Final Project Paper

Geog 560

Jake Kruse

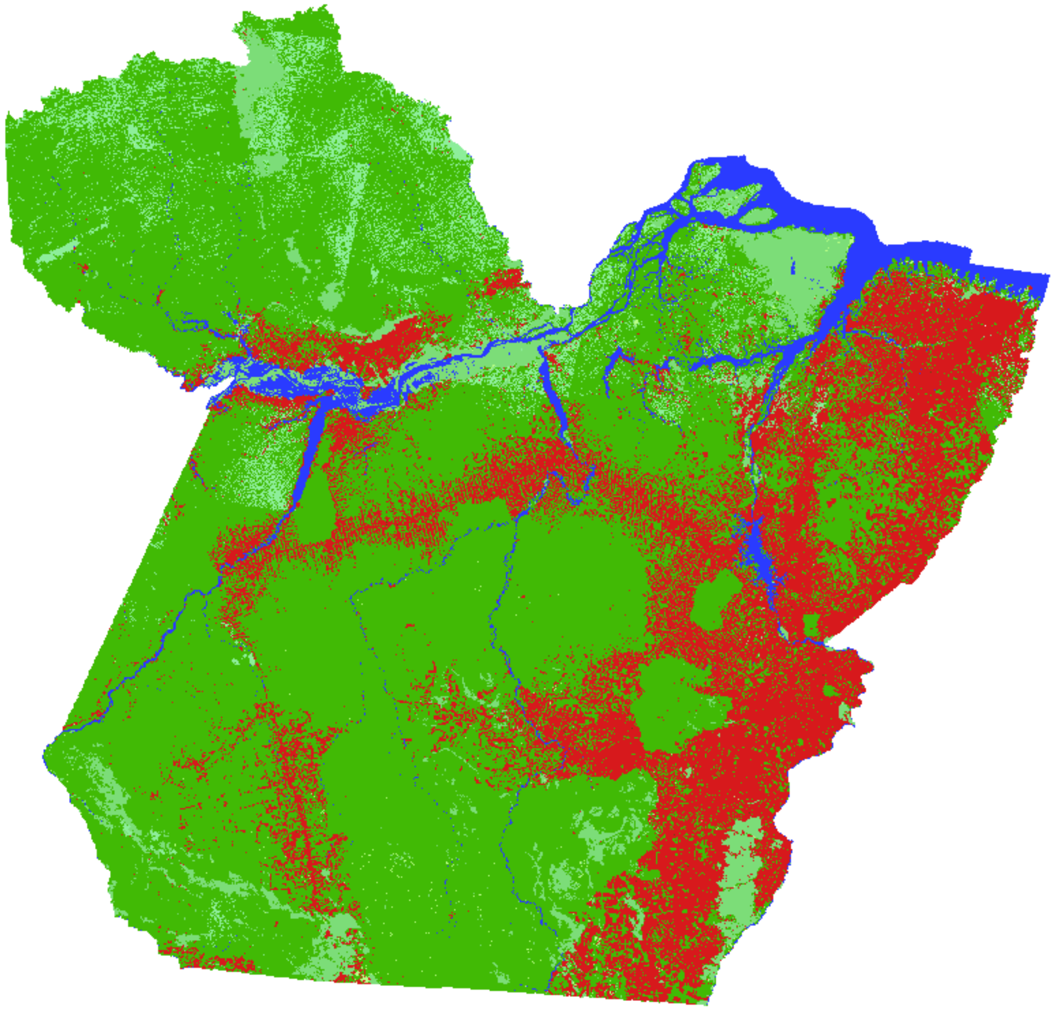
## Introduction

Deforestation in the Brazilian Amazon has occurred over centuries, but it has dramatically increased in the last decade (Fearnside and Barbosa 2004). The state of Pará is one of the most problematic states in Brazil in terms of deforestation, and therefore assessing deforestation risk in Pará is crucial to succeed in preserving forest. Between 2001 and 2018, Pará lost an area of 79,000 km2 which is 37% of the area deforested in the Brazil Amazon for the same period (INPE, 2018). To better understand the variables of deforestation and to develop a model that predicts deforestation in that region, random forest classification was performed on a dataset for the state of Para.

