Creating a Complex Reforestation Strategy
By Zachary Stein

Intro:
I have been working during the fall semester of 2018 at the World Resources Institute (WRI) and the Climate Institute (CI) in an internship that is analyzing carbon taxes in Latin America. They plan to study the various studies in order to make a policy recommendation to Peru. Peru does not yet have carbon taxes, but they are planning to implement them, and they want to know what the best strategy will be for accomplishing this.

The purpose of these carbon taxes is to reduce the amount of carbon dioxide released into the atmosphere. The way they are meant to accomplish this is by encouraging the more judicious use of fossil fuels and a switch to renewable energy resources. Another way to reduce the amount of carbon dioxide that a nation emits is to absorb carbon dioxide from the atmosphere. One of the easiest ways to do this is through reforestation. Not only will the new trees absorb carbon dioxide, but they will also have other positive environmental effects that can benefit an area.

While there have been many studies on reforestation, there has not been much from a data science perspective. For that reason, I tried to make a method for determining what could be the best way to reforest an area.

Background:
Currently, climate change is an increasingly serious issue. Human beings have released multiple types of greenhouse gases (GHGs) in the atmosphere (e.g. Carbon dioxide, methane, etc.). Carbon dioxide, while not the most effective GHG, is released in the greatest volume. This is due to various types of fossil fuels being burned that release carbon dioxide as a waste gas: Gasoline, diesel, coal, kerosene, LPG, and natural gas account for almost all of this. These fossil fuels are consumed to generate power, for heating, for transportation, and in certain industrial processes (e.g. Ammonia production for making nitrogen fertilizers).

The amount of carbon dioxide that has been released has greatly increased over time. Over 85% of all carbon dioxide that has been released by human beings has been released since 1950, and 30% of carbon dioxide has been released since 2000.[1] The more carbon dioxide that is released, the greater the amount of climate change that human beings are seeing (Rising sea levels, drought, changes in local and global temperatures, etc.). Various efforts have been taken to address this, mainly through climate change agreements, such as the Kyoto Protocol and the Paris Climate Change agreement. The basis of these agreements is a reduction in the amount of carbon dioxide released by the signatories.[2]

Many countries from Latin America have participated in various climate change projects and/or agreements. As part of their participation, they have implemented carbon taxes on various sources of carbon dioxide. This varies by country, but it might manifest as taxes on all types of fossil fuels (Colombia),[3] or taxes on fossil fuels used by industries and powerplants if they exceed a certain amount of power (Chile).[4]

The goal of these taxes is to reduce the amount of carbon dioxide released by the country, by encouraging people and organizations to reduce the amount of fossil fuels that they require, as well as encouraging the greater use of renewable resources (Many of these countries also import fossil fuels). However, another way to reduce carbon dioxide emissions is to capture carbon from the atmosphere.

Plants absorb carbon dioxide from the atmosphere and converting it into the various organic molecules that make up their structure. For this reason, deforestation has been a major source of carbon dioxide emissions into the atmosphere, and reversing the process through reforestation is a growing topic of discussion for how to reduce the effects of climate change. Additionally, there are multiple other benefits that come from reforestation. Because trees absorb water from the ground and respire it into the surrounding air, they are a source of precipitation in an area. They also hold onto soil with their roots, which translates into flooding, soil erosion, river damage, and drought when an area is deforested. By reforesting, the environmental health of a large area can be restored, with numerous cost savings and benefits to people living in an area.

Reforestation used to consist of planting large numbers of one species of tree in an area.[5] Naturally, this was a great reduction from the ecological variety that used to occur in an area, so now there is more of a focus on planting multiple species throughout an area in order to create a more varied environment for recovery.[6] However, there has not been extensive publication in the literature on analysis methods for determining the best way to reforest an area. The WRI and CI are interested in possible ways for accomplishing this, and for that reason I have used my data science training in order to create an experimental way for carrying this out.

Goals:

There are multiple goals that I want to carry out:

1. What mix of species is best for creating a mixed ecosystem? Is there a range for each species that works well?
2. We have to keep in mind that with an initial seeding or planting, some species might end up dying while others will thrive. This will result in a change to the mix of the ecosystems. If this happens, will the reforested area become a problem area?
3. Can we also determine which species are important for the success of a reforestation effort?

Methods:

The initial data was acquired from WRI. They have looked at reforestation in Colombia and Mexico. This data describes the species that are used, and the amount of carbon they take in for the first 20 years in terms of tC/ha/y (Their carbon absorption rate can change after 20 years). This data was acquired from one of the previous interns, who put it together from the Global Emissions and Removals Databases.

With the species of trees and the amount of carbon dioxide that they can absorb, I then had to determine how I would actually go about analyzing the best way to reforest a data. There was not a lot of data on this, so I had to think about what would be the most useful for researchers. This will be done with the following parameters: Min_Per (Minimum Percentage), Max_Per (Maximum Percentage), Low_Cutoff, and High_Cutoff. Low_Cutoff and High_Cutoff are the minimum and maximum amounts of percentage of land that can be devoted to each species of tree. Min_Per and Max_Per are the min and maximum percentages of land for each species of tree that will be used by the study.

Because the reforested area might be near the coast, in an area with lakes, and/or an area with rivers, I also included percentage of the area that has salt water (SWater) and fresh water (FWater). Because mangrove trees and shrubs grow in salty water, I assume that the percentage of land that is exposed to salt water has mangrove trees and shrubs in even proportions (For example, if 20% of the land is exposed to or has salt water, then 10% will be mangrove trees and 10% will be shrubs). Because mangrove trees and shrubs are dependent on salt water, and other species of trees will not grow well on or near salty water, mangrove trees are treated as if they are stable over time.

