

Building Footprint Identification with Airborne LiDAR: A Final Project for FOR 796

Lucas K Johnson

Abstract

Airborne LiDAR has emerged as a uniquely valuable and information-rich source of remote sensing (RS) data for high-resolution forest assessment and mapping. However, LiDAR data is limited in its ability to distinguish man-made structures from natural ones, necessitating the addition of external forest masks in LiDAR-based forest mapping projects. These external masks are often either too inaccurate or too expensive, leaving room for an efficient middle way. In this report I aimed to produce cost-efficient models that can identify buildings in 30m pixels. I fit a simple logistic regression model and two machine-learning models with a set of 39 LiDAR-derived predictors and open source building footprint data. Results indicated that both random forests and stochastic gradient boosting machines can predict the presence of buildings with a high degree of accuracy ($AUC = 0.96$) given the input data used in this report. With further assessment and tuning, these models can likely be used to efficiently produce accurate forest masks anywhere high-quality airborne LiDAR data has been acquired.

1. Introduction

Forest mapping and monitoring is becoming increasingly important as federal, state, and global agencies look towards natural solutions to mitigate a warming climate and the myriad resulting challenges. Field sampling programs, like the United States Department of Agriculture's Forest Inventory and Analysis program (FIA) (Gray et al. 2012), provide unbiased estimates of forest structure over large areas, but lack the fine spatial resolution to understand and manage

forests at relevant scales. Thus, high-resolution forest mapping is needed to inform decision-makers where forest resources should be managed or preserved.

Airborne LiDAR has been established as the most valuable remote sensing data for the purposes of forest structure mapping (Huang et al. 2019; Hurtt et al. 2019; Chen and McRoberts 2016). However, due to the nature of this data and its near-singular ability to characterize three-dimensional height-profiles, LiDAR cannot inherently distinguish between man-made structures and woody vegetation. To address this challenge, auxiliary landcover or forest canopy masks are often applied to LiDAR-modeled surfaces in attempt remove erroneous predictions in buildings from those in forested areas (Huang et al. 2019). However, it is well documented that landcover maps are not 100% accurate, and significant quantities of forest can be contained in non-forested classes (Johnson et al. 2014; Perry et al. 2008; Meneguzzo, Liknes, and Nelson 2012). Additionally, high-resolution tree-canopy delineation surfaces are expensive to produce often relying on expert interpretation and iterative tuning (J. P. M. O’Neil-Dunne et al. 2013; J. O’Neil-Dunne, MacFaden, and Royar 2014).

In this report I attempt to find a middle way by producing models at reduced cost that can predict the presence of buildings in a mixed-use landscape with a high-degree of accuracy at a 30m resolution. To do this I train a simple logistic model, a random forest model, and a stochastic gradient boosting machine to classify the presence (1) or absence (0) of any buildings in a given map pixel. Building indicator response data was derived from an open-source building footprint dataset developed by Microsoft (Microsoft 2018; Team 2018). The predictor data used to train these models are LiDAR-derived grid metrics, commonly used in models of forest-structure, aiding in the reduction of cost. If these models prove to be successful classifiers of building presence, the same predictors will afford future forest-structure modelers double the benefit.

2. Methods

2.1. Building and LiDAR Data

Rasterized building footprint data served as the response data in this study (Pourpeikari Heris et al. 2020; Heris et al. 2020; Team 2018; Microsoft 2018). The raw raster contains counts of the number of buildings intersecting each pixel. This data was converted to a Boolean raster where 1s represent the presence of any buildings, and 0s represent a complete lack of buildings. This data was chosen for its availability across the entire country, its high accuracy ($> 99\%$ positive predictive value), and its 30m resolution (Heris et al. 2020).

The raw LiDAR data originates from a single acquisition covering the city of Buffalo and larger portions of Erie, Genesee, and Livingston counties in western New York (New York Office of Information Technology Services 2019). This particular data was selected due to its known ability to characterize three dimensional height-profiles at high-resolution, and the range of landcover conditions (urban, forest, cropland). The data was made available by the New York State GIS Program Office. The raw LiDAR data was height-normalized and converted into a set of 39 predictors (Table 1) chosen for their prevalence in models of forest structure (Hawbaker et al. 2010; Huang et al. 2019; Pflugmacher et al. 2014).

