Predicting forest carbon stocks in the contiguous U.S.

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

PLACEHOLDER

1 Introduction

An important step in addressing the risks of climate change is to protect and properly manage our natural resources. Forests are one such resource with a portfolio of benefits from temperature and moisture attenuation to carbon capture and storage [source]. This resource has been directly impacted by humans – through urbanization and agricultural expansion – and indirectly through droughts, fires, and other risks posed by climate change. Monitoring our forests and properly accounting for their extent and health could help improve their management and strengthen their protection. We can grow this pool through afforestation and deferred deforestation, oftentimes supported through carbon offset payments by sectors emitting carbon to the landowners growing these forests.

In order for a payment system like this to work, we need precise, low-cost methods for estimating the amount of carbon stored in a forest. These methods can be used to establish baseline estimates of forest biomass, providing a counterfactual scenario to how forest growth would have changed in the absence of offset payments. These methods can also be used to monitor changes in forest biomass after offset payments (the difference between these scenarios is precisely the amount of carbon offset by payments).

Decades of research and data collection by US government departments offers an opportunity for such low-cost estimation methods. In this project I propose to train a model to use remote sensing data collected by NASA satellites to predict forest biomass, collected through surveys by the USDA FIS. The model will aim to predict 5-year survey-period forest biomass at the county-level across the contiguous United States (CONUS). I will also aim to generate bootstrapped standard errors for my predictions to properly account for uncertainty in forest size, as provided by sample errors in the USDA FIS dataset. Standard errors can be a critical ingredient in accurately pricing carbon offsets.

1.1 Related work

Given the economic and biologic importance of forests, there is already much established literature on forest biomass prediction.

Han et al. and Li et al. both evaluate several machine learning algorithms to estimate forest biomass, using Sentilel-1 and Landsat 8 satellites **??**. [[SUMMARY OF APPROACHES]]. [[SUMMARY OF FINDINGS]]. Bjork et al. adopts deeper learning techniques for a similar aim: to estimate forest biomass from radar backscatter **?**. [[SUMMARY OF APPROACH]]. [[SUMMARY OF FINDINGS]].

Saarela et al. and Naik et al. evaluate newer remote sensing layers for incremental benefits to forest biomass estimates **??**. Saarela uses LiDAR and field data to estimate aboveground biomass and associated uncertainty, while Naik uses multispectral remote sensing data [[SUMMARY OF APPROACHES]]. [[SUMMARY OF FINDINGS]].

And Vorster et al. contributes to this body of research by exploring how allometric equations can improve estimates of uncertainty in biomass estimates **?**. [[SUMMARY OF APPROACH]], [[SUMMARY OF FINDINGS]].

2 Dataset

In this research, I compile a dataset of county-level biomass estimates from the USDA Forest Service and spatio-temporal aggregated satellite imagery data from NASA's Moderate Resolution Imaging Spectoradiometer (MODIS). I describe both data sources and their processing in more detail below.

Since the late 1930s, the USDA Forest Service has surveyed US forests at the county level through the Forest Inventory and Analysis program (USDA FS FIA). The FIA surveys a forest over 5-7 year periods through a random sampling procedure, and provides county-level 5-year-period forest biomass estimates to the public through an online portal called The Design and Analysis Toolkit for Inventory and Monitoring (DATIM) **?**. In this analysis I use the FIA's estimate of aboveground forest biomass measured in short tons of carbon.

The NASA MODIS satellite-based sensor is on board two satellites tracking land and ocean surface climate measurements, launched in the late 1990s and early 2000s respectively. For broader scale use, NASA provides Level 3 gridded data products built from lower-level products relying on the raw sensor readings. In this analysis, I use monthly binned L3 readings at a 5600m grid scale, accessed through NASA's EarthDataSearch platform [[CITE]]. In this analysis I consider mean monthly land surface daytime temperature and mean monthly normalized difference vegetation index (NDVI).

I create a uniform analysis dataset for this research by aggregating NASA MODIS data across space and time. The ultimate objective is to provide annual forecasts of forest biomass, so I only consider aggregating MODIS data at the annual level, and then map these annual readings to FIA reports by taking the average annual readings across each county-level 5-7-year reporting period. The annual aggregations I consider are mean, min, max, range, IQR, standard deviation, consecutive months below the mean reading, months below the 25th percentile reading, and months above the 75th percentile reading. I also aggregate these annual aggregations to the county level, and consider the mean, range, and standard deviation of these values across each county.

