A robust, generalizable model for canopy detection from generic aerial LiDAR in the contiguous United States

S. Jones[1a*], R. El Kadiri[1b] and H.G. Momm[1b]

[1]Department of Geosciences, Middle Tennessee State University, Murfreesboro, Tennessee

*corresponding author: rj3h@mtmail.mtsu.edu

Abstract:

LiDAR point clouds have long been used to perform land cover classification, and in particular has been extremely helpful for developing canopy detection models, which differentiate tree canopies from other types of landcover. However, canopy detection from LiDAR requires investigators to either provide data to train the model, use vendor classification codes vary in quality and availability, or rely on rough heuristics to estimate canopy cover. By analyzing land cover signatures in 10 watersheds across the contiguous United States that are physiographically and ecologically distinct and have LiDAR coverage that has been collected during various seasons and is of variable quality, we have developed a pretrained canopy detection model that can identify canopy cover with high accuracy (98.5%, Cohen's $\kappa = 0.958$) in watersheds that the model has not yet encountered and with LiDAR of variable quality, allowing general canopy detection with no need for parameter tweaking or other user input. This work also establishes a framework for the design of generalized models that perform more granular classification using LiDAR data.

Keywords:

Canopy; general; LiDAR; machine learning; point cloud; land cover; classification

One Sentence Summary:

Tree canopy signatures from LiDAR are consistent across a diverse assortment of watersheds and LiDAR missions, and this has been exploited to create a generalized, pretrained canopy detection model.

[a] Responsible for design, implementation and testing of model, collection and processing of data, and drafting of manuscript.

[b] Responsible for supervision of model design, manuscript drafting and revision.

1 Introduction

Land cover classification is one of oldest and most widely used applications of geospatial technologies. Traditional models have utilized aerial or satellite imagery, particularly hyperspectral imagery, in conjunction with machine learning algorithms, to make assess and classify land cover, but advancements in and proliferation of aerially-based light detection and ranging (LiDAR) technology over the last two decades has allowed the use of LiDAR for land cover classification, either as an enhancement to aerial imagery or on its own. Raw LiDAR data consists of a set of points where each point is associated with geographic data, including elevation, and auxiliary data, often including return intensity and return number. These data can be analyzed in its “point cloud” form but are more often rasterized first so they can be analyzed using more conventional raster-based algorithms. Rasterization of the geographic data encoded in a point cloud is used to produce elevation models of various aspects of the earth’s surface, depending on the parameters used to produce the raster.

LiDAR-based classification models have achieved remarkable classification quality, even with very granular classification schemes that attempt to recognize many feature classes. The fact that LiDAR captures physical structure of sensed objects has extended the limits of classification studies by enabling tree crown delineation or using tree structure to identify tree species, and fusion of LiDAR and hyperspectral imagery represent the state-of-the-art for land cover classification by allowing the simultaneous analysis of the spectral and elevation data using machine learning algorithms.

Though custom-trained models are extremely accurate, they are limited by two things. First, machine learning algorithms require input of training data, and such training datasets can be time-consuming to produce and are prone to sampling bias or training error, particularly when they are created by inexperienced investigators. Second, the trained model is only applicable to the data on which it was trained. A model trained, for example, in an urban area, will likely not produce satisfactory results in an agricultural or forested area or even in a different urban area. It may not even provide satisfactory results if a different dataset with the same geographic extent is used because of differences in collection methodologies such as sensor strength.

Thus, these site- and dataset specific models are powerful, but require expert knowledge to train and create, and a new model must be created for each new study area or dataset. On the other hand, a so-called “general” model could be applied to any dataset in any study area by exploiting land cover signals that are consistent regardless of study area and data collection methodology, and so can be used without the creation of training data. Necessarily, non-normalized data such as LiDAR return intensity must be excluded from a general model because such data is a function of sensor strength and sensitivity rather than purely of land cover, and so there will be little correlation between land cover type and intensity across different missions even though a correlation might exist within a single mission. Excluding this

data allows generalizability, but also discards the information contained within it and thus limits potential model accuracy and granularity.

Canopy detection models, which seek only to differentiate tree canopy cover from all other land cover types, are among the most simple classification models, but they are also widely used to track forest loss/gain, quantify reforestation efforts, and assess riparian health. Canopy models can be created by training a machine learning algorithm in the same manner as one would for more complex classification schemes, but canopy models can be approximated by using simple height filters. Some geospatial tools also allow generation of canopy models by using vendor-supplied classification codes, though these are not always available and are typically created using proprietary black-box methods when they are. Both methods are simple but rely on heuristics that might not be appropriate for all study areas or data that might not even be provided.

In light of this, we have a created a true generalized canopy detection model that relies on land cover signals from LiDAR data rather than heuristic height approximation and achieves classification quality that is comparable with site-specific models that are trained on data specific to that study area. We believe it is applicable in any watershed in the contiguous United States, and that it can be applied to nearly all commercially available aerial LiDAR regardless of point spacing or sensor characteristics. Furthermore, the creation of this model demonstrates that canopy cover has a distinct and consistent signal in LiDAR data across the contiguous United States, and as such represents the first step toward creating more granular generalized models: if other land cover types have distinct and consistent signals, they can be incorporated into a generalized model as well.