The LiDAR predictors, in raster stack form, were overlaid with the Boolean building raster to create stack of data where each pixel contained a set of 39 predictors and one building indicator response variable. A stratified random sample was conducted on the raster stack, with the building indicator providing the levels of stratification. 3,500 pixels were selected from each stratum resulting in 7,000 observations for model training and testing. This final dataset was converted to a 7000x40 (rows, columns) data frame. The *lidR* (Roussel and Auty 2020; Jean-Romain et al. 2020) and *raster* (Hijmans 2021) packages were used for height-normalization and dataset generation. Additionally, the first seven principle components were derived from the final dataset to produce an alternative dataset without multicollinearity predictor dataset (R Core Team

2021). This alternative dataset accounted for $\geq 95\%$ of the information in the raw predictors and existed as a 7000x7 data frame.

2.2. Models

Three candidate classification models were fit to a random 70% (calibration data; $n = 3500$) of the observations, with the remaining 30% reserved for model performance assessment (holdout data; $n = 1500$). The first candidate model was a simple logistic regression model (R Core Team 2021), and was trained on the principle components variant of the calibration data. The second candidate model was a random forest (RF hereafter) trained with the ranger R package (Wright and Ziegler 2017). The third candidate model was a stochastic gradient boosting machine (LGB hereafter) trained with the LightGBM R package (Ke et al. 2021). The hyperparameters for both the RF and LGB models were selected using a standard grid search where each combination of hyperparameters were compared against each other using the cross-entropy loss function (CEL; Equation (1)) computed from a random five-fold cross-validation with the calibration dataset. CEL is computed as follows:

$$\text{CEL} = \sum_{i=1}^n -\log(\hat{y}_i) \quad (1)$$

Where n is the number of observations in the fold, and \hat{y}_i is the predicted probability of the true class.

Positive prediction thresholds for all models were chosen using the optimal ROC coordinates for the fully tuned models fit to the calibration data. Each of the models were assessed against the holdout dataset and compared to one another using overall accuracy, specificity, sensitivity, and AUC. Additionally ROC curves were plotted for each model's results on the holdout set. The caret and pROC R packages were used to compute these accuracy metrics (Kuhn 2021; Robin et al. 2011).

Table 1: Definitions of predictors used for model fitting.

Predictor	Definition
H0, H10, ... H100, H95, H99	Decile heights of returns, in meters, as well as 95th and 99th percentile return heights.
D10, D20... D90	Density of returns above a certain height, as a proportion. After return height is divided into 10 equal bins ranging from 0 to the maximum height of returns, this value reflects the proportion of returns at or above each breakpoint.
ZMEAN, ZMEAN_C	Mean height of all returns (ZMEAN) and all returns above 2.5m (ZMEAN_C)
Z_KURT, Z_SKEW	Kurtosis and skewness of height of all returns
QUAD_MEAN, QUAD_MEAN_C	Quadratic mean height of all returns (QUAD_MEAN) and all returns above 2.5m (QUAD_MEAN_C)
CV, CV_C	Coefficient of variation for heights of all returns (CV) and all returns above 2.5m (CV_C)
L2, L3, L4, L_CV, L_SKEW, L_KURT	L-moments and their ratios as defined by Hosking (1990), calculated for heights of all returns
CANCOV	Ratio of returns above 2.5m to all returns (Pflugmacher et al. 2012)
HVOL	CANCOV * ZMEAN (Pflugmacher et al. 2012)
RPC1	Ratio of first returns to all returns (Pflugmacher et al. 2012)

Table 2: Model accuracy metrics computed against holdout partition (n = 1500).