## Methods

PRODES is a dataset produced by the Global Forest Watch; it details the extent of deforestation, reforestation, and other broad land-class categories (Image 1). Rural land use, population, human develop index (HDI), road and dam infrastructure, and river data was obtained from the SICAR dataset, which is maintained by the government of Brazil. Using ArcMap, land use and classification, population, and HDI layers for 2017 were converted to raster files. Euclidean distance to rivers, roads, and dams was stored in a unique raster layer for each feature, respectively. Finally, 500 and 1000 random points were generated in ArcMap for the training and testing data sets, respectively. All raster layers were stacked in R, and the random points were used to extract the raster layer values at each training and testing location. The land use classification model was a random forest algorithm with 500 trees, and two variables were tried at each split.

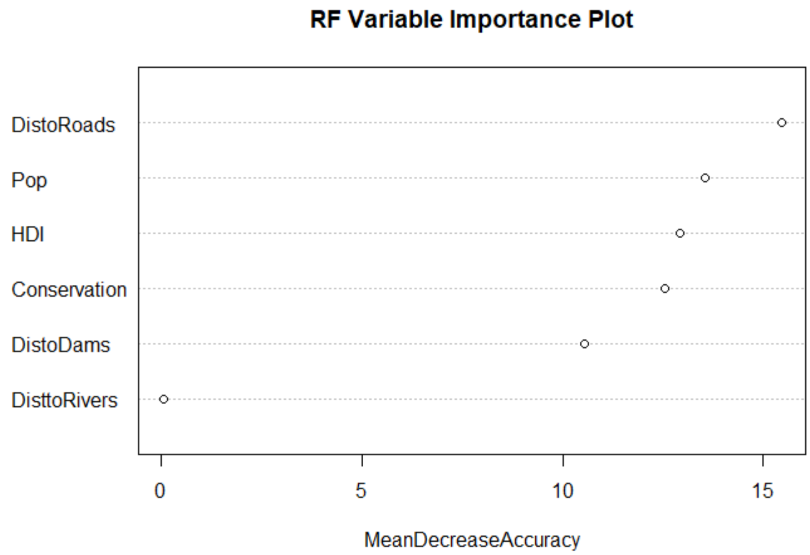
**Image 1**. State of Para, Brazil. Deforestation, river and vegetation are in red, blue, and shades of green, respectively.



**Results**

With the random forest model overall accuracy was 74%, and incorrect classification was 26%. The following Variable Importance Plot shows the variables from least important to most important (higher MeanDecreaseAccuracy is greater importance).

**Image 2**. Random Forest Variable Importance Plot.




For the number of variables tried at each node, two was the number that most reduced error. The error rate for each class was as follows (1, 2, 3, 4, 5, 6, and 7 were Deforestation, New Growth, New Regrowth, New Regrowth Two, New Forest, River, and Forest categories, respectively)

**Image 3**. OOB and class error for random forest model.

OOB estimate of error rate: 24.8%

Confusion matrix:

| | 1 | 3 | 5 | 6 | 7 | class.error |
|---|---|---|---|---|---|---|
| 1 | 60 | 0 | 0 | 1 | 50 | 0.45945946 |
| 3 | 0 | 0 | 0 | 0 | 9 | 1.00000000 |
| 5 | 4 | 0 | 3 | 0 | 13 | 0.85000000 |
| 6 | 5 | 0 | 0 | 2 | 8 | 0.86666667 |
| 7 | 31 | 0 | 1 | 0 | 305 | 0.09495549 |

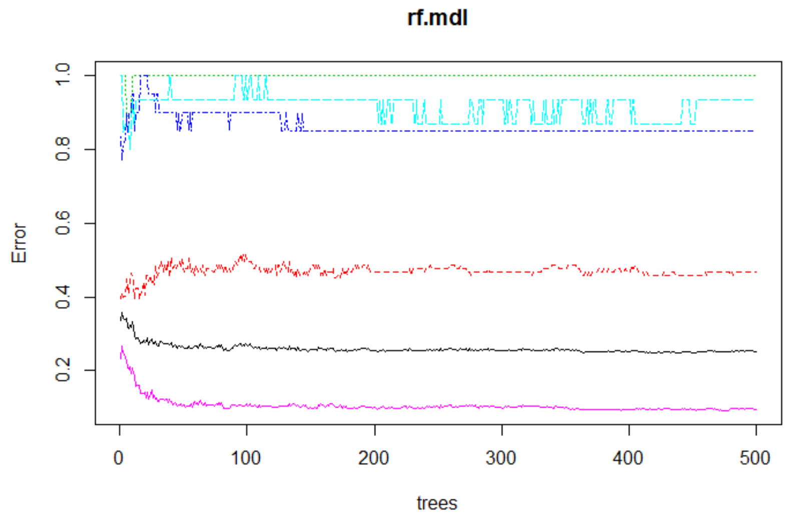
The following table mirrors the results of the variable importance plot, but also shows the related MeanDecreaseGini. In particular, Distance to Dams played a more important role in decreasing node impurity than suggested by the variable importance plot alone.