2 Methods

2.1 Model Design

LiDAR data in LAS format covering 10 watersheds across the contiguous United States were downloaded. The individual LAS files were then mosaiced together on a per-watershed basis using LAStools, and a suite of tools, including WhiteboxTools, laspy, GDAL, Orfeo, scipy and custom Python tools were used to process the LAS files into various raster products. These derived raster products were manually inspected along with aerial imagery to create training and validation data for each of the 10 watersheds (Table 4). 80% of the manually classified pixels were used as training data for a decision tree machine learning algorithm. A decision tree scheme was chosen over other machine learning algorithms because its output is easily interpretable, does not require normalized input and can be translated into a set of if-else statements, allowing the model to be distributed without a need for specific software to run the model.

The resultant decision tree was then validated against the remaining 20% of the data from the 7 trained watersheds, as well as 100% of the data from the 3 naïve watersheds. Input data was inversely weighted to amount of data for each watershed; that is, all watersheds were weighted equally regardless of differences in the number of contributing pixels. Raw accuracy scores as well as Cohen's κ statistics were calculated individually for each watershed in this study. Additionally, classification models specific for each watershed but with identical parameterization were also created for comparison against the general model. The general and specific models used identical input raster types with the exception of intensity rasters, which were included in specific models and not the general model.

2.2 Derived Raster Products

A summary of the raster products generated from the collected LiDAR data is given in Table 5. The details of each product are given in subsequent sections. Because the classification model described in this paper is intended to be general, no data products generated made use of vendor-supplied classification codes. For rasters created by interpolating LiDAR point cloud values, a triangular irregular network (TIN) algorithm was used. All rasters created used a pixel size of 1m x 1m, regardless of source LiDAR point density.

2.2.1 Digital surface model

A digital surface model (DSM) is created by interpolating the height values from only the first returns in a LiDAR point cloud. Thus, the DSM represents the elevation of the earth's surface including above-ground structures such as buildings and tree canopies.

Though the DSM was used to generate other data products included in the classification model, the DSM itself was excluded as input to the classification model. This was done because non-normalized elevations have little to no predictive power for landcover type.

2.2.2 Digital elevation model

A digital elevation model (DEM) is similar to a DSM but is created by interpolating the height values from only the last (rather than first) returns in a LiDAR point cloud. Thus, the DEM represents the elevation of the earth's surface excluding some above-ground structures tree canopies. Some definitions of the DEM also require removal of built structures, but because this study did not utilize vendor-supplied classification codes, buildings could not be removed from the DEM directly, and so our generated DEM products do include built structures.

The DEM was excluded as direct input to the classification model for the same reasons that the DSM was excluded.

2.2.3 Digital height model

A digital height model (DHM), sometimes referred to as a normalized digital surface model (nDSM) or height-above-ground model, is the result of the raster subtraction of the DEM from the DSM. The DHM represents the distance between the ground and any overhead structures, such as tree canopies or powerlines. Because building edges often produce multiple returns, building edges frequently appear in DHMs. The DHM is often used as a rough proxy for canopy height because its signature is dominated by tree, while building edges, powerlines, and other anthropogenic structures are minor contributors.

The DHM was excluded as direct input to the classification model but a modified version of it, described in the next section, was included.

2.2.4 Filtered digital height model

Though the DHM roughly represents canopy height, the anthropogenic structures such as building edges and powerlines contribute significantly to the DHM in urban areas. Because of this, investigators have proposed methods to distinguish the signature of trees from built structures through various means, such as filtering the linear signatures from LiDAR data that tend to be the result of powerline or building edge return splitting (Goodwin, et al. 2009). Though many other methods exist to filter built structures from a DHM, we elected to use a simple density filter due to its ease of implementation, quick runtime and high selectivity for canopy signals. The filter works such that, for every target pixel in the DHM, a 3 by 3 window is extracted and if f over 30% of the pixels (3 or more) in the window are nonzero, then the target pixel's value is retained. Otherwise, the target pixel is set to 0. This filter tends to remove linear features, such as building edges and powerlines, while retaining more rounded features such as tree canopies. The result is a filtered DHM (fDHM).

Though the unfiltered DHM was excluded as a direct input to the classification model to limit false classification of building edges and powerlines as trees, the filtered DHM was included as direct input.

2.2.5 Intensity raster

A raster representing the average return intensity at each pixel was created by interpolating LiDAR return intensity values. Because intensity values are not normalized and are a function of sensor type, pulse strength, and aircraft altitude in addition to the land cover type being sensed, the intensity rasters were not used as general model input, though these rasters were included in individual (specific) watershed models. Intensity rasters were also used qualitatively to aid manual land cover classification.

2.2.6 Return raster

A raster representing the average number of returns in a pixel was created by interpolating the return number of the last return for each LiDAR pulse. Like the intensity raster, the number of returns is a function of variables beyond just the land cover type and so was excluded from being used as classification model input, but it was used to aid manual land cover classification.

2.2.7 Slope rasters

For each elevation-based model (DSM, DEM, DMM, fDHM) a slope raster was created representing the maximum slope between a given pixel and its 8 neighbors. The slope is given in degrees. These products were included as direct input to the classification model.

2.2.8 Roughness rasters

A “roughness” raster was created for each elevation-based model (DSM, DEM, DMM, fDHM) by applying a filter that returns the maximum difference in values between a target pixel and its 8 neighbors. These products were included as direct input to the classification model.

2.2.9 Laplacian filtered rasters

Laplacian filters are commonly used to emphasize areas of rapid change in imagery. This emphasis on rapid change has led to widespread use of the Laplace filter for edge detection, but a Laplace-filtered image can be thought of more generally as a textural derivative of the parent image. Laplacian filters can be expressed as a convolution operation, and the convolution matrix used in this study is $\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$. A Laplacian-filtered raster was created for each elevation-based model (DSM, DEM, DMM, fDHM), and these products were included as direct input to the classification model.

