The University of North Carolina at Chapel Hill

COMP 590-800
Data Science for Earth
Final Paper

Authors
Jonah Soberano
Vraj Patel
Varun Tanna

Supervisors
Prof. Brian White
Carly Richardson
Izzy Hinks

April 30th, 2020

Classifying Forest Degradation in the Amazon Rainforest

Jonah Soberano soberano@live.unc.edu
Vraj Patel patelvap@live.unc.edu
Varun Tanna varunt@live.unc.edu

Abstract

This is a final course project for the special topics course titled "Data Science for Earth" at the University of North Carolina at Chapel Hill, taught by Associate Professor in the Marine Science's Department, Brian White. Using Google's Earth Engine (GEE) editor and its public libraries of Landsat imagery data, this paper seeks to address issues with current classification models of forest growth and forest loss in the Amazon rainforest. With GEE's in-editor supervised classification models, we attempt to create a model that improves upon and gives new insight into the growing issues of forest degradation in the Amazon rainforest. We recognize that our final analysis has room for improvement, but due to time and academic constraints of the time (restricted collaboration due to the 2020 COVID-19 pandemic), this paper serves its purpose within the scope of the classroom, in that we have gained invaluable experience working with satellite imagery data, when exposure to this in the classroom is limited.

1. Introduction

1.1. Motivation

The Amazon rainforest is well-known as an acting carbon-sink, effectively absorbing copious, but difficult to adequately quantify amounts of carbon dioxide (CO_2) out of the earth's atmosphere[1]. For those unfamiliar with its significance, CO_2 is a major greenhouse gas that when released, creates a layer of gas in the atmosphere that traps heat and contributes to global warming and climate change.

The main motivation behind our research includes the most recent 2019 ecological disaster in the Amazon rainforest, where Brazil saw a 77% increase in forest fires from the previous year (an estimated total 40,000 fires)[2], largely attributed to Brazil's president's loosening of environmental regulations and relaxed enforcement of the illegal and unsustainable practice of slash-and-burn farming, where farmers and cattle ranchers burn remains of forests or crops to fertilize the soil. Additionally, it is important to note that these fires were largely in regions where fires previously occurred, so overall forest cover of old-growth rainforests were not largely affected.

1.2. Additional Research

This move away from deforestation has been the general trend over the past decade in the Amazon an overall 80% decrease in deforestation of the Amazon Rainforest over

[1] Edna Rodig, *From small-scale forest structure to Amazon-wide carbon estimates* (London: Nature Communication 2019)
[2] Alexandria Symonds *Amazon Rainforest Fires: Here's What's Really Happening* (New York: New York Times 2019)

2008-2018[3]. However, forest *degradation* in the form of forest fires and logging have continued to not only negatively affect the world's carbon stock, but also continues to negatively affect the ecological health of surrounding forests, biodiversity, and endangered species' habitats.

The same study found that carbon stocks in single-burned forests decrease by 50% of their estimated original carbon stock; a second occurance of fires in a given forest decreases the percent of remaining carbon stock to less than 20% and forests burned greater than two times yield estimated decreases down to less than 10% of original carbon stock[3]. This study shows that even though deforestation rates are declining and farmers are reusing land, that forest degradation in the form of slash and burn farming isn't sustainable and it still contributes to the growing issue of climate change due to greenhouse gases. The area of increased forest fire prevalence is known as the Amazon arc of deforestation and this occurs at the southern border of the Amazon Rainforest, predominantly located in Brazil, who is estimated to profit $8.3 billion annually from the ecological services provided by the rainforest[4].

1.3. Our Hypothesis

After diving into this research, it becomes clear that there has always been an issue regarding weighing the economic and environmental benefits and impacts when it comes to the Amazon Rainforest. That is why we sought to create a supervised classification model over the same time period 2008-2018 that is able to take in spectral data and change over a time period, in order to 1) validate and visualize the forest loss and forest gain in these past 10 years leading up to the 2019 ecological disaster, and 2) to ultimately create a model that is be able to detect these changes from 2018-today. We hypothesize that if we are able to test and train this model, we will be able to classify future levels of degradation as well as correlate spikes in yearly degradation with regional and human activities like El Nino events (leading to decrease crop yields in Southern Hemisphere) and government reported statistics of forest fire counts. This'll ideally show us the impact that El Nino and fire counts contribute to forest loss or gain.