**Image 3**. Mean Decrease in Gini for random forest predictors.

| | 1 | 3 | 5 | 6 | 7 | MeanDecreaseAccuracy | MeanDecreaseGini |
|---|---|---|---|---|---|---|---|
| Conservation | 31.463493 | 8.5624445 | 7.063604 | 4.943061 | 8.931247 | 25.941838 | 18.72294 |
| DistoDams | 8.490599 | 3.6964919 | 7.874043 | 7.688526 | 17.212361 | 21.627525 | 43.08646 |
| DisttoRivers | 2.112955 | 3.6001237 | 6.103307 | -1.154069 | 1.229954 | 3.502156 | 37.96136 |
| HDI | 12.481033 | 4.3412165 | 11.921208 | 7.891663 | 24.414232 | 28.510208 | 35.43684 |
| Pop | 15.206022 | 4.4846147 | 5.451287 | 5.049083 | 22.027004 | 26.509618 | 38.67089 |
| DistoRoads | 35.785123 | -0.1467599 | 5.565957 | 4.038259 | 14.989321 | 32.326826 | 56.85889 |

Based on the following graph of the model error as number of trees increased, error seems to hold steady after approximately 100 trees.

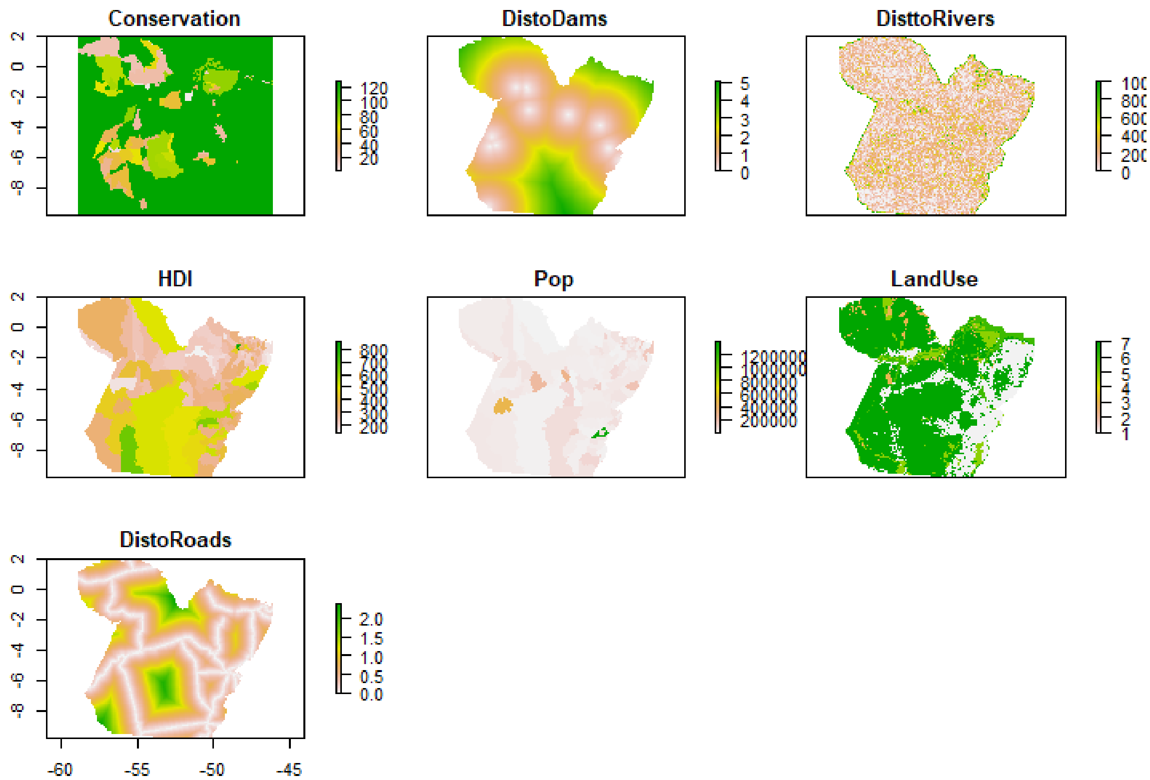
**Image 4**. Decrease in model error with number of trees.