2.2.10 Haralick textures

Haralick textures, or Haralick features, is the name given to a set of indicators derived from a matrix describing frequency of occurrence of specific value-pairs within subsets of an image. This frequency matrix is called the gray level co-occurrence matrix (GLCM). These derived features are referred to as textures because they tend to quantify the relationships between a given pixel and its neighbors.

From the DHM, we generated 8 Haralick textures: energy, entropy, correlation, inverse difference moment, inertia, cluster shade, cluster prominence and Haralick correlation. However, Haralick textures

were ultimately excluded as input to the classification model due to the great computational resources required relative to the low boost in model performance compared to models that excluded the textures.

2.3 Study Areas

The 10 areas of interest selected for this study were chosen for their diverse geomorphological, anthropogenic and ecological signatures as well as availability of LiDAR data. Seven of these study areas were selected to train the model ("trained" watersheds), which was validated against the three remaining watersheds ("naïve" watersheds). Characteristics each watershed can be found in Table 1, and an overview of each study area is given in subsequent sections. Characteristics of LiDAR data used to generate data products for each watershed can be found in Table 3.

2.3.1 Trained Watersheds

The following watersheds were used to train the classification model. 80% of the manually classified pixels from these watersheds were used for training, with the remaining reserved for validation.

2.3.1.1 010500021301 – Penobscot, ME

Branch Lake watershed (USGS hydrologic unit code: 010500021301) is located approximately 10 miles to the northeast of the coastal town of Penobscot, Maine, and drains to the eponymous Branch Lake. The watershed has a total area of 80.0 square kilometers. The majority of the watershed consists of coniferous forest and open water. Limited agricultural land, open fields and exposed rock are present as well. Overall development in the watershed is low. Relief within the watershed is moderate. Approximately 10.4 square kilometers, or 13.0%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 4/19/2015 and 5/12/2015 with a nominal pulse spacing of 0.70m.

2.3.1.2 030902040303 – Naples, FL

Strand State Preserve watershed (USGS hydrologic unit code: 030902040303) is located approximately 20 miles to the east of the town of Naples in southwestern Florida. The watershed is named for the Fakahatchee Strand State Preserve, a Florida State Park. The watershed has a total area of 74.7 square kilometers. The watershed is dominated by herbaceous forest, shrubs and wetlands. There is very little development within the watershed. Relief within the watershed is very limited. Approximately 8.1 square kilometers, or 10.8%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 6/12/2007 and 3/8/2008 with a nominal pulse spacing of 0.67m.

2.3.1.3 0070801050901 – Ames, IA

Walnut Creek watershed (USGS hydrologic unit code: 0070801050901) is located approximately 5 miles to the south of the city of Ames, Iowa. The watershed has a total area of 50.7 square kilometers. The watershed is dominated by agricultural land, with sporadic deciduous tree clusters and riparian buffers. Development is limited to sporadic housing and connecting roads. Relief within the watershed is very limited. Approximately 8.8 square kilometers, or 17.3%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected of a great length of time: between 4/7/2007 and 5/27/2010. The collected lidar has a nominal pulse spacing of 0.88m.

2.3.1.4 080102040304 – Trenton, TN

North Fork Forked Deer River Middle watershed (USGS hydrologic unit code: 080102040304) partially encompasses the town of Trenton in western Tennessee. The watershed has a total area of 161.1 square kilometers. Land cover in the watershed consists primarily of agricultural fields and thick deciduous-dominated riparian buffers. There are sporadic townships throughout the watershed. Relief within the watershed is modest. Approximately 10.3 square kilometers, or 6.4%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 12/2/2011 and 1/4/2012 with a nominal pulse spacing of 0.70m.

2.3.1.5 130202090102 – Datil, NM

Ox Spring Canyon watershed (USGS hydrologic unit code: 130202090102) is located approximately 10 miles north of the town of Datil, New Mexico. The watershed drains a portion of the northern face of Madre Mountain, part of the Datil mountain range. The watershed has a total area of 66.1 square kilometers. Land cover in the watershed is a mix of sand and sandy loam, including dry streambeds, and sparse, shrubby ligneous cover at lower elevations and somewhat denser mixed deciduous and coniferous forest at higher elevations. During the winter, snow may be present at higher elevations. Development in the watershed is nonexistent apart from occasional dirt roads. Relief within the watershed is significant as the watershed encompasses an area that includes mountain peaks at its head and flat scrubland near its terminal point. Approximately 7.7 square kilometers, or 11.6%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 11/2/2017 and 2/8/2018 with a nominal pulse spacing of 0.71m.

2.3.1.6 140801040103 – Telluride, CO

Mineral Creek watershed (USGS hydrologic unit code: 140801040103) is located approximately 5 miles south of the town of Telluride, Colorado. The watershed is situated within the Rock Mountains and has a total area of 138.6 square kilometers. Land cover in the watershed is diverse and consists of bare rock,

grass fields, alluvial deposits, melt-fed lakes and mixed coniferous-deciduous forest. Some amount of snow and ice is typically present perennially, though snow cover can be ubiquitous during the winter. Development in the watershed is limited to footpaths. Relief within the watershed is significant. Approximately 13.2 square kilometers, or 9.6%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 9/6/2017 and 10/24/2017 with a nominal pulse spacing of 0.35m.