1.4. The Data

Our initial endeavors in data collection began with Kaggle. By navigating Kaggle, we found a csv dataset that gave us initial impressions on forest degradation in the Amazon Rainforest through plots that were easy to understand and visualize. We found numerous datasets, the first, being a data set that contained the number of forest fires per state per month, by region in the Amazon Rainforest[5]. This data allowed us to conduct preliminary analysis on one key source of forest degradation in the Amazon Rainforest. From this data we were able to plot the rate of forest fires over time and confirm empirically the steady increase of man-made forest-fires in the Amazon as well as underscore states in Brazil that were most affected by deforestation and to study more in depth with data we obtained later.

We then explored another, broader data set, that detailed the square kilometerage of deforestation of each Brazilian state that was a part of the Amazon Rainforest, el niño and

[3] Danielle Rappaport, *Quantifying long-term changes in carbon stocks and forest structure from Amazon forest degradation* (Berkeley: Environmental Research Letters 2018)

[4] Jon Stand, *Spatially explicit valuation of the Brazilian Amazon Forest's Ecosystem Services* (United Kingdom: Nature 2018)

[5] Modelli, Gustavo. "Forest Fires in Brazil." Kaggle, 24 Aug. 2019, www.kaggle.com/gustavomodelli/forest-fires-in-brazil.

la niña years of occurrences and levels of severity, as well as more detailed data on forest fires that included the latitude and longitude of forest fires per state as well as the month it occured[6]. We were able to continue our initial analysis using this data as well as attempt to analyze something novel in how el niño and la niña affect levels of forest degradation in the Amazon.

After our initial data analysis we were ready to move onto what would be our primary source of data. The primary source of our data comes from Google's Earth Engine (GEE)[7]. GEE provides satellite imaging data from NASA's Landsat and Sentinel-2 satellites in a manner that facilitates gathering and analysis. GEE furthermore provides a console that allows for simple data collection through the vast Earth Engine API and allows us to use Google's Cloud Computing Servers to perform the complex data analysis of machine learning models required for handling satellite image data.

1.5. The Plan

In Section 1, we outlined our motivations behind this classification and we described how this data set was acquired. In Section 2, we further analyzed the data and discussed our methods for model construction. Section 3 presents our full final models, and additional applications of the model. Finally Section 4 is the conclusion and summary of our results.

2. Data Exploration and Model Construction

[6] Netto, Mariana Boger. "Brazilian Amazon Rainforest Degradation 1999-2019." *Kaggle*, 27 Dec. 2019, www.kaggle.com/mbogernetto/brazilian-amazon-rainforest-degradation.

[7] "FAQ – Google Earth Engine." *Google*, Google, earthengine.google.com/faq/.

2.1. Preliminary Data Visualization

Using the second set of data acquired from Kaggle, we were able to create Figure 1 in Python, which shows the heavily decreased but sustained deforestation rate between 2008-2018 of the different independent states of Brazil. With this data and visualization we were able to determine forest degradation in the Amazon Rainforest has diminished significantly culminating in all time low in 2012 to what is now a steady increase. Using Tableau with the first set of data acquired from Kaggle, we were able to plot the quantity of forest fires by year and