Model	AUC	Overall Accuracy	Sensitivity	Specificity
Logistic	0.87	0.79	0.83	0.75
RF	0.96	0.90	0.92	0.89
LGB	0.96	0.90	0.90	0.90

3. Results

The RF and LGB models were significantly better than the Logistic model across all accuracy metrics (Table 2). While the RF and LGB models shared the same AUC and Overall Accuracy results, the RF model was slightly more specific while the LGB model was slightly more sensitive. However, all three candidate models performed quite well with all AUC values ≥ 0.87 , and all overall accuracies ≥ 0.79 . The ROC curves plotted in Figure 1 display similar patterns.

4. Discussion

It is unsurprising that the two machine learning models (RF and LGB) outperformed the simple Logistic regression model given the constraints on multicollinearity required for Logistic regression. In particular, the Logistic model might have been improved by including more principle components, as in this case we limited the input data to the first seven principle components which accounted for 95% of the information in the raw predictors. This may have been an unfair disadvantage as the RF and LGB models were given the opportunity to leverage 100% of the information in the predictor space.

There are a few ways we could further improve the models described in this report. First, more extensive grid searches could have been performed to find better hyperparameters. There is a trade-off here between the combination of

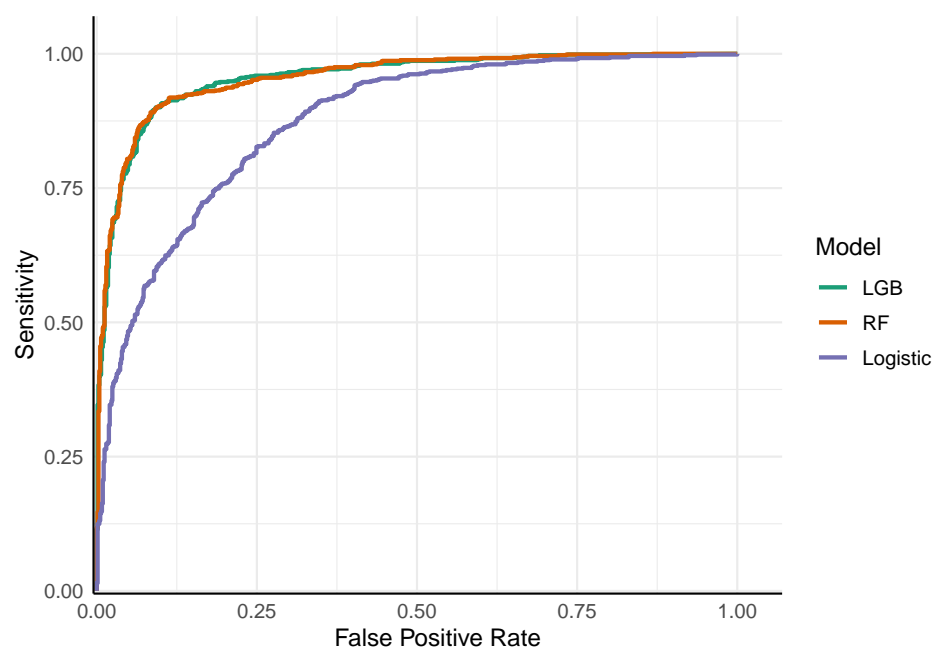


Figure 1: ROC curves for three models tested against the holdout partition ($n = 1500$).

time and performance gains, with eventually diminishing returns on the time invested. The tuning performed in this report produced models that were good enough for our purposes. Additionally, we might be able to produce an even better model by using stacked ensembles, which often serve to reduce predictive error especially when the error from component models is dominated by variance (Dormann et al. 2018). Noisy data, which we can assume categorizes our LiDAR and building data, often yields models with variance dominated error (Dormann et al. 2018). One potential source of error in our models is the temporal match between predictor and response data used to fit the models. The building data, though published in 2018, has no associated time-of-acquisition requiring us to hope that the building classifications describe conditions close to those represented in the 2019 LiDAR acquisition (Heris et al. 2020).