2.3.1.7 180500020905 – San Francisco, CA

Lobos Creek-Frontal San Francisco Bay Estuaries watershed (USGS hydrologic unit code: 180500020905) encompasses part of the city of San Francisco, California. The watershed has a total area of 25.3 square kilometers. Land cover in the watershed largely consists of high-density urban development, including some high-rise buildings. However, a significant amount of plant cover is present as the watershed encompasses The Presidio and part of Golden Gate Park, two large urban parks, in addition to numerous smaller green spaces. A significant number of individual trees are also present along streets throughout the watershed. Thus, herbaceous and woody plant cover comprise an appreciable portion of the land cover in the watershed. San Francisco is a famously hilly city, and relief within the watershed is appreciable though moderate compared to the drastic relief seen in some of the other watersheds included in this study. Approximately 1.3 square kilometers, or 5.1%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 4/23/2010 and 7/14/2010 with a nominal pulse spacing of 0.6m.

2.3.2 Naïve Watersheds

The following watersheds were used for model validation only – no data from these watersheds were used for model training.

2.3.2.1 080902030201 – New Orleans, LA

Morrison Canal watershed (USGS hydrologic unit code: 080902030201) is encompasses part of the city of New Orleans, Louisiana. The has a total area of 61.2 square kilometers. Land cover in the watershed consists of medium density urban development, industrial yards, mowed fields, wetlands and open water. Limited agglomerations of trees are also present. Relief within the watershed is minimal. Approximately 4.0 square kilometers, or 6.6%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 1/23/2017 and 4/24/2017 with a nominal pulse spacing of 0.70m.

2.3.2.2 100301011309 – Helena, MT

Last Chance Gulch watershed (USGS hydrologic unit code: 100301011309) encompasses part of the city of Helena, Montana. The watershed has a total area of 55.5 square kilometers. Though a portion of the watershed is within the city limits of Helena and thus has land use typical of a small American city, most of the watershed is within a rugged, undeveloped mountain drainage to the southwest. Land cover in this drainage consists of open grassy areas punctuated by sparse shrub cover on south-facing slopes and dense mixed coniferous-deciduous tree cover on northerly slopes. Relief within the watershed is significant. Approximately 3.2 square kilometers, or 5.8%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 5/8/2012 and 5/9/2012 with a nominal pulse spacing of 1.46m.

2.3.2.3 102901110304 – Freeburg, MO

Loose Creek-Maries River watershed (USGS hydrologic unit code: 102901110304) encompasses part of the town of Freeburg, Missouri. The watershed has a total area of 77.8 square kilometers. The watershed's land cover is dominated by agricultural land use and open fields punctuated by relatively large stands of deciduous trees. Sporadic rural development is also present. Relief within the watershed is moderate. Approximately 6.1 square kilometers, or 6.8%, of the watershed was manually classified. The LiDAR data utilized in this study for the watershed was collected between 4/14/2010 and 5/7/2010 with a nominal pulse spacing of 1.33m.

3 Results and Discussion

The decision tree for the final model is represented graphically in Figure 3. Though a total of 26 data products (Table 5) were generated from LiDAR data, only five (fDHM roughness, DSM roughness, Laplace filter of DSM, slope DEM, slope of DSM) appear in the final decision tree. The fDHM roughness is by far the most import input to the model, with a normalized feature importance of 0.9468. The DSM roughness, Laplace filter of DSM, slope of DEM and slope of DSM have normalized feature importances of 0.0421, 0.0047, 0.0046 and 0.0018, respectively. This suggests that nearly all of the model's classification power is comes from the fDHM roughness and DSM roughness, while the Laplace filter of DSM, slope of DEM and slope of DSM provide only minor contributions.

A summary of model performance is given in Table 1. Average general model quality is 96.3% (Cohen's κ = 0.901) within the 7 watersheds used to train the model. Average general model accuracy for the three naïve watersheds is slightly higher at 98.5% (Cohen's κ = 0.958). Most watersheds show, at most, marginal improvement when classification is performed using a specific model rather than the general model. The majority of misclassification results from false positives (classification of non-tree pixels as tree pixels). This primarily occurs at building edges and power lines, which split LiDAR returns, resulting

in a canopy-like signal. Though filtering the DHM removes many of these false signals, it does not remove all of them. Filtering also removes some true canopy signals, resulting in false negatives (classification of tree pixels as non-tree pixels). This is a smaller contributor to model misclassification than false positives. Model misclassification is summarized as a confusion matrix in Table 2.

Overall model performance is excellent both on datasets from which subsets were used to train the model and on entirely naïve datasets even though the study areas that comprise the naïve data are physiographical and developmentally dissimilar from those that comprise the trained datasets. Furthermore, two of the three LiDAR datasets for the naïve watersheds have a point density that is significantly lower than the point density in the trained watersheds. These two facts together suggest that the canopy signal in the contiguous United States is consistent regardless of study area of LiDAR sampling density. Thus, the model presented in this paper can identify canopy coverage in the contiguous United States from LiDAR alone and without any user input regardless of the study area or if the LiDAR sampling density is suboptimal. Additionally, the negligible improvement when using a water-shed specific classification model that also includes intensity data shows that general model performance rivals that of custom models, even though the general model is more limited in the diversity of data that can be utilized.