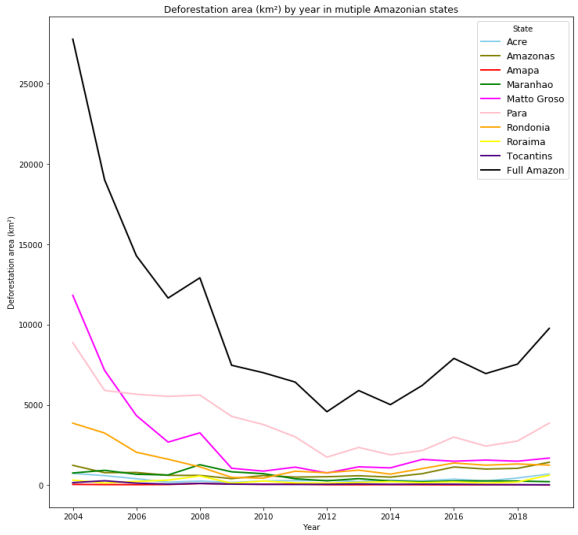

Figure 1. Deforestation area (km^2) by year and Amazonian state

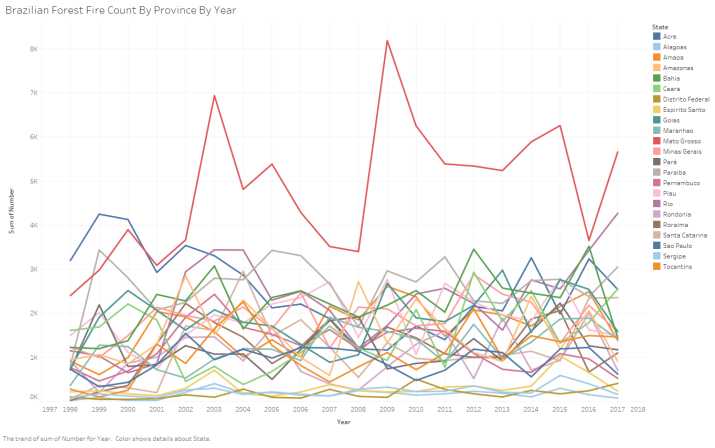

Figure 2. Fire count by year and Amazonian state

Amazonian state in Figure 2. This data shows us that forest fires are not the primary determining factor of forest degradation in the Amazon Rainforest as the ebbs and flows in figure 2 do not correspond with the upwards and downwards trends in figure 1. From this discrepancy it is more than likely that other factors, such as those political, play a more determining factor as to the level of deforestation given a particular year.

2.2. Methods for Model Construction

Using GEE and following a tutorial[8], we initially were able to develop a supervised random forest classification model on a single, arbitrary square of coordinate data in the Amazon Rainforest, as seen in Figure 3. Our first step is to access the desired imagery; in this case it was spectral Top of Atmosphere (TOA) data from Landsat 5 in 2008 and Landsat 7 in 2018, in which we want to classify change in forest cover over this 10-year time. We followed common practices when removing cloud cover from these images to better access their imagery.

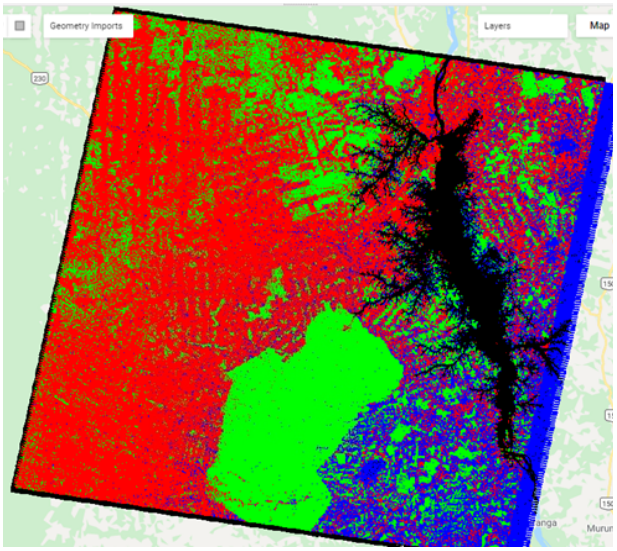

Figure 3. Single Square of Random Forest Classification 2008-2018

[8] Matt Strimas-Mackey. "Land-cover Change Mapping with Earth Engine." Github, 3, Feb 2017 https://github.com/mstrimas/ee-land-change/blob/master/ee-land-change.md

Next, we defined a training set, and in utilizing GEE's editor capabilities, we delineated "known" regions of our classes, and used these regions and the Landsat image of 2008 to train a model by taking random points in our defined regions, and getting the spectral data associated with this class. Our classes are defined as forest, non-forest, forest gain, and forest loss and once the model is run on the testing set, these regions will be classified in green, black, blue, and red accordingly.

Finally, our testing set is defined by the combined spectral images from 2008 and 2018 that emphasizes the change in forest cover, both gain and loss. This random forest model used 30 trees as the parameter, and the model yielded an overall training accuracy of 97.6% and a testing accuracy of 70.71%, found using confusion and validation matrix accordingly.

2.3. Improving our Model

In order to improve this model, we realized that we needed to expand our analysis to a larger region as well as validate our classification model via an already established dataset of forest loss and gain so that we may have a tested and more authoritative model.