I used various data analysis methods that required the program R. There are two parts to it: 'Pre-Processing' and 'Analyze the Data':

Pre-Processing

1. First, I set the directory for the where the data is located and call on the libraries that will be needed, and the number of random datasets (nrun) is set. Then the data is read in.
2. Next, I create two empty data sets for handling the data in the program after it has been run. The first is 'RanData' and the second is 'CalcData' (What these are used for will be explained a bit later in the paper).
3. This is followed by calculating the percentage of land taken up by mangrove trees and shrubs. Half of the salt water land is taken up by mangrove trees, and the other half of the salt water land is taken up by mangrove shrubs.
4. The amount of leftover land is calculated by subtracting the percentage of saltwater and freshwater from 1.
5. Then, a for loop is run in order to finish prepping the data (The length of the for loop is set at nrun in step 1). In the for loop, a random percentage is chosen using the minimum and maximum percentages for each tree. Then, this is adjusted based on the amount of land that is left (e.g. If there is no fresh water rivers or lakes, as well as no salt water areas, then if the percentage is 23%, it stays 23%. If there is an amount of salt and fresh water equal to 25%, then only 75% of the land can grow trees, and the random percentages are adjusted accordingly). Then, I calculate if the amount of land occupied by a tree species is outside of the ideal range. If it is outside of the ideal range, then it is assigned a value of 0, otherwise it is assigned a value of 1.
6. All of the 0 and 1 values that were generated are added together. If 2 or more are 0, then we say that the forest will be non-viable, and assign a value of 0 to a new variable called 'Viable.' Otherwise, we assign a value of 1 to Viable.
7. I then put the data on the percentage of land occupied by each tree species in the RanData data set. The data on Viable and whether or not the tree species is outside of the viable range is placed into CalcData.
8. Finally, the names of each column are added to each dataset.

Analyze the Data

1. Now, the data can be analyzed. It is set into train and test sets.

2. Then, I call on "Jags-Ydich-XmetMulti-Mlogistic.R." This also requires “DBDA2E-utilities.R” to be in the same directory in order for it to run properly.
3. This is then run through a variety of Bayesian functions. They calculate factors such as the intercept and the ROC values.
4. Then, a variety of graphs are produced using mcmc.
5. Also, the amount of carbon dioxide that will be captured over a 20 year period is calculated for each randomization. So the area occupied by each species of tree is multiplied by the carbon absorption rate, multiplied by 20, and then added together for all species.
6. Together, all of this displays plots on which species of trees are the most important for whether or not a reforested area will be viable in the long term, using intercept and ROC plots.

In order to run this, I depend on two different techniques: Markov Chain Monte Carlo (MCMC) and JAGS. JAGS automatically builds MCMC for complex hierarchical models. MCMC samples probability distribution in order to make an inference. By randomly sampling random values from a distribution with a lot of variables, it is able to make inferences about their importance to the system.[7] In this way, I can see which range of trees are best suited to an environment and how well they interact with each other.

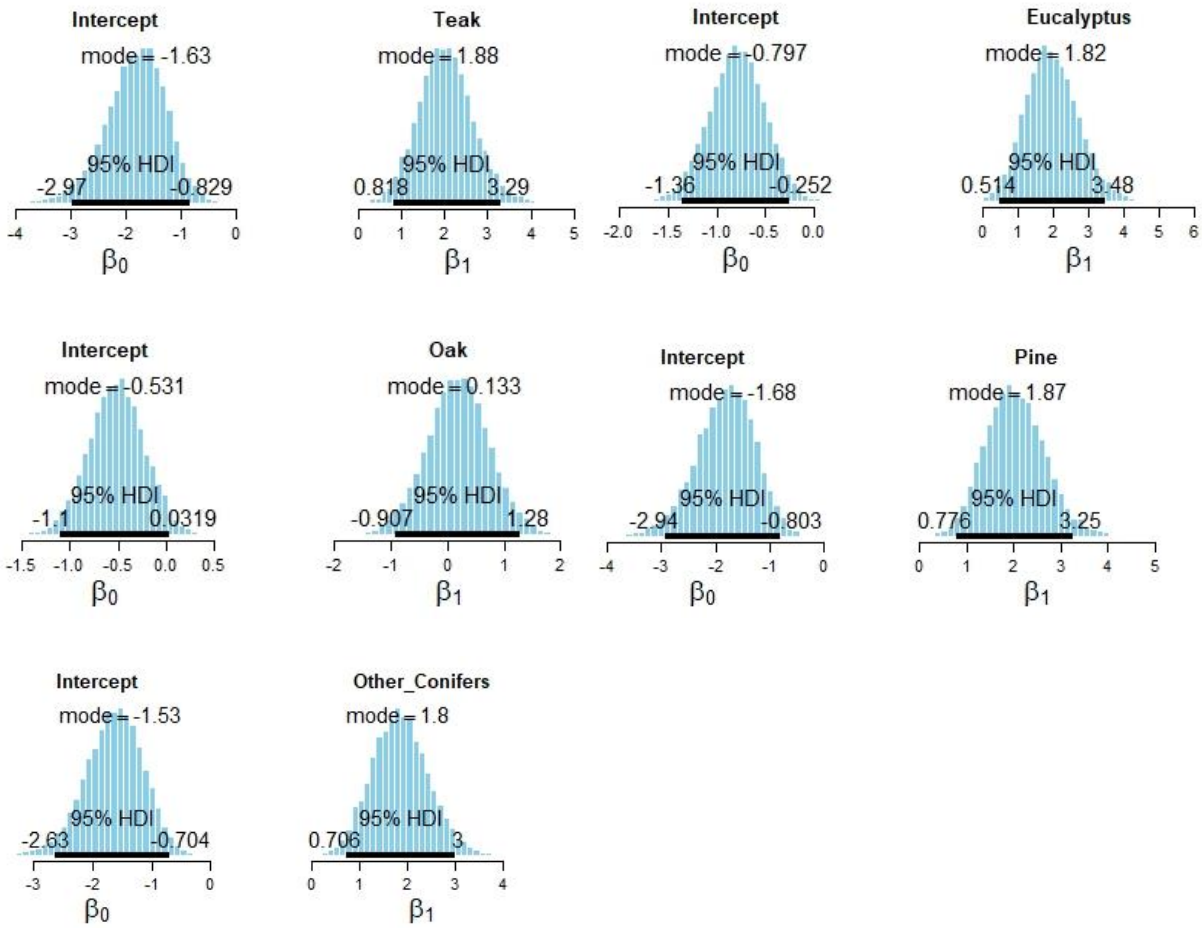
Main:
For this, I want to determine which mixture of species is best for an area. Ecologists who study a particular environment can make estimates for which species of trees are best for a particular area, and what is the range for the percentage of the space they should occupy. By taking random values for each species and comparing it to their ideal range, I can test which species is most important for an environment, and which combinations are most ideal for the environment. In addition, the initial planting of trees can change over time, as tree species can unexpectedly flourish or die off. By estimating which combinations are successful or not, I can make predictions for how a reforested area will do over time.