General, pretrained models such as the one presented here are of great use to investigators seeking to quickly quantify land cover. Canopy coverage quantification is, as an example, is critical to the monitoring of forest changes or riparian buffer health. By generalizing canopy detection, our model lowers the level of technical ability and data quality needed to produce accurate canopy cover models. Generalization, however, imposes limitations on the kinds of data that can be used in the model. Most previously published, site-specific models that make use of LiDAR data use return intensity values, which vary depending on the type of land cover that is struck by a given pulse. Often a significant portion of land cover variance is explained by the return intensities and as a result some of these models can accurately differentiate multiple land cover classes or even tree species. However, return intensities are not normalized, and are a function of not just the land cover being sensed but the original signal intensity, receiver sensitivity, flight altitude and other mission-specific characteristics as well. Because of this, return intensities cannot be included in a pretrained general model as there is no general correspondence between landcover and signal intensity. This is also true of return numbers: higher intensity pulses and pulses with larger footprints will result in more “pulse splitting”, where a multi-level object such as a tree creates multiple returns as different levels reflect different parts of the pulse footprint[3]. By restricting model input, the generalizability of the model is increased at the expense potential model granularity. However, comparison of our general model to the site-specific models generated for this paper using

[3] Though we have excluded ordinal return numbers as model input, our model does make use of first and last return labeling as we believe this is a coarse enough distinction that it is uniformly captured by the vast majority of LiDAR missions.

intensity values shows that inclusion of intensity data provides little advantage for canopy detection. Still, developments in LiDAR data standards, such as intensity normalization, may allow the use of additional LiDAR data fields in a general model and thus allow finer land cover classification, broadening the usefulness of general models.

We also recognize that the physiographic diversity of the entire contiguous United States cannot be captured from sampling just ten watersheds. However, the remarkable consistency of the canopy signals for very physiographical disparate areas suggests that canopy signals are relatively consistent regardless of study area.

4 Conclusions

This study demonstrated the existence of a consistent canopy signal in LiDAR data across the contiguous United States and at various sampling densities, and that this can be exploited to create a general canopy detection model that is functional regardless of study area or LiDAR quality. The model presented performed with high accuracy on both the datasets used to train the model and on naïve datasets that the model had never encountered. This general model can be used without any parameter adjustment or input of training data, allowing investigators with limited knowledge of machine learning or land cover classification to perform high quality canopy delineations. The ability of the model to provide accurate classification without any training data input or parameter adjustment also suggests that a general classification model can be included in remote sensing software packages as a preset, providing an alternative to existing preset tools that rely on vendor-supplied classification codes or height-based heuristics.

Though our model is currently limited in its classification ability beyond simple binary canopy detection, future normalization of LiDAR parameters or data fusion may enable the design of general models that can perform more granular land cover classification.

5 Data and Materials Availability

All code used in the analysis for this paper is publicly available at https://github.com/rsjones94/riparian-id. The LiDAR and manual landcover classification data are available upon request.

6 Tables and Figures

Table 1. Information related to each study area used to train or validate the model. Model quality was quantified for both the general model and for custom models for each watershed trained using only data from the corresponding

watershed. The changes in quality metrics were quantified as the difference between the model metric for the model specific to a given watershed and the model metric for the general model as applied to that watershed.

Watersheds used to train and validate model												
HUC12	Name	Location	Type	Size (sq km)	Manually classified area (sq km)	Fraction manually classified	Cohen's κ (general)	Accuracy (general)	Cohen's κ (specific)	Accuracy (specific)	Change in κ	Change in Accuracy
010500021301	Branch Lake	Penobscot, ME	Rugged coastal	80.0	10.4	13.0%	0.939	97.1%	0.982	99.1%	0.04	2.1%
030902040303	Strand State Preserve	Naples, FL	Flat coastal	74.7	8.1	10.8%	0.855	95.2%	0.969	99.1%	0.11	3.8%
070801050901	Walnut Creek	Ames, IA	Agricultural	50.7	8.8	17.3%	0.978	99.6%	0.953	99.2%	-0.03	-0.5%
080102040304	North Fork Forked Deer River Middle	Trenton, TN	Agricultural	161.1	10.3	6.4%	0.900	95.6%	0.961	98.3%	0.06	2.7%
130202090102	Ox Spring Canyon	Datil, NM	Mountains and desert	66.1	7.7	11.6%	0.823	94.8%	0.891	97.0%	0.07	2.2%
140801040103	Mineral Creek	Telluride, CO	Mountainous	136.8	13.2	9.6%	0.870	94.3%	0.885	94.9%	0.01	0.6%
180500020905	Lobos Creek-Frontal San Francisco Bay Estuaries	San Francisco, CA	Urban	25.3	1.3	5.1%	0.945	97.3%	0.972	98.6%	0.03	1.3%
TOTAL (unweighted)				**594.7**	**59.6**	**10.0%**	**0.901**	**96.3%**	**0.945**	**98.0%**	**0.04**	**1.8%**
Watersheds used for validation only												
080902030201	Morrison Canal	New Orleans, LA	Urban	61.2	4.0	6.6%	0.951	97.7%	0.955	97.8%	0.00	0.2%
100301011309	Last Chance Gulch	Helena, MT	Mixed urban/arid forest	55.5	3.2	5.8%	0.929	98.0%	0.936	98.2%	0.01	0.2%
102901110304	Loose Creek-Maries River	Freeburg, MO	Agricultural	77.8	6.1	7.8%	0.995	99.8%	0.995	99.8%	0.00	0.0%
TOTAL (unweighted)				**194.5**	**13.3**	**6.8%**	**0.958**	**98.5%**	**0.962**	**98.6%**	**0.00**	**0.1%**

Table 2. Confusion matrix for the model as applied to the trained and naïve watersheds.