To achieve this, we delineated arbitrary points around the Amazon Rainforest to create a polygon that will serve as the new geometry of our image collections rather than using a single image as we did before. The satellite imagery we used comes from image collections of Landsat 5 from Jan. 1, 2008 to Dec. 31, 2008 and of Landsat 8 from Jan. 1, 2018 to Dec. 31, 2018. We were able to convert these image collections into a singular image by removing the cloud cover masks and then combining the images to form a median image. We then applied the random forest model in GEE to produce

Figure 4 which shows the random forest classification model with 100 different trees. Clearly, it is overfitted, so in order to determine the number of trees to use, we plotted the errors for models run with several different numbers of trees as seen in Figure 5. This process allowed us to determine the optimal number of trees that we should use with our final model without overfitting (i.e. the last arbitrarily large drop in error) at 9 trees which yields an accuracy of 95.1%. To further improve our model we chose to use new delineated regions based on Hansen data[9], an already existing image of global forest change from 2000-2018 that shows the forest, non-forest, forest gain and forest loss over this time period. Using this as a more accurate training set, we produced more accurate and representative results using supervised training and validation.

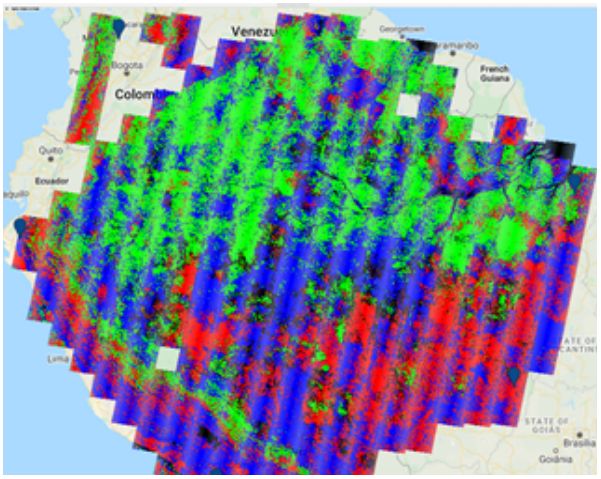

Figure 4. Larger Region of Random Forest Classification Using 100 Trees

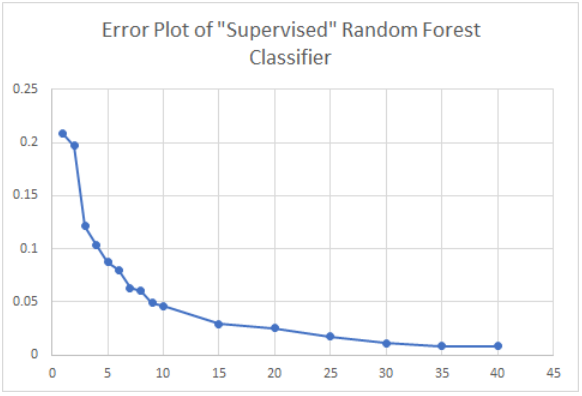

Figure 5. Error plot of supervised random forest classifier

[9] "Global Forest Change 2000–2014Data Download." *Global Forest Change*, earthenginepartners.appspot.com/science-2013-global-forest/download_v1.2.html.

3. Final Models

3.1. Random Forest Classifier 2008-2018

Our final model is shown below in Figure 6, showing a denser, more robust, and more accurate classification of the Amazon Rainforest. The aforementioned Amazon arc of deforestation (i.e. the area in red on the southern border of the rainforest) is also present. This model uses the previously optimal 9 trees in its random forest, yielding a 97.35% training accuracy and 81.7% testing accuracy, as shown in Figure 7 along with the classifier's corresponding confusion matrix.