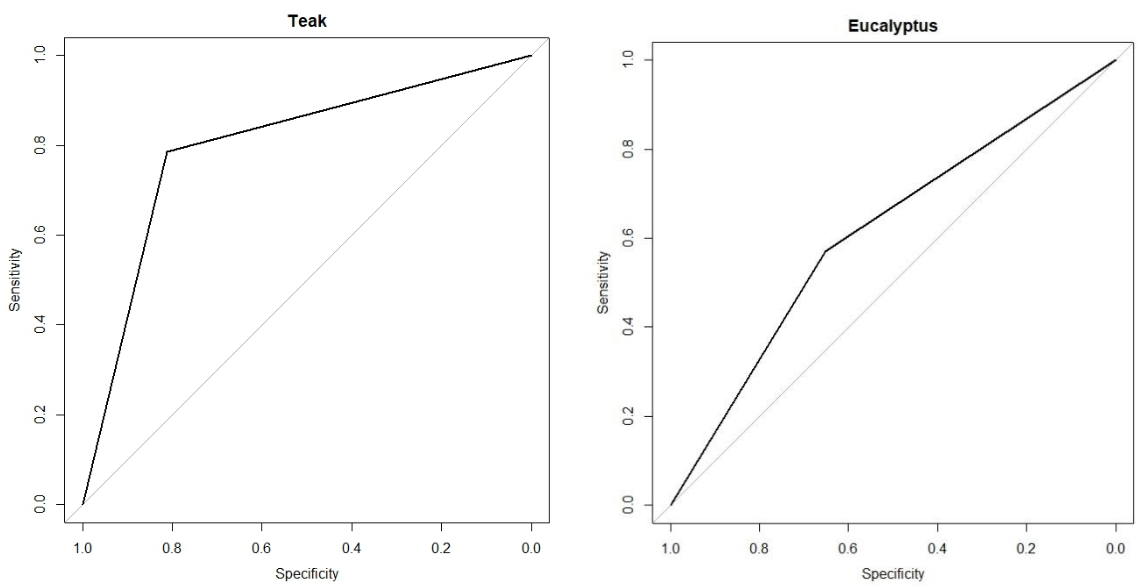
I carried out three different test data sets in order check on how the program performed and what were the outputs. All of them were for 5,000 ha of land, and the carbon was calculated for a 20-year period.

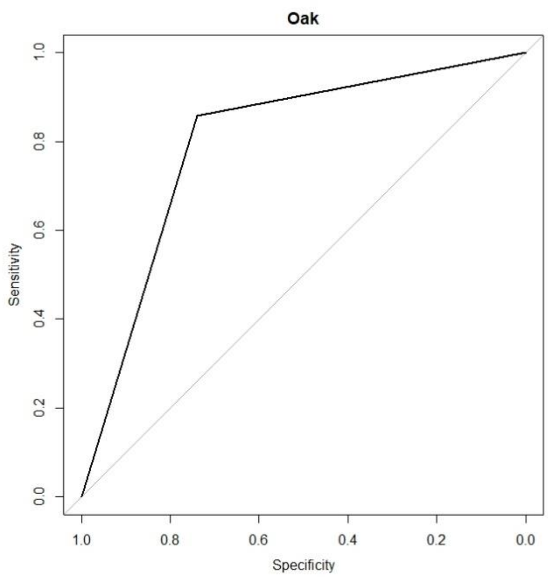
Test 1

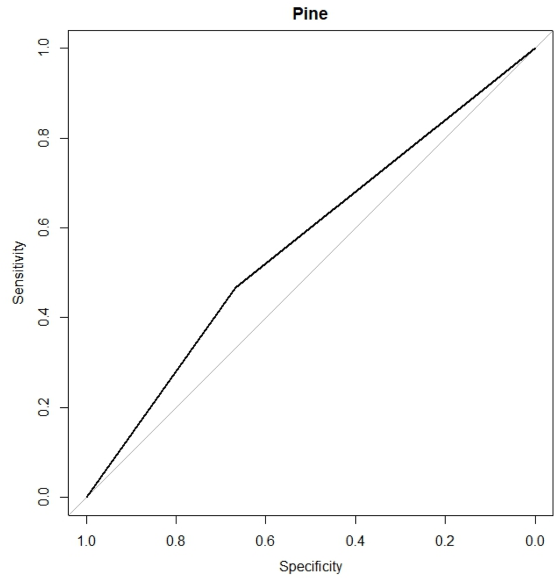
Input	Teak	Eucalyptus	Oak	Pine	Other_Conifers	SWater	FWater
Carbon_Absorption	8.4	11.1	5	5	5.9	0.11	0.05
Min_Per	0.01	0.01	0.01	0.01	0.01		
Max_Per	0.2	0.2	0.2	0.2	0.2		
Low_Cutoff	0.02	0.05	0.03	0.07	0.07		
High_Cutoff	0.19	0.11	0.12	0.18	0.18		