Since the predictors used to train these models are primarily used for predicting forest structure, these models provide a relatively simple option to mask man-made structures from LiDAR-derived maps of forest structure like above-ground biomass or canopy height. Further investigation is required to assess the transferrability of these models trained in one region (western New York) with one LiDAR acquisition to others. If separate models are required for each distinct LiDAR acquisition or region, the relative benefit of the models developed in this report would only be slightly diminished, as these models would still serve as a reference point for other applications. Finally, a true accuracy assessment, using time-relevant reference data of higher quality than the building data used herein should be conducted to assess the suitability of these models in real-world applications (Stehman and Foody 2019).

References

- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r*. <https://github.com/rstudio/rmarkdown>.
- Chen, Qi, and Ronald McRoberts. 2016. “Statewide Mapping and Estimation of Vegetation Aboveground Biomass Using Airborne Lidar.” In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. <https://doi.org/10.1109/igarss.2016.7730157>.
- Dormann, Carsten F., Justin M. Calabrese, Gurutzeta Guillera-Arroita, Eleni Matechou, Volker Bahn, Kamil Bartoń, Colin M. Beale, et al. 2018. “Model Averaging in Ecology: A Review of Bayesian, Information-Theoretic, and Tactical Approaches for Predictive Inference.” *Ecological Monographs* 88 (4): 485–504. <https://doi.org/10.1002/ecm.1309>.
- Gray, Andrew N, Thomas J Brandeis, John D Shaw, William H McWilliams, and Patrick Miles. 2012. “Forest Inventory and Analysis Database of the United States of America (FIA).” *Biodiversity and Ecology* 4: 225–31. <https://doi.org/10.7809/b-e.00079>.
- Hawbaker, Todd J., Terje Gobakken, Adrian Lesak, Eric Trømborg, Kirk Contrucci, and Volker Radeloff. 2010. “Light Detection and Ranging-Based Measures of Mixed Hardwood Forest Structure.” *Forest Science* 56 (3): 313–26. <https://doi.org/10.1093/forestscience/56.3.313>.
- Henry, Lionel, and Hadley Wickham. 2021. *Tidysselect: Select from a Set of Strings*. <https://CRAN.R-project.org/package=tidysselect>.
- Heris, Mehdi P., Nathan Leon Foks, Kenneth J. Bagstad, Austin Troy, and Zachary H. Ancona. 2020. “A Rasterized Building Footprint Dataset for the United States” 7 (1). <https://doi.org/10.1038/s41597-020-0542-3>.
- Hijmans, Robert J. 2021. *Raster: Geographic Data Analysis and Modeling*. <https://CRAN.R-project.org/package=raster>.
- Hosking, J. R. M. 1990. “L-Moments: Analysis and Estimation of Distri-

- butions Using Linear Combinations of Order Statistics.” *Journal of the Royal Statistical Society. Series B (Methodological)* 52 (1): 105–24. <http://www.jstor.org/stable/2345653>.
- Huang, Wenli, Katelyn Dolan, Anu Swatantran, Kristofer Johnson, Hao Tang, Jarlath O’Neil-Dunne, Ralph Dubayah, and George Hurtt. 2019. “High-Resolution Mapping of Aboveground Biomass for Forest Carbon Monitoring System in the Tri-State Region of Maryland, Pennsylvania and Delaware, USA.” *Environmental Research Letters* 14 (9): 095002. <https://doi.org/10.1088/1748-9326/ab2917>.
- Hurtt, G, M Zhao, R Sahajpal, A Armstrong, R Birdsey, E Campbell, K Dolan, et al. 2019. “Beyond MRV: High-Resolution Forest Carbon Modeling for Climate Mitigation Planning over Maryland, USA.” *Environmental Research Letters* 14 (4): 045013. <https://doi.org/10.1088/1748-9326/ab0bbe>.
- Jean-Romain, Roussel, David Auty, Nicholas C. Coops, Piotr Tompalski, Tristan R. H. Goodbody, Andrew Sánchez Meador, Jean-François Bourdon, Florian de Boissieu, and Alexis Achim. 2020. “lidR: An r Package for Analysis of Airborne Laser Scanning (ALS) Data.” *Remote Sensing of Environment* 251: 112061. <https://doi.org/10.1016/j.rse.2020.112061>.
- Johnson, Kristofer D, Richard Birdsey, Andrew O Finley, Anu Swantaran, Ralph Dubayah, Craig Wayson, and Rachel Riemann. 2014. “Integrating Forest Inventory and Analysis Data into a LIDAR-Based Carbon Monitoring System.” *Carbon Balance and Management* 9 (1). <https://doi.org/10.1186/1750-0680-9-3>.
- Ke, Guolin, Damien Soukhavong, James Lamb, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2021. *lightgbm: Light Gradient Boosting Machine*. <https://CRAN.R-project.org/package=lightgbm>.
- Kuhn, Max. 2021. *Caret: Classification and Regression Training*. <https://CRAN.R-project.org/package=caret>.
- Meneguzzo, Dacia M., Greg C. Liknes, and Mark D. Nelson. 2012. “Mapping Trees Outside Forests Using High-Resolution Aerial Imagery: A Comparison