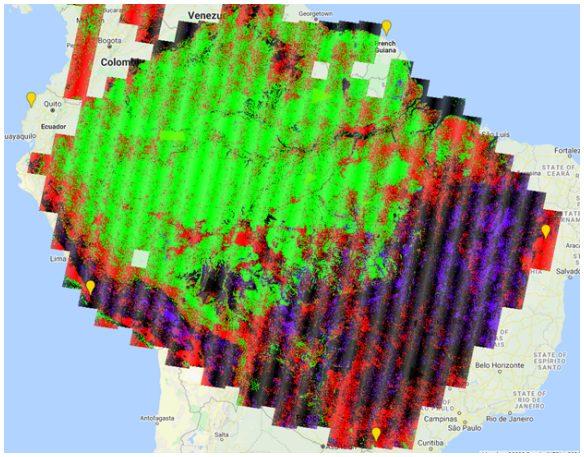

Figure 6. Random forest classifier 2008-2018

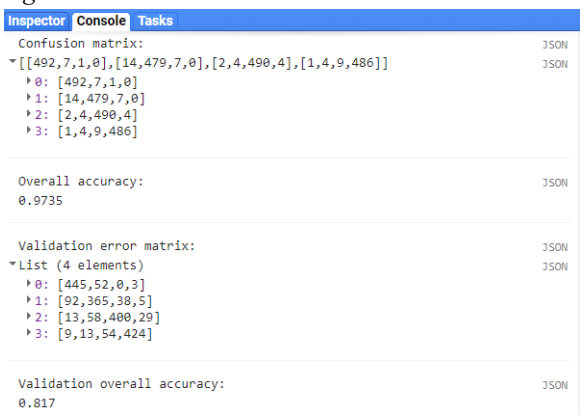

Figure 7. Random forest classifier confusion matrix and validation matrix 2008-2018

From the validation error matrix it is determinable that the model did well in giving us true positives, while refraining from giving us either false positives or false negatives. Out of the 2000 points tested, 1634 were classified correctly with there being less than 239 false positives and less than 127 false negatives for each of forest, non-forest, forest loss and forest gain. Although the most mishaps at 92 happened when a forest loss region incorrectly classified as forested, it was not statistically significant enough to see a trend. Thus, it was concluded that the model performed well on each forest type, rather than excelling on one and faltering on the others.

3.2. Same Model 2018-today

For an additional model, we explored our ability to isolate regions that were affected by the forest fires in summer of 2019. To do this, we created new polygons; however, instead of tracing Hansen's image, we traced our own image from section 3.1. We worked with image collections from Landsat 8's Top of Atmosphere dataset, obtaining data from Jan. 1, 2018 - Dec. 31, 2018 and Oct. 1, 2019 - Dec. 31, 2019 to isolate the time following the dry season fires. While we do obtain a higher accuracy in the confusion matrix, this model in theory, should not work, because we are not training it on regions we know experienced forest loss or

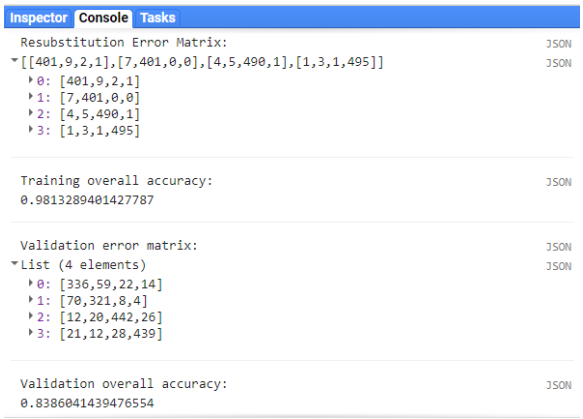

Figure 8. Random forest classifier confusion and validation matrix 2018-present

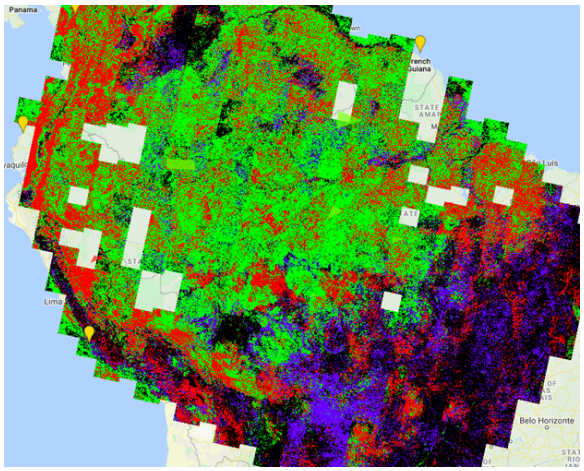

Figure 9. Random forest classifier 2018-present

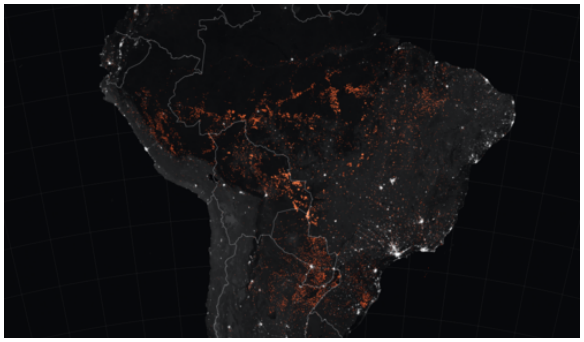

Figure 10. Wikipedia[10] map of 2019 forest fires

forest gain within this time frame. Figures 9 and 10 show our model above in comparison to Wikipedia's map of forest fires during this time in 2019. Our model surprisingly isn't too far off, aligning with the spread of forest fires along the Amazon's arc of deforestation. For accuracy of this model, see Figure 8, as it shows our confusion matrix yielding 98.13% training accuracy and 83.86% testing accuracy.