		True Class			
		Trained		Naïve	
		Trees	Other	Trees	Other
Predicted Class	Trees	6270072 (97.7%)	322156 (5.9%)	9057777 (99.5%)	129929 (3.1%)
	Other	149893 (2.3%)	5167202 (94.1%)	43050 (0.5%)	4039993 (96.9%)

Table 3. Information related to LiDAR characteristics for each study area.

Watersheds used to train and validate model							
HUC12	Location	Mission Begin	Mission End	Custodian	Vendor	LAS Version	Nominal Pulse Spacing (m)
010500021301	Penobscot, ME	4/19/2015	5/12/2015	United States Geological Survey	Quantum Spatial	1.4	0.70
030902040303	Naples, FL	6/12/2007	3/8/2008	United States Geological Survey	Woolpert	1.1	0.67
070801050901	Ames, IA	4/7/2007	5/27/2010	United States Geological Survey	Sanborn	1.1	0.88
080102040304	Trenton, TN	12/2/2011	1/4/2012	US Army Corps of Engineers	Laser Mapping Specialists	1.2	0.70
130202090102	Datil, NM	11/2/2017	2/8/2018	United States Geological Survey	Merrick-Surdex Joint Venture	1.4	0.71
140801040103	Telluride, CO	9/16/2017	10/24/2017	United States Geological Survey	Woolpert	1.4	0.35
180500020905	San Francisco, CA	4/23/2010	7/14/2010	United States Geological Survey	Earth Eye	1.2	0.60
Watersheds used for validation only							
080902030201	New Orleans, LA	1/23/2017	4/24/2017	United States Geological Survey	Fugro Earthdata	1.4	0.70
100301011309	Helena, MT	5/8/2012	5/9/2012	United States Geological Survey	Sanborn	1.3	1.46
102901110304	Freeburg, MO	4/14/2010	5/7/2010	United States Geological Survey	Photo Science	1.2	1.33

Table 4. Training and validation data classification scheme used in this study.

Category	Reclassification	Classified Area (sq km)
Forest	Trees	41.16
Linear tree stands	Trees	0.58
Individual trees/small clusters	Trees	0.43
Building tops	Other	0.66
Building edges	Other	0.08
Dirt/bare field	Other	16.22
Crops	Other	3.22
Rough vegetation	Other	1.10
Other impervious surfaces	Other	0.78
Water	Other	4.53
Snow	Other	0.08
Bare rock	Other	2.79
Sand	Other	0.30
Wetlands	Other	0.78
Power Lines	Other	0.07
Utility easement	Other	0.04
Canyon and cliff edges	Other	0.01

Table 5. Summary of LiDAR-derived raster products used in the model. A total of 26 products were generated. Of these, only five derived products appear in the final decision tree.

Raster	Used to train model	Appears in decision tree	Notes
Digital surface model	No	No	
Digital elevation model	No	No	
Digital height model	No	No	
Filtered digital height model	Yes	No	
Intensity raster	No	No	
Return raster	No	No	
Slope rasters	Yes	Yes (from DEM and DSM)	Generated for surface, elevation, height and filtered height models.
Roughness rasters	Yes	Yes (from fDHM and DSM)	Generated for surface, elevation, height and filtered height models.
Laplacian filtered rasters	Yes	Yes (from DSM)	Generated for surface, elevation, height and filtered height models.
Haralick textures	No	No	Generated for digital height model only. 8 textures total.

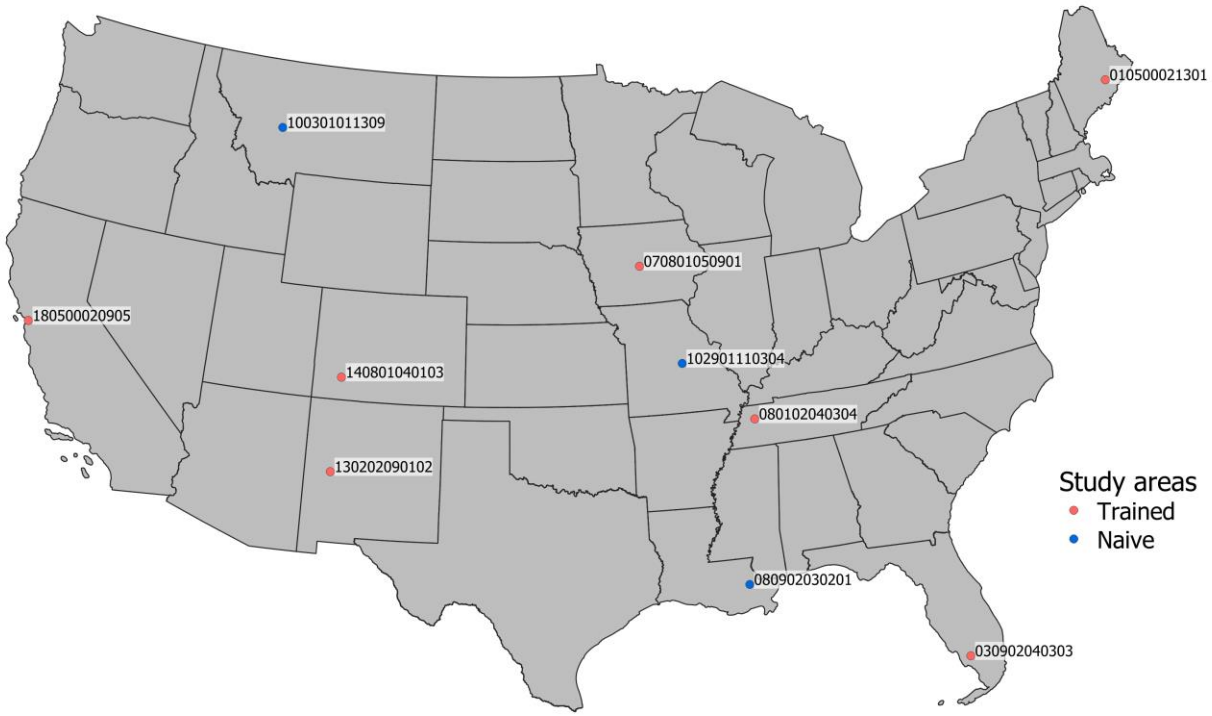

Figure 1. Distribution of watersheds analyzed in this study.