**Discussion**

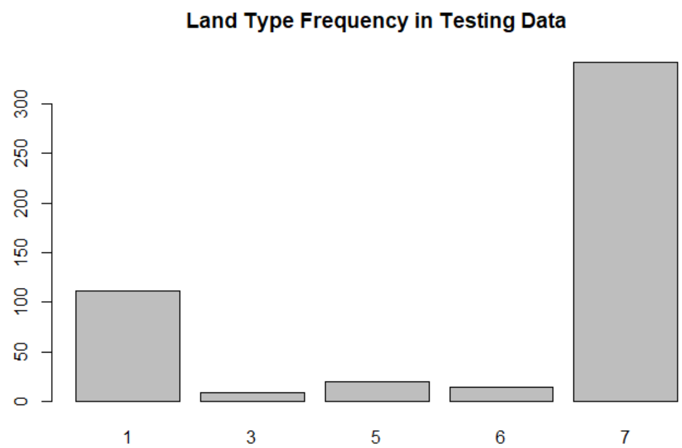
At first it was unclear why randomForest was not working on the raster stack, and the head, summary, and plot functions were used to examine the data. The plot function (Image 5) showed that indeed all the data sets were converted to raster and stacked in a way that conserved the original range of information for each layer. Although the land classes were stored as integers, they needed to be converted to characters for randomForest classification to work in R.

**Image 5**. Plot function of raster stack.




One of the problems was 500 and 1000 random sample points did not seem to capture enough points in all the classes, as they were predominately the land type (LandUse) classes Deforestation and Forest (Image 6). Increasing the number of points would have helped, but the available computers did not have the processing power to handle more.

**Image 6.** Bar plot for counts in each class sampled for testing dataset. 1, 3, 5, 6, and 7 were Deforestation, New Regrowth, New Forest, River, and Forest categories, respectively. Two classes were not captured in the sample, but they were relatively small classes.




Distance to rivers was the least important variable in the random forest model, but that clashes with domain knowledge. Instead of using all rivers, using only rivers large enough to support industrial shipping should have been included. It is also interesting to note that conservation areas did not have more weight as a predictor, suggesting that conservation status in an area does not mean that no deforestation occurs.

In summary, 74% accuracy is good accuracy for a map produced in two weeks from publicly available data. All variables were shown to be important expect Distance to Rivers, but that would change if only industrially usable rivers had been used as predictors instead of all rivers. Future efforts would include increasing the training and testing set sizes, and using a more systematic coding approach to calibrate the model parameters.

**References**

Aide, T. M., Clark, M. L., Grau, H. R., López-Carr, D. , Levy, M. A., Redo, D. , Bonilla-Moheno, M. ,

Riner, G. , Andrade-Núñez, M. J. and Muñiz, M. (2013), Deforestation and Reforestation of Latin America and the Caribbean (2001–2010). Biotropica, 45: 262-271. doi:10.1111/j.1744-7429.2012.00908.x CMMC (Comissão Mista Permanente sobre Mudanças Climáticas) (2013). Legislação Brasileira Sobre Mudanças Climáticas.

Fearnside, P. (2017). Deforestation of the Brazilian Amazon. Oxford Research Encyclopedia of Environmental Science. doi:10.1093/acrefore/9780199389414.013.102

National Institute of Space Research (INPE). "PRODES deforestation." Accessed through Global Forest Watch on 5/1/2019. www.globalforestwatch.org

Cadastro Ambiental Rural (CAR). "SICAR." Accessed through Deutsche Gesellschaft fur Internationale Zusammenarbeit (GIZ) on 5/1/2019. www.giz.de/en/worldwide/34060.html