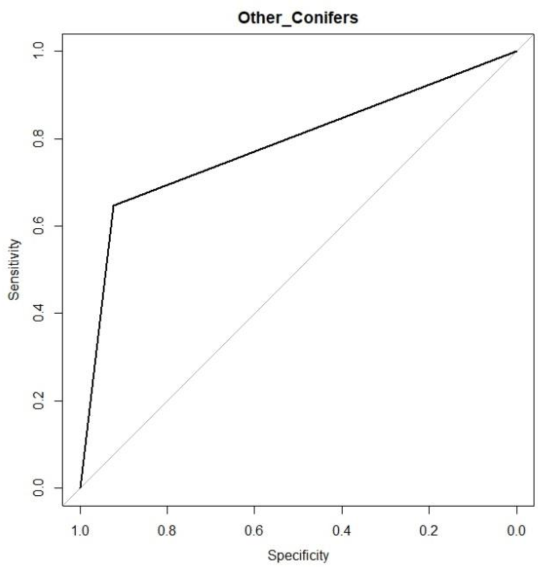
	Teak	Eucalyptus	Oak	Pine	Other_Conifers
Hit Rates	0.53	0.63	0.6	0.567	0.767


Intercept
mode = -1.63
95% HDI
-2.97
-0.829
β_0
Teak
mode = 1.88
95% HDI
0.818
3.29
β_1
Intercept
mode = -0.797
95% HDI
-1.36
-0.252
β_0
Eucalyptus
mode = 1.82
95% HDI
0.514
3.48
β_1
Intercept
mode = -0.531
95% HDI
-1.1
0.0319
β_0
Oak
mode = 0.133
95% HDI
-0.907
1.28
β_1
Intercept
mode = -1.68
95% HDI
-2.94
-0.803
β_0
Pine
mode = 1.87
95% HDI
0.776
3.25
β_1
Intercept
mode = -1.53
95% HDI
-2.63
-0.704
β_0
Other_Conifers
mode = 1.8
95% HDI
0.706
3
β_1


