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 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 airbone 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 (Chen and McRoberts, 2016; Huang et al., 2019; Hurtt et al., 2019). 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; Meneguzzo et al., 2012; Perry et al., 2008). Additionally, high-resolution tree-canopy delineation surfaces are expensive to produce often relying on expert interpretation and iterative tuning (O'Neil-Dunne et al., 2014, 2013).

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 (Heris et al., 2020; Microsoft, 2018; Pourpeikari Heris et al., 2020; Team, 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 (Jean-Romain et al., 2020; Roussel and Auty, 2020) and raster (Hijmans, 2021) packages were

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)

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. 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% (training 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, and was trained on the principle components variant of the training data. The second candidate model was a random forest (RF herafter) trained with the ranger R package (Wright and Ziegler, 2017). The third candidate model was a stochastic gradient boosting

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	0.89	0.90
LGB	0.96	0.90	0.90	0.90

machine (LGB hereafter) trained with the LightGBM R package (Ke et al., 2021). The hyperparemeters for both the RF and LGB models were selected using a standard grid search where each combination of hyperparameters were compared against eachother using the cross-entropy loss function (CEL; Equation (1)) computed from a random five-fold cross-validation with the training dataset. CEL is computed as follows:

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

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

Postitive prediction thresholds for all models were chosen using the optimal ROC coordinates for the fully tuned models fit to the training 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).

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, Overall Accuracy, and specificity, the LGB model was slightly more sensitive than the RF model. 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

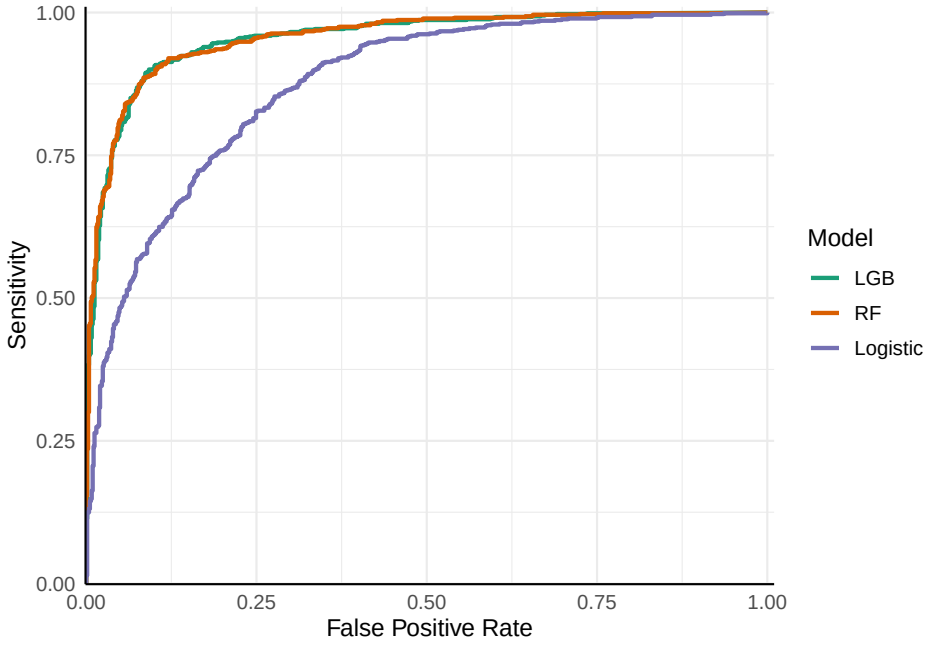

Figure 1: ROC curves for three models tested against the holdout partition (n = 1500).

this case I 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 I 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 time and performance gains, with eventually diminishing returns on the time invested. The relatively limited tuning performed in this report produced models that were good enough for my purposes. Additionally, I 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).

The models developed in this report can be used to produce Boolean building presence maps, which can be leveraged to mask away buildings from maps

of forest structure (e.g. canopy-height, aboveground biomass), aiding in the production of more accurate representations of forest area and conditions. Since the predictors used to train these models can be used for both modeling forest structure and building presence, they offer an efficient way to improve the accuracy of forest structure maps with the same dataset. Further investigation is required to assess the transferrability of these models trained in one region 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 diminsihed, 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 mapping applications (Stehman and Foody, 2019).

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., Iannone, R., 2021. Rmarkdown: Dynamic documents for r.