Moreover, the validation error matrix in Figure 8 provides insight into how the model performed for each type of forest class (in the order of forest, forestLoss, non-forest, and forestGain). Out of the 1,834 points, 1,538 were properly classified. This included no more than 133 false positives and 163 false negatives for each category type. There were no significant trends in the

[10] Wikipedia contributors. "2019 Amazon rainforest wildfires" *Wikipedia*. (accessed April 30, 2020)

confusion matrix, which boded well for the outlook of the model.

4. Summary

4.1. Reviewing Our Models

In this paper, we created two supervised machine learning models to produce images that classify the Amazon Rainforest's forest change over two given time periods. Our best model was in section 3.2, yielding 98.13% accuracy. With a high accuracy we are able to see that the model performed well in translating to the testing data as there were no significant trends in the sensitivity and specificity of the model.

The random forest classifier provided a great deal of accuracy in predicting forest coverage as it is an ensemble model which allows for multiple trees to be considered in making a decision. Since it is an ensemble model it definitely provided greater accuracy, highlighted by the training accuracies being greater than 97% and the testing accuracies being greater than 81% for both of the models in 3.1 and 3.2. Additionally, after monitoring the error rates for various amounts of trees, we were able to account for overfitting in the model. This allowed our validation set to be tested by an appropriate classifier which wasn't completely overfitted using the training data.

Finally, it is important to note the shortcomings of our models. Working with this supervised machine learning classifier, there are apparent limitations with its applications, as testing polygons need to be drawn around regions of known change. Documenting some changes in regions to train a larger model would allow for more accurate and dependable models than we currently have, and our model can definitely be improved upon to extend beyond the Amazon rainforest, using more detailed imagery for a more detailed analysis, or in quantifying the change that we were able to classify.

4.2. Conclusion

Both of our models were effective in what they sought to do. We recognize that mapping global forest change is work that already exists, however, further analysis of the changes in forests, as we did in our model, is an effective and necessary step towards solving the forest degradation crisis that affects the biodiversity, ecological health, and carbon stocks of the Amazon Rainforest. Once the Amazon Rainforest is effectively monitored, and the remaining carbon stock in the forest is thoroughly estimated, it would soon be feasible to instantiate the widespread use of carbon as a stock and to provide economic incentives to preserve what's remaining of the Amazon Rainforest.

4.3. Future Work

We achieved the goal outlined in our hypothesis of creating a machine learning model that outlined how the Amazon Rainforest has degraded over a ten year period. With this model we believe that it is now possible to predict quantities of degradation in the future in the Amazon Rainforest. We also believe that we can continue improving this model by potentially finding correlations or lack of correlations between forest degradation and various natural phenomena such as natural disasters, El Niño, and La Niña. This will allow us to enhance our model by including variations in how natural causes may affect increased forest degradation and forest growth.

Additionally, the techniques used to obtain our current model are easily applicable to other forest ecosystems as Landsat satellite imagery is not limited to one location, but instead stores imagery of

the entire world. Therefore, these techniques can be applied to other large forest ecosystems such as Siberia or Alaska. Lastly, one of our initial aims was to expand this research over a much larger period of time to analyze the rates of degradation from the 20th century to current times. While we were successful in classifying these regions with their spectral data, it proved more difficult to obtain rates of degradation or of deforestation by region than we initially anticipated.

4.4. Github Code

While this course was Python-focused, learning to work with Landsat data with Python in Pangeo and Google Colaboratory, for this project, however, we primarily used Google Earth Engine and coded primarily in Javascript. We did use some Python for our initial model, so we included that notebook with the rest of our code in the following Github repository: https://github.com/jonahsoberano/gee-amazon-rainforest-classification

Acknowledgements

We thank Brian White for teaching this course on data analytics in application to Earth Science. We would not have gained experience working with GEE and satellite imagery data had this course not been offered. We also thank our courses Teaching Assistants Carly Richardson and Izzy Hinks for assisting and giving feedback throughout the process.