Teak
Sensitivity
Specificity
Eucalyptus
Sensitivity
Specificity

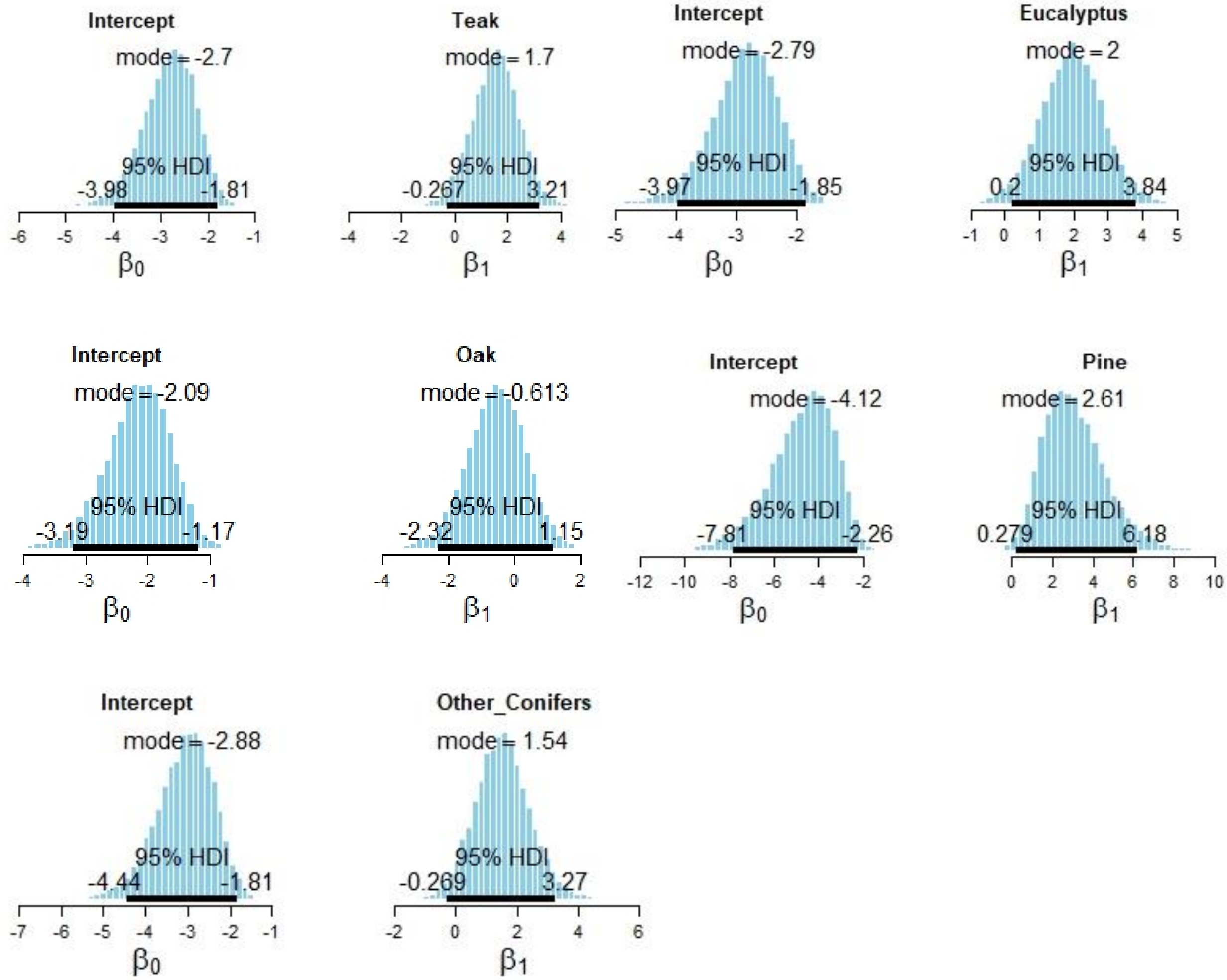




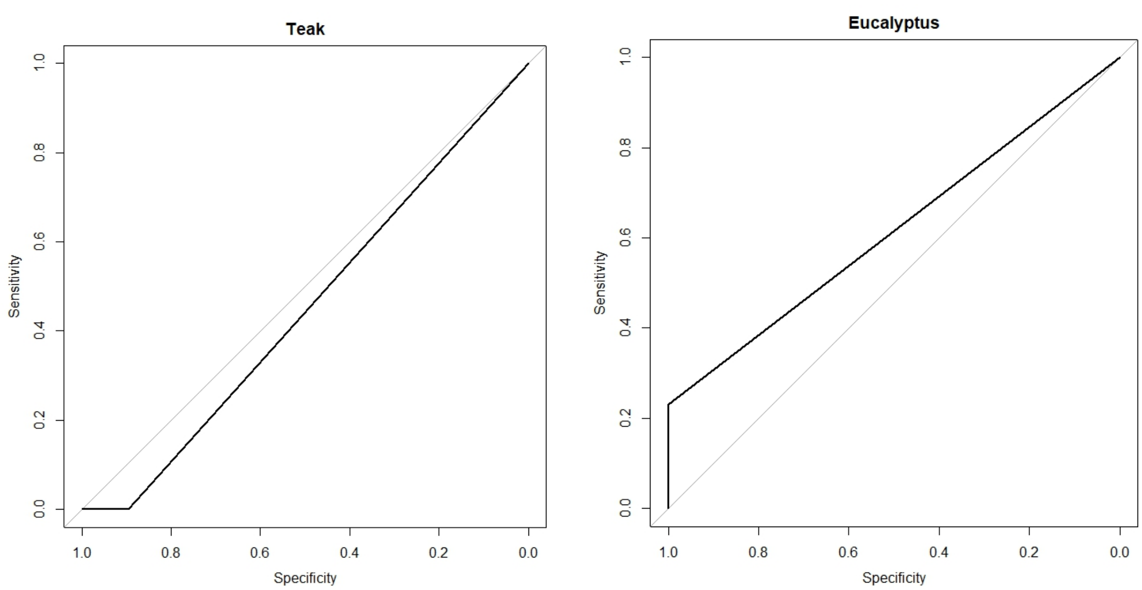


Test 2

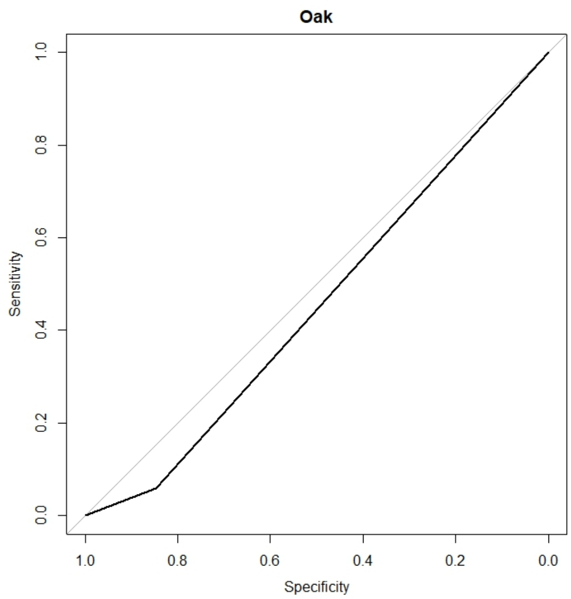
	Teak	Eucalyptus	Oak	Pine	Other_Conifers	SWater	FWater
Carbon_Absorption	8.4	11.1	5	5	5.9	0.11	0.05
Min_Per	0.01	0.01	0.01	0.01	0.01		
Max_Per	0.2	0.2	0.2	0.2	0.2		
Low_Cutoff	0.015	0.07	0.02	0.05	0.03		
High_Cutoff	0.09	0.11	0.15	0.17	0.13		