Chen, Q., McRoberts, R., 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, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wüest, R.O., Hartig, F., 2018. Model averaging in ecology: A review of bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs 88, 485–504. https://doi.org/10.1002/ecm.1309

Gray, A.N., Brandeis, T.J., Shaw, J.D., McWilliams, W.H., Miles, P., 2012. Forest inventory and analysis database of the united states of america (FIA). Biodiversity and Ecology 4, 225–231. https://doi.org/10.7809/b-e.00079

Hawbaker, T.J., Gobakken, T., Lesak, A., Trømborg, E., Contrucci, K., Radeloff, V., 2010. Light Detection and Ranging-Based Measures of Mixed Hardwood Forest Structure. Forest Science 56, 313–326. https://doi.org/10.1093/forestscience/56.3.313

Henry, L., Wickham, H., 2021. Tidyselect: Select from a set of strings.

Heris, M.P., Foks, N.L., Bagstad, K.J., Troy, A., Ancona, Z.H., 2020. A rasterized building footprint dataset for the united states 7. https://doi.org/10.1038/s41597-020-0542-3

Hijmans, R.J., 2021. Raster: Geographic data analysis and modeling.

Hosking, J.R.M., 1990. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society. Series B (Methodological) 52, 105–124.

Huang, W., Dolan, K., Swatantran, A., Johnson, K., Tang, H., O'Neil-Dunne, J., Dubayah, R., Hurtt, G., 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, 095002. https://doi.org/10.1088/1748-9326/ab2917

Hurtt, G., Zhao, M., Sahajpal, R., Armstrong, A., Birdsey, R., Campbell, E., Dolan, K., Dubayah, R., Fisk, J.P., Flanagan, S., Huang, C., Huang, W., Johnson, K., Lamb, R., Ma, L., Marks, R., O'Leary, D., O'Neil-Dunne, J., Swatantran, A., Tang, H., 2019. Beyond MRV: High-resolution forest carbon modeling for climate mitigation planning over maryland, USA. Environmental Research Letters 14, 045013. https://doi.org/10.1088/1748-9326/ab0bbe

Jean-Romain, Roussel, Auty, D., Coops, N.C., Tompalski, P., Goodbody, T.R.H., Meador, A.S., Bourdon, J.-F., de Boissieu, F., Achim, A., 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, K.D., Birdsey, R., Finley, A.O., Swantaran, A., Dubayah, R., Wayson, C., Riemann, R., 2014. Integrating forest inventory and analysis data into a LIDAR-based carbon monitoring system. Carbon Balance and Management 9. https://doi.org/10.1186/1750-0680-9-3

Ke, G., Soukhavong, D., Lamb, J., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2021. Lightgbm: Light gradient boosting machine.

Kuhn, M., 2021. Caret: Classification and regression training.

Meneguzzo, D.M., Liknes, G.C., Nelson, M.D., 2012. Mapping trees outside forests using high-resolution aerial imagery: A comparison of pixel- and object-based classification approaches. Environmental Monitoring and Assessment 185, 6261–6275. https://doi.org/10.1007/s10661-012-3022-1

Microsoft, 2018. US Building Footprints.

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, J.P., MacFaden, S.W., Royar, A.R., 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, 12837–12865.

O'Neil-Dunne, J.P., MacFaden, S.W., Royar, A.R., Pelletier, K.C., 2013. An object-based system for LiDAR data fusion and feature extraction. Geocarto International 28, 227–242.

Perry, C.H., Woodall, C.W., Liknes, G.C., Schoeneberger, M.M., 2008. Filling the gap: Improving estimates of working tree resources in agricultural landscapes. Agroforestry Systems 75, 91–101. https://doi.org/10.1007/s10457-008-9125-6

Pflugmacher, D., Cohen, W.B., Kennedy, R.E., Yang, Z., 2014. Using landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics. Remote Sensing of Environment 151, 124–137. https://doi.org/10.1016/j.rse.2013.05.033

Pourpeikari Heris, M., Foks, N., Bagstad, K.J., Troy, A., 2020. A national dataset of rasterized building footprints for the u.s. https://doi.org/10.5066/P9J2Y1WG

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: An open-source package for r and s+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.

Roussel, J.-R., Auty, D., 2020. Airborne LiDAR data manipulation and visualization for forestry applications.

Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products 231, 111199. https://doi.org/10.1016/j.rse.2019.05.018

Team, B.M., 2018. Computer Generated Building Footprints for the United States.

Wickham, H., 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., Müller, K., 2021. Dplyr: A grammar of data manipulation.

Wright, M.N., Ziegler, A., 2017. Ranger: A fast implementation of random forests for high dimensional data in c++ and r. Journal of Statistical Software, Articles 77, 1–17. https://doi.org/10.18637/jss.v077.i01

Xie, Y., Allaire, J.J., Grolemund, G., 2018. R markdown: The definitive guide. Chapman; Hall/CRC, Boca Raton, Florida.

Xie, Y., Dervieux, C., Riederer, E., 2020. R markdown cookbook. Chapman; Hall/CRC, Boca Raton, Florida.

Zhu, H., 2021. kableExtra: Construct complex table with 'kable' and pipe syntax.