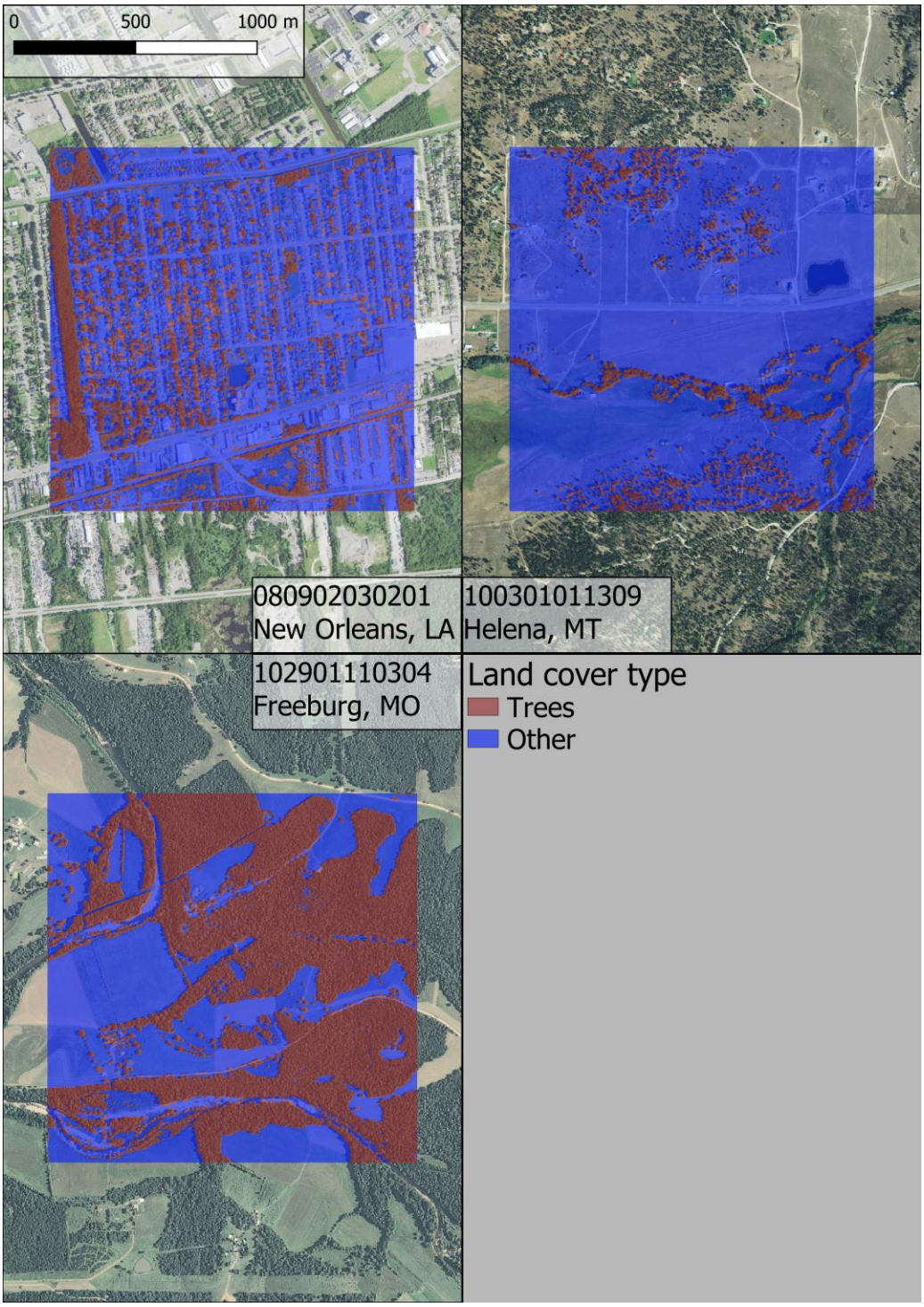

Figure 2. Sample classifications for three of the naïve study areas. Note that the orthoimagery for the New Orleans watershed is not perfectly contemporaneous with the LiDAR input; some trees in the study area were removed between LiDAR and orthoimagery data collection.