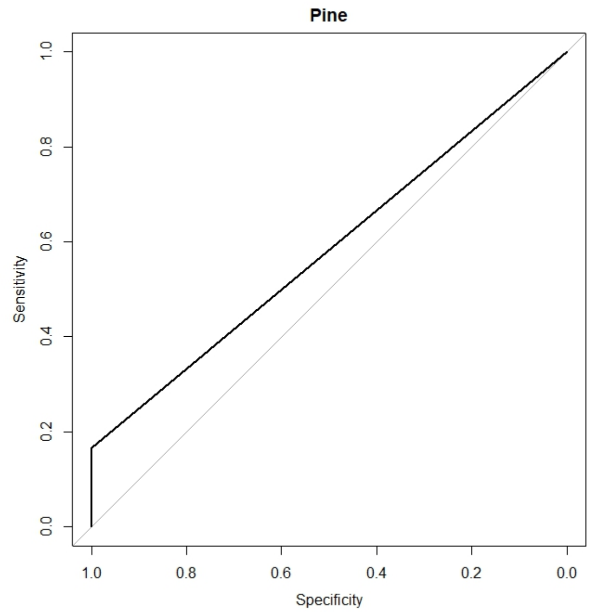


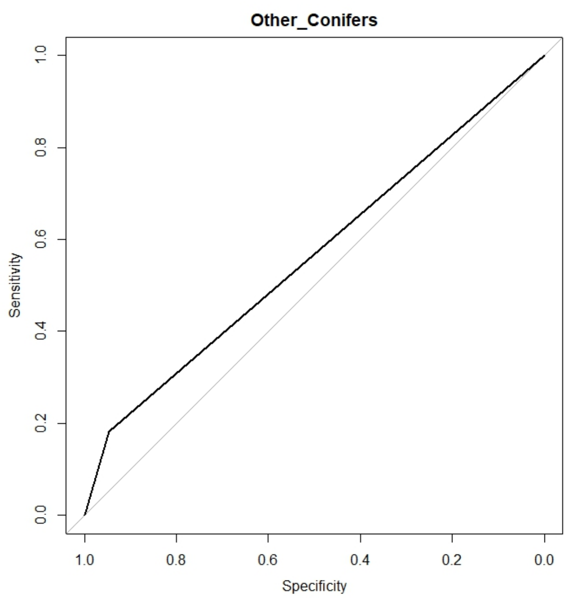
Intercept
mode = -2.7
95% HDI
-3.98
-1.81
β0
Teak
mode = 1.7
95% HDI
-0.267
3.21
β1
Intercept
mode = -2.79
95% HDI
-3.97
-1.85
β0
Eucalyptus
mode = 2
95% HDI
0.2
3.84
β1
Intercept
mode = -2.09
95% HDI
-3.19
-1.17
β0
Oak
mode = -0.613
95% HDI
-2.32
1.15
β1
Intercept
mode = -4.12
95% HDI
-7.81
-2.26
β0
Pine
mode = 2.61
95% HDI
0.279
6.18
β1
Intercept
mode = -2.88
95% HDI
-4.44
-1.81
β0
Other_Conifers
mode = 1.54
95% HDI
-0.269
3.27
β1