- of Pixel- and Object-Based Classification Approaches.” *Environmental Monitoring and Assessment* 185 (8): 6261–75. <https://doi.org/10.1007/s10661-012-3022-1>.
- Microsoft. 2018. “US Building Footprints.” Microsoft.
- New York Office of Information Technology Services. 2019. “LIDAR collection (QL2) for Erie, Genesee, and Livingston Counties New York Lidar; Classified Point Cloud.”
- O’Neil-Dunne, Jarlath P. M., Sean W. MacFaden, Anna R. Royar, and Keith C. Pelletier. 2013. “An Object-Based System for LiDAR Data Fusion and Feature Extraction.” *Geocarto International* 28 (3): 227–42. <https://doi.org/10.1080/10106049.2012.689015>.
- O’Neil-Dunne, Jarlath, Sean MacFaden, and Anna Royar. 2014. “A Versatile, Production-Oriented Approach to High-Resolution Tree-Canopy Mapping in Urban and Suburban Landscapes Using GEOBIA and Data Fusion.” *Remote Sensing* 6 (12): 12837–65. <https://doi.org/10.3390/rs61212837>.
- Perry, C. H., C. W. Woodall, G. C. Liknes, and M. M. Schoeneberger. 2008. “Filling the Gap: Improving Estimates of Working Tree Resources in Agricultural Landscapes.” *Agroforestry Systems* 75 (1): 91–101. <https://doi.org/10.1007/s10457-008-9125-6>.
- Pflugmacher, Dirk, Warren B. Cohen, Robert E. Kennedy, and Zhiqiang Yang. 2014. “Using Landsat-Derived Disturbance and Recovery History and Lidar to Map Forest Biomass Dynamics.” *Remote Sensing of Environment* 151: 124–37. <https://doi.org/10.1016/j.rse.2013.05.033>.
- Pourpeikari Heris, Mehdi, Nathan Foks, Kenneth J Bagstad, and Austin Troy. 2020. “A National Dataset of Rasterized Building Footprints for the u.s.” U.S. Geological Survey. <https://doi.org/10.5066/P9J2Y1WG>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. “pROC: An

- Open-Source Package for r and s+ to Analyze and Compare ROC Curves.” *BMC Bioinformatics* 12: 77.
- Roussel, Jean-Romain, and David Auty. 2020. *Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*. <https://cran.r-project.org/package=lidR>.
- Stehman, Stephen V., and Giles M. Foody. 2019. “Key Issues in Rigorous Accuracy Assessment of Land Cover Products” 231 (September): 111199. <https://doi.org/10.1016/j.rse.2019.05.018>.
- Team, Bing Maps. 2018. “Computer Generated Building Footprints for the United States.” Microsoft.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Wright, Marvin N., and Andreas Ziegler. 2017. “Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r.” *Journal of Statistical Software, Articles* 77 (1): 1–17. <https://doi.org/10.18637/jss.v077.i01>.
- Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>.
- Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>.
- Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.