3 Technical approach

In this paper I aim to train a model with the lowest prediction error as determined by root mean squared error (RMSE). This metric best aligns with the ultimate objective of this model, which is to make the best county-level predictions of forest biomass from remote sensing data. For robustness, I also evaluate mean absolute error (MAE) and the coefficient of determination (R^2) across all models. Among a subset of the highest performing models, I use bootstrapping techniques to estimate standard errors of mean predictions across these models.

3.1 Model selection

I use cross validation to evaluate all models and arrive at the performance metrics described above. I choose to hold out 10% of my data for final testing, and to use the remaining 90% for model selection. Because of the potential spatial and temporal correlation of my data, I evaluate these models across three different splitting regimes: First, within each fold of my cross validation, I hold out [[25%]] of the county-years from each state at random and evaluate the performance of out-of-county-year prediction. Second, I hold out the last forest biomass report for each county and evaluate the performance of future predictions using a model trained on past data (each county has between 1 and 4 total reports). And third, I perform leave-one-state-out cross validation where, in each fold, I use all available time and county data from all but one state and evaluate the performance of out-of-state prediction.

In selecting the highest-performing model, I consider three categories of models: baseline models, machine learning models, and neural net models. I first train two baseline models. These models serve as a standard against which I can aim to improve predication error through the use of machine learning and neural net models. First, I consider using the simple global mean at each prediction location. And second, I consider a simple linear regression of all predictor variables on county-level forest biomass predictions.

I next optimize 5 machine learning models: Regularized linear regression (with Elastic Net), Simple Decision Trees, AdaBoost, XGBoost, and Random Forests. I optimize the hyperparameters associated with each of these models through a grid search approach. And lastly, I implement a fully connected neural net model, selecting optimization parameters, as well as the number and size of hidden layers, through a grid search approach.

My model selection is based on the highest performing model among those described in this section.

3.2 Bootstrapping prediction error

With the [[3]] highest performing models, I also bootstrap mean prediction values for each county and the standard errors of those mean predictions. I do this by generating new training datasets by taking random draws of a normal distribution centered at the USDA FIS provided estimate, with standard deviation equal to the standard error provided by the USDA FIS. I then evaluate the same scores I describe above, but can generate ranges of uncertainty about each. These results can be used to further support the final model selection.

4 Preliminary results

After tuning each of the models, I find that [[MODEL]] results in the lowest RMSE in [[SPLITTING REGIME]]. This performance is [[XXX]] percentage points lower than the next highest model, [[MODEL]], within this splitting regime. Overall, I find that [[MODEL]] performs best overall across all splitting regimes. Results are included in Table **??**

Broader impact

This research contributes to the larger field of forest biomass quantification. On the whole, improved approaches to forest monitoring will support humans' care for this natural resource.

While a central component of this paper focuses on estimate uncertainty, it is often the case for downstream users to use estimates without consideration of their uncertainty. When these forest biomass estimates are used in carbon markes or other in governmental or corporate statements without uncertainty qualifications, this can lead and has led to inflated claims of climate impact (a.k.a., "greenwashing").

Additionally, this research and the models in this paper directly benefit those countries and regions with better surveyed data. While remote sensing layers tend to have complete coverage of the Earth, the surveyed forest biomass as well as other potential predictors like soil types and crop layers tend to only be availble in wealthier countries. This translates to wealthier nations having access to more precise models and downstream economic and enviornmental benefits. That said, the contiguous United States has a variety of climactic regions and may be representative of large parts of the world, meaning these models could support developing countries.

Beyond what is mentioned above, leveraging biases in the data is not applicable to this research.

Split	Model	RMSE	MAE	R^2
County-year	Global mean			
County-year	Linear regression			
County-year	Elastic Net			
County-year	Decision Tree			
County-year	AdaBoost			
County-year	XGBoost			
County-year	Random Forest			
County-year	FC NNet			
Out-of-state	Global mean			
Out-of-state	Linear regression			
Out-of-state	Elastic Net			
Out-of-state	Decision Tree			
Out-of-state	AdaBoost			
Out-of-state	XGBoost			
Out-of-state	Random Forest			
Out-of-state	FC NNet			
Forward-in-time	Global mean			
Forward-in-time	Linear regression			
Forward-in-time	Elastic Net			
Forward-in-time	Decision Tree			
Forward-in-time	AdaBoost			
Forward-in-time	XGBoost			
Forward-in-time	Random Forest			
Forward-in-time	FC NNet			

Table 1: Performance results of models across splitting regimes