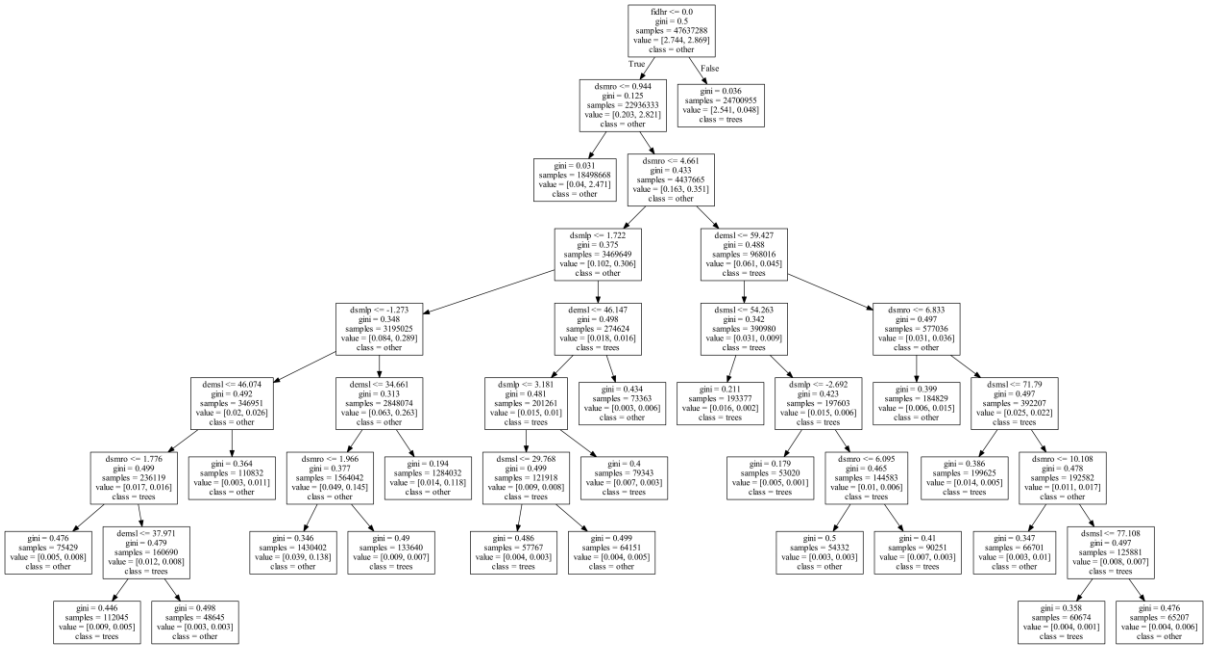

Figure 3. Decision tree of the final model. fidhr = filtered digital height model roughness. dsmro = digital surface model roughness. dsmlp = Laplace filter of digital surface model. demsl = slope of digital elevation model. dsmsl = slope of digital surface model.

5

7 Acknowledgements

We give our thanks to Connor Reed for his assistance with the creation of our manually classified land cover dataset.

8 Funding

This work was supported by the Natural Resources Conservation Service (agreement number NR194741XXXXC005).

9 References

Goodwin, Nicholas R., Nicholas Coops, Andreas Christen, and James A. Voogt. 2009. "Characterizing urban surface cover and structure with airborne lidar technology." *Canadian Journal of Remote Sensing* 35 (3): 297-309.

Holmgren, Johan, and Eva Lindberg. 2019. "Tree crown segmentation based on a tree crown density model derived from Airborne Laser Scanning." *Remote Sensing Letters* 10 (12): 1143-1152.

Kupidura, Przemysław. 2019. "The Comparison of Different Methods of Texture Analysis for Their Efficacy for Land Use Classification in Satellite Imagery." *Remote Sensing* 11 (10): 1233.

MacFaden, Sean W., Jacqueline W.T. Lu, and Andrew Rundle. n.d. "High-resolution tree canopy mapping for New York City using LIDAR and object-based image analysis." *Journal of Applied Remote Sensing* 6 (1): 3567.

Weinstein, Ben G., Sergio Marconi, Stephanie Bohlman, Alina Zare, and Ethan White. 2019. "Individual Tree-Crown Detection in RGB Imagery Using Semi-Supervised Deep Learning Neural Networks." *Remote Sensing* 11 (11): 1309.

Zhen, Zhen, Lindi J. Quackenbush, and Lianjun Zhang. 2016. "Trends in Automatic Individual Tree Crown Detection and Delineation—Evolution of LiDAR Data." *Remote Sensing* 8 (4): 333.

Potential works:

- Weinstein, B.G.; Marconi, S.; Bohlman, S.; Zare, A.; White, E. Individual Tree-Crown Detection in RGB Imagery Using Semi-Supervised Deep Learning Neural Networks. Remote Sens. 2019, 11, 1309 **AND** Geographic Generalization in Airborne RGB Deep Learning Tree Detection Ben Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, Ethan P White bioRxiv 790071; doi: https://doi.org/10.1101/790071

- Related to https://github.com/weecology/DeepForest. Covers a generalizable RGB-based model for crown-detection. Essentially our work, except imagery based and goes a step further (detecting crowns). Maybe we can implement crown detection?

- https://pypi.org/project/forestutils/
  - Tree extraction directly from point clouds. I believe it needs full color point clouds, which come from either photogrammetry or full-color (i.e., expensive) LiDAR scans
- A Segmentation Method for Tree Crown Detection and Modelling from LiDAR Measurements - https://link.springer.com/chapter/10.1007/978-3-642-31149-9_7
  - 2012 conference paper on the inverted watershed (raster-based) method of crown detection
- Trends in Automatic Individual Tree Crown Detection and Delineation—Evolution of LiDAR Data
- https://www.researchgate.net/publication/259128823_High-resolution_tree_canopy_mapping_for_New_York_City_using_LIDAR_and_object-based_image_analysis]
  - Characterizing urban surface cover and structure with airborne lidar technology
  - Contains schematic for image filter that remove linear structures (building edge detection)
- The Comparison of Different Methods of Texture Analysis for Their Efficacy for Land Use Classification in Satellite Imagery