Teak
Sensitivity
Specificity
Eucalyptus
Sensitivity
Specificity

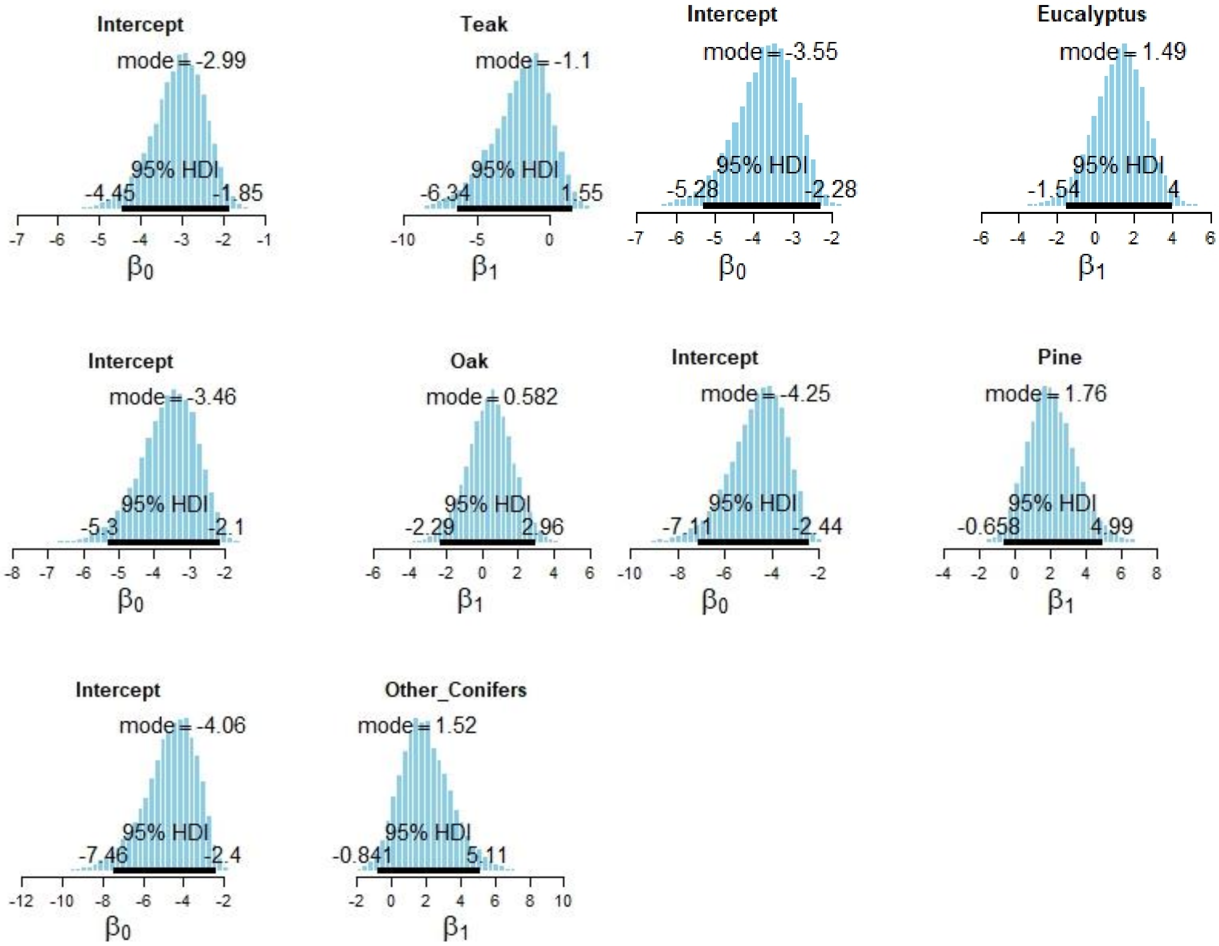


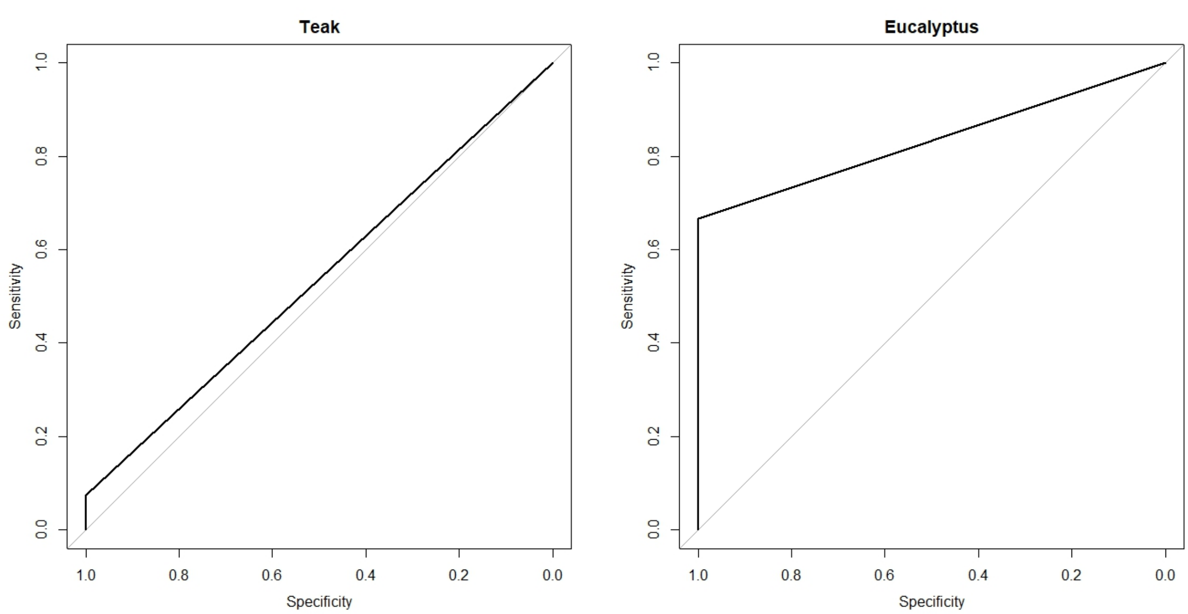


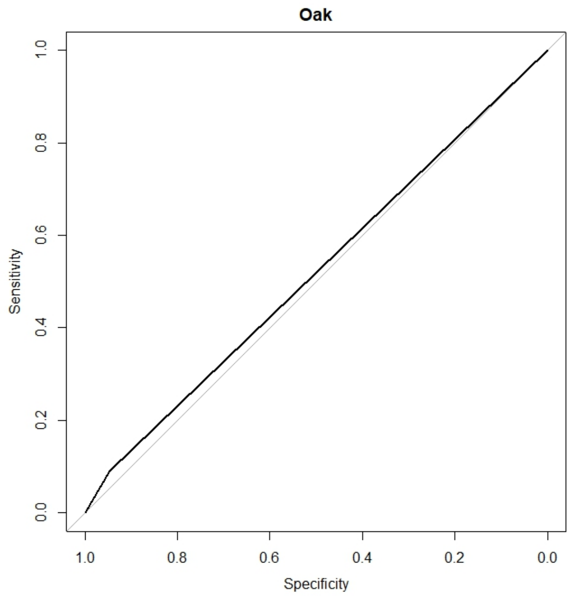


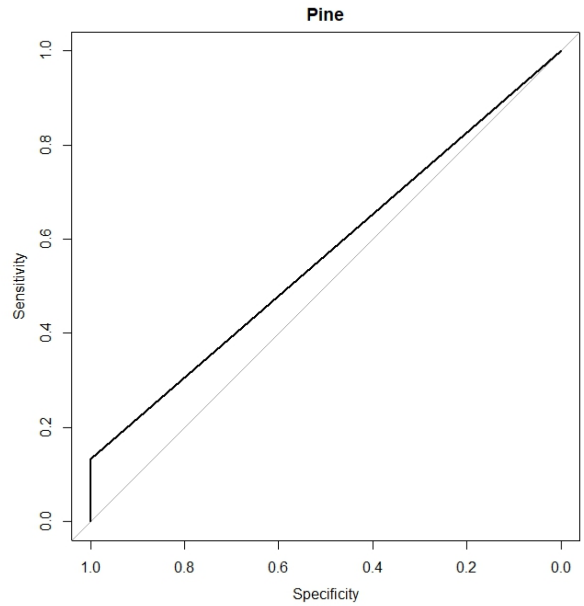
Test 3

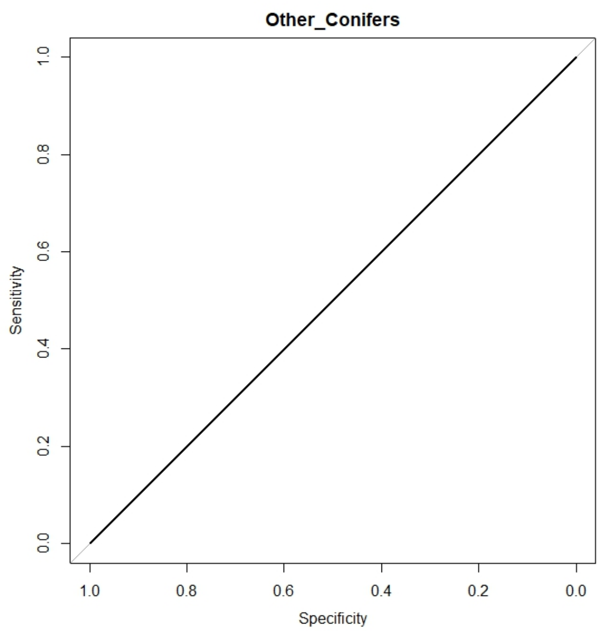
	Teak	Eucalyptus	Oak	Pine	Other_Conifers	SWater	FWater
Carbon_Absorption	8.4	11.1	5	5	5.9	0	0.09
Min_Per	0.01	0.01	0.01	0.01	0.01		
Max_Per	0.2	0.2	0.2	0.2	0.2		
Low_Cutoff	0.03	0.06	0.05	0.035	0.055		
High_Cutoff	0.13	0.1	0.15	0.139	0.19		


Intercept
mode = -2.99
95% HDI
-4.45
-1.85
β0
Teak
mode = -1.1
95% HDI
-6.34
1.55
β1
Intercept
mode = -3.55
95% HDI
-5.28
-2.28
β0
Eucalyptus
mode = 1.49
95% HDI
-1.54
4
β1
Intercept
mode = -3.46
95% HDI
-5.3
-2.1
β0
Oak
mode = 0.582
95% HDI
-2.29
2.96
β1
Intercept
mode = -4.25
95% HDI
-7.11
-2.44
β0
Pine
mode = 1.76
95% HDI
-0.658
4.99
β1
Intercept
mode = -4.06
95% HDI
-7.46
-2.4
β0
Other_Conifers
mode = 1.52
95% HDI
-0.841
5.11
β1


Teak
Sensitivity
Specificity
Eucalyptus
Sensitivity
Specificity







All of these intercept and ROC graphs show which tree species will be viable for each type of environment. In some environments, certain species of trees will be more important than others.

