine Learning Methods Without Tears: A Primer for Ecologists.” *The Quarterly Review of Biology* 83 (2):171-193.

Peñuelas, J., F. Baret, and I. Filella. 1995. “Semi-empirical indices to assess carotenoids / chlorophyll a ratio from leaf spectral reflectance.” *Photosynthetica* 31 (2):221-230.

Plourde, L. C., S. V. Ollinger, M.-L. Smith, and M. E. Martin. 2007. “Estimating species abundance in a northern temperate forest using spectral mixture analysis.” *Photogrammetric Engineering and Remote Sensing* 73 (7):829-840.

Polikar R. 2006. “Ensemble based systems in decision making.” *IEEE Circuits and Systems Magazine* 6(3):21-45.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. “Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction.” *Ecosystems* 9: 181-199.

Pretzsch H., E. Dauber, and P. Biber. 2013. “Species-Specific and Ontogeny-Related Stem Allometry of Forest Trees: Evidence from Extensive Stem Analyses.” *Forest Science* 59 (3): 290-302.

Python Software Foundation. Python Language Reference, version 2.7.

Rennie J. D. M., L. Shih, J. Teevan, and D. R. Karger. 2003. “Tackling the Poor Assumptions of Naive Bayes Text Classifiers.” *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*:1-8.

Roujean J.-L. and F.-M. Breon. 1995. “Estimating PAR absorbed by vegetation from bidirectional reflectance measurements.” *Remote Sensing of Environment* 51 (3):375-384. doi:10.1016/0034-4257(94)00114-3.

Schardt, M., M. Ziegler, A. Wimmer, R. Wack, and J. Hyyppä. 2002. “Assessment of Forest Parameters by Means of Laser Scanning.” *The International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Science* 36: 272−276.

Schreier, H., J. Lougheed, C. Tucker, and D. Leckie. 1985. "Automated Measurements of Terrain Reflection and Height Variations Using an Airborne Infrared-Laser System." *International Journal of Remote Sensing* 6 (1):101-113.

van Aardt, J. A. N., R. H. Wynne, and J. A. Scrivani. 2008. "LiDAR-based Mapping of Forest Volume and Biomass by Taxonomic Group Using Structurally Homogenous Segments." *Photogrammetric Engineering & Remote Sensing* 74 (8):1033-1044.

van der Walt, S., S. C. Colbert, and G. Varoquaux. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation*." Computing in Science & Engineering* 13:22-30.

van Ewijk, K. Y., C. F. Randin, P. M. Treitz, and N. A. Scott. 2014. "Predicting fine-scale tree species abundance patterns using biotic variables derived from LiDAR and high spatial resolution imagery." *Remote Sensing of Environment* 150:120-131. doi: 10.1016/ j.rse.2014.04.026.

Vapnik V. N. 1982. *Estimation of dependencies based on empirical data.* Translated by S. Kotz, *Springer series in statistics.* New York, USA: Springer-Verlag.

Vauhkonen, J., I. Korpela, M. Maltamo, and T. Tokola. 2010. "Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics." *Remote Sensing of Environment* 114 (6):1263-1276. doi: 10.1016/ j.rse.2010.01.016.

Verrelst, J., M. E. Schaepman, B. Koetz, and M. Kneubühler. 2008. "Angular sensitivity analysis of vegetation indices derived from CHRIS/PROBA data." *Remote Sensing of Environment* 112(5): 2341–2353.

Vogelmann J. E., B. N. Rock, and D. M. Moss. 1993. "Red edge spectral measurements from sugar maple leaves." International Journal of Remote Sensing 14 (8):1563-1575. doi: 10.1080/01431169308953986.

Whitney G. G. 1976. "The Bifurcation Ratio as an Indicator of Adaptive Strategy in Woody Plant Species." *Bulletin of the Torrey Botanical Club* 103 (2)"67-72.

Wulder, M. A., C. W. Bater, N. C. Coops, T. Hilker, and J. C. White. 2008. "The role of LiDAR in sustainable forest management." *Forestry Chronicle* 84 (6):807-826.

Yao, W., P. Krzystek, and M. Heurich. 2012. "Tree species classification and estimation of stem volume and DBH based on single tree extraction by exploiting airborne full-waveform LiDAR data." *Remote Sensing of Environment* 123:368-380. doi: 10.1016/ j.rse.2012.03.027.

Yu, X., J. Hyyppä, M. Holopainen, and M. Vastaranta. 2010. "Comparison of Area-Based and Individual Tree-Based Methods for Predicting Plot-Level Forest Attributes." *Remote Sensing* 2 (6):1481-1495. doi: 10.3390/rs2061481.

Yu, X., J. Hyyppä, M. Vastaranta, M. Holopainen, and R. Viitala. 2011. “Predicting individual tree attributes from airborne laser point clouds based on the random forests technique.” *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (1):28-37. doi: 10.1016/ j.isprsjprs.2010.08.003.

P. J. Zarco-Tejada, A. Berjón, R. López-Lozano, J. R. Miller, P. Martın, V. Cachorro, M. R. González, and A. de Frutos. 2005. “Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinued canopy.” *Remote Sensing of Environment* 99:271-287. doi:10.1016/j.rse.2005.09.002.

Zhang, H. 2004. “The optimality of naïve Bayes.” *Proceedings of the 17th International FLAIRS Conference:*1-6.

Ørka, H. O., E. Næsset, and O. M. Bollandsås. 2007. “Utilizing Airborne Laser Intensity for Tree Species Classification.” *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 36 (Part 3/W52): 300–304.