Comparison of Output Data

The following is a comparison of the output data files. The first one is the percentage of the area covered by each species of tree, plus the total amount of carbon absorbed by the forest. The second one is whether each species will be viable for that percentage of area, and whether the forest as a whole will be viable.

A	Teak	Eucalyptus	Oak	Pine	Other_Conifers	Mangrove_Tree	Mangrove_Shrub	Carbon_Absorb
1	0.135709182	0.035448886	0.149480895	0.213935542	0.305425494	0.055	0.055	105.5806474
2	0.183641195	0.056814445	0.191233028	0.256804762	0.151506569	0.055	0.055	108.6760818
3	0.179521118	0.157899653	0.143649934	0.185904165	0.17302513	0.055	0.055	121.1156461
4	0.112632239	0.038283334	0.299301454	0.182226756	0.207556217	0.055	0.055	102.5955709
5	0.17544669	0.234988646	0.239275623	0.119334688	0.070954353	0.055	0.055	128.4061681
6	0.124644889	0.105841083	0.260333217	0.177429168	0.171751644	0.055	0.055	111.0099941
7	0.159397571	0.07260497	0.209607663	0.1854203	0.212969496	0.055	0.055	110.0602921
8	0.036064917	0.164231074	0.220427739	0.207693958	0.211582311	0.055	0.055	112.827087
9	0.217821131	0.185000965	0.184159942	0.057184494	0.195833469	0.055	0.055	127.436957
10	0.166279174	0.379910981	0.149582192	0.099866476	0.044361177	0.055	0.055	144.9846247
11	0.195236737	0.304893874	0.113313676	0.033223785	0.193331928	0.055	0.055	140.4831254
12	0.038858571	0.169457913	0.125955986	0.199484075	0.306243455	0.055	0.055	115.3586304
13	0.224870966	0.211309206	0.119123253	0.101021764	0.18367481	0.055	0.055	130.9070954
14	0.213923694	0.196939865	0.108251603	0.233630355	0.087254483	0.055	0.055	126.6740554
15	0.018185505	0.178987605	0.248075508	0.202748266	0.192003116	0.055	0.055	113.0591583
16	0.024864064	0.253131681	0.161398872	0.249942669	0.150662713	0.055	0.055	121.8147503
17	0.201968891	0.184623444	0.187167889	0.167232672	0.099007104	0.055	0.055	124.5700727
18	0.050383554	0.311098141	0.219771485	0.143032566	0.115714255	0.055	0.055	129.9929114
19	0.181750735	0.184743164	0.049860919	0.279752456	0.143892726	0.055	0.055	124.0177851
20	0.300713724	0.123008671	0.13389364	0.249509514	0.032874451	0.055	0.055	122.5773313
21	0.339169443	0.099057726	0.056581013	0.048160645	0.297031173	0.055	0.055	127.0251258

A	Teak	Eucalyptus	Oak	Pine	Other_Conifers	Viable
1	1	0	0	0	0	0
2	1	1	0	0	1	1
3	1	0	0	0	1	0
4	1	0	0	0	0	0
5	1	0	0	1	1	1
6	1	1	0	1	1	1
7	1	1	0	0	0	0
8	1	0	0	0	0	0
9	0	0	0	1	0	0
10	1	0	0	1	1	1
11	0	0	1	1	0	0
12	1	0	0	0	0	0
13	0	0	1	1	0	0
14	0	0	1	0	1	0
15	0	0	0	0	0	0
16	1	0	0	0	0	0
17	0	0	0	1	1	0
18	1	0	0	0	0	0
19	1	0	1	0	1	1
20	0	0	0	0	1	0
21	0	1	1	1	0	1

Conclusions:

The results that I get so far are promising. For different environments, I am seeing that certain trees are a poor fit (Such as teak and oak for the second set of test data) or are a good fit (Such as eucalyptus for the third set of test data) for different types of environments. I can also see what mix of species will or will not be viable for the environment, as well as whether or not this will create a viable mixed forest. Lastly, I can also calculate the amount of carbon dioxide that it will absorb over a 20-year period. If I can continue to develop this, then I think that I can create a viable data analysis tool for people who are looking for better ways to deforest an area.

One way this can be improved on in the future is to use the results to change the acceptable range for each species of tree. The results show which species of tree is best suited to the conditions. The goal is to make sure all of the species are best suited to the local conditions. For species that have a good ROC curve and intercept value, we can leave them alone for the time being. For species that have a bad ROC curve and intercept value, we can change the range of the cutoff and see if that will improve the results. In this manner, the best range of values for each species of tree can be determined.

References:

1. Hignett K. (2018, May). “Here’s how much carbon dioxide emissions have increased since the industrial age.” *Newsweek (Tech & Science).*
2. Harrington R and Gould S. (2017, June). “The US will join Syria and Nicaragua as the only nations that aren’t part of the Paris agreement.” *Business Insider*.
3. “Colombia: An Emissions Trading Case Study”. *Carbon Trust, EDF, and IETA.* January 2018.
4. “Carbon Tax Guide: A Handbook for Policy Makers (Appendix: Carbon Tax Case Studies).” *World Bank Group*. March 2017.
5. Moss C. (2015, May) “Smart reforestation must go beyond carbon: expert.” *Forest News*.
6. Locatelli B., Catterall C., et. Al. (2015 July). “Tropical reforestation and climate change: beyond carbon.” *Reforestation Ecology, 23(4), 337-343.*
7. Kruschke J. (2010) *Doing Bayesian Data Analysis: A tutorial with R